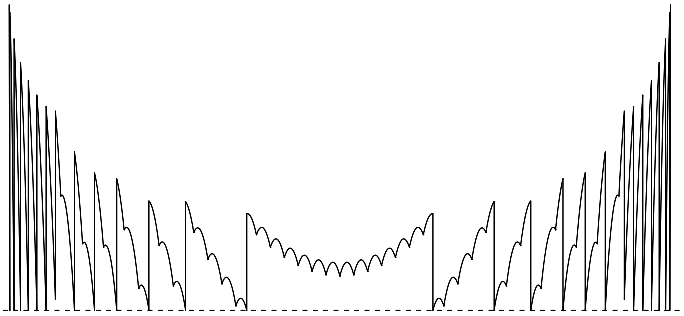
Confidence and Prediction under Covariates and Prior Information

Dissertationsschrift zur Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften
an der Julius-Maximilians-Universität Würzburg



vorgelegt von

Kristina Lurz

aus Bamberg

Würzburg 2015

Eingereicht am 20.05.2015

1. Gutachter: Prof. Dr. Rainer Göb

2. Gutachter: Prof. Dr. Sven Knoth

Tag der Disputation: 12.11.2015

Abstract

The purpose of confidence and prediction intervals is to provide an interval estimation for an unknown distribution parameter or the future value of a phenomenon. In many applications, prior knowledge about the distribution parameter is available, but rarely made use of, unless in a Bayesian framework. This thesis provides exact frequentist confidence intervals of minimal volume exploiting prior information. The scheme is applied to distribution parameters of the binomial and the Poisson distribution. The Bayesian approach to obtain intervals on a distribution parameter in form of credibility intervals is considered, with particular emphasis on the binomial distribution. An application of interval estimation is found in auditing, where two-sided intervals of Stringer type are meant to contain the mean of a zero-inflated population. In the context of time series analysis, covariates are supposed to improve the prediction of future values. Exponential smoothing with covariates as an extension of the popular forecasting method exponential smoothing is considered in this thesis. A double-seasonality version of it is applied to forecast hourly electricity load under the use of meteorological covariates. Different kinds of prediction intervals for exponential smoothing with covariates are formulated.

Zusammenfassung

Konfidenz- und Prognoseintervalle dienen der Intervallschätzung unbekannter Verteilungsparameter und künftiger Werte eines Phänomens. In vielen Anwendungen steht Vorinformation über einen Verteilungsparameter zur Verfügung, doch nur selten wird außerhalb von bayesscher Statistik davon Gebrauch gemacht. In dieser Dissertation werden exakte frequentistische Konfidenzintervalle unter Vorinformation kleinsten Volumens dargelegt. Das Schema wird auf Verteilungsparameter für die Binomial- und die Poissonverteilung angewandt. Der bayessche Ansatz von Intervallen für Verteilungsparameter wird in Form von Vertrauensintervallen behandelt, mit Fokus auf die Binomialverteilung. Anwendung findet Intervallschätzung in der Wirtschaftsprüfung, wo zweiseitige Intervalle vom Stringer-Typ den Mittelwert in Grundgesamtheiten mit vielen Nullern enthalten sollen. Im Zusammenhang mit Zeitreihenanalyse dienen Kovariaten der Verbesserung von Vorhersagen zukünftiger Werte. Diese Arbeit beschäftigt sich mit exponentieller Glättung mit Kovariaten als eine Erweiterung der gängigen Prognosemethode der exponentiellen Glättung. Eine Version des Modells, welche doppelte Saison berücksichtigt, wird in der Prognose des stündlichen Elektrizitätsbedarfs unter Zuhilfenahme von meteorologischen Variablen eingesetzt. Verschiedene Arten von Prognoseintervallen für exponentielle Glättung mit Kovariaten werden beschrieben.

Acknowledgments

I am sincerely grateful to Prof. Dr. Rainer Göb for supervising this thesis and sharing his expertise and experience with me, which considerably shaped my research and work. I appreciate that he attached great importance that I keep the balance between theory and the practical purpose of our work. I am very thankful that he gave me the possibility to visit conferences, where I both could learn a lot and meet colleagues, some of whom I can now call good friends.

Thank you, Dr. Antonio Pievatolo, for a fruitful collaboration, inspiring discussions and your hospitality during a both productive and enjoyable research stay at CNR-IMATI in Milan. Thank you, Dr. Fabrizio Ruggeri and all colleagues from CNR-IMATI, for considerably contributing to the latter.

Several topics of my research would not have been covered if I had not had the financial assistance to go on the above mentioned and another research stay or attend congresses. I would like to thank all institutions and programs that financially supported these undertakings.

I am grateful to Dr. Ljubica Lozo for her guidance within a very beneficial mentoring program. Thank you, Dr. Bernhard Hein, for preparing me for the business life.

Besides, I would like to thank those who helped proofreading this thesis: Dr. Frederike Rüppel, Dr. Thomas Christ, Gregor Nagel and Matthias Krebs.

A special thanks goes to my family and friends for their continuous support, patience and trust during the last couple of years.

Finally, thank you, Matthias, for always being there for me in the first year out of many.

Contents

List of Figures

List of Tables

1 Introduction

Estimates of quantities and predictions of future values are matters of every-day life. We encounter them when enquiring about the rate of adverse reactions to medications, watching the next week's weather forecast or listening to experts' opinions on next year's economic growth, and we base decisions on them: We decide to take the drug when the benefits sufficiently outweigh the side-effects, take an umbrella if it is likely to rain or invest in a stock that is likely to increase in value. In industrial contexts, exploiting good estimates is just as useful. In an incoming goods inspection, the estimated defective rate determines the acceptance or rejection of a lot. Demand forecasts support inventory control.

The numbers we most frequently encounter are *point estimates* or *point forecasts*. They are appealing for obvious reasons: One particular number is easy to communicate and leaves little room for doubt on how to take action. In reality, however, mere point forecasts may lead one to believe that there is a degree of certainty in the estimates which can rarely be found in practice. A way out are *interval estimates*, which are commonly called *confidence intervals* in the context of estimating distribution parameters and *prediction intervals* in connection with forecasting. The term *confidence interval* was shaped by Neyman (1934), but the concept goes back both to him and Fisher (1930). According to Neyman's (1934) definition, a confidence interval B is an interval in which the unknown population parameter θ is assumed to be lying. The probability that this statement is wrong is supposed to be less than $1-\gamma$, that is, $\gamma \in (0;1)$, the predetermined *confidence level*, is intended to be a lower bound for the probability $\mathrm{P}_\theta(\theta \in B)$. Interval estimates are more difficult to communicate because they make one aware of how uncertain the outcome or statement about a phenomenon can be. Uncertainty is not desirable, but rarely avoidable, which is why giving in to the illusion of precision of a point estimate can be fallacious.

Various distinctions have to be made with respect to estimation problems: The target of the estimation procedure can be an unknown distribution parameter underlying the investigated phenomenon, in which case there is an *estimation* problem, or the future value of a time series, in which case the purpose is *prediction*. We can take a *frequentist's* *approach* or a *Bayesian approach*. Exploiting Bayes' theorem requires knowledge about the *a priori* probabilities related to the target parameter. It had been the common approach to perform interval estimation before Fisher (1930) proposed an equivalent following the frequentist *repeated sampling principle*, which does not rely on prior distributions. Interval estimates can be of *exact* or *approximate* type. The common definition of *exact* in the context of estimation is that the actual *coverage probability* $\mathrm{P}_\theta(\theta \in B)$ – the probability that θ is contained in B given θ is the true value of the distribution parameter – is greater or equal γ for an arbitrary θ from the parameter space Θ, that is, $\mathrm{P}_\theta(\theta \in B) \geq \gamma$ holds for all θ. Otherwise, the latter condition holds only approximately. Usually, the actual coverage probability is intended to be close to the nominal confidence level, but strict exactness $\mathrm{P}_\theta(\theta \in B) = \gamma$ can only hold in the case of continuous distributions, as was already pointed out by Fisher (1930). In the case of discrete distributions, the coverage probability is usually greater than γ for a large range of values from Θ.

Confidence and prediction intervals are regarded the more useful the more precise they are, where preciseness is usually measured by means of the interval length: Narrower intervals mean a more accurate estimation than wider intervals because they allow a more detailed statement on the quantity of interest. Smaller intervals usually involve larger sample sizes. They are more likely to lead to a decision than larger intervals. The interval length is therefore an important quality characteristic of a confidence or prediction interval. Exploiting additional information, if available, can help increase accuracy and by that lead to more economic estimation procedures.

In this thesis, the focus is on estimation and prediction procedures under prior information and covariates with particular emphasis on interval estimation. On the one hand, confidence intervals for unknown distribution quantities are considered, where *prior knowledge* on an unknown distribution parameter is supposed to enhance the estimation procedure. On the other hand, *covariates* are made use of to improve the prediction of future values of a time series.

Initially, the focus of this thesis lay on the very practical topic of statistical auditing, which will be addressed in Chapter 5: the precise estimation of the average misstatement in accounting or auditing populations. The necessity for statistical sampling and evaluation procedures in the auditing field arises from the usually large sizes of the populations. Although they are finite, their complete investigation is normally too costly. The particular feature of this field of application lies in the availability of prior information. From previous audits, experiences with organisations from the same industrial sector or tests of the internal control systems, the auditor is usually not completely ignorant about the misstatement rate in the books. It is natural to attempt to exploit this knowledge in order to achieve a more precise estimation, which is closely associated with the sample size. Because it is a main driver of the audit costs, the sample size is intended to be kept low. The *Stringer bound* is a commonly used upper bound for the mean misstatement in audit populations proposed by Stringer (1963) and is supposed to support audit decisions based on statistical procedures. It has so far been used in its one-sided version with upper limit, which can lead to the decision of accepting the population in case the bound is not larger than the tolerable misstatement, and otherwise leads to indifference. To enable that the sampling and evaluation procedure can also entail a rejection of the population, a two-sided procedure is required. It is not strict model-based theory that drove the development of the one-sided Stringer bound. Its conservativeness under reasonable choices of the confidence level has been shown either asymptotically or in simulation studies. We will prove the conservativeness of the two-sided interval in several important special cases and support its conservativeness in more complex situations in a simulation study. Auditing is not the only field of application of this particular confidence interval. Since it is strong especially in populations that contain many values of zero – so-called *zeroinflated populations* – potential other applications are, for example, estimating accident costs in insurance or measuring entities where measurement imprecision causes small signals to be cumulated on zero.

Since it processes a well-known statistical procedure, namely that of estimating a probability, the Stringer bound provides the opportunity to fine-tune the interval estimation by means of improved two-sided confidence intervals for a binomial proportion. With the help of prior knowledge imposed on the binomial probability parameter p in the flexible form of a beta distribution, exact confidence intervals can be obtained that are of minimum weighted volume. This method to compute confidence intervals is an instance of a theory that generalises an approach by von Collani & Dumitrescu (2001) and von Collani et al. (2001). To a great extent, they built on the work of Neyman (1937) who established the duality of confidence regions with prediction regions. The theory as well as the application to confidence limits for a binomial proportion, including a comparison to existing confidence limits for a proportion, will be presented in Chapter 2. An efficient computational algorithm is described and has been implemented in R. The majority of the chapter was published in the *Metrika* journal in 2014 (Göb & Lurz 2014).

Although they make use of prior information, the intervals in Chapter 2 are of frequentist type. Less frequentist statistics than Bayesian statistics is however the typical framework to include prior knowledge. The Bayesian equivalent to confidence intervals are credibility intervals. They serve as a summary of the posterior distribution and in the form of highest posterior density (HPD) intervals maximise the posterior density. A popular credibility interval for the binomial parameter makes use of the beta distribution as prior distribution, for it is the conjugate prior in the binomial case. Comparing these intervals to the frequentist confidence intervals exploiting prior knowledge is an obvious undertaking. Therefore, Chapter 3 is devoted to understand Bayesian HPD intervals and to compare them with frequentist minimum volume confidence intervals under prior information. We acknowledge that especially for many Bayesians the mere attempt to compare these two substantially different approaches is already reprehensible, but encouraged by Bayarri & Berger (2004), Fraser (2011) and the *matching prior* idea, both approaches are united in their own Chapter 3 and some theoretical findings are taken from the comparison in the binomial case.

The theory for minimum volume confidence intervals presented in Chapter 2 is of general type and applicable to more than just binomial confidence limits. Another application of the theory in the field of estimating a distribution parameter is provided in Chapter 4, where the theory is used to develop confidence intervals for the parameter of another discrete distribution, the *Poisson distribution*. The conjugate prior again serves as the prior information distribution, hence the gamma distribution is exploited to express prior knowledge on the mean = variance of the Poisson distribution. The general theory is basically applicable to the Poisson distribution, but several difficulties arise in the Poisson case in contrast to the binomial case, mainly due to the unboundedness of the parameter space of the Poisson parameter.

The second part of the thesis, consisting of Chapters 6 and 7, is concerned with the problem of predicting time-dependent observations. Time series methodology is applied to obtain both point estimates as well as prediction intervals for future outcomes of a series of a time-indexed phenomenon. The popular forecasting method of *exponential smoothing* initialised by Brown (1959) and Holt (1957) and extended by Wang (2006) from the purely history-based – what Chatfield (2001) calls *univariate* method – to the *multivariate* procedure including further explanatory variables, is considered in Chapter 6. *Exponential smoothing with covariates (ESCov)* with a *single source of error (SSOE) state-space model* as the underlying statistical model is formulated for multiple seasonalities. The methodology is applied in an electricity load study with Italian data to forecast the hourly electricity consumption with the help of meteorological covariates, more precisely, temperature. Large parts of the chapter can be found published in the *ASMBI* journal (Göb et al. 2013a,b). Chapter 7 is devoted to prediction intervals for

ESCov. So-called *plug-in* prediction intervals, which do not take the uncertainty in the estimation of parameters into account, easily turn out to show undercoverage. We therefore formulate a method exploiting linear model theory that is supposed to make up for this. The theory as well as the simulation study and application on daily electricity load data has been published in *QREI* journal (Göb et al. 2014). Further prediction interval types are considered, among them an empirical interval based on the Camp-Meidell inequality that uses findings from a book chapter of *Frontiers in Statistical Quality Control 11*, see Göb & Lurz (2015).

Structure of the Thesis

The thesis is structured as follows: Minimum volume confidence intervals under prior information are considered in Chapter 2. The general theory is introduced and applied to obtain shortest confidence intervals for a probability. Bayesian credibility intervals for a probability are investigated in Chapter 3. The highest posterior density intervals from Chapter 3 are being looked at from a frequentist point of view and a connection between them and the frequentist intervals from Chapter 2 is established. The method of minimum volume confidence intervals for a distribution parameter from Chapter 2 is transferred to the expectation of a Poisson distribution in Chapter 4. In Chapter 5, confidence intervals for the mean in zero-inflated populations are examined, with particular focus on auditing populations. In Chapter 6, the time series method exponential smoothing with covariates is presented and applied with the purpose of forecasting hourly electricity load. Chapter 7 is devoted to prediction intervals for exponential smoothing with covariates.

Proofs

For the sake of readability, proofs are provided in the appendices of each chapter.

Software

The numerical results of Chapters 2 to 7 have been achieved by means of R code that has been developed by the author of this thesis. It is available, upon request, from her.

Notation

Throughout the thesis, uppercase bold letters denote matrices (e. g. $\mathbf{M}$), lowercase bold letters denote vectors (e. g. $\boldsymbol{x}$), lowercase letters, not bold, denote scalars (e. g. α). The symbol $^\top$ is the symbol for transposition of matrices. $E[\cdot]$ is the expectation, $V[\cdot]$ the variance and $\mathrm{Cov}[\cdot]$ the covariance. $\mathbf{I}$ denotes the identity matrix, P the probability operator and $\mathbb{1}$ the indicator function.

Statistical Distributions

Beta Distribution A continuous random variable X has the four-parametric beta distribution $\mathrm{Beta}(p_0, p_1, a, b)$ on the support $[p_0; p_1]$ if it has the density function

$$f_X(x) \quad = \quad \frac{1}{B(a,b)(p_1 - p_0)} \left(\frac{x - p_0}{p_1 - p_0}\right)^{a-1} \left[1 - \frac{x - p_0}{p_1 - p_0}\right]^{b-1} \qquad \text{for } x \in (p_0; p_1) \quad (1.1)$$

and $f_X(x) = 0$ for $x \in \mathbb{R}\backslash(p_0; p_1)$. Here, $a, b > 0$ are shape parameters and

$$B(s,t) \quad = \quad \frac{\Gamma(s)\Gamma(t)}{\Gamma(s+t)} \quad = \quad \int_0^1 p^{s-1}(1-p)^{t-1}\,\mathrm{d}p \quad = \quad B(t,s) \quad \text{for } s,t > 0 \qquad (1.2)$$

is the symmetric beta function and

$$\Gamma(z) \quad = \quad \int_0^\infty t^{z-1} \exp(-t)\,\mathrm{d}t \qquad (1.3)$$

the Gamma function. See Abramowitz & Stegun (1972, Sections 6.1 & 6.2) for the formulas for the gamma and beta functions.

The most popular type of beta distribution is the two-parametric beta distribution with density

$$f_X(x) \quad = \quad \frac{1}{B(a,b)} y^{a-1}(1-y)^{b-1} \qquad \text{for } x \in (0;1) \qquad (1.4)$$

and $f_X(x) = 0$ for $x \in \mathbb{R}\backslash(0;1)$. We denote the two-parametric beta distribution as $\mathrm{Beta}(a,b)$ in this thesis to spare us the lengthy notation $\mathrm{Beta}(0,1,a,b)$. See Johnson et al. (1994) for a definition and details on the beta distribution.

Binomial Distribution A discrete random variable X has the binomial distribution $Bi(n,p)$ with sample size $n \in \mathbb{N}$ and probability parameter $p \in [0;1]$ if it has the probability mass function

$$f_X(x) \quad = \quad \mathrm{P}(X = x) \quad = \quad \binom{n}{x} p^x (1-p)^{n-x} \qquad \text{for } x = 0, 1, \ldots, n. \qquad (1.5)$$

See Johnson et al. (1992) for a definition of the binomial distribution and further details.

Chi-square Distribution A continuous random variable X has the chi-square distribution $\chi^2(\nu)$ with $\nu \in \mathbb{N}$ degrees of freedom if it has the density function

$$f_X(x) \quad = \quad \frac{1}{2^{\nu/2}\Gamma\left(\frac{\nu}{2}\right)} x^{\nu/2-1} \exp\left(-\frac{x}{2}\right) \qquad \text{for } x \geq 0 \tag{1.6}$$

and $f_X(x) = 0$ for $x < 0$.

See Abramowitz & Stegun (1972, 26.4.1) for a definition of the distribution function of $\chi^2(\nu)$.

***F*-distribution** A continuous random variable X has the F-distribution with $\nu_1, \nu_2 \in \mathbb{N}$ degrees of freedom if it has the density function

$$f_X(x) \quad = \quad \frac{\nu_1^{\frac{1}{2}\nu_1} \nu_2^{\frac{1}{2}\nu_2}}{B\left(\frac{1}{2}\nu_1, \frac{1}{2}\nu_2\right)} x^{\frac{1}{2}(\nu_1-2)} (\nu_2 + \nu_1 x)^{-\frac{1}{2}(\nu_1+\nu_2)} \qquad \text{for } x \geq 0 \tag{1.7}$$

and $f_X(x) = 0$ for $x < 0$. See Abramowitz & Stegun (1972, 26.6.1) for a definition of the density function of the F-distribution.

Gamma Distribution A continuous random variable X has the gamma distribution $\text{Gamma}(\vartheta, \kappa)$, $\vartheta, \kappa > 0$, on the support $(0; +\infty)$ if it has the density function

$$f_X(x) \quad = \quad \frac{x^{\kappa-1}}{\vartheta^\kappa \Gamma(\kappa)} \exp\left(\frac{-x}{\vartheta}\right) \quad \text{for } x > 0 \tag{1.8}$$

and $f_X(x) = 0$ for $x \leq 0$. See Bowman & Shenton (1988) for a definition and details on the gamma distribution.

Normal Distribution A continuous random variable X is normally distributed with mean μ and variance σ^2, i. e. $X \sim N(\mu, \sigma^2)$, if it has the density function

$$f_X(x) \quad = \quad \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \qquad \text{for } x \in \mathbb{R}. \tag{1.9}$$

See Abramowitz & Stegun (1972, 26.2.9) for a definition of the density function of the normal distribution.

Poisson Distribution A discrete random variable X is distributed according to the Poisson distribution $Po(\lambda)$, with $\lambda > 0$, if it has the probability mass function

$$f_X(x) \quad = \quad \mathrm{P}(X = x) \quad = \quad \frac{\lambda^x}{x!} \exp(-\lambda) \qquad \text{for } x = 0, 1, \ldots \tag{1.10}$$

See Johnson et al. (1992) for a definition of the Poisson distribution and further details.

Student's t-Distribution A continuous random variable X has the central t-distribution $t(\nu)$ with ν degrees of freedom if it has the density function

$$f_X(x) \quad = \quad \frac{1}{\sqrt{\nu} B\left(\frac{1}{2}, \frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \mathrm{d}x, \tag{1.11}$$

where $B(s, t)$ is the beta function as defined in Eq. (1.2). See Abramowitz & Stegun (1972, 26.7.1) for a definition of the distribution function of the distribution $t(\nu)$.

Uniform Distribution A continuous random variable X has the uniform distribution or equidistribution $\mathrm{Unif}(u, v)$ with support $[u; v]$ if it has the density function

$$f_X(x) \quad = \quad \frac{1}{v - u} \quad \text{for } x \in [u; v] \tag{1.12}$$

and $f_X(x) = 0$ for $x \in \mathbb{R}\backslash[u; v]$. The uniform distribution can be obtained as a special case of the beta distribution by setting $a = 1 = b$ and $p_0 = u, p_1 = v$ in Eq. (1.1). See Abramowitz & Stegun (1972, 26.1.34) for a definition of the uniform distribution.

2 Design and Analysis of Shortest Two-sided Confidence Intervals for a Probability under Prior Information

2.1 Introduction

About the same time when Neyman (1934) introduced the general concept of a *confidence interval* for a distribution parameter θ, Clopper & Pearson (1934) presented a method for determining confidence intervals for a probability p. A random variable X that is distributed according to the binomial distribution $Bi(n,p)$ with $p \in [0;1]$ and sample size $n \in \mathbb{N}$ has the probability mass function

$$f_X(x) \quad = \quad \mathrm{P}(X = x) \quad = \quad \binom{n}{x} p^x (1-p)^{n-x} \quad \text{for } x = 0, 1, \ldots, n, \tag{2.1}$$

and the distribution function $F_X(x)$ or operating characteristic (OC) function $L_{n,c}(p)$

$$F_X(x) \quad = \quad \mathrm{P}(x \leq c) \quad = \quad \sum_{x=0}^{c} \binom{n}{x} p^x (1-p)^{n-x} \quad = \quad L_{n,c}(p) \quad \text{for } c = 0, 1, \ldots, n. \tag{2.2}$$

The definition and properties of the binomial distribution can be found in Johnson et al. (1992).

Let $X = x$ be the number of realised binomial successes in a sample of size n. Let $\gamma \in (0;1)$ be a confidence level. The two-sided Clopper & Pearson (1934) confidence interval for a binomial probability p is given by

$$\begin{aligned} B &= [p_{L,CP}; p_{U,CP}] \qquad\qquad (2.3) \\ &= \left[z_{\mathrm{Beta}(x,n-x+1)}\left(\frac{1-\gamma}{2}\right); z_{\mathrm{Beta}(x+1,n-x)}\left(\frac{1+\gamma}{2}\right) \right], \\ &= \left[\frac{x}{x + (n-x+1) z_{F_{2(n-x+1),2x}}\left(\frac{1+\gamma}{2}\right)}; \frac{x+1}{(x+1) + (n-x) z_{F_{2(n-x),2(x+1)}}\left(\frac{1-\gamma}{2}\right)} \right] \end{aligned}$$

where $p_{L,CP}$ is equal to the solution in p of the equation

$$P(X \geq x) \quad = \quad \sum_{k=x}^{n} \binom{n}{k} p^k (1-p)^{n-k} \quad = \quad \frac{1-\gamma}{2}$$

and $p_{U,CP}$ is equal to the solution in p of the equation

$$P(X \leq x) \quad = \quad \sum_{k=0}^{x} \binom{n}{k} p^k (1-p)^{n-k} \quad = \quad \frac{1-\gamma}{2},$$

respectively, see e. g. Agresti & Coull (1998). $z_{\text{Beta}(a,b)}(\alpha)$ is the $100\alpha\,\%$-quantile of the two-parametric beta distribution $\text{Beta}(a,b)$ on $[0;1]$ with parameters $a,b>0$, and $z_{F_{d_1,d_2}}(\alpha)$ is the $100\alpha\,\%$-quantile of the F-distribution with $d_1, d_2 \in \mathbb{N}$ degrees of freedom. The relation between the beta distribution and the F-distribution utilised in Eq. (2.3) can be established by exploiting their relationships with the regularised incomplete beta function, see Abramowitz & Stegun (1972, 26.6.2 & 26.5.1).

One-sided versions of a confidence interval of level $\gamma \in (0;1)$ for a binomial proportion if x successes out of n are observed are given by

$$\begin{aligned}
[0;\ p_{U,CP}] \quad &= \quad \left[0; z_{\text{Beta}(x+1,n-x)}\left(\gamma\right)\right] \\
&= \quad \left[0; \frac{x+1}{(x+1)+(n-x)z_{F_{2(n-x),2(x+1)}}\left(1-\gamma\right)}\right] \\
&\quad\ \text{(one-sided interval with upper bound)},
\end{aligned} \tag{2.4}$$

$$\begin{aligned}
[p_{L,CP};\ 1] \quad &= \quad \left[z_{\text{Beta}(x,n-x+1)}\left(1-\gamma\right); 1\right] \\
&= \quad \left[\frac{x}{x+(n-x+1)z_{F_{2(n-x+1),2x}}\left(\gamma\right)}; 1\right] \\
&\quad\ \text{(one-sided interval with lower bound)},
\end{aligned} \tag{2.5}$$

where $p_{U,CP}$, $p_{L,CP}$ are solutions in p of the equations

$$\sum_{k=0}^{x} \binom{n}{k} p^k (1-p)^{n-k} \quad = \quad 1-\gamma \quad \Leftrightarrow \quad \sum_{k=x+1}^{n} \binom{n}{k} p^k (1-p)^{n-k} \quad = \quad \gamma,$$

and

$$\sum_{k=x}^{n} \binom{n}{k} p^k (1-p)^{n-k} \quad = \quad 1-\gamma \quad \Leftrightarrow \quad \sum_{k=0}^{x-1} \binom{n}{k} p^k (1-p)^{n-k} \quad = \quad \gamma,$$

respectively, see e. g. Agresti & Coull (1998).

The intervals by Clopper & Pearson (1934) from Eq. (2.3) and those from Eqs. (2.4)–(2.5) are exact in the sense that a prescribed confidence level γ is preserved, i. e. $\text{P}_p(p \in B) \geq \gamma$

for any value $p \in [0; 1]$. However, the two-sided version of it is unnecessarily wide, and the actual coverage probability $\mathrm{P}_p(p \in B)$ considerably exceeds the prescribed level γ for a wide range of values $p \in [0; 1]$. Neyman (1937) established the duality of confidence regions and acceptance or prediction regions. The criterion of Neyman's (1937) definition of shortest confidence intervals is not geometrical volume. Instead, it is required that for any true parameter value which determines the actual probability measure and any other value differing from the true value, the probability of covering the non-true value is a minimum. For discrete distributions, this construction ends up in randomised confidence intervals. Tables of randomised Neyman-shortest confidence intervals for a binomial parameter were calculated by Blyth & Hutchinson (1960). However, randomised procedures are rarely used by practitioners.

Without explicit reference to Neyman's (1937) work, Sterne (1954) used the relation between confidence and prediction regions to calculate confidence regions for p which are throughout smaller in a geometric sense than the Clopper & Pearson (1934) intervals. Crow (1956) seems to be the first author who explicitly considered the concept of the total geometric volume of confidence regions, demonstrating that Sterne's (1954) regions were the smallest in this sense. Crow (1956) realised that Sterne's (1954) regions were not always intervals. By a modification of Sterne's (1954) method, Crow (1956) obtained shortest intervals without increasing the total volume of the regions. A survey and classification of various approaches is provided by Blyth & Still (1983).

Crow's (1956) approach is refined by von Collani & Dräger (2001) who account for prior knowledge on p, expressed by a rectangular distribution of p with support $[p_0; p_1] \subset [0; 1]$. Narrow intervals $[p_0; p_1]$ express a high degree of prior knowledge and lead to shorter confidence intervals. This prior information approach is adopted to the estimation of parameters of parametric families of distributions in general by von Collani & Dumitrescu (2001) and von Collani et al. (2001).

Volume minimising confidence intervals have not yet become a customary tool of statistical field work, presumably for two reasons: i) Various authors have produced extensive tables of shortest intervals, e. g. von Collani & Dräger (2001), but efficient numerical routines for real time computation are not readily available. As a consequence, statistical software packages like SAS, Statistica, SPSS or Minitab throughout offer only the Clopper & Pearson (1934) intervals as exact solutions. ii) The methods suggested hitherto have not accounted for prior knowledge on p, except von Collani & Dräger's (2001) recent approach which may be considered as cumbersome because of the sharp cut-off between the region $[p_0; p_1]$ of equally likely p and the complementary region $[0; 1] \setminus [p_0; p_1]$ of p

out of consideration. Using prior knowledge is a critical issue in many applications, in particular in auditing or industrial quality control where small or very small probabilities nonconforming are a certainty.

The present study presents and analyses a more flexible approach to expressing prior knowledge on p. The study is organised in the following sections: Section 2.2 explains the general concept of minimum volume confidence intervals for a distribution parameter and the connection between prediction regions and confidence regions. Section 2.3 introduces a model imposing prior information on the parameter p of the binomial distribution in terms of a beta distribution. Important properties of functions necessary for the computation of the prediction and confidence intervals are discussed in Sections 2.4 and 2.5. In Section 2.6, the prediction intervals and confidence intervals for a probability p are presented. Section 2.7 compares minimum volume confidence intervals without prior information with some other approaches to confidence intervals for a probability. Section 2.8 illustrates the effect of prior information on the confidence intervals. Section 2.9 investigates the probability of being indifferent if a decision making process is based on the two-sided minimum volume confidence intervals under prior information. Finally, Section 2.10 outlines the algorithm used for an efficient computation of the prediction and confidence intervals.

2.2 Prediction Regions and Confidence Regions

The subsequent framework for parameter measurement and prediction generalises the approach of von Collani & Dumitrescu (2001) and von Collani et al. (2001).

Consider two random variables $X\colon \Omega \to R_1$, $Y\colon \Omega \to R_2$ with ranges $R_1 \subset \mathbb{R}^{m_1}$, $R_2 \subset \mathbb{R}^{m_2}$. In the empirical interpretation, X is an observable datum and Y is the parameter of the distribution of X. Two empirical interests may occur: 1) Given the parameter value Y, determine a *prediction region* for the event X. 2) Having observed X, provide a *confidence region* for the parameter Y.

Let $\mathcal{A}_1$, $\mathcal{A}_2$ be σ-fields in R_1, R_2, respectively. Let $f_{X,Y}$ be the joint density of X, Y with respect to a product measure $\mu_1 \otimes \mu_2$ on the product field $\mathcal{A}_1 \otimes \mathcal{A}_2$, and let f_X, f_Y be the respective marginal densities. For sets $B \in \mathcal{A}_1$ let

$$\mathrm{P}_y(B) \;=\; \mathrm{P}(B|Y=y) \;=\; \int_B f_{X|Y=y}(x)\,\mathrm{d}\mu_1(x) \tag{2.6}$$

be the conditional probability under $Y = y$. For sets $A \in \mathcal{A}_1 \otimes \mathcal{A}_2$ let $A_x = \{y|(x,y) \in A\}$, $A_y = \{x|(x,y) \in A\}$ be the projections for fixed $x \in R_1$, $y \in R_2$, respectively. In the model considered by von Collani & Dumitrescu (2001) and von Collani et al. (2001), f_Y is the density of an equidistribution on a finite rectangle.

Let $0 < \gamma < 1$. A set $A \in \mathcal{A}_1 \otimes \mathcal{A}_2$ is called a *level* γ *measurement and prediction space for* $X|Y$ (*level* γ *MPS for* $X|Y$) if the projection A_X constitutes a confidence interval for the unknown value of Y, i. e.

$$\gamma \quad \leq \quad \mathrm{P}_y(y \in A_X) \quad = \quad \mathrm{P}_y(X \in A_y) \qquad \text{for all } y \in R_2. \tag{2.7}$$

The right-hand equality in Eq. (2.7) shows that for each $y \in R_2$, A_y is a *level* γ *prediction region* or *level* γ *acceptance region* for X, i. e. $\gamma \leq \mathrm{P}_y(X \in A_y)$. Although the model reflects prior knowledge on the parameter Y via the density f_Y, the confidence region characterised by Eq. (2.7) follows the frequentist approach. The inequality $\mathrm{P}_y(y \in A_X) \geq \gamma$ is stipulated pointwise for each parameter value y. A Bayesian credibility region B_x rather requires $\int_{B_x} f_{Y|X=x}(y)\,\mathrm{d}y \geq \gamma$ pointwise for each observation x.

von Collani & Dumitrescu (2001) and von Collani et al. (2001) evaluate the quality of a level γ MPS A by the weighted volume

$$V(A) \quad = \quad \int_{R_1}\int_{A_x} \mathrm{d}\nu(y) f_X(x)\,\mathrm{d}\mu_1(x) \quad = \quad \int_{R_2}\int_{A_y} f_X(x)\,\mathrm{d}\mu_1(x)\,\mathrm{d}\nu(y), \tag{2.8}$$

where ν is the Borel measure, i. e. $\int_{A_x} \mathrm{d}\nu(y) = \nu(A_x)$ is the geometric volume of A_x. Under a prescribed level γ, it is desired to use a *minimum volume* MPS, i. e. an MPS $A^\star$ with $V(A^\star) \leq V(A)$ for all level γ MPS A.

As shown by Theorem 2.2, the problem of determining minimum volume level γ MPSs is closely related to determining level γ prediction regions consisting of largest *prediction likelihood ratios*

$$Q_y(x) \quad := \quad \frac{f_{X|Y=y}(x)}{f_X(x)} \quad = \quad \frac{f_{X,Y}(x,y)}{f_X(x) f_Y(y)} \tag{2.9}$$

in x for pointwise prescribed y. For $y \in R_2$, $t \in \mathbb{R}$, define the segments of largest prediction likelihood ratios

$$D_{\geq t}(y) \quad := \quad \{x | Q_y(x) \geq t\}\,, \qquad \text{for } y \in R_2 \tag{2.10}$$

and $D_{=t}(y)$, $D_{\leq t}(y)$ analogously. The subsequent proposition considers the coverage probability of $D_{>t}(y)$ with respect to X as a function of the bound t.

Proposition 2.1 (Coverage Probability of Likelihood Segments)**.** *Let $y \in R_2$ and let the function $G_y\colon \mathbb{R} \to [0;1]$ be defined by $G_y(t) = P_y(X \in D_{>t}(y)) = P_y(Q_y(X) > t)$. Then we have:*

G_y is decreasing, right-continuous, with $G_y(t^-) = P_y(X \in D_{\geq t}(y))$, $G_y(t^-) - G_y(t) = P_y(X \in D_{=t}(y))$ for $t \in \mathbb{R}$. G_y is continuous in t iff $P_y(X \in D_{=t}(y)) = 0$.

Proposition 2.1 is obvious since $1 - G_y(t) = \mathrm{P}_y(Q_y(X) \leq t)$ is the distribution function of $Q_y(X)$. The subsequent theorem explains the relation between segments of largest prediction likelihood ratios and minimum volume MPSs.

Theorem 2.2 (Minimum Volume Level γ MPS). *Let $0 < \gamma < 1$, for $y \in R_2$ let $s_y := \inf\{t | G_y(t) \leq \gamma\}$, and for each $y \in R_2$ let a set $E(y) \subset D_{=s_y}(y)$ with $P_y(X \in D_{>s_y}(y) \cup E(y)) \leq \gamma$. Let $A^\star := \{(x, y) | y \in R_2, x \in D_{>s_y}(y) \cup E(y)\}$.*

Then we have:

a) $G_y(s_y^-) = P_y(X \in D_{\geq s_y}(y)) \geq \gamma \geq P_y(X \in D_{>s_y}(y)) = G_y(s_y)$ for all $y \in R_2$.

b) For any level γ MPS $A \in \mathcal{A}_1 \otimes \mathcal{A}_2$ we have $V(A^\star) \leq V(A)$.

c) If μ_1 is the Lebesgue measure, and if for all $y \in R_2$ the set $D_{=s_y}(y)$ is a countable union of intervals with disjoint interior, then $A^\star$ is a level γ MPS with $\gamma = P_y(X \in A^\star_y)$ for all $y \in R_2$, and hence by assertion b) a minimum volume MPS.

Proof. See Appendix 2.A, Section 2.A.1. □

von Collani et al. (2001) conjecture that the equation $G_y(s) = \mathrm{P}_y(Q_y(X) \geq s) = \gamma$ in s has always a solution if μ_1 is the Lebesgue measure. This conjecture does not hold in general, as can be demonstrated by counterexamples. However, in most all practically relevant cases the condition of assertion c) of Theorem 2.2 is fulfiled.

The minimum volume objective is accompanied by intuitive requirements on the structure of the prediction and confidence regions. A basic requirement is the *convexity* of prediction and confidence regions. In the univariate case, the convexity of a region is equivalent to the region being an interval. The subsequent proposition shows that the interval property and the monotonicity of the interval bounds of prediction and confidence regions are closely related.

Proposition 2.3 (Interval Property). *Let the ranges $R_1, R_2 \subset \mathbb{R}$, let $A \in \mathcal{A}_1 \otimes \mathcal{A}_2$ with $A_x \neq \emptyset$ for each $x \in R_1$ and with $A_y \neq \emptyset$ for each $y \in R_2$. We use the following definitions:*

1) For a set $B \subset \mathbb{R}$, and a real $z \in \mathbb{R}$ define $z < B$ iff $z < z'$ for all $z' \in B$. Analogously, define $z > B$.

2) A has nondecreasing bounds with respect to x *if the following holds: For $x_1, x_2 \in$*

R_1, $x_1 < x_2$, all $y \in R_2$, we have that $y < A_{x_1}$ implies $y < A_{x_2}$, and that $y > A_{x_2}$ implies $y > A_{x_1}$.

3) *The property of* nondecreasing bounds with respect to y *is defined analogously to 2) by interchanging the positions of "x" and "y".*

4) *A has* nonincreasing bounds with respect to x *if the following holds: For $x_1, x_2 \in R_1$, $x_1 < x_2$, all $y \in R_2$, we have that $y > A_{x_1}$ implies $y > A_{x_2}$, and that $y < A_{x_2}$ implies $y < A_{x_1}$.*

5) *The property of* nonincreasing bounds with respect to y *is defined analogously to 4) by interchanging the positions of "x" and "y".*

Then we have:

a) *The following two sets of conditions are equivalent:*

   a.i) *For each $x \in R_1$, A_x is an interval, and A has the nondecreasing bounds property with respect to x.*

   a.ii) *For each $y \in R_2$, A_y is an interval, and A has the nondecreasing bounds property with respect to y.*

b) *The following two sets of conditions are equivalent:*

   b.i) *For each $x \in R_1$, A_x is an interval, and A has the nonincreasing bounds property with respect to x.*

   b.ii) *For each $y \in R_2$, A_y is an interval, and A has the nonincreasing bounds property with respect to y.*

c) *The following two sets of conditions c.i) and c.ii) are equivalent:*

   c.i) *The sets A_x and A_y fulfil one of the following two conditions:*

      c.i.1) *For each $x \in R_1$, A_x is an interval, and for each $y \in R_2$, A_y is an interval.*

      c.i.2) *There is at least one $x \in R_1$ such that A_x is not an interval, and at least one $y \in R_2$ such that A_y is not an interval.*

   c.ii) *A has either the nondecreasing bounds property simultaneously with respect to x and y, or the nonincreasing bounds property simultaneously with respect to x and y.*

PROOF. See Appendix 2.A, Section 2.A.2. □

If the intervals A_x or A_y, respectively, are all uniformly open or all uniformly closed, the above defined nondecreasing and nonincreasing bounds conditions can be replaced by the simpler conditions that $\inf A_x$, $\sup A_x$ ($\inf A_y$, $\sup A_y$) are nondecreasing in x (in y), or nonincreasing in x (in y), respectively.

2.3 The Beta Prior Model for Inference on a Probability p

We consider an instance of the model established by the preceding section applied to the probability parameter $y = p$ of a binomial distribution.

1) The univariate random variable Y with values in $R_2 = [p_0; p_1] \subset [0; 1]$ represents a random probability parameter which varies according to a beta distribution $\mathrm{Beta}(p_0, p_1, a, b)$ with shape parameters $a, b > 0$ on the support $[p_0; p_1]$, with density

$$f_Y(y) \;=\; \frac{1}{B(a,b)(p_1 - p_0)} \left(\frac{y - p_0}{p_1 - p_0}\right)^{a-1} \left[1 - \frac{y - p_0}{p_1 - p_0}\right]^{b-1} \tag{2.11}$$

for $p_0 < y < p_1$, $f_Y(y) = 0$ for $y \in \mathbb{R} \setminus (p_0; p_1)$, where

$$B(s,t) \;=\; \frac{\Gamma(s)\Gamma(t)}{\Gamma(s+t)} \;=\; \int_0^1 p^{s-1}(1-p)^{t-1}\,\mathrm{d}p \;=\; B(t,s) \quad \text{for } s,t > 0 \tag{2.12}$$

is the symmetric beta function. The measure μ_2 is the Lebesgue measure on R_2.

2) Given a value $Y = y$ of the probability, the random sum X of n binary occurrences with values 0 or 1 is conditionally distributed by $Bi(n, y)$. The range of X is $R_1 = \{0, \ldots, n\}$ and μ_1 is the counting measure on R_1.

The prior information model used by von Collani & Dräger (2001), i. e. equidistribution of Y on the support $[p_0; p_1]$, is a special case of the assumptions 1) and 2), in particular a special case of Eq. (2.11) with $a = 1 = b$.

From assumptions 1) and 2), the unconditional density $f_X(x)$ of X is $f_X(x) = w_{p_0,p_1,a,b}(x)$ where

$$w_{p_0,p_1,a,b}(x) \;= \\ \frac{\binom{n}{x}}{B(a,b)(p_1 - p_0)} \int_{p_0}^{p_1} \left(\frac{y - p_0}{p_1 - p_0}\right)^{a-1} \left[1 - \frac{y - p_0}{p_1 - p_0}\right]^{b-1} y^x (1-y)^{n-x}\,\mathrm{d}y \tag{2.13}$$

for $x \in \{0, \ldots, n\}$. In the important special case $p_0 = 0$, $p_1 = 1$ we have

$$w_{0,1,a,b}(x) \;=\; \frac{B(x+a, n-x+b)}{B(a,b)} \binom{n}{x} \;=\; \frac{B(x+a, n-x+b)}{xB(x, n-x+1)B(a,b)}\,. \tag{2.14}$$

The random probability Y has expectation and variance

$$\mu_Y \;=\; E[Y] \;=\; p_0 + \frac{(p_1 - p_0)a}{a+b}, \;\; \sigma_Y^2 \;=\; V[Y] \;=\; \frac{(p_1 - p_0)^2 ab}{(a+b+1)(a+b)^2}\,. \tag{2.15}$$

From the general formula (2.8) we obtain the weighted volume of an MPS A as

$$V(A) \quad = \quad \sum_{x=0}^{n} \int_{A_x} \mathrm{d}\nu(y) w_{p_0,p_1,a,b}(x) \quad = \quad \sum_{x=0}^{n} \nu(A_x) w_{p_0,p_1,a,b}(x), \tag{2.16}$$

where ν is the Lebesgue measure, and A_x is the confidence region for y formed under the observation x. Considered in the volume context, the values $w_{p_0,p_1,a,b}(x)$ are called *volume weights.*

The beta distribution model has several appealing characteristics, which made it the preferred distribution for expressing prior information on a probability $y = p$, particularly in Bayesian statistics: flexibility; sparse parametrisation; the property of being the conjugate prior for the binomial distribution in the case $[p_0; p_1] = [0; 1]$; the potential to express various density shapes like bathtub, inverted bathtub, strictly decreasing, strictly increasing or constant (equidistribution). In stochastic modelling of sampling inference on a probability p under prior information, the most prevalent version of the beta distribution is the support $[p_0; p_1] = [0; 1]$, see for instance Hald (1981) in acceptance sampling in quality control, Godfrey & Andrews (1982) or Berg (2006) in audit sampling. The generalised beta distribution with supports differing from $[0; 1]$ is often used in risk analysis, particularly in project risk analysis, see Kendrick (2009).

In the case of repetitive sampling, the parameters of the prior information distribution may be estimated from historical data, for instance in audit sampling or quality control, where data from past inspections may be exploited. However, often appropriate reference data are not available. In this case, the features of the distribution have to be elicited from expert opinions in interviews or panels. The process of eliciting distributions from experts has received considerable interest in the literature, with particular emphasis on the beta distribution, see Corless (1972), Hogarth (1975), Kadane et al. (1980), Chaloner & Duncan (1983), O'Hagan (1998), Walls & Quigley (2001), for instance. Software assisted approaches were considered by Blocher & Robertson (1976) or Garthwaite & O'Hagan (2000). A customary algorithm for determining the beta parameters has the following steps: i) Specify the support $[p_0; p_1]$ of Y. ii) Specify the mean μ_Y. iii) Specify a quantile p_ρ, i. e. a value p_ρ with $F_Y(p_\rho) = \rho$. iv) Solve for the parameters a and b. The topic of eliciting a beta distribution will be resumed in Section 3.7.

In many applications, very small probabilities $p = y$ are to be expected, in particular in auditing or quality control, where $p = y$ is a misstatement rate of account entries or a proportion of nonconforming product units. Empirical studies on audit populations confirm that high misstatement rates rarely occur, see Johnson et al. (1981) and Ham et al. (1985). For this case, an appropriate model is a prior density f_Y decreasing on

$[0;1]$ with a large probability mass close to 0. A simple instance of this shape is a beta distribution $\mathrm{Beta}(0,1,1,b)$ on the support $[p_0;p_1] = [0;1]$ with first shape parameter $a = 1$. The parameter b can be determined by specifying either the mean μ_Y or a quantile p_ρ.

A reduction of the support $[p_0;p_1]$ to a proper subset of $[0;1]$ should be handled with care. The support reduction leads to a corresponding cut-off in the obtained confidence intervals so that values outside $[p_0;p_1]$ are excluded. A misspecified support has a devastating effect on the coverage probability: Values $p \notin [p_0;p_1]$ are covered with probability zero. The effect of a misspecified beta prior on the full unit interval $[0;1]$ is much less problematic: The minimum volume objective is failed, but the prescribed confidence remains unaffected for every $p \in [0;1]$. An appropriately chosen beta distribution on $[0;1]$ may have nearly the same volume reducing effect as a cut-off, without involving problematic effects on the coverage.

2.4 Prediction Likelihood Maximisation and the Interval Property of Confidence Regions for a Probability p

Blyth & Still (1983) list desirable properties of confidence regions for a probability $p = y$. A basic requirement is that the confidence regions and the corresponding prediction regions should be intervals. Conditions ensuring the interval property have been discussed above in Proposition 2.3.

On the other hand, if interest is in minimum volume regions, the basic Theorem 2.2 suggests to consider prediction regions which are subsets of the areas $D_{\geq s}(y)$ of largest prediction likelihood ratio, see the defining Eq. (2.10). Under the model of Section 2.3 we obtain from Eq. (2.9) the prediction likelihood ratio

$$Q_{p_0,p_1,a,b,y}(x) \quad = \quad \frac{\binom{n}{x} y^x (1-y)^{n-x}}{w_{p_0,p_1,a,b}(x)} \quad = \quad \frac{y^x (1-y)^{n-x}}{v_{p_0,p_1,a,b}(x)} \quad \text{for } x \in \{0,\ldots,n\} \tag{2.17}$$

where the *relative volume weights* are defined by

$$\begin{aligned} v_{p_0,p_1,a,b}(x) \quad &= \quad \frac{w_{p_0,p_1,a,b}(x)}{\binom{n}{x}} \\ &= \quad \frac{1}{B(a,b)(p_1-p_0)} \int_{p_0}^{p_1} \left(\frac{y-p_0}{p_1-p_0}\right)^{a-1} \left[1 - \frac{y-p_0}{p_1-p_0}\right]^{b-1} y^x (1-y)^{n-x} \, \mathrm{d}y \end{aligned} \tag{2.18}$$

for $x \in \{0,\ldots,n\}$. For technical analysis, the weights $w_{p_0,p_1,a,b}$, $v_{p_0,p_1,a,b}$ and the likelihood ratio $Q_{p_0,p_1,a,b,y}$ can be considered as functions on the continuous interval $[0;n]$.

The interest in obtaining prediction intervals and the interest in prediction regions of largest prediction likelihood ratio are not conflicting on principal. Proposition 2.6 in the subsequent section shows that the prediction likelihood ratio $Q_{p_0,p_1,a,b,y}(x)$ as a function of x is either increasing or decreasing or of inverted bathtub shape. Hence the areas of largest values of $Q_{p_0,p_1,a,b,y}(x)$ are always intervals.

2.5 Properties of Weights and Prediction Likelihood Ratios

We investigate the essential quantities of the beta prior model introduced by Sections 2.3 and 2.4 as functions of the number x of outcomes “1” in a sample of size n: the volume weights $w_{p_0,p_1,a,b}(x)$ defined by Eq. (2.13), the relative volume weights $v_{p_0,p_1,a,b}(x)$ defined by Eq. (2.18) and the prediction likelihood ratio $Q_{p_0,p_1,a,b,y}(x)$ defined by Eq. (2.17). The obtained propositions are helpful for the discussion of the structure of confidence regions in Section 2.6.

Propositions 2.4 and 2.5 consider the volume weights $w_{p_0,p_1,a,b}(x)$ introduced by Eq. (2.13). The weights determine the influence of the confidence region A_x formed under the observation x onto the total volume $V(A)$ of the MPS, see Eq. (2.16). Proposition 2.4 considers the case of an equidistributed probability parameter Y on an arbitrary interval $[p_0; p_1] \subset [0; 1]$. Proposition 2.5 considers the case of a beta distributed probability parameter Y on the full unit interval $[p_0; p_1] = [0; 1]$. The combined case, i. e. a beta distribution on a proper subinterval $[p_0; p_1] \neq [0; 1]$, is omitted. The analysis of this case implies technical involvements which are not in reasonable proportion to the importance of the case, see the concluding remarks in Section 2.3. The proofs of Propositions 2.4 and 2.5 are based on elementary calculation, exploiting the properties of the beta and gamma function with respect to Eq. (2.14).

Proposition 2.4 (Volume Weights under Equidistribution)**.** *Let $a = 1 = b$, and consider the weights $w_{p_0,p_1,a,b}(x) = w_{p_0,p_1,1,1}(x)$ for $x = 0, \dots, n$.*

a) For $x \in \{0, \dots, n-1\}$ we have

$$\begin{aligned} w_{p_0,p_1,1,1}(x+1) &= w_{p_0,p_1,1,1}(x) \\ &\quad + \binom{n}{x+1} \frac{1}{n-x} \left\{ p_0^{x+1}(1-p_0)^{n-x} - p_1^{x+1}(1-p_1)^{n-x} \right\}. \end{aligned} \tag{2.19}$$

b) In the case $p_0 = 0$, $p_1 = 1$ the weights are constant on $\{0, \dots, n\}$ with $w_{p_0,p_1,1,1}(x) = \frac{1}{n+1}$.

In the case $0 = p_0 < p_1 < 1$ the weights are strictly decreasing on $\{0, \dots, n\}$.

In the case $0 < p_0 < p_1 = 1$ *the weights are strictly increasing on* $\{0, \ldots, n\}$.

In the case $0 < p_0 < p_1 < 1$, *let*

$$x_0 \quad := \quad \frac{-\ln\left(\frac{p_1}{p_0}\right) - n\ln\left(\frac{1-p_1}{1-p_0}\right)}{\ln\left(\frac{p_1}{p_0}\right) - \ln\left(\frac{1-p_1}{1-p_0}\right)}.$$

Then we have

$$w_{p_0,p_1,1,1}(x+1) \begin{cases} > w_{p_0,p_1,1,1}(x) & \text{if } x < x_0, \\ = w_{p_0,p_1,1,1}(x) & \text{if } x = x_0, \\ < w_{p_0,p_1,1,1}(x) & \text{if } x > x_0. \end{cases}$$

Proposition 2.5 (Volume Weights under Beta Distribution on Full Support)**.** *Let* $p_0 = 0$, $p_1 = 1$, *and consider the weights* $w_{p_0,p_1,a,b}(x) = w_{0,1,a,b}(x)$ *for* $x = 0, \ldots, n$.

a) *For* $x \in \{0, \ldots, n-1\}$ *we have*

$$\frac{w_{p_0,p_1,a,b}(x+1)}{w_{p_0,p_1,a,b}(x)} \quad = \quad \frac{(x+a)(n-x)}{(x+1)(n-x+b-1)}. \tag{2.20}$$

b) *In the case* $a = 2 - b > 1$ *the weights are strictly increasing on* $\{0, \ldots, n\}$.

In the case $a = 2 - b < 1$ *the weights are strictly decreasing on* $\{0, \ldots, n\}$.

In the case $a + b \neq 2$, *let* $x_0 := \frac{n(1-a)+b-1}{2-a-b}$. *Then we have in the case* $a + b < 2$

$$w_{0,1,a,b}(x+1) \begin{cases} > w_{0,1,a,b}(x) & \text{if } x > x_0, \\ = w_{0,1,a,b}(x) & \text{if } x = x_0, \\ < w_{0,1,a,b}(x) & \text{if } x < x_0. \end{cases}$$

In the case $a + b > 2$ *we have*

$$w_{0,1,a,b}(x+1) \begin{cases} > w_{0,1,a,b}(x) & \text{if } x < x_0, \\ = w_{0,1,a,b}(x) & \text{if } x = x_0, \\ < w_{0,1,a,b}(x) & \text{if } x > x_0. \end{cases}$$

Proposition 2.6 considers the relative volume weights $v_{p_0,p_1,a,b}(x)$ introduced by Eq. (2.18), which appear as the denominator of the prediction likelihood ratio $Q_{p_0,p_1,a,b,y}(x)$, see Eq. (2.17). It is shown that the relative weight as a function of $x \in [0; n]$ is either decreasing, increasing or has a bathtub shape.

Proposition 2.6 (Properties of Relative Volume Weights)**.** *Consider the relative volume weights* $v_{p_0,p_1,a,b}(x)$ *defined for* $x \in [0; n]$ *by Eq.* (2.18).

a) Let $0 < p_0 < p_1 < 1$*. Then* $v_{p_0,p_1,a,b}$ *has on* $[0;n]$ *the derivatives*

$$v^{(m)}_{p_0,p_1,a,b}(x) = \frac{1}{B(a,b)(p_1-p_0)} \cdot \int_{p_0}^{p_1} \left(\frac{y-p_0}{p_1-p_0}\right)^{a-1} \left(1-\frac{y-p_0}{p_1-p_0}\right)^{b-1} \ln\left(\frac{y}{1-y}\right)^m y^x(1-y)^{n-x}\,dy. \tag{2.21}$$

In particular, $v'_{p_0,p_1,a,b}$ *is strictly increasing on* $[0;n]$*, and there exists a value* $0 \le x_{p_0,p_1,a,b} \le n$ *such that* $v_{p_0,p_1,a,b}$ *is strictly decreasing on* $[0;x_{p_0,p_1,a,b}]$ *and strictly increasing on* $[x_{p_0,p_1,a,b};n]$*.*

b) Let $p_0 = 0$*,* $p_1 = 1$*. Then the first derivative of* $v_{p_0,p_1,a,b} = v_{0,1,a,b}$ *on* $[0;n]$ *is*

$$\begin{aligned} v'_{0,1,a,b}(x) &= B(x+a,n-x+b)\Big[\psi(x+a)-\psi(n-x+b)\Big] \\ &= v_{0,1,a,b}(x)\Big[\psi(x+a)-\psi(n-x+b)\Big] \quad \text{for } x \in [0;n], \end{aligned} \tag{2.22}$$

where ψ *is the digamma function, see Abramowitz & Stegun (1972, Chapter 6). The derivative* $v'_{0,1,a,b}$ *has at most one change of sign* $x_{0,1,a,b}$ *from* $-$ *to* $+$ *on* $[0;n]$*, and* $v_{0,1,a,b}$ *is strictly decreasing on* $[0;x_{0,1,a,b}]$ *and strictly increasing on* $[x_{0,1,a,b};n]$*.*

c) Let $0 < p_0 < 1 = p_1$*. Then there is a value* $x_{p_0,1,a,b} \in [0;n]$ *such that* $v_{p_0,1,a,b}$ *is decreasing on* $[0;x_{p_0,1,a,b}]$ *and increasing on* $[x_{p_0,1,a,b};n]$*.*

d) Let $p_0 = 0 < p_1 < 1$*. Then there is a value* $x_{0,p_1,a,b} \in [0;n]$ *such that* $v_{0,p_1,a,b}$ *is decreasing on* $[0;x_{0,p_1,a,b}]$ *and increasing on* $[x_{0,p_1,a,b};n]$*.*

PROOF. See Appendix 2.A, Section 2.A.3. □

Proposition 2.7 considers the prediction likelihood ratio $Q_{p_0,p_1,a,b,y}(x)$ defined by Eq. (2.17). It is shown that the prediction likelihood ratio as a function of $x \in [0;n]$ is either decreasing, increasing, or has an inverted bathtub shape.

Proposition 2.7 (Likelihood Ratio)**.** *For* $0 \le p_0 < p_1 \le 1$*,* $p \in (p_0;p_1)\cap(0;1)$ *consider the likelihood ratios* $Q_{p_0,p_1,a,b,y}(x) = y^x(1-y)^{n-x}/v_{p_0,p_1,a,b}(x)$ *defined for* $x \in [0;n]$*, see Eq. (2.17).*

a) For $x \in [0;n]$ *we have*

$$Q'_{p_0,p_1,a,b,y}(x) = Q_{p_0,p_1,a,b,y}(x)\left(\ln\left(\frac{y}{1-y}\right) - \frac{v'_{p_0,p_1,a,b}(x)}{v_{p_0,p_1,a,b}(x)}\right). \tag{2.23}$$

b) Let $0 < p_0 < p_1 < 1$ *or* $0 = p_0$*,* $1 = p_1$*. Then the function*

$$[0;n] \ni x \mapsto \frac{v'_{p_0,p_1,a,b}(x)}{v_{p_0,p_1,a,b}(x)}$$

is strictly increasing. Let the quantity $x_{p_0,p_1,a,b,y} \in [0;n]$ *be defined as follows:*

b.i) *In the case* $v'_{p_0,p_1,a,b}(0)/v_{p_0,p_1,a,b}(0) \geq \ln\left(\frac{y}{1-y}\right)$*:* $x_{p_0,p_1,a,b,y} = n$.

b.ii) *In the case* $v'_{p_0,p_1,a,b}(n)/v_{p_0,p_1,a,b}(n) \leq \ln\left(\frac{y}{1-y}\right)$*:* $x_{p_0,p_1,a,b,y} = 0$.

b.iii) *In the case* $v'_{p_0,p_1,a,b}(0)/v_{p_0,p_1,a,b}(0) < \ln\left(\frac{y}{1-y}\right) < v'_{p_0,p_1,a,b}(n)/v_{p_0,p_1,a,b}(n)$*:* $x_{p_0,p_1,a,b,y}$ *is the unique change of sign from* $-$ *to* $+$ *of* $Q'_{p_0,p_1,a,b,y}$ *on* $[0;n]$.

Then $Q_{p_0,p_1,a,b,y}$ *is strictly increasing on* $[0;x_{p_0,p_1,a,b,y}]$ *and strictly decreasing on* $[x_{p_0,p_1,a,b,y};n]$.

c) *Let* $0 < p_0 < 1 = p_1$. *Then there is a value* $x_{p_0,1,a,b,y} \in [0;n]$ *such that* $Q_{p_0,1,a,b,y}$ *is increasing on* $[0;x_{p_0,1,a,b,y}]$ *and decreasing on* $[x_{p_0,1,a,b,y};n]$.

d) *Let* $p_0 = 0 < p_1 < 1$. *Then there is a value* $x_{0,p_1,a,b,y} \in [0;n]$ *such that* $Q_{0,p_1,a,b,y}$ *is increasing on* $[0;x_{0,p_1,a,b,y}]$ *and decreasing on* $[x_{0,p_1,a,b,y};n]$.

PROOF. See Appendix 2.A, Section 2.A.4. □

2.6 Prediction Intervals and Confidence Intervals for a Probability p

Proposition 2.7 in the preceding section implies that prediction regions of largest prediction likelihood ratio are always intervals. Together with Theorem 2.2, this result suggests a close affinity between the interval property and volume minimisation. However, Proposition 2.3 shows that the interval property of the prediction regions alone does not imply the interval property of confidence regions. Crow (1956) gave five examples of minimum volume MPSs with prediction intervals consisting of segments of largest prediction likelihood ratio, but corresponding confidence regions not being intervals.

However, the interval property of prediction and confidence regions is essential for communicating the methodology to practitioners. Though shortest intervals are not always shortest regions, we follow Crow (1956) and stipulate the interval property both for prediction regions and for confidence regions. In particular, we restrict attention to MPSs with the following characteristics:

P1) The prediction regions are nonempty and of the form

$$A_y = \Big\{c_L(y), c_L(y)+1, \ldots, c_U(y)\Big\} \quad \text{for each } y \in [p_0;p_1],$$

where the *prediction limits* $c_L(y)$, $c_U(y)$ are increasing in $y \in [p_0;p_1]$.

P2) The confidence intervals are nonempty closed intervals of the form

$$A_x = [y_L(x);y_U(x)] \subset [p_0;p_1] \quad \text{for each } x \in \{0,\ldots,n\},$$

where the *confidence limits* $y_L(x)$, $y_U(x)$ are increasing in $x \in \{0, \dots, n\}$.

Among the MPSs satisfying the properties P1) and P2), we search for an MPS $A^\star$ of minimum volume $V(A^\star)$, see (2.16) for the formula of the volume. The subsequent proposition shows how appropriately prescribed prediction limits or appropriately prescribed confidence limits can determine the entire MPS with properties P1) and P2).

Proposition 2.8 (Confidence Limits and Prediction Limits)**.** *Let A be an MPS with nonempty projections A_x and A_y for all $x \in \{0, \dots, n\}$ and all $y \in [p_0; p_1]$.*

- *a) Let A satisfy the property P1), and for the projections A_x which contain at least two points let the lower prediction limit $c_L\colon [p_0; p_1] \to \{0, \dots, n\}$ be left-continuous in $\sup A_x$ and let the upper prediction limit $c_U\colon [p_0; p_1] \to \{0, \dots, n\}$ be right-continuous in $\inf A_x$. Then A satisfies the property P2).*
- *b) If A satisfies the property P2), then A satisfies the property P1).*

Proof. See Appendix 2.A, Section 2.A.5. □

Section 2.2 establishes the duality between level γ prediction and level γ confidence regions. For prediction intervals of type P1) and confidence intervals of type P2), the defining characteristic (2.7) amounts to

$$\begin{aligned} \gamma \quad &\leq \quad \mathrm{P}_y\Big(y_L(X) \leq y \leq y_U(X)\Big) \quad = \quad \mathrm{P}_y\Big(c_L(y) \leq X \leq c_U(y)\Big) \\ &= \quad L_{n,c_U(y)}(y) - L_{n,c_L(y)-1}(y) \quad \text{for each } y \in [p_0; p_1], \end{aligned} \tag{2.24}$$

where the *binomial OC function* $L_{n,c}(y)$ is defined by Eq. (2.2).

Proposition 2.9 describes the minimum content of any level γ prediction interval of type P1) and corresponding confidence interval of type P2).

Proposition 2.9 (Minimum Content of Level γ Prediction Interval)**.** *Let $0 < \gamma < 1$. For $x = 1, \dots, n$, let $p_{x,\gamma}$ be the unique solution of the equation $L_{n,x-1}(p) \overset{!}{=} \gamma$, and let $p_{0,\gamma} = 0.0$. For $x = 0, \dots, n-1$, let $\widetilde{p}_{x,\gamma}$ be the unique solution of the equation $L_{n,x}(p) \overset{!}{=} 1 - \gamma$, and let $\widetilde{p}_{n,\gamma} = 1.0$. Let c_L, c_U be level γ prediction limits with corresponding confidence limits y_L, y_U as characterised by Eq. (2.24). Then the following assertions hold:*

- *a) The sequences $(p_{x,\gamma})_{x\in\{0,\dots,n\}}$ and $(\widetilde{p}_{x,\gamma})_{x\in\{0,\dots,n\}}$ are strictly increasing.*
- *b) In the case of $\gamma \geq 0.5$ we have $p_{x,\gamma} < \widetilde{p}_{x,\gamma}$ for $x \in \{0, \dots, n\}$.*
- *c) Let $x = 0, \dots, n$. We have $(p_{x,\gamma}; \widetilde{p}_{x,\gamma}) \cap [p_0; p_1] \subset [y_L(x); y_U(x)]$ and $c_L(y) \leq x \leq c_U(y)$ for $y \in (p_{x,\gamma}; \widetilde{p}_{x,\gamma}) \cap [p_0; p_1]$.*

Proof. See Appendix 2.A, Section 2.A.6. □

Proposition 2.9 turns out to be very helpful for the efficient computation of shortest confidence intervals. The minimum intervals $[p_{x,\gamma}; \widetilde{p}_{x,\gamma}] \cap [p_0; p_1]$ are contained in any closed level γ confidence interval A_x, but they are in general too short and not yet of level γ. They can, however, be used as starting intervals to be extended to the left and to the right to the proper shortest intervals.

Two more results play a role for the computation of the intervals, see Propositions 2.10 and 2.11. The following describes the monotonicity of the prediction region coverage.

Proposition 2.10 (Monotonicity of Prediction Region Coverage)**.** *For $0 \le x_1 \le x_2 \le n$, $y \in [0;1]$, let $\Delta_{n,x_1,x_2}(y) = L_{n,x_2}(y) - L_{n,x_1-1}(y)$. In the case of $x_1 > 0$ let*

$$p_{x_1,x_2} \quad := \quad \left[1 + \left(\frac{(n-x_2)\cdot\ldots\cdot(n-x_1)}{x_1\cdot\ldots\cdot x_2}\right)^{\frac{1}{x_2-x_1+1}}\right]^{-1}.$$

Then we have:

a) In the case of $x_1 = 0$, $\Delta_{n,x_1,x_2} = L_{n,x_2}$ is strictly decreasing on $[0;1]$ if $x_2 < n$, and constant with $\Delta_{n,x_1,x_2} = L_{n,n}(y) = 1$ if $x_2 = n$.

b) In the case of $x_1 > 0$, Δ_{n,x_1,x_2} is strictly increasing on $[0; p_{x_1,x_2}]$ and strictly decreasing on $[p_{x_1,x_2}; 1]$.

Proof. See Appendix 2.A, Section 2.A.7. □

Proposition 2.10, assertion b), includes the special case $x_2 = n$, when the prediction region coverage Δ_{n,x_1,x_2} is strictly increasing on $[0;1]$ if $x_1 > 0$.

The following proposition deals with the comparison of the prediction likelihood ratios of two prediction points $x_1 \neq x_2$.

Proposition 2.11 (Comparison of Likelihood Ratios)**.** *Consider the likelihood ratios $Q_{p_0,p_1,a,b,y}(x) = y^x(1-y)^{n-x}/v_{p_0,p_1,a,b}(x)$ defined for $x \in [0;n]$, see Eq. (2.17). Let $0 \le x_1 < x_2 \le n$, $y \in (0;1)$. Then we have*

$$Q_{p_0,p_1,a,b,y}(x_1) \begin{cases} < Q_{p_0,p_1,a,b,y}(x_2) & \text{if } y > q_{x_1,x_2}, \\ = Q_{p_0,p_1,a,b,y}(x_2) & \text{if } y = q_{x_1,x_2}, \\ > Q_{p_0,p_1,a,b,y}(x_2) & \text{if } y < q_{x_1,x_2}, \end{cases} \tag{2.25}$$

where

$$q_{x_1,x_2} \quad := \quad \left[\left(\frac{v_{p_0,p_1,a,b}(x_1)}{v_{p_0,p_1,a,b}(x_2)}\right)^{\frac{1}{x_2-x_1}} + 1\right]^{-1}. \tag{2.26}$$

Proof. See Appendix 2.A, Section 2.A.8. □

2.7 Numerical Comparison with other Confidence Intervals without Prior Information

In this section, we analyse confidence intervals for a probability p without prior information on a numerical basis. The absence of prior information is expressed by the beta shape parameters $a = 1 = b$, and the support parameters $p_0 = 0$, $p_1 = 1$. The minimum length intervals conforming to the design principles established by Section 2.6 are compared with five other confidence intervals for a probability p:

i) Blaker's (2000) interval is based on inverting a specific two-sided test for p. Agresti & Min (2001) point out some similarity of Blaker's construction principle with the principle behind the classical shortest volume intervals of Sterne (1954) and Crow (1956).

ii) The interval by Clopper & Pearson (1934), see Eq. (2.3).

iii) The well-known textbook interval with endpoints $\widehat{p} \mp z_{N(0,1)}((1-\gamma)/2)\sqrt{\widehat{p}(1-\widehat{p})/n}$ around the estimator $\widehat{p} = x/n$, where $z_{N(0,1)}(\rho)$ is the $100\rho\,\%$-quantile of the standard normal distribution $N(0,1)$. This interval is often addressed as *Wald's interval.*

iv) Wilson's (1927) score interval. A detailed analysis of the score interval is provided by Krishnamoorthy & Peng (2007).

v) The interval suggested by Agresti & Coull (1998).

The intervals of type i) and ii) are exact in the sense of Eq. (2.24), i. e. the prescribed nominal confidence level γ never exceeds the actual coverage probability. In this sense, the intervals of type iii), iv), and v) are not exact.

For sample sizes $n = 5, \ldots, 30$ and the nominal confidence level $\gamma = 0.95$, Table 2.1 displays the ratios $V(A_i)/V(A)$ of the weighted lengths $V(A_i)$, $i = 1, \ldots, 5$, relative to the weighted length $V(A)$ of the minimum length interval. See Eq. (2.16) for the definition of the weighted length. Blaker's interval comes very close to the minimum length. The Clopper & Pearson interval exceeds the minimum length by more than $5\,\%$, even for larger sample size. The non-exact intervals are considerably shorter than the minimum length interval.

The comparison by weighted length is fallacious with respect to non-exact intervals. They achieve a lower length by offending against the nominal confidence level. This effect is visible from Fig. 2.1, where, under the nominal confidence level $\gamma = 0.95$, the actual coverage $\mathrm{P}_p(p \in A)$ is displayed as a function of $p \in [0; 1]$ for sample size $n = 10$.

Table 2.1: Ratios $V(A_i)/V(A)$ of weighted volumes $V(A_i)$, $i = 1, \dots, 5$, relative to the weighted volume $V(A)$ of the minimum length interval under $\gamma = 0.95$.

n	Blaker	Clopper & Pearson	Wald	score	Agresti & Coull
5	1.00001	1.08095	0.73346	0.88998	0.93756
6	1.00001	1.10377	0.78611	0.91775	0.96686
7	1.00001	1.09158	0.80706	0.91566	0.96413
8	1.00002	1.07251	0.81635	0.90673	0.95391
9	1.00073	1.09336	0.85159	0.93061	0.97744
10	1.00002	1.06829	0.84769	0.91482	0.95906
11	1.00002	1.07576	0.86681	0.92626	0.96946
12	1.00047	1.08752	0.88763	0.94099	0.98345
13	1.00283	1.08582	0.89601	0.94369	0.98498
14	1.00215	1.06098	0.88360	0.92589	0.96517
15	1.00253	1.06501	0.89374	0.93278	0.97133
16	1.00104	1.06775	0.90218	0.93842	0.97612
17	1.00066	1.06756	0.90758	0.94110	0.97803
18	1.00041	1.05638	0.90309	0.93396	0.96970
19	1.00113	1.06095	0.91160	0.94053	0.97566
20	1.00009	1.06164	0.91643	0.94347	0.97793
21	1.00003	1.06811	0.92594	0.95158	0.98543
22	1.00004	1.05698	0.91988	0.94374	0.97658
23	1.00003	1.05394	0.92056	0.94302	0.97509
24	1.00050	1.05944	0.92847	0.94980	0.98142
25	1.00051	1.06301	0.93452	0.95475	0.98591
26	1.00003	1.06381	0.93795	0.95719	0.98776
27	1.00002	1.05390	0.93175	0.94987	0.97960
28	1.00053	1.05277	0.93314	0.95037	0.97952
29	1.00056	1.05174	0.93447	0.95085	0.97950
30	1.00054	1.05390	0.93852	0.95422	0.98240

Again, the behaviour of the minimum length interval and Blaker's interval is nearly indiscernible. The Clopper & Pearson interval is extremely conservative, the nominal level is exceeded for all p. The minimum length interval and Blaker's interval exploit the confidence level in a much more economic way, but still exhibit considerable exceedances over the bound. Among the non-exact intervals, the Wald interval is not a competitive choice since the nominal confidence bound is drastically violated. The score interval and the Agresti & Coull interval behave much better, but still exhibit considerable shortfalls. Agresti & Coull (1998) argue that their interval provides a good balance between the two conflicting interests of reducing the average length and maintaining the nominal confidence level. However, the user has to be aware that the Agresti & Coull interval suffers from considerable violations of the nominal level particularly for very small and very large p.

2.8 Numerical Analysis of the Effect of Prior Information

This section analyses the effect of prior information on minimum volume confidence intervals for a probability p. In the scheme established by Sections 2.3 to 2.6, prior information is expressed by the distribution $\mathrm{Beta}(p_0, p_1, a, b)$ of the random probability Y, where $a, b > 0$ are the beta shape parameters, and where $[p_0; p_1] \subset [0; 1]$ is the support interval.

For the confidence level of $\gamma = 0.95$, Fig. 2.2 provides the relative weighted volume or relative average length $V(A)/V(A_{\mathrm{CP}})$ of minimum volume prior information intervals $A_x = [y_L(x); y_U(x)]$ relative to the Clopper & Pearson (1934) intervals $A_{\mathrm{CP},x}$, where the weighted volume is defined by Eq. (2.16). The absence of prior information is expressed by $Y \sim \mathrm{Beta}(0, 1, 1, 1) = \mathrm{Beta}(1, 1)$, i. e. a uniform distribution $\mathrm{Unif}(0, 1)$ on the full support $[p_0; p_1] = [0; 1]$. The prior information distributions $Y \sim \mathrm{Beta}(0, 0.2, 1, 1)$ and $Y \sim \mathrm{Beta}(0, 0.1, 1, 1)$ with $a = b = 1$ describe uniform distributions $\mathrm{Unif}(0, 0.2)$ and $\mathrm{Unif}(0, 0.1)$ on the support intervals $[p_0; p_1] = [0; 0.2]$ and $[p_0; p_1] = [0; 0.1]$, respectively. The distributions $\mathrm{Beta}(0, 1, 1, b) = \mathrm{Beta}(1, b)$ with $p_0 = 0, p_1 = 1, a = 1$, are beta distributions with strictly decreasing densities on $[0; 1]$. In these cases, small values of p come with a high probability. Finally, the bathtub shaped prior $\mathrm{Beta}(0, 1, 0.5, 0.5) = \mathrm{Beta}(0.5, 0.5)$, the Jeffreys prior (Jeffreys 1946), is considered where the two extremes of very small and very large p have high probability.

Corresponding to the results of Table 2.1, Fig. 2.2 shows that the gain in length of the minimum volume confidence interval without prior information is visible, but not

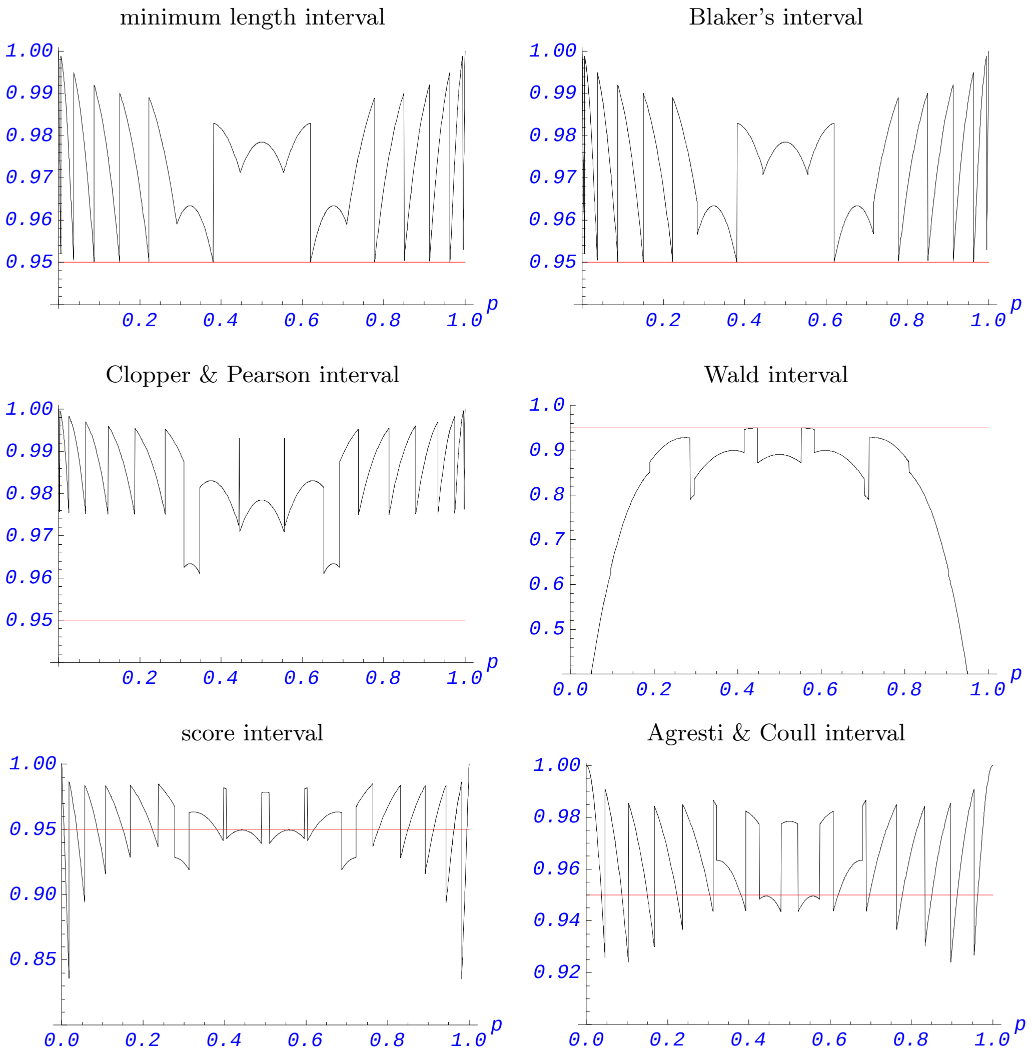


Figure 2.1: Actual coverage $\mathrm{P}_p(p \in A)$ as a function of $p \in [0;1]$ for sample size $n = 10$ under the nominal confidence level $\gamma = 0.95$.

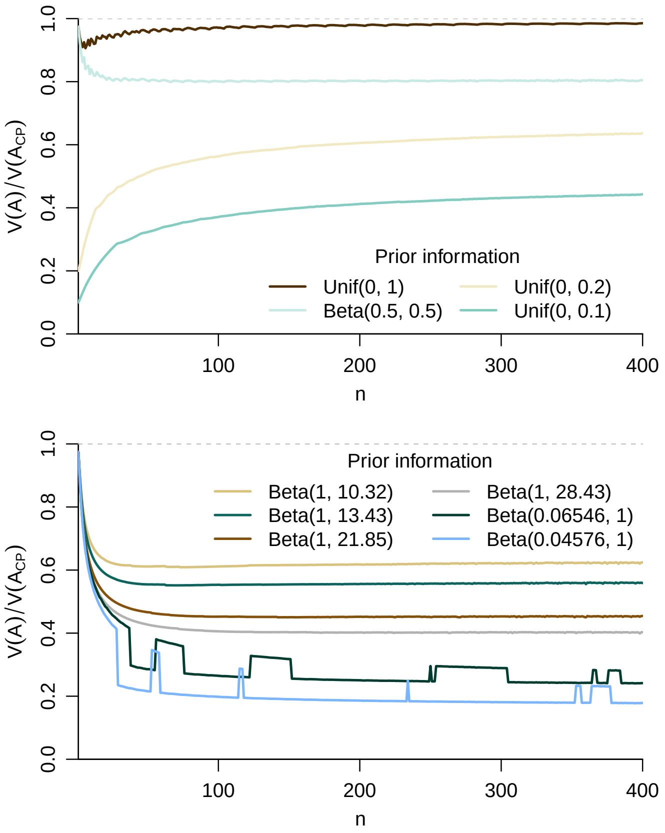


Figure 2.2: Average length of minimum volume prior information intervals relative to the average length of Clopper & Pearson intervals as a function of the sample size n for confidence level $\gamma = 0.95$. $\text{Beta}(1, 10.32)$ and $\text{Beta}(0.06546, 1)$ have 90 %-quantiles 0.2, $\text{Beta}(1, 13.43)$ has 95 %-quantile 0.2, $\text{Beta}(1, 21.85)$ and $\text{Beta}(0.04576, 1)$ have 90 %-quantiles 0.1, $\text{Beta}(1, 28.43)$ has 95 %-quantile 0.1.

striking. However, the length decreases considerably when imposing narrower prior information on Y. The second part of Fig. 2.2 compares the effect of distributions $\text{Beta}(1, b)$ and $\text{Beta}(a, 1)$ on the full support $[p_0; p_1] = [0; 1]$ with choices of b and a corresponding to specified quantiles of the beta distribution.

The modification of the quantile level from 90 % to 95 % with an unchanged quantile point at 0.1 or 0.2, respectively, has little effect on the interval length. However, the effect of changes in the quantile point from 0.2 down to 0.1 is considerable.

Table 2.2 shows the required minimum sample sizes if an upper limit of 0.1, 0.075 or 0.05 is imposed for the average length. If, for instance, an average length of at most 0.075 is intended and 0.2 can be specified as the 90 %-quantile, the sample size is about 39 % of the sample size required when using the Clopper & Pearson interval. If Y is assumed to be uniformly distributed between 0 and 0.1, the sample size can be reduced

Table 2.2: Required minimum sample sizes, $\gamma = 0.95, Y \sim \text{Beta}(p_0, p_1, a, b)$.

	Average length below		
Prior information on $p = y$	0.1	0.075	0.05
n.a.: Clopper & Pearson	254	445	983
$Y \sim \text{Unif}(0,1)$	245	432	964
$Y \sim \text{Beta}(0.5, 0.5)$	166	290	642
$Y \sim \text{Unif}(0, 0.2)$	79	163	407
$Y \sim \text{Beta}(1, 10.3)$, i. e. $0.2 = 90\,\%$ point	98	174	390
$Y \sim \text{Beta}(1, 13.4)$, i. e. $0.2 = 95\,\%$ point	81	142	316
$Y \sim \text{Beta}(1, 21.9)$, i. e. $0.1 = 90\,\%$ point	57	96	209
$Y \sim \text{Beta}(1, 28.4)$, i. e. $0.1 = 95\,\%$ point	49	79	167
$Y \sim \text{Unif}(0, 0.1)$	1	50	167

by about 89 %.

We present the coverage probability functions together with the corresponding MPSs for a selection of prior information for the minimum volume confidence interval. In Fig. 2.3, the MPS and coverage probability functions are shown under two uniform prior distributions $\text{Beta}(0,1,1,1) = \text{Unif}(0,1)$ and $\text{Beta}(0.1, 0.7, 1, 1) = \text{Unif}(0.1, 0.7)$, the $\text{Beta}(0.5, 0.5)$ distribution (bathtub shaped), the $\text{Beta}(0.2, 1)$ and the $\text{Beta}(1, 5)$ distribution (both monotonously decreasing), and the $\text{Beta}(7, 3)$ distribution (left-skewed). The $\text{Beta}(0.2, 1)$ distribution, for example, causes smaller upper bounds for some $x \in \{0, \dots, 10\}$ than the $\text{Unif}(0,1)$ prior information distribution. To nevertheless maintain the prescribed coverage of at least $\gamma = 0.95$, the confidence intervals are extended sufficiently far to the left for some $x \in \{0, \dots, 10\}$. A similar behaviour is observed for the $\text{Beta}(1, 5)$ distribution. The left-skewed $\text{Beta}(7, 3)$ prior information distribution shows the opposite behaviour. Lower bounds are increased, if possible, and the upper bounds increased to compensate.

While the coverage probability plots for the symmetric distributions $\text{Unif}(0,1)$ and $\text{Beta}(0.5, 0.5)$ show a symmetric behaviour, the coverage probability plots in the case of the asymmetric prior information distributions $\text{Beta}(0.2, 1)$, $\text{Beta}(1, 5)$ and $\text{Beta}(7, 3)$ turn out to be asymmetric. In all cases, however, the minimal coverage probability of at least γ is ensured.

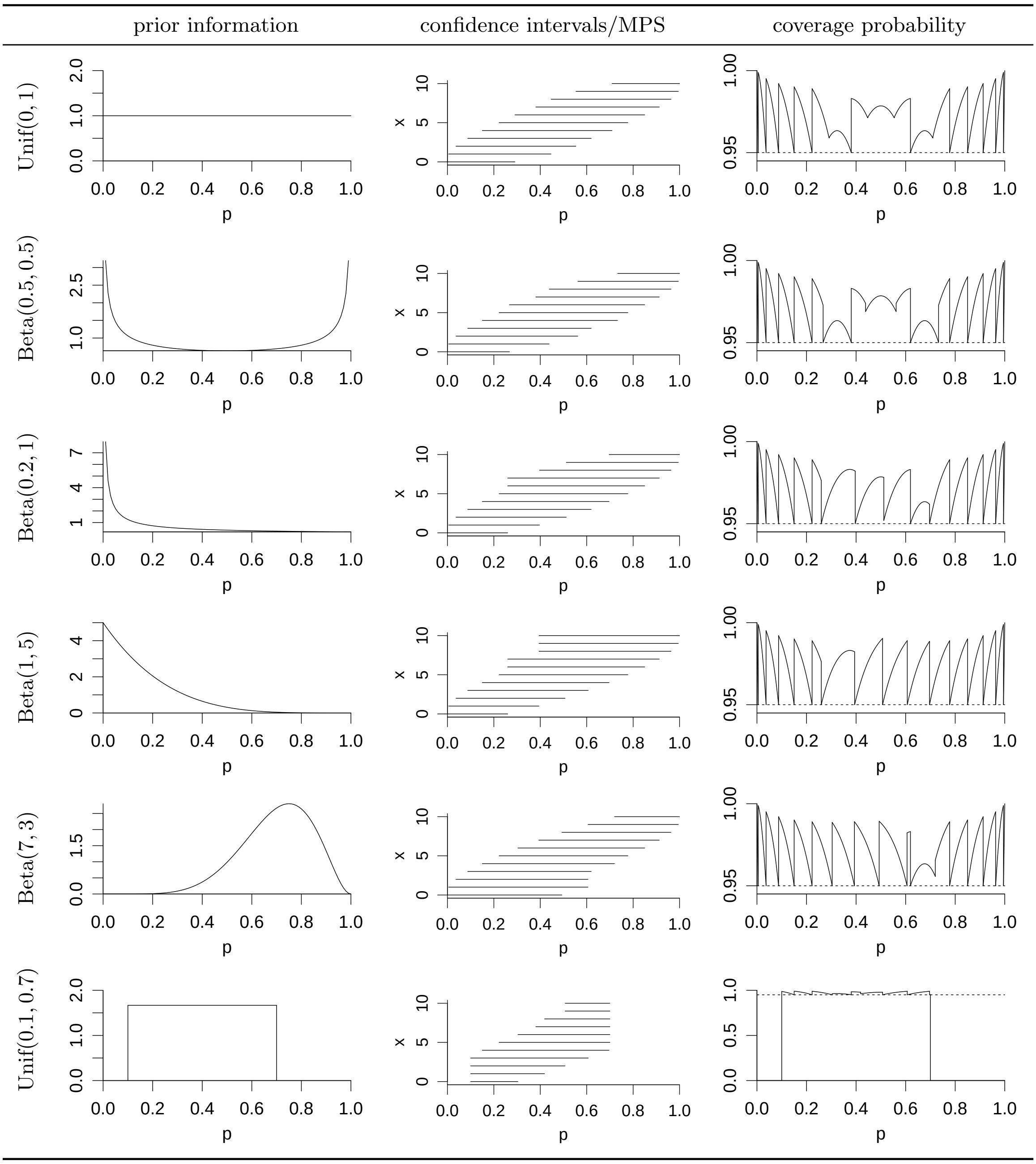


Figure 2.3: Level 95 % minimum volume MPS for a binomial proportion p and coverage probability for a selection of prior information. Sample size $n = 10$.

2.9 Indifference Probability

2.9.1 Definition and Monotonicity Characteristics

Two-sided confidence intervals are often used to support a decision making process on whether a true, unknown distribution parameter exceeds a certain critical threshold, falls below it or none of both. A confidence interval is usually preferred to another one if it is more likely to lead to a (correct) decision. The situation when it does not lead to a decision, is called *indifference*. We describe indifference in the context of confidence intervals for a binomial probability.

Let A be an MPS with nonempty projections A_y and $A_X = [y_L(X); y_U(X)] \subset [p_0; p_1] \subset [0; 1]$ for all $y \in [0; 1]$, $X \in \{0, \ldots, n\}$, where $n \in \mathbb{N}$ is the sample size. The confidence limits $y_L(x)$ and $y_U(x)$ are increasing in $x \in \{0, \ldots, n\}$, compare property P2) in Section 2.6. Let $\gamma \in (0; 1)$ be the confidence level and $q \in [p_0; p_1]$ be a critical threshold. Then a decision maker is indifferent with respect to the threshold q if $q \in A_x$ under an observed $x \in \{0, \ldots, n\}$. The *indifference probability* $I_q(y)$ as a function of y is given by

$$I_q(y) \quad = \quad \mathrm{P}_y\left(q \in A_X\right) \quad = \quad \mathrm{P}_y(y_L(X) \leq q \leq y_U(X)). \tag{2.27}$$

From Proposition 2.8 we can infer that the projection A_y is an interval $A_y = \{c_L(y), c_L(y)+1, \ldots, c_U(y)\}$ with prediction limits $c_L(y)$, $c_U(y)$ increasing in $y \in [p_0; p_1]$. In particular, the set $A_q = \{x \in \{0, \ldots, n\} | y_L(x) \leq q \leq y_U(x)\}$ with $q \in [p_0; p_1]$ is an interval. Consequently, we have

$$I_q(y) \quad = \quad \mathrm{P}_y\left(c_L(q) \leq X \leq c_U(q)\right) \quad = \quad L_{n,c_U(q)}(y) - L_{n,c_L(q)-1}(y), \tag{2.28}$$

compare Eqs. (2.24) and (2.2).

The following corollary is concerned with the monotonicity of the indifference probability function $I_q(y)$. It is a consequence of Proposition 2.10 with $x_1 = c_L(q)$ and $x_2 = c_U(q)$.

Corollary 2.12 (Monotonicity of Indifference Probability)**.** *Let A be an MPS with nonempty projections A_x and A_y for all $x \in \{0, \ldots, n\}$ and all $y \in [p_0; p_1]$. Let $q \in [p_0; p_1]$ be a critical value with prediction region $A_q = \{c_L(q), \ldots, c_U(q)\} \subset \{0, \ldots, n\}$. Let $I_q(y)$ be the indifference probability function for q under A. In the case $c_L(q) > 0$ let*

$$p_{c_L(q),c_U(q)} \quad := \quad \left[1 + \left(\frac{(n - c_U(q)) \cdot \ldots \cdot (n - c_L(q))}{c_L(q) \cdot \ldots \cdot c_U(q)}\right)^{\frac{1}{c_U(q)-c_L(q)+1}}\right]^{-1}.$$

Then we have:

a) *In the case $c_L(q) = 0$, $I_q(y)$ is strictly decreasing on $[p_0; p_1]$ if $c_U(q) < n$, and constant with $I_q(y) = 1$ if $c_U(q) = n$.*

b) *In the case $c_L(q) > 0$, $I_q(y)$ is strictly increasing on $[p_0; p_{c_L(q),c_U(q)}]$ and strictly decreasing on $[p_{c_L(q),c_U(q)}; p_1]$.*

Corollary 2.12 is concerned with the case that the critical value $q \in [0; 1]$ is an element of the interval $[p_0; p_1]$. The cases $p_0 > q$ and $p_1 < q$ signify that the unknown probability $y = p$, which is considered to be in the interval $[p_0; p_1]$ with probability 1, is below or above the critical threshold q a priori, which leads to a decision with probability 1 and hence to an indifference probability of 0.

From Section 2.10 it will become clear that in the case of an MPS for a binomial probability, there is a partition $0 = \pi_1 < \ldots < \pi_{s^*} = 1$ of $[0; 1]$, such that the prediction regions $A_y = \{c_L(y), \ldots, c_U(y)\}$ are constant for $y \in [\pi_i; \pi_{i+1})$, $i = 1, \ldots, s^* - 1$. In consequence, two critical values $c_1, c_2 \in [p_0; p_1]$ have equal indifference probability functions if $c_1, c_2 \in [\pi_i; \pi_{i+1})$.

2.9.2 Numerical Analysis of the Indifference Probability

We examine the indifference probability of the minimum volume confidence interval and the Clopper & Pearson interval for a probability under various prior information and sample sizes $n = 50, 100, 150$ for critical thresholds $q = 0.05$ and $q = 0.1$ and a confidence level of $\gamma = 0.95$. We consider the following types of prior information distributions: 1) the uniform distribution $\text{Beta}(0, 1, 1, 1) = \text{Unif}(0, 1)$, 2) the uniform distribution $\text{Beta}(0, 0.1, 1, 1) = \text{Unif}(0, 0.1)$, 3) the uniform distribution $\text{Beta}(0, 0.05, 1, 1) = \text{Unif}(0, 0.05)$, 4) the $\text{Beta}(1, 10.32)$ distribution, 5) the $\text{Beta}(1, 21.85)$ distribution, 6) the $\text{Beta}(1, 44.89)$ distribution, which has 0.05 as the 90 %-quantile, 7) the $\text{Beta}(0.06546, 1)$ distribution.

The indifference probability in dependence of the true probability parameter $p = y$ corresponding to the confidence intervals is illustrated in Fig. 2.4. Among the investigated prior information types, many show similar behaviour. For example, for $n = 100$ and $q = 0.1$, the indifference probability function is identical for $\text{Unif}(0, 1)$, $\text{Beta}(1, 10.32)$ and $\text{Beta}(0.06546, 1)$. The $\text{Unif}(0, 0.05)$ indifference probability plot is omitted in the plots corresponding to the critical threshold $q = 0.1$. That is because the upper bounds of the intervals obtained under this prior information are by definition below $q = 0.1$ and hence always lead to a conclusion.

In all plots, the indifference probability at the critical threshold q is at least the prescribed confidence level γ. That this is necessarily the case follows from the validity of Eq. (2.7) when setting $y = q$ in Eq. (2.28).

The indifference probability graphs that increase as late as possible to the maximum $\geq \gamma$ if p approaches the critical threshold q and quickly decrease afterwards are preferable. In the case $n = 50, q = 0.05$, all minimum volume confidence intervals are equivalent and better than the Clopper & Pearson interval with the exception of the one with prior information Unif$(0, 0.05)$, which does not lead to a decision, irrespective of which is the true probability parameter $p \in [0; 1]$. In the case $n = 100, q = 0.05$, the Clopper & Pearson method and the minimum volume confidence interval using the uniform prior on $[0; 1]$ increase early when p approaches q, but descend quickly afterwards. All other priors have preferable indifference probability functions if $p < q$, but are mostly worse if $p > q$. The remaining plots in Fig. 2.4 have to be interpreted in a similar fashion. In several cases, as $n = 50, q = 0.1$, the use of prior information positively influences the indifference probability function for $p < 0.1$ and the use of meaningful prior information can be useful if p is considered to be rather close to 0. In other cases, as $n = 150, q = 0.05$, the indifference probability does not show significant changes in behaviour among the investigated confidence intervals with the exception of the Unif$(0, 0.5)$ prior. A difference in behaviour, however, can sometimes be spotted if the true p is larger than the critical threshold. The descend in indifference probability is slower for Beta$(1, b)$ prior information distributions with b increasing. For example, Beta$(1, 44.89)$ decreases slower than Beta$(1, 21.85)$ in all investigated cases apart from $n = 50, q = 0.05$, in which case they have equal indifference probability functions.

Considering only the right-skewed prior information distributions Beta$(1, 10.32)$, Beta$(1, 21.85)$ and Beta$(1, 44.89)$, it seems unnecessary in the investigated examples to take into account the more extreme priors in terms of skewness like Beta$(1, 21.85)$ or Beta$(1, 44.89)$, for they show similar behaviour as Beta$(1, 10.32)$ for small true p and Beta$(1, 10.32)$ shows the preferable faster descend for $p > q$.

Figure 2.4 also reveals that from the point of view of the indifference probability, it does not make sense to use the uniform prior information distribution on $[0; 0.05]$ with the sharp cut-off if $q = 0.05$ is the critical threshold as well as the prior Unif$(0, 0.1)$ if $q = 0.1$ is the critical threshold because they lead to an indifference with probability 1 for a wide range of values $p \in [0; 1]$.

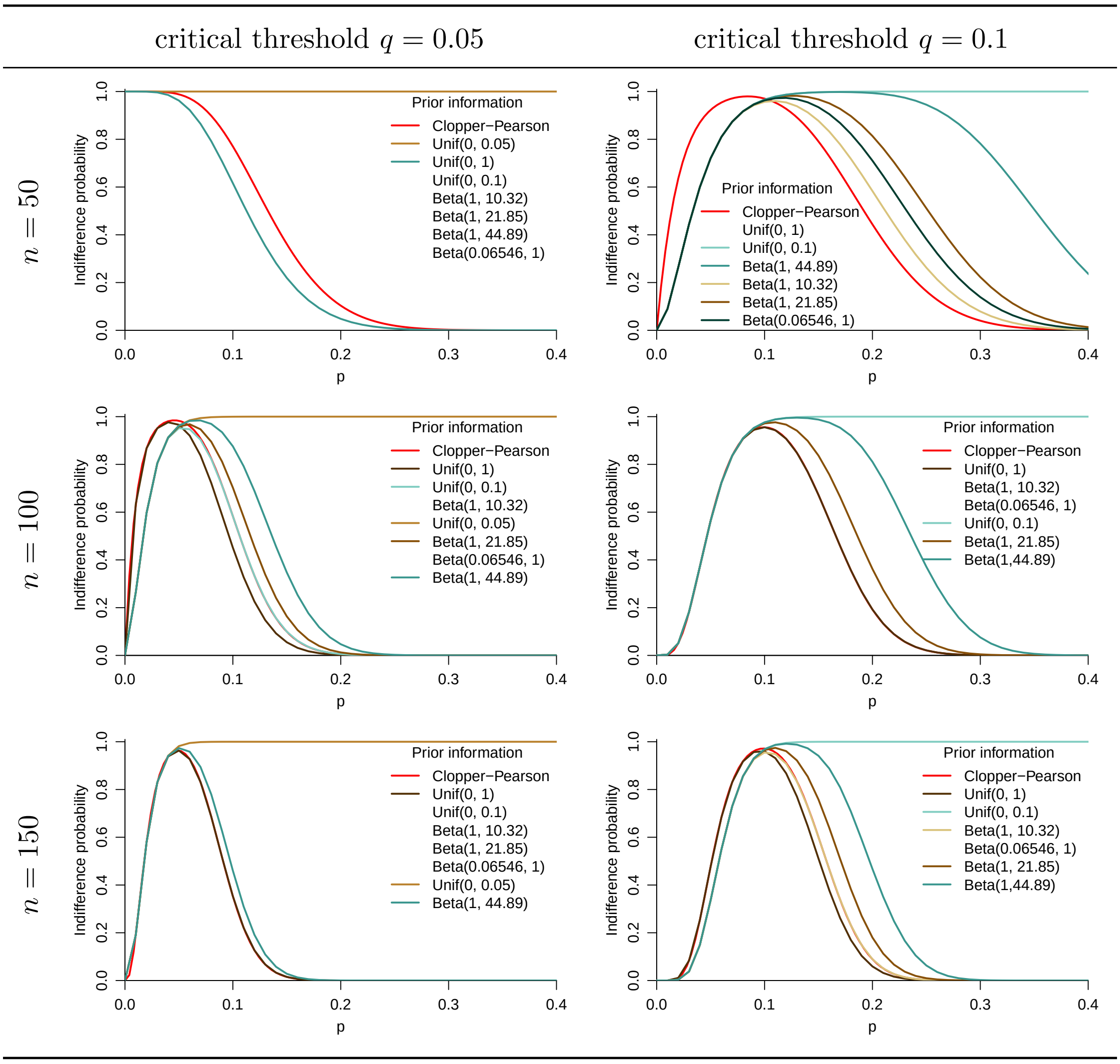


Figure 2.4: Indifference probability of the minimum volume confidence interval in dependence of the true probability $y = p$ under a selection of prior information for critical thresholds $q = 0.05, 0.1$ and sample sizes $n = 50, 100, 150$. Confidence level $\gamma = 95\,\%$.

2.10 Computational Algorithm

The minimum volume confidence intervals for a probability under prior information cannot be obtained by means of an explicit formula. Their computation requires a numerical algorithm, of which we describe the major steps in Section 2.10.1, and the detailed steps in Section 2.10.2. The difficulties associated with the computation and the necessity of applying a specific numerical algorithm as well as further remarks regarding the implementation can be found in Section 2.10.3.

2.10.1 Main Steps of the Algorithm

The algorithm to compute the minimum volume confidence intervals for a probability under prior information determines for a given sample size n, confidence level $\gamma \in (0;1)$ and prior information on $p = y$ the minimum volume confidence intervals for the unknown probability p. The confidence intervals are determined by making use of the relation between confidence intervals $[y_L(x); y_U(x)] \subset [0;1]$ and prediction intervals $\{c_L(y), \ldots, c_U(y)\} \subset \{0, \ldots, n\}$. The objective of the algorithm is to find an appropriate partition $0 = \pi_1 < \ldots < \pi_{s^*} = 1$ of $[0;1]$ of prediction regions A_y that are constant for $y \in [\pi_i; \pi_{i+1}), i = 1, \ldots, s^* - 1$.

To obtain the numerical results of Sections 2.7 to 2.9, the algorithm has been implemented by the author of this thesis in the statistical software R.

Algorithm 1 (Computation of Minimum Volume Confidence Intervals)**.**

Step I: *For $x = 0, \ldots, n$, determine the minimum intervals $A_x^{(0)} = [p_{x,\gamma}; \widetilde{p}_{x,\gamma}] \subset [0;1]$ according to Proposition 2.9.*

Step II: *For $y \in [0;1]$, determine the minimum prediction regions $A_y^{(0)}$ belonging to the minimum intervals $A_x^{(0)}$, $x = 0, \ldots, n$. There is a partition $0 = \pi_1 < \ldots < \pi_s = 1$ of $[0;1]$ such that $A_y^{(0)}$ is invariant for $y \in [\pi_i; \pi_{i+1}), i = 1, \ldots, s-1$.*

Step III: *Expand any minimum prediction region $A_y^{(0)}$ according to the criterion of highest prediction likelihood ratio with the help of Propositions 2.10 and 2.11. Stop the procedure as soon as the resulting prediction regions $A_y^{(1)}$ fulfil the condition $P_y(X \in A_y^{(1)}) \geq \gamma$ for $y \in [0;1]$.*

Step IV: *Reduce the prediction regions $A_y^{(1)}$ to the optimal prediction regions $A_y^* \subset A_y^{(1)}$ while ensuring the condition $P_y(X \in A_y^*) \geq \gamma$ for $y \in [0;1]$. If possible, drop predictions with lowest prediction likelihood ratio first. Again, Propositions 2.10 and 2.11 are used.*

Step V: *Determine the minimum volume level* γ *confidence intervals* $A_x^* = [y_L(x)^*; y_U(x)^*]$ *by*

$$\begin{aligned} y_L(x)^* &:= \min\{\max\{\min\{y|x \in A_y^*\} \cup p_0\} \cup p_1\}, \\ y_U(x)^* &:= \max\{\min\{\max\{y|x \in A_y^*\} \cup p_1\} \cup p_0\}. \end{aligned}$$

2.10.2 Detailed Description of the Algorithm

In the following we give a more detailed description of the algorithm used for the computation of the minimum volume confidence intervals for a probability. Let $n \in \{1, 2, 3, \ldots\}$, $0 < \gamma < 1$.

Algorithm 2 (Computation of Minimum Volume Confidence Intervals in Detail)**.**

Step I: Minimum confidence intervals *By Proposition 2.9 and the well-known relation* $L_{n,x}(p) = F_{n-x,x+1}(1-p)$ *between the binomial OC* $L_{n,c}(p)$ *and the distribution function* $F_{n-x,x+1}(1-p)$ *of the beta distribution* $Beta(n-x, x+1)$*, determine the minimum confidence intervals* $[p_{x,\gamma}; \widetilde{p}_{x,\gamma}]$*, where*

$$p_{x,\gamma} = \begin{cases} 0.0 & \text{for } x = 0, \\ 1 - z_{Beta(n-x+1,x)}(\gamma) & \text{for } x = 1, \ldots, n, \text{ and} \end{cases}$$

$$\widetilde{p}_{x,\gamma} = \begin{cases} 1 - z_{Beta(n-x,x+1)}(1-\gamma) & \text{for } x = 0, \ldots, n-1, \\ 1.0 & \text{for } x = n. \end{cases}$$

Here, $z_{Beta(a,b)}(\alpha)$ *is the* $100\alpha\,\%$*-quantile of the beta distribution* $Beta(a, b)$ *on* $[0; 1]$*.*

Step II: Minimum prediction intervals *Let* $H = \{p_{x,\gamma}|x = 0, \ldots, n\} \cup \{\widetilde{p}_{x,\gamma}|x = 0, \ldots, n\}$, $|H| =: s$*, and let* $0 = \pi_1 < \ldots < \pi_s = 1$ *be an ordering of the elements of* H*. For* $y \in [\pi_i; \pi_{i+1})$*,* $i = 1, \ldots, s-1$*, let*

$$\begin{aligned} c_L^{(0)}(y) &:= \max\Big\{\{x \in \{1, \ldots, n\}|\widetilde{p}_{x-1,\gamma} \le \pi_i\} \cup \{0\}\Big\}, \\ c_U^{(0)}(y) &:= \max\Big\{x \in \{0, \ldots, n\}|p_{x,\gamma} \le \pi_i\Big\} \end{aligned}$$

and obtain $A_y^{(0)} := \Big\{c_L^{(0)}(y), \ldots, c_U^{(0)}(y)\Big\}$*. The sets* $A_y^{(0)}$ *are invariant for* $y \in [\pi_i; \pi_{i+1}), i = 1, \ldots, s-1$*.*

Step III: Expansion of the prediction regions

Step III.a) *Consider the function* $\Delta_{n,x_1,x_2}(y)$ *from Proposition 2.10. Let* $0 = \pi_1 < \ldots < \pi_s = 1$ *be an ordering of the elements of* H *with invariant prediction* *regions* $A_y^{(0)} = \left\{c_L^{(0)}(y),\ldots,c_U^{(0)}(y)\right\}$ *on* $[\pi_i;\pi_{i+1}) \subset [0;1], i = 1,\ldots,s-1$. *If the condition* $\Delta_{n,c_L^{(0)}(y),c_U^{(0)}(y)}(y) \geq \gamma$ *is fulfiled for* $y \in [\pi_i;\pi_{i+1})$*, define* $A_y^{(1)} := \left\{c_L^{(0)}(y),\ldots,c_U^{(0)}(y)\right\}$*. Make use of Proposition 2.10.*

Step III.b) *If the condition* $\Delta_{n,c_L^{(0)}(y),c_U^{(0)}(y)}(y) \geq \gamma$ *is fulfiled for* $y \in [0;1]$*, i. e.* $A_y^{(1)}$ *found for all* $y \in [0;1]$*, stop the procedure and go to step IV.*

Step III.c) *For every* $i = 1,\ldots,s-1$*, do the following: If* $\Delta_{n,c_L^{(0)}(y),c_U^{(0)}(y)}(y) < \gamma$ *for at least one* y *from an interval* $[\pi_i;\pi_{i+1})$*, determine a decomposition* $[\pi_{i_0};\pi_{i_1}),\ldots,[\pi_{i_{m-1}},\pi_{i_m})$ *of* $[\pi_i;\pi_{i+1})$ *with* $[\pi_{i_0};\pi_{i_1}) \cup \ldots \cup [\pi_{i_{m-1}},\pi_{i_m}) = [\pi_i;\pi_{i+1})$*, where* $[\pi_{i_0};\pi_{i_1}),\ldots,[\pi_{i_{m-1}},\pi_{i_m})$ *are pairwise disjunct, such that for all* $y \in [\pi_{i_{k-1}};\pi_{i_k})$ *either* $\Delta_{n,c_L^{(0)}(y),c_U^{(0)}(y)}(y) \geq \gamma$ *or* $\Delta_{n,c_L^{(0)}(y),c_U^{(0)}(y)}(y) \leq \gamma$ *is fulfiled: Calculate the maximum point* $p_{c_L^{(0)}(y),c_U^{(0)}(y)}$ *of the prediction region coverage function from Proposition 2.10. If* $p_{c_L^{(0)}(y),c_U^{(0)}(y)} \leq \gamma$*, define* $[\pi_{i_0};\pi_{i_1}) := [\pi_i;\pi_{i+1})$*, i. e.* $m = 1$*. Otherwise calculate the roots of* $\Delta_{n,c_L^{(0)}(y),c_U^{(0)}(y)}(y) - \gamma$*. According to Proposition 2.10,* $\Delta_{n,c_L^{(0)}(y),c_U^{(0)}(y)}(y) - \gamma$ *has at most two roots* $r_1 < r_2$*. If* $r_1, r_2 \notin [\pi_i;\pi_{i+1})$*, define* $[\pi_{i_0};\pi_{i_1}) = [\pi_i;\pi_{i+1})$*, i. e.* $m = 1$*. Otherwise if* $r_1, r_2 \in [\pi_i;\pi_{i+1})$*, define the decomposition* $\underbrace{[\pi_i}_{\pi_{i_0}};\underbrace{r_1}_{\pi_{i_1}}) \cup \underbrace{[r_1}_{\pi_{i_1}}\underbrace{r_2}_{\pi_{i_2}}) \cup \underbrace{[r_2}_{\pi_{i_2}},\underbrace{\pi_{i+1}}_{\pi_{i_3}}) = [\pi_i;\pi_{i+1})$*, i. e.* $m = 3$*. Otherwise if* $r_1 \in [\pi_i;\pi_{i+1}), r_2 \notin [\pi_i;\pi_{i+1})$ *or* $r_2 \in [\pi_i;\pi_{i+1}), r_1 \notin [\pi_i;\pi_{i+1})$*, define* $\underbrace{[\pi_i}_{\pi_{i_0}};\underbrace{r_1}_{\pi_{i_1}}) \cup [\pi_{i_1};\pi_{i_2}) = [\pi_i;\pi_{i+1})$*, i. e.* $m = 2$*.*

Redefine $H := H \cup \{\pi_{i_1},\ldots,\pi_{i_{m-1}}\}$ *and* $s := |H|$*.*

Step III.d) *Let* $0 = \pi_1 < \ldots < \pi_s = 1$ *be an ordering of the elements of* H*. If for an interval* $[\pi_i;\pi_{i+1})$ *the condition* $\Delta_{n,c_L^{(0)}(y),c_U^{(0)}(y)}(y) \leq \gamma$ *is fulfiled, add a prediction point to the prediction region* $A_y^{(0)} = \left\{c_L^{(0)}(y),\ldots,c_U^{(0)}(y)\right\}$*. In the case* $c_L^{(0)}(y) = 0$*, add* $c_U^{(0)}(y)+1$*, i. e. redefine* $A_y^{(0)} := \left\{c_L^{(0)}(y),\ldots,c_U^{(0)}(y)+1\right\}$*. In the case* $c_U^{(0)}(y) = n$*, add* $c_L^{(0)}(y) - 1$*, i. e. redefine* $A_y^{(0)} := \left\{c_L^{(0)}(y)-1,\ldots,c_U^{(0)}(y)\right\}$*. In the case* $0 < c_L^{(0)}(y) < c_U^{(0)}(y) < n$ *apply the principle of the greatest prediction likelihood ratio* $Q_{p_0,p_1,a,b,y}(x)$*, see Eq. (2.17), to the expansion of prediction regions: If* $Q_{p_0,p_1,a,b,y}\left(c_L^{(0)}(y)-1\right) \geq Q_{p_0,p_1,a,b,y}\left(c_U^{(0)}(y)+1\right)$ *for all* $y \in [\pi_i;\pi_{i+1})$*, add* $c_L^{(0)}(y)-1$*, i. e. redefine* $A_y^{(0)} := \left\{c_L^{(0)}(y)-1,\ldots,c_U^{(0)}(y)\right\}$*. If* $Q_{p_0,p_1,a,b,y}\left(c_L^{(0)}(y)-1\right) \leq Q_{p_0,p_1,a,b,y}\left(c_U^{(0)}(y)+1\right)$ *for all* $y \in [\pi_i;\pi_{i+1})$*, add* $c_U^{(0)}(y)+1$*, i. e. redefine* $A_y^{(0)} := \left\{c_L^{(0)}(y),\ldots,c_U^{(0)}(y)+1\right\}$*. Otherwise calculate the section point* $q_{c_L^{(0)}(y)-1,c_U^{(0)}+1}$ *of the two prediction likelihood ratio* *functions for* $c_L^{(0)}(y)-1$ *and* $c_U^{(0)}(y)+1$ *with Proposition 2.11. If*

$$Q_{p_0,p_1,a,b,y}\left(c_L^{(0)}(y)-1\right)\leq Q_{p_0,p_1,a,b,y}\left(c_U^{(0)}(y)+1\right)$$

for all $y\in\left[\pi_i;q_{c_L^{(0)}(y)-1,c_U^{(0)}(y)+1}\right)$, *add* $c_U^{(0)}(y)+1$, *i. e. redefine* $A_y^{(0)}:=\left\{c_L^{(0)}(y),\ldots,\ c_U^{(0)}(y)+1\right\}$ *for* $y\in\left[\pi_i;q_{c_L^{(0)}(y)-1,c_U^{(0)}(y)+1}\right)$. *Else if*

$$Q_{p_0,p_1,a,b,y}\left(c_L^{(0)}(y)-1\right)\geq Q_{p_0,p_1,a,b,y}\left(c_U^{(0)}(y)+1\right)$$

for all $y\in[\pi_i;q_{c_L^{(0)}(y)-1,c_U^{(0)}(y)+1})$, *add* $c_L^{(0)}(y)-1$, *i. e. redefine* $A_y^{(0)}:=\left\{c_L^{(0)}(y)-1,\ldots,c_U^{(0)}(y)\right\}$ *for* $y\in[\pi_i;q_{c_L^{(0)}(y)-1,c_U^{(0)}(y)+1})$. *Proceed with* $\left[q_{c_L^{(0)}(y)-1,c_U^{(0)}(y)+1};\pi_{i+1}\right)$ *in a similar fashion.*

Redefine $H:=H\cup\left\{q_{c_L^{(0)}(y)-1,c_U^{(0)}(y)+1}\right\}$ *and* $s:=|H|$.

Return to step III.a).

Step IV: Reduction of the prediction regions

Step IV.a) *Let* $0=\pi_1<\ldots<\pi_s=1$ *be an ordering of the elements of* H. *Let* $A_y^{(1)}=\left\{c_L^{(1)}(y),\ldots,c_U^{(1)}(y)\right\}$ *be the prediction regions, which are invariant for* $y\in[\pi_i;\pi_{i+1})\subset[0;1],i=1,\ldots,s-1$, *and for which* $\Delta_{n,c_L^{(1)}(y),c_U^{(1)}(y)}(y)\geq\gamma$ *is fulfiled.*

If for all $y\in[\pi_i;\pi_{i+1})$ *we have* $c_U^{(1)}(y)=0$ *or* $c_L^{(1)}(y)=n$ *or* $c_L^{(1)}(y)=c_U^{(1)}(y)$, *no reduction of the prediction region is possible. Set* $A_y^*:=A_y^{(1)}$.

Step IV.b) *For every* $i=1,\ldots,s-1$, *do the following: If* $\Delta_{n,c_L^{(1)}(y)+1,c_U^{(1)}(y)}(y)>\gamma$ *for at least one* y *from* $[\pi_i;\pi_{i+1})$, *determine a decomposition* $[\pi_{i_0};\pi_{i_1}),\ldots,[\pi_{i_{m-1}},\pi_{i_m})$ *of* $[\pi_i;\pi_{i+1})$ *with* $[\pi_{i_0};\pi_{i_1})\cup\ldots\cup[\pi_{i_{m-1}},\pi_{i_m})=[\pi_i;\pi_{i+1})$, *where* $[\pi_{i_0};\pi_{i_1}),\ldots,[\pi_{i_{m-1}},\pi_{i_m})$ *are pairwise disjunct, such that for all* $y\in[\pi_{i_{k-1}};\pi_{i_k})$ *either* $\Delta_{n,c_L^{(0)}(y)+1,c_U^{(0)}(y)}(y)\geq\gamma$ *or* $\Delta_{n,c_L^{(0)}(y)+1,c_U^{(0)}(y)}(y)\leq\gamma$ *is fulfiled: Calculate the maximum point* $p_{c_L^{(0)}(y)+1,c_U^{(0)}(y)}$ *of the prediction region coverage function from Proposition 2.10. If* $p_{c_L^{(0)}(y)+1,c_U^{(0)}(y)}\leq\gamma$, *define* $[\pi_{i_0};\pi_{i_1}):=[\pi_i;\pi_{i+1})$, *i. e.* $m=1$. *Otherwise calculate the roots of* $\Delta_{n,c_L^{(0)}(y)+1,c_U^{(0)}(y)}(y)-\gamma$. *According to Proposition 2.10,* $\Delta_{n,c_L^{(0)}(y)+1,c_U^{(0)}(y)}(y)-\gamma$ *has at most two roots* $r_1<r_2$. *If* $r_1,r_2\notin[\pi_i;\pi_{i+1})$, *define* $[\pi_{i_0};\pi_{i_1}):=[\pi_i;\pi_{i+1})$, *i. e.* $m=1$. *Else if* $r_1,r_2\in[\pi_i;\pi_{i+1})$, *define the decomposition* $\big[\underbrace{\pi_i}_{\pi_{i_0}};\underbrace{r_1}_{\pi_{i_1}}\big)\cup\big[\underbrace{r_1}_{\pi_{i_1}}\ \underbrace{r_2}_{\pi_{i_2}}\big)\cup\big[\underbrace{r_2}_{\pi_{i_2}},\underbrace{\pi_{i+1}}_{\pi_{i_3}}\big)=[\pi_i;\pi_{i+1})$, *i. e.* $m=3$. *Else if* $r_1\in[\pi_i;\pi_{i+1}),r_2\notin[\pi_i;\pi_{i+1})$ *or*

$r_2 \in [\pi_i;\pi_{i+1}), r_1 \notin [\pi_i;\pi_{i+1})$, define $[\underbrace{\pi_i}_{\pi_{i_0}};\underbrace{r_1}_{\pi_{i_1}}) \cup [\pi_{i_1};\pi_{i_2}) = [\pi_i;\pi_{i+1})$, i. e. $m = 2$.

Redefine $H := H \cup \{\pi_{i_1},\dots,\pi_{i_{m-1}}\}$ and $s := |H|$.

Step IV.c) *Repeat step IV.b) applied to $\Delta_{n,c_L^{(1)}(y),c_U^{(1)}(y)-1}(y)$ instead of $\Delta_{n,c_L^{(1)}(y)+1,c_U^{(1)}(y)}(y)$.*

Step IV.c) *If for all $y \in [\pi_i;\pi_{i+1})$, $i = 1,\dots,s-1$, we have $\Delta_{n,c_L^{(1)}(y)+1,c_U^{(1)}(y)}(y) \le \gamma$ and $\Delta_{n,c_L^{(1)}(y),c_U^{(1)}(y)-1}(y) \le \gamma$, define $A_y^* := A_y^{(1)}$.*

Step IV.d) *If the conditions $\Delta_{n,c_L^{(1)}(y)+1,c_U^{(1)}(y)}(y) \le \gamma$ and $\Delta_{n,c_L^{(1)}(y),c_U^{(1)}(y)-1}(y) \le \gamma$ are fulfiled for every $y \in [0;1]$, i. e. A_y^* found for all $y \in [0;1]$, stop the procedure and go to step V.*

Step IV.e) *If for all $y \in [\pi_i;\pi_{i+1})$, $i = 1,\dots,s-1$, we have $\Delta_{n,c_L^{(1)}(y)+1,c_U^{(1)}(y)}(y) \ge \gamma$ and $\Delta_{n,c_L^{(1)}(y),c_U^{(1)}(y)-1}(y) \le \gamma$, define $A_y^{(1)} := \left\{c_L^{(1)}(y)+1,\dots,c_U^{(1)}(y)\right\}$.*

Step IV.f) *If for all $y \in [\pi_i;\pi_{i+1})$, $i = 1,\dots,s-1$, we have $\Delta_{n,c_L^{(1)}(y),c_U^{(1)}(y)-1}(y) \ge \gamma$ and $\Delta_{n,c_L^{(1)}(y)+1,c_U^{(1)}(y)}(y) \le \gamma$, define $A_y^{(1)} = \left\{c_L^{(1)}(y),\dots,c_U^{(1)}(y)-1\right\}$.*

Step IV.g) *If for all $y \in [\pi_i;\pi_{i+1})$ we have $\Delta_{n,c_L^{(1)}(y)+1,c_U^{(1)}(y)}(y) \ge \gamma$ and $\Delta_{n,c_L^{(1)}(y),c_U^{(1)}(y)-1}(y) \ge \gamma$, apply the principle of greatest prediction likelihood ratio $Q_{p_0,p_1,a,b,y}(x)$, see Eq. (2.17), to the reduction procedure. If*

$$Q_{p_0,p_1,a,b,y}(c_L^{(1)}(y)+1) \ge Q_{p_0,p_1,a,b,y}(c_U^{(1)}(y)-1)$$

for all $y \in [\pi_i;\pi_{i+1})$, define $A_y^{(1)} := \left\{c_L^{(1)}(y),\dots,c_U^{(1)}(y)-1\right\}$. If

$$Q_{p_0,p_1,a,b,y}(c_L^{(1)}(y)+1) \le Q_{p_0,p_1,a,b,y}(c_U^{(1)}(y)-1)$$

for all $y \in [\pi_i;\pi_{i+1})$, define $A_y^{(1)} := \left\{c_L^{(1)}(y)+1,\dots,c_U^{(1)}(y)\right\}$.

Otherwise calculate the section point $q_{c_L^{(1)}(y),c_U^{(1)}(y)}$ of the two prediction likelihood ratio functions for $c_L^{(1)}(y)$ and $c_U^{(1)}(y)$ with Proposition 2.11.

Redefine $H := H \cup \left\{q_{c_L^{(1)}(y),c_U^{(1)}(y)}\right\}$ and $s := |H|$.

Return to step IV.a).

Step V: Minimum volume confidence intervals

Let $0 = \pi_1 < \ldots < \pi_s = 1$ be an ordering of the elements of H. Let $A_y^* = \{c_L^*(y),\dots,c_U^*(y)\}$ be the optimal prediction regions, which are invariant for $y \in [\pi_i;\pi_{i+1}) \subset [0;1], i = 1,\dots,s-1$, and fulfil $\Delta_{n,c_L^*(y),c_U^*(y)} \ge \gamma$. The minimum volume level γ confidence intervals for a given realisation $x \in \{0,\dots,n\}$ is given *by the closed interval* $[y_L(x)^*; y_U(x)^*]$, *where*

$$\begin{aligned} y_L(x)^* &= \min\left\{\max\left\{\min_{i\in\{1,\ldots,s-1\}}\left\{\pi_i \;\middle|\; c_L^*(\pi_i) \le x \le c_U^*(\pi_i)\right\} \cup p_0\right\} \cup p_1\right\}, \\ y_U(x)^* &= \max\left\{\min\left\{\max_{i\in\{1,\ldots,s-1\}}\left\{\pi_{i+1} \;\middle|\; c_L^*(\pi_i) \le x \le c_U^*(\pi_i)\right\} \cup p_1\right\} \cup p_0\right\}. \end{aligned}$$

Mind that the algorithm is finite due to the fact that a prediction region A_y for $y \in [0;1]$ contains at most $n+1$ prediction points.

2.10.3 Computational Challenges

The necessity of following an algorithm as the one presented in the previous Sections 2.10.1 and 2.10.2 arises from difficulties related to the discreteness of the considered problem. Against intuition, the regions $A_y^{(1)} \subset \{0,\ldots,n\}$ of largest prediction likelihood ratio (see step III in Algorithm 1) are not always increasing in $p = y \in [0;1]$. Despite a success probability p_2 larger than p_1, i. e. $p_2 > p_1$, it can happen that $x_2 = \max A_{p_2}^{(1)} < \max A_{p_1}^{(1)} = x_1$.

Consider, for example, the prior information Beta$(0.15, 0.5)$ and the success probabilities $p_1 = 0.115$ and $p_2 = 0.15$ under the confidence level $\gamma = 0.95$ and sample size $n = 25$. The set of prediction points of maximal prediction likelihood ratio for $p_1 = 0.115$ ensuring a coverage probability of at least 95 % at $p_1 = 0.115$ is $\{1,\ldots,8\}$. For $p_2 = 0.15$, the region of largest prediction likelihood ratio is given by $\{1,\ldots,7\}$. Consequently, the upper bounds of the regions of largest prediction likelihood ratio are not increasing in $y = p$. In fact, a prediction region of $\{1,\ldots,7\}$ would not be sufficient for $p_1 = 0.115$: Consider the two cases $X_1 \sim Bi(25, p_1)$ and $X_2 \sim Bi(25, p_2)$. We have $P_{p_1}(X_1 \in \{1,\ldots,7\}) = 0.947 < \gamma$ and $P_{p_1}(X_1 \in \{1,\ldots,8\}) = 0.952 \ge \gamma$, hence 8 is indispensible in the prediction region, whereas $P_{p_2}(X_2 \in \{1,\ldots,7\}) = 0.957 \ge \gamma$, i. e. 8 is not needed in the prediction region here.

The example is illustrated by the top part of Fig. 2.5, which shows the upper and lower bounds of the regions of maximal prediction likelihood ratio. If confidence regions were constructed from these prediction regions, the confidence region under $X = 8$ would result to $(0.1138; 0.1176) \cup (0.1613; 0.5154)$ and consequently not be an interval. If the additional condition of increasing prediction bounds in $y = p$ is imposed, the bottom part of Fig. 2.5 is obtained. The confidence region under $X = 8$ then changes to the interval $(0.1138; 0.5154)$, compare Proposition 2.3 for the theoretical justification of this behaviour.

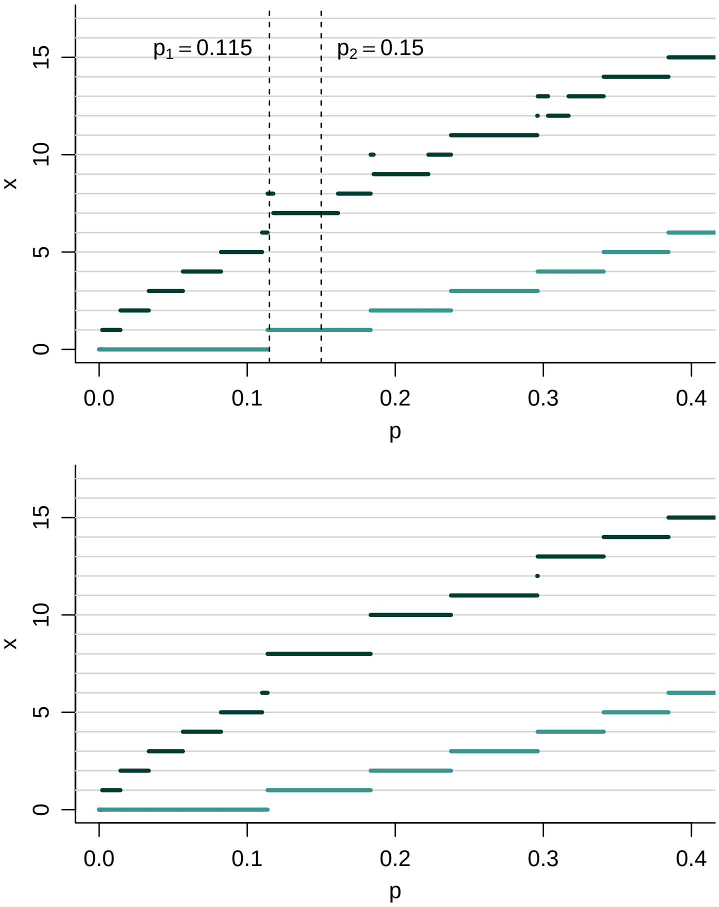


Figure 2.5: Bounds of the prediction likelihood ratio maximising prediction regions (top); bounds of the prediction likelihood ratio maximising prediction regions under additional monotonicity condition (bottom). Sample size $n = 25$; prior information Beta$(0.15, 0.5)$; confidence level $\gamma = 0.95$.

Due to the a priori lacking monotonicity of the bounds of regions of largest prediction likelihood ratio, numerical algorithms relying on monotonicity, as e. g. the bisection (interval halving) method, cannot be applied for the calculation of the minimum volume confidence intervals.

Consider another possibility of calculating the prediction intervals as follows: Let $\Delta = 1/m$ be a prescribed granularity deviding the interval $[0; 1]$ in m equally sized intervals. To calculate the prediction regions at the sequence of points $0, \Delta, 2\Delta, 3\Delta, \ldots, (m-1)\Delta, 1$ in $[0; 1]$ and to derive the confidence intervals from the resulting $(m + 1)$-element set of prediction regions is not only very inefficient. It furthermore produces results with an unsatisfying precision. Since with the above $m + 1$ points only a finite number of $p = y \in [0; 1]$ is taken into account, discontinuities in the prediction regions are in danger of being missed. In the above example illustrated in Fig. 2.5, if the prediction regions had been calculated with a granularity of $\Delta = 0.01$, the prediction region $\{1, \ldots, 7\}$ would have been obtained for $p = 0.11, 0.12, \ldots, 0.16$. Instead of at around $0.1128 \approx 0.11$, the lower bound of the confidence interval would have been ≥ 0.16 because it would have been ignored that for p between 0.1138 and 0.1176 the prediction point 8 is contained in the prediction region. In consequence, these discontinuities and jumps in the prediction regions – although seemingly harmless – can significantly alter the result and should consequently not be missed due to unsatisfactory precision.

Remark 2.13 (Implementation in R)**.** *In R, the root of a function, as requested in steps III.c) and IV.b) of Algorithm 2, can be found with the help of the function* `uniroot` *from the* `stats`*-package, which is contained in the standard installation of R (R Core Team 2014). The function searches between two points that are known to have function values of opposite sign for the point where the function value equals zero.*

The calculation of the prediction likelihood ratio $Q_{p_0,p_1,a,b,y}(x)$, Eq. (2.17), requires numerical integration in the calculation of the relative volume weights from Eq. (2.18). We have made good experiences with the function `distrExIntegrate` *from the* `distrEx`*-package (Ruckdeschel et al. 2006).*

2.11 Conclusion and Outlook

We have suggested a general scheme for minimum volume confidence regions for a distribution parameter under prior information, and we have applied this scheme to obtain shortest two-sided confidence intervals for the probability parameter p of a binomial distribution. Prior information on p has been specified by a beta distribution. The numerical algorithm developed, which has also been presented in this chapter, manages to produce the shortest confidence intervals in real time. Appropriate prior information has a demonstrable effect on the average length and on the sample size required to achieve a prescribed precision. The intervals are competitive in terms of coverage probability. They are less conservative than the Clopper & Pearson exact confidence intervals, but in contrast to several approximative confidence intervals always exact. The indifference probability for decisions made on the basis of the minimum volume interval has been explored and a more favourable behaviour could be observed under stricter prior information distributions.

Several aspects remain to be studied in more detail: i) The sensitivity of the average interval length with respect to changes in prior information: The numerical results presented in Section 2.8 indicate that minor changes have minor effects only, and that in this sense the method is robust against misspecified prior information. However, this impression has to be substantiated by more extensive numerical studies. ii) Asymptotics: Figure 2.2 indicates stability for large n, the underlying asymptotics remain to be studied analytically. iii) Blaker's interval seems to be very close to the minimum volume interval if considered without prior information. Ways of including prior information into Blaker's approach remain to be studied.

2.A Appendix

2.A.1 Proof of Theorem 2.2

The proof of assertion a) of Theorem 2.2 is obvious.

Proof of Assertion b) of Theorem 2.2. Let A be an arbitrary level γ MPS. Let $y \in R_2$ be fixed. We have

$$\begin{aligned} \mathrm{P}_y(X \in A_y \cap A_y^\star) + \mathrm{P}_y(X \in A_y \setminus A_y^\star) &= \mathrm{P}_y(X \in A_y) \geq \gamma \\ &\geq \mathrm{P}_y(X \in A_y^\star) = \mathrm{P}_y(X \in A_y \cap A_y^\star) + \mathrm{P}_y(X \in A_y^\star \setminus A_y), \end{aligned}$$

and hence $\mathrm{P}_y(X \in A_y^\star \backslash A_y) \leq \mathrm{P}_y(X \in A_y \backslash A_y^\star)$. We have $A_y^\star \backslash A_y \subset D_{\geq s_y}(y)$ and $A_y \backslash A_y^\star \subset D_{\leq s_y}(y)$, thus $\inf_{x \in A_y^\star \backslash A_y} Q_y(x) \geq \sup_{x \in A_y \backslash A_y^\star} Q_y(x)$. If $\inf_{x \in A_y^\star \backslash A_y} Q_y(x) = 0$, then $s_y = 0$, hence $D_{>s_y}(y) = \{x | f_{X|Y=y}(x) > 0\}$, and thus $G_y(s_y) = \mathrm{P}_y(X \in D_{>s_y}(y)) = 1 > \gamma$ contradictory to the assumptions of Theorem 2.2. Hence $\inf_{x \in A_y^\star \backslash A_y} Q_y(x) > 0$. We obtain the inequalities

$$\begin{aligned} \inf_{x \in A_y^\star \backslash A_y} Q_y(x) \int_{A_y^\star \backslash A_y} f_X(x)\, \mathrm{d}\mu_1(x) &\leq \int_{A_y^\star \backslash A_y} Q_y(x) f_X(x)\, \mathrm{d}\mu_1(x) \\ &= \int_{A_y^\star \backslash A_y} f_{X|Y=y}(x)\, \mathrm{d}\mu_1(x) = \mathrm{P}_y(X \in A_y^\star \setminus A_y) \leq \mathrm{P}_y(X \in A_y \setminus A_y^\star) \\ &= \int_{A_y \backslash A_y^\star} f_{X|Y=y}(x)\, \mathrm{d}\mu_1(x) = \int_{A_y \backslash A_y^\star} Q_y(x) f_X(x)\, \mathrm{d}\mu_1(x) \\ &\leq \sup_{x \in A_y \backslash A_y^\star} Q_y(x) \int_{A_y \backslash A_y^\star} f_X(x)\, \mathrm{d}\mu_1(x) \\ &\leq \inf_{x \in A_y^\star \backslash A_y} Q_y(x) \int_{A_y \backslash A_y^\star} f_X(x)\, \mathrm{d}\mu_1(x). \end{aligned}$$

Hence $\int_{A_y^\star \backslash A_y} f_X(x)\, \mathrm{d}\mu_1(x) \leq \int_{A_y \backslash A_y^\star} f_X(x)\, \mathrm{d}\mu_1(x)$, and thus we have $\int_{A_y^\star} f_X(x)\, \mathrm{d}\mu_1(x) \leq \int_{A_y} f_X(x)\, \mathrm{d}\mu_1(x)$.

The latter inequality has been proven for arbitrary $y \in R_2$. Hence by formula (2.8) we obtain $V(A^\star) \leq V(A)$. This completes the proof of assertion b) of Theorem 2.2.

Proof of Assertion c) of Theorem 2.2. Consider $y \in R_2$ where $D_{=s_y}(y) = \bigcup_I B_i$ is a countable union of intervals B_i with disjoint interior. We can assume $G_y(s_y^-) - G_y(s_y) = \mathrm{P}_y(X \in D_{=s_y}(y)) > 0$. From properties of the Lebesgue-Borel integral it follows that for any $0 \leq \alpha_i \leq \int_{B_i} f_X \,\mathrm{d}\mu_1$ there is a subinterval $B_{i,\alpha_i} \subset B_i$ with $\int_{B_{i,\alpha_i}} f_X \,\mathrm{d}\mu_1 = \alpha_i$. Let $\rho = [\gamma - G_y(s_y)]/[G_y(s_y^-) - G_y(s_y)]$. For $i \in I$, let $\alpha_i = \rho \int_{B_i} f_X \,\mathrm{d}\mu_1$. Let $E(y) = \bigcup_I B_{i,\alpha_i}$. Then

$$\mathrm{P}_y(X \in E(y)) \quad = \quad \sum_I \int_{B_{i,\alpha_i}} f_X \,\mathrm{d}\mu_1 \quad = \quad \rho \mathrm{P}_y(X \in D_{=s_y}(y)) \quad = \quad \gamma - G_y(s_y),$$

hence $\mathrm{P}_y(X \in D_{>s_y}(y) \cup E(y)) = G_y(s_y) + \gamma - G_y(s_y) = \gamma$.

2.A.2 Proof of Proposition 2.3

We prove assertion a) of Proposition 2.3. The proof of assertion b) is completely analogous.

The sets of conditions a.i) and a.ii) are completely symmetric. Hence it suffices to prove the implication from a.i) to a.ii). Assume that the conditions of a.i) are valid.

To prove the validity that A_y is an interval, assume $y \in R_2$ such that A_y is not an interval. Then there are $x_1, x', x_2 \in A_y$, $x_1 < x' < x_2$, with $x_1, x_2 \in A_y$, $x' \notin A_y$. Hence $y \notin A_{x'}$, i. e. either $y < A_{x'}$, or $y > A_{x'}$. Consider the former case. By the assumption of a.i) about the nondecreasing bound property of A with respect to x it follows $y < A_{x_2}$, hence $x_2 \notin A_y$, contradictory to the assumptions on x_2. The second case $y > A_{x'}$ is treated analogously. It follows that A_y is an interval.

To prove the validity that A has the nondecreasing bounds property with respect to y, consider $y_1, y_2 \in R_2$, $y_1 < y_2$, $x \in R_1$ with $x < A_{y_1}$. Assume $x' \in A_{y_2}$ with $x \geq x'$. Then $x' \leq x < A_{y_1}$. Hence $x' \in A_{y_2} \setminus A_{y_1}$. Since $x \notin A_{y_1}$ we have $y_1 \notin A_x$. Because of the stipulated interval property of A_x in a.i), we have either $y_1 < A_x$ or $y_1 > A_x$. Assume $y_1 > A_x$. Then, since A has the nondecreasing bound property with respect to x, also $y_1 > A_{x'}$, hence $y_2 > A_{x'}$ and therefore $x' \notin A_{y_2}$, in contradiction to the above result $x' \in A_{y_2} \setminus A_{y_1}$. Thus $y_1 < A_x$. From $x' \leq x < A_{y_1}$ we obtain with the nondecreasing bound property with respect to x for all $x'' \in A_{y_1}$ that $y_1 < A_{x''}$, hence in particular $x'' \notin A_{y_1}$, in contradiction to the assumption $x'' \in A_{y_1}$. This final contradiction results from assuming an $x' \in A_{y_2}$ with $x \geq x'$. Hence $x < A_{y_2}$. This proves that the lower A_y bounds are nondecreasing in y.

The proof that the upper A_y bound is nondecreasing in y proceeds analogously.

This proves the validity of a.ii).

Assertion c) is a summary of assertions a) and b).

2.A.3 Proof of Proposition 2.6

Proof of Assertion a) of Proposition 2.6. Let $0 < p_0 < p_1 < 1$. We have

$$\frac{\mathrm{d}^m}{\mathrm{d}t^m} t^x(1-t)^{n-x} \quad = \quad \ln\left(\frac{t}{1-t}\right)^m t^x(1-t)^{n-x} \quad \text{for } p_0 \leq t \leq p_1.$$

The function $[p_0; p_1] \ni t \mapsto \ln(t/(1-t))^m$ is a continuous function on $[p_0; p_1]$. Hence

$$\sup_{p_0 \leq t \leq p_1} \left| \ln\left(\frac{t}{1-t}\right)^m t^x(1-t)^{n-x} \right| \quad \leq \quad \sup_{p_0 \leq t \leq p_1} \left| \ln\left(\frac{t}{1-t}\right)^m \right| \quad < \quad +\infty.$$

Hence, by applying the well-known theorem on differentiation under the integral sign to Section 2.4, we find that $v_{p_0,p_1,a,b}(x)$ can be differentiated with respect to x with the result given by Eq. (2.21). By Eq. (2.21) we have $v''_{p_0,p_1,a,b}(x) > 0$ on $[0; n]$. Hence $v'_{p_0,p_1,a,b}$ is strictly increasing on $[0; n]$. The remainder of assertion a) follows obviously.

Proof of Assertion b) of Proposition 2.6. The first partial derivatives of the symmetric beta function are given by the equation $\frac{\partial}{\partial s} B(s,t) = B(s,t)\Big[\psi(s) - \psi(s+t)\Big]$, where ψ is the digamma function, see Abramowitz & Stegun (1972, Eqs. 6.2.2 and 6.3.1). With Eqs. (2.14) and (2.18) we obtain the derivative (2.22). Since $B(x+a, n-x+b) > 0$, the sign of $v'_{0,1}(x)$ is determined by the sign of $\psi(x+1) - \psi(n-x+1)$. The integral representation of Abramowitz & Stegun (1972, Eq. 6.3.21) shows that ψ is strictly increasing on $(0; +\infty)$, hence $[0; n] \ni x \mapsto \psi(x+a) - \psi(n-x+b)$ is strictly increasing.

Proof of Assertion c) of Proposition 2.6. Let $0 < p_0 < 1$. For $p_0 < q \leq 1$, $x \in [0; n]$ let

$$h_{p_0,q,a,b}(x) \quad := \quad \frac{1}{B(a,b)(1-p_0)} \int_{p_0}^{q} \left(\frac{y-p_0}{1-p_0}\right)^{a-1} \left(1 - \frac{y-p_0}{1-p_0}\right)^{b-1} y^x(1-y)^{n-x}\,\mathrm{d}y. \tag{2.29}$$

Evidently, $\lim_{q \uparrow 1} h_{p_0,q,a,b}(x) = h_{p_0,1,a,b}(x) = v_{p_0,1,a,b}(x)$, $x \in [0; n]$. Analogously to the proof of a) of Proposition 2.6 we find

$$h^{(m)}_{p_0,q,a,b}(x) \quad = \quad \frac{1}{B(a,b)(1-p_0)} \cdot \tag{2.30}$$

$$\int_{p_0}^{q} \left(\frac{y-p_0}{1-p_0}\right)^{a-1} \left(1 - \frac{y-p_0}{1-p_0}\right)^{b-1} \ln\left(\frac{y}{1-y}\right)^m y^x(1-q)^{n-x}\,\mathrm{d}y$$

for $m = 0, 1, \ldots$, and we see that there exists a value $0 \leq z_{p_0,q,a,b} \leq n$ such that $h'_{p_0,q,a,b}(x) < 0$ for $0 < x < z_{p_0,q,a,b}$, and $h'_{p_0,q,a,b}(x) > 0$ for $z_{p_0,q,a,b} < x < n$. From Eq. (2.30) we obtain for fixed $x \in [0; n]$, $m = 0, 1, \ldots$,

$$\frac{\mathrm{d}}{\mathrm{d}q} h^{(m)}_{p_0,q,a,b}(x) \quad = \quad \frac{\ln\left(\frac{q}{1-q}\right)^m}{B(a,b)(1-p_0)} \left(\frac{q-p_0}{1-p_0}\right)^{a-1} \left(1 - \frac{q-p_0}{1-p_0}\right)^{b-1} q^x(1-q)^{n-x} \tag{2.31}$$

for $p_0 < q < 1$, with the recursion

$$\frac{\mathrm{d}}{\mathrm{d}q} h^{(m)}_{p_0,q,a,b}(x) \quad = \quad \ln\left(\frac{q}{1-q}\right) \frac{\mathrm{d}}{\mathrm{d}q} h^{(m-1)}_{p_0,q,a,b}(x) \quad \text{for } p_0 < q < 1,\ m \geq 1. \tag{2.32}$$

From Eq. (2.31) we find in particular for $m = 1, 2, \ldots$

$$\frac{\mathrm{d}}{\mathrm{d}q} h^{(m)}_{p_0,q,a,b}(x) \ > \ 0 \quad \text{for } x \in (0; n),\ p_0 < q < 1,\ q > 0.5. \tag{2.33}$$

Let $p_0 < q_1 < q_2 < \ldots < 1$ with $\lim_k q_k = 1$, $q_k > 0.5$. From Eq. (2.33) we see that the sequence $(z_{p_0,q_k,a,b})_{k\in\mathbb{N}}$ is decreasing. Since $0 \leq z_{p_0,q_k,a,b} \leq n$, $(z_{p_0,q_k,a,b})_{k\in\mathbb{N}}$ converges to a value $x_{p_0,1,a,b} \in [0; n]$. Let $0 < u < w < x_{p_0,1,a,b}$. Then $0 < u < w < z_{p_0,q_k,a,b}$ for all $k \in \mathbb{N}$, thus $h_{p_0,q_k,a,b}(u) > h_{p_0,q_k,a,b}(w)$ for all $k \in \mathbb{N}$, and hence by convergence $v_{p_0,1,a,b}(u) \geq v_{p_0,1,a,b}(w)$. Let $x_{p_0,1,a,b} < u < w < n$. Then there is a $k_0 \in \mathbb{N}$ with $z_{p_0,q_k,a,b} < u < w < n$ for all $k \geq k_0$. Thus $h_{p_0,q_k,a,b}(u) < h_{p_0,q_k,a,b}(w)$ for all $k \geq k_0$, and hence by convergence $v_{p_0,1,a,b}(u) \leq v_{p_0,1,a,b}(w)$.

The proof of assertion d) of Proposition 2.6 is completely analogous to the proof of assertion c).

2.A.4 Proof of Proposition 2.7

Elementary calculus provides the derivative (2.23) in assertion a).

Proof of Assertion b) of Proposition 2.7. Consider assertion b) with the case $0 < p_0 < p_1 < 1$. Let $0 < x < n$ and let the measure μ_x on the Borel field in $[p_0; p_1]$ be defined by

$$\mu_x(D) \quad = \quad \frac{1}{B(a,b)(p_1-p_0)} \int_D \left(\frac{t-p_0}{p_1-p_0}\right)^{a-1} \left(1 - \frac{t-p_0}{p_1-p_0}\right)^{b-1} t^x(1-t)^{n-x}\,\mathrm{d}t.$$

For $x \in (0; n)$ we obtain from Eq. (2.18) $\mu_x([p_0; p_1]) = v_{p_0,p_1,a,b}(x)$. With Eq. (2.21) we obtain for $x \in (0; n)$

$$\frac{\mathrm{d}}{\mathrm{d}x} \frac{v'_{p_0,p_1,a,b}(x)}{v_{p_0,p_1,a,b}(x)} \quad = \quad \frac{\int_{[p_0;p_1]} \ln\left(\frac{t}{1-t}\right)^2 \mathrm{d}\mu_x(t)\mu_x([p_0;p_1]) \ - \ \left[\int_{[p_0;p_1]} \ln\left(\frac{t}{1-t}\right) \mathrm{d}\mu_x(t)\right]^2}{v_{p_0,p_1,a,b}(x)^2}.$$

The strict Cauchy-Schwarz inequality $[\int_B |g_1 g_2| \,\mathrm{d}\mu_x]^2 < \int_B g_1^2 \,\mathrm{d}\mu_x(t) \int_B g_2^2 \,\mathrm{d}\mu_x(t)$ holds for $g_1(t) = \ln(t/(1-t))$, $g_2(t) = 1$, since these functions are not linearly dependent, see Hewitt & Stromberg (1969). Hence

$$\begin{aligned}\left[\int_{[p_0;p_1]} \ln\left(\frac{t}{1-t}\right) \mathrm{d}\mu_x(t)\right]^2 &\leq \left[\int_{[p_0;p_1]} |g_1(t)g_2(t)| \,\mathrm{d}\mu_x(t)\right]^2 \\ &< \int_{[p_0;p_1]} g_1(t)^2 \,\mathrm{d}\mu_x(t) \int_{[p_0;p_1]} g_2(t)^2 \,\mathrm{d}\mu_x(t) \\ &= \int_{[p_0;p_1]} \ln\left(\frac{t}{1-t}\right)^2 \mathrm{d}\mu_x(t)\ \mu_x([p_0;p_1]).\end{aligned}$$

Hence $\frac{\mathrm{d}}{\mathrm{d}x} \frac{v'_{p_0,p_1,a,b}(x)}{v_{p_0,p_1,a,b}(x)} > 0$ on $[0;n]$, and $[0;n] \ni x \mapsto v'_{p_0,p_1,a,b}(x)/v_{p_0,p_1,a,b}(x)$ is strictly increasing. The remainder of assertion b) follows from Eq. (2.23) for the derivative $Q'_{p_0,p_1,a,b,y}$ on $[0;n]$.

Consider assertion b) with the case $0 = p_0$, $1 = p_1$. From (2.22) we obtain

$$\frac{v'_{0,1,a,b}(x)}{v_{0,1,a,b}(x)} \quad = \quad \psi(x+a) - \psi(n-x+b) \quad \text{for } x \in [0;n].$$

The proof of assertion b) of Proposition 2.6 demonstrates that $[0;n] \ni x \mapsto \psi(x+a) - \psi(n-x+b)$ has at most one change of sign on $[0;n]$ which is from $-$ to $+$, if existing. Again, the remainder of assertion b) follows from Eq. (2.23) for the derivative $Q'_{p_0,p_1,a,b,p}$ on $[0;n]$.

Proof of Assertion c) of Proposition 2.7. For $p_0 < q \leq 1$, $x \in [0;n]$, consider $h_{p_0,q,a,b}(x)$ as defined by Eq. (2.29). Proceding analogously to the proof of assertion b) of Proposition 2.7, we find that $\frac{\mathrm{d}}{\mathrm{d}x} \frac{h'_{p_0,q,a,b}(x)}{h_{p_0,q,a,b}(x)} > 0$ on $[0;n]$, and $[0;n] \ni x \mapsto h'_{p_0,q,a,b}(x)/h_{p_0,q,a,b}(x)$ is strictly increasing for $p_0 < q < 1$. In analogy to the definition of $x_{p_0,q,a,b,y}$ in assertion b), define $0 \leq z_{p_0,q,a,b,y} \leq n$ by comparing $h'_{p_0,p_1,a,b}/h_{p_0,p_1,a,b}$ with $\ln(y/(1-y))$. From Eqs. (2.29) and (2.30) we find

$$h^{(m)}_{p_0,q,a,b}(x) \ < \ \ln\left(\frac{q}{1-q}\right)^m h_{p_0,q,a,b}(x) \quad \text{for } x \in [0;n],\ p_0 < q < 1. \tag{2.34}$$

With the recursion (2.32) and the inequality (2.34) we obtain for fixed $x \in [0;n]$

$$\frac{\mathrm{d}}{\mathrm{d}q} \frac{h'_{p_0,q,a,b}(x)}{h_{p_0,q,a,b}(x)} \quad = \quad \frac{\mathrm{d}}{h_{p_0,q,a,b}(x)^2} \left(\ln\left(\frac{q}{1-q}\right) h_{p_0,q,a,b}(x) - h'_{p_0,q,a,b}(x) \right) \quad > \quad 0 \tag{2.35}$$

for $p_0 < q < 1$, $q > 0.5$. Let $p_0 < q_1 < q_2 < \ldots < 1$ with $\lim_k q_k = 1$, $q_k > 0.5$. From Eq. (2.35) we see that the sequence $(z_{p_0,q_k,a,b,y})_{k \in \mathbb{N}}$ is decreasing. Since $0 \leq z_{p_0,q_k,a,b,y} \leq$

n, $(z_{p_0,q_k,a,b,y})_{k\in\mathbb{N}}$ converges to a value $x_{p_0,1,a,b,y} \in [0;n]$. The remainder of the proof proceeds analogously to the proof of assertion c) of Proposition 2.6.

The proof of assertion d) of Proposition 2.7 is completely analogous to the proof of assertion c).

2.A.5 Proof of Proposition 2.8

Consider the assumptions of assertion a). By Proposition 2.3, $A_x = \{y \in [p_0;p_1] | c_L(y) \leq x \leq c_U(y)\}$ is an interval for all $x \in \{0,\ldots,n\}$ with endpoints $\inf A_x$, $\sup A_x$ increasing in x. We have to prove that the endpoints $\inf A_x$, $\sup A_x$ are elements of A_x for $x \in \{0,\ldots,n\}$. Let $x \in \{0,\ldots,n\}$. If A_x is a singleton, then clearly $\inf A_x, \sup A_x \in A_x$. Let A_x have an open interior and let (y_l) be a sequence from A_x with $y_1 > y_2 > \ldots$, $\lim_l y_l = \inf A_x$. Then $c_L(y_l) \leq x \leq c_U(y_l)$ for all l. Hence by convergence $c_L(\inf A_x) \leq c_L(\inf {A_x}^+) \leq x \leq c_U(\inf {A_x}^+) = c_U(\inf A_x)$, hence $\inf A_x \in A_x$, and thus $\inf A_x = \min A_x$. The proof of $\sup A_x = \max A_x \in A_x$ is completely analogous.

Assertion b) is an application of Proposition 2.3.

2.A.6 Proof of Proposition 2.9

The proof makes use of the subsequent proposition on the binomial OC $L_{n,c}(y)$ defined by Eq. (2.2), see Uhlmann (1982).

Proposition 2.14 (Binomial OC)**.** *Let $n \in \mathbb{N}$, $c \in \mathbb{N}_0$, $c \leq n$.*
For $y \in (0;1)$ we have $L'_{n,c}(y) = -n\binom{n-1}{c}y^c(1-y)^{n-c-1} = -(n-c)\binom{n}{c}y^c(1-y)^{n-c-1}$. In case of $c < n$, $L_{n,c}$ is strictly decreasing on the interval $[0;1]$ with $L_{n,c}(0) = 1$, $L_{n,c}(1) = 0$. In case of $c = n$ we have $L_{n,c}(y) = L_{n,n}(y) = 1$ for $y \in [0;1]$.

Assertion a) of Proposition 2.9 follows directly from the definition of the quantities $p_{x,\gamma}$, $\widetilde{p}_{x,\gamma}$, and the monotonicity properties of the binomial OC explained by Proposition 2.14. For the proof of assertion b), let $\gamma \geq 0.5$. Let $0 < x < n$. Then

$$L_{n,x}(\widetilde{p}_{x,\gamma}) = 1-\gamma \leq \gamma = L_{n,x-1}(p_{x,\gamma}) < L_{n,x}(p_{x,\gamma}),$$

hence $\widetilde{p}_{x,\gamma} > p_{x,\gamma}$ since $L_{n,x}$ is strictly decreasing on $[0;1]$. In the case of $x = 0$, we have by definition $p_{x,\gamma} = 0.0 < \widetilde{p}_{x,\gamma}$, and similarly in the case of $x = n$ by definition $p_{x,\gamma} < 1.0 = \widetilde{p}_{x,\gamma}$.

For the proof of assertion c), let $x \in \{0, \ldots, n\}$, $y \in (p_{x,\gamma}; \widetilde{p}_{x,\gamma}) \cap [p_0; p_1]$. Assume $x \leq c_L(y) - 1$. Then we have $x \leq n-1$ and we find

$$\begin{array}{rcccl} L_{n,c_U(y)}(y) - L_{n,c_L(y)-1}(y) & \leq & 1 - L_{n,c_1(y)-1}(y) & \leq & 1 - L_{n,x}(y) \\ & < & 1 - L_{n,x}(\widetilde{p}_{x,\gamma}) & = & \gamma \end{array}$$

in contradiction to the property (2.24). Now assume $x \geq c_U(y) + 1$. Then we have $x \geq 1$ and we find

$$\begin{array}{rcccl} L_{n,c_U(y)}(y) - L_{n,c_L(y)-1}(y) & \leq & L_{n,c_U(y)}(y) & \leq & L_{n,x-1}(y) \\ & < & L_{n,x-1}(p_{x,\gamma}) & = & \gamma \end{array}$$

in contradiction to the property (2.24). This proves $c_L(y) \leq x \leq c_U(y)$, and hence also $y_L(x) \leq y \leq y_U(x)$.

2.A.7 Proof of Proposition 2.10

Assertion a) is obvious from Proposition 2.14. For the proof of assertion b), let $x_1 > 0$. By Proposition 2.14 we have for $y \in (0; 1)$

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}y}\Delta_{n,x_1,x_2}(y) &= -n\binom{n-1}{x_2} y^{x_2}(1-y)^{n-x_2-1} + n\binom{n-1}{x_1-1} y^{x_1-1}(1-y)^{n-x_1} \\ &= n\binom{n-1}{x_1-1} y^{x_2}(1-y)^{n-x_2-1}\left(\left(\frac{1}{y}-1\right)^{x_2-x_1+1} - \frac{(n-x_2)\ldots(n-x_1)}{x_1\ldots x_2}\right). \end{aligned}$$

This proves assertion b).

2.A.8 Proof of Proposition 2.11

We prove Proposition 2.11 about the comparison of two likelihood ratios. We have

$$\begin{aligned} Q_{p_0,p_1,a,b,y}(x_1) \quad - \quad & Q_{p_0,p_1,a,b,y}(x_2) \\ &= \frac{y^{x_1}(1-y)^{n-x_1}}{v_{p_0,p_1,a,b(x_1)}} - \frac{y^{x_2}(1-y)^{n-x_2}}{v_{p_0,p_1,a,b(x_2)}} \\ &= y^{x_2}(1-y)^{n-x_2}\left(\frac{y^{x_1}(1-y)^{n-x_1}}{y^{x_2}(1-y)^{n-x_2}}\frac{1}{v_{p_0,p_1,a,b(x_1)}} - \frac{1}{v_{p_0,p_1,a,b(x_2)}}\right) \\ &= y^{x_2}(1-y)^{n-x_2}\left(y^{-(x_2-x_1)}(1-y)^{x_2-x_1}\frac{1}{v_{p_0,p_1,a,b(x_1)}} - \frac{1}{v_{p_0,p_1,a,b(x_2)}}\right) \\ &= y^{x_2}(1-y)^{n-x_2}\left(\left(\frac{1-y}{y}\right)^{x_2-x_1}\frac{1}{v_{p_0,p_1,a,b(x_1)}} - \frac{1}{v_{p_0,p_1,a,b(x_2)}}\right) \end{aligned}$$

and hence

$$\begin{aligned}
& & Q_{p_0,p_1,a,b,y}(x_1) &\lesseqqgtr Q_{p_0,p_1,a,b,y}(x_2) \\
&\Leftrightarrow & Q_{p_0,p_1,a,b,y}(x_1) - Q_{p_0,p_1,a,b,y}(x_2) &\lesseqqgtr 0 \\
&\Leftrightarrow & \left(\frac{1-y}{y}\right)^{x_2-x_1} \frac{1}{v_{p_0,p_1,a,b(x_1)}} &\lesseqqgtr \frac{1}{v_{p_0,p_1,a,b(x_2)}} \\
&\Leftrightarrow & \left(\frac{1}{y}-1\right)^{x_2-x_1} &\lesseqqgtr \frac{v_{p_0,p_1,a,b(x_1)}}{v_{p_0,p_1,a,b(x_2)}} \\
&\Leftrightarrow & \frac{1}{y} &\lesseqqgtr \left(\frac{v_{p_0,p_1,a,b(x_1)}}{v_{p_0,p_1,a,b(x_2)}}\right)^{1/(x_2-x_1)} + 1 \\
&\Leftrightarrow & y &\gtreqqless \left(\left(\frac{v_{p_0,p_1,a,b(x_1)}}{v_{p_0,p_1,a,b(x_2)}}\right)^{1/(x_2-x_1)} + 1\right)^{-1} .
\end{aligned}$$

This proves Proposition 2.11.

3 Bayesian Credibility Intervals for a Probability

3.1 Introduction

Bayesian statistics is the classical framework under which prior information is made use of. Chapter 2, although the approach of interval estimation of a binomial proportion processes prior information as well, describes frequentist confidence intervals. In the Bayesian framework, credibility intervals instead of confidence intervals are used for inference about distribution parameters. That both approaches – although they show some similarities – are markedly different, becomes obvious in the present chapter.

Bayesian analysis dates back to the work of Thomas Bayes in the 18th century (Bayes & Price 1763). The basic idea of Bayesian analysis is to make use of prior information with respect to a certain unknown parameter θ. Prior information is expressed in terms of a *prior distribution* of θ with density $\pi(\theta)$, which quantifies the degree of personal belief in the likelihood of the event of interest (Berger 1985). Given a prior distribution, the information about the unknown distribution parameter is updated once data has been observed and is quantified in the *posterior distribution*.

Credibility regions are a popular way to extract information about the posterior distribution. For the binomial distribution, they are frequently encountered in the literature as a way to estimate the probability parameter by means of an interval. Not rarely, they end up being analysed from a frequentist viewpoint in these studies. Brown et al. (2001) investigate the credibility interval for a binomial probability under the use of the Jeffreys prior, a *non-informative prior*. See Section 3.7 for further consideration of non-informative priors. They argue that in view of the frequentist behaviour of the Jeffreys interval, it should be considered for practical use. In their study, they do not consider other priors than the Jeffreys prior, in particular no informative priors. Another study that examines Bayesian credibility intervals under the frequentist viewpoint was executed by Agresti & Min (2005). They advocate the use of non-informative priors like the Jeffreys prior if the coverage properties over the whole parameter space are of interest. If the coverage properties are of little concern and the focus is on the average length of the intervals, they find informative priors advantageous.

The frequentist and the Bayesian approaches of interval estimation of a parameter are conceptually different. For example, the confidence level $\gamma \in (0;1)$ and the credibility level $\beta \in (0;1)$ have to be interpreted differently. While a frequentist confidence interval of level γ is constructed such that in the long run the quantity of interest lies between the confidence limits in approximately $\gamma \cdot 100\,\%$ times, the Bayesian interpretation is that given this particular case with the observed data and chosen prior, the probability that the quantity of interest lies in the credibility interval is at least $\beta \cdot 100\,\%$, see Thatcher (1964). Despite these conceptual differences, many authors deem it worth comparing both approaches, often under $\beta = \gamma$. Thatcher (1964) argues that there are relationships between the two solutions produced by the frequentist and Bayesian approaches. According to Bayarri & Berger (2004), the debate about which of the two is superior, is far from over and statisticians are encouraged to "readily use both Bayesian and frequentist ideas" (Bayarri & Berger 2004, p. 58). This openness towards the respective other approach cannot be expected in general. Bayesians might not understand why their procedures should fulfil certain frequentist requirements and might argue that their approach is completely valid if one believes in the prior distribution chosen. On the other hand, frequentists might argue that Bayesian procedures lack some generally indispensible properties.

Bayarri & Berger (2004) give a definition of the frequentist principle as follows:

> "Frequentist principle. In repeated practical use of a statistical procedure, the long-run average actual accuracy should not be less than (and ideally should equal) the long-run average reported accuracy." (Bayarri & Berger 2004, p. 60)

The frequentist usually observes the pointwise coverage probability function. By ensuring that a certain prescribed coverage probability is met for each possible parameter value – which means one imagines the same experiment with that arbitrary fixed parameter value being carried out a number of times – the certain confidence level is also ensured if the confidence procedure is repeatedly used in varying circumstances (Bayarri & Berger 2004). Bayarri & Berger (2004) argue that a practical frequentist would not put so much emphasis on a couple of single parameter values for which the coverage probability is unsatisfying if the performance of the confidence procedure in the neighbourhood of those single points is satisfying. Evaluating the performance of an interval by taking not only the pointwise coverage into account, but in some way averaging over either various parameter values or sample sizes, would be a step of a pure frequentist towards a practical frequentist by using the Bayesian idea on the usefulness of averaging (Bayarri & Berger 2004).

On the other hand, Bayarri & Berger (2004, pp. 60–62) describe, in which sense frequentist principles are important to everybody and should also be considered by followers of the Bayesian principle. They reason that, for example, a 90 % credibility interval which only contains the unknowns about 70 % of the times gives the impression that something is wrong.

One idea of common ground between both approaches – while acknowledging that this means somewhat combining two very different viewpoints – is to compare the solutions obtained by the frequentist and Bayesian approaches by means of *probability matching priors*. Probability matching priors are prior distributions used in the Bayesian framework that lead to inferences that are of approximate frequentist validity, see Scricciolo (1999).

In the definition of Datta & Sweeting (2005)

> "A *probability matching prior* (PMP) is a prior distribution under which the posterior probabilities of certain regions coincide with their coverage probabilities, either exactly or approximately." (Datta & Sweeting 2005, p. 91)

The simplest example, and one of few in which the equality in Datta & Sweeting's (2005) definition is exact, is a credibility interval for the expected value μ in the $N(\mu, 1)$ distribution under a uniform prior for μ. Instead of exact matching, approximate matching of posterior probabilities can be demanded, see Datta & Sweeting (2005). Frequentist validity is mostly rated by means of coverage probabilities. The search for matching priors is an approach to derive non-informative prior distributions and a way to further justify the use of well-known priors of that kind, see Scricciolo (1999).

To the pioneers in terms of probability matching priors count Lindley (1958) and Welch & Peers (1963). Other important references in the context of probability matching priors are Tibshirani (1989), Datta & Ghosh (1995), Scricciolo (1999), Rousseau (2000), Datta et al. (2000), Bayarri & Berger (2004) and Datta & Sweeting (2005). Among authors advocating the use of Bayesian and frequentist ideas at the same time are Marchand et al. (2008), Fraser et al. (2010), Fraser (2011), Wasserman (2011) and Marchand & Strawderman (2013). A review of non-informative priors, of which the probability matching prior idea is one way towards identifying such, can be found in Kass & Wasserman (1996). Scricciolo (1999) distinguish in their review between first- and second-order probability matching priors, which are, among others, defined by means of quantiles or by matching the true coverage probability, and priors matching posterior and frequentist distribution functions.

Matching priors can be a topic both in parametric and predictive inference, depending on whether confidence or prediction limits are sought after. Acting in the context of prediction limits is Thatcher (1964), who establishes a connection between the Bayesian and the frequentist approach by investigating Bayesian priors that lead to the same prediction limits for a binomial success that could be obtained by the frequentist way. Different priors for the upper and lower prediction limits are applied. A more recent reference focusing on matching priors in prediction is Datta et al. (2000).

In this chapter, the probability matching prior idea is picked up in the sense that Bayesian credibility intervals for inference on a binomial probability are evaluated by means of the coverage probability, a frequentist measure. The coverage properties on the whole parameter space $[0;1]$ are taken into account. Other than finding priors in the Bayesian setting that would have approximately frequentist validity under $\beta = \gamma$, the level β is allowed to be larger than γ to ensure the desired properties in terms of the coverage probability.

The outline of this chapter is as follows: Bayesian credibility intervals in general (Section 3.2) and in particular for a binomial proportion (Section 3.4) are briefly reviewed. Section 3.3 considers the concept of Bayesian measurement and prediction spaces. In Section 3.5, the probability matching prior idea is taken up by presenting a way to compare frequentist confidence intervals and Bayesian credibility intervals for a probability under prior information. The findings are applied in Section 3.6. Elicitation of prior information with particular focus on the binomial case is discussed in Section 3.7.

3.2 Bayesian Credibility Intervals

This section introduces in short some aspects about Bayesian credibility intervals. The principles of Bayesian statistics are based on Bayes' famous theorem about the conditional probabilities of events. We provide an instance of the theorem.

Theorem 3.1 (Bayes). *Let prior knowledge on the random variable Y be expressed by the density function $f_Y(y)$. Let $f_{X|Y=y}(x)$ denote the conditional density of the random* *variable X under $Y = y$ and $f_X(x)$ the marginal density of X. Then the posterior density function of Y if $X = x$ is observed is given by*

$$f_{Y|X=x}(y) \quad = \quad \frac{f_{X|Y=y}(x) f_Y(y)}{f_X(x)} \quad \propto \quad f_{X|Y=y}(x) f_Y(y). \tag{3.1}$$

In Theorem 3.1, $\propto$ stands for the proportionality between the left-hand side and right-hand side.

Bayes' theorem describes the updated knowledge on the parameter Y after the value $X = x$ has been observed, where the prior belief about Y had been expressed in form of the prior density $f_Y(y)$. In Eq. (3.1), the denominator $f_X(x)$ is constant under fixed $X = x$. Since for this reason the denominator does not change the theorem's crucial statement of a proportionality between $f_{Y|X=x}(y)$ and $f_{X|Y=y}(x)f_Y(y)$, it is often designated the name *normalising constant*, see e. g. Congdon (2006). The Bayes theorem is therefore sufficiently explained by the statement that the posterior is proportional to the product of the prior and the likelihood.

Credibility sets, often also called *credible sets*, determine intervals in which an unknown distribution parameter lies with a prescribed probability. In the case of a univariate distribution parameter and a connected credibility set, they are intervals, and the terms *credibility interval* or *credible interval* are used. Credibility sets can be seen as easily reportable summaries of the posterior distribution, see Berger (1985, p. 145).

Definition 3.2 (Credibility Interval)**.** *Let Y be an unknown parameter with prior density function $f_Y(y)$ and $f_{X|Y=y}(x)$ be the conditional density of X given $Y = y$ and let $\beta \in (0; 1)$. Let $X = x$ be observed. Then A_x is a $\beta \cdot 100\%$ credibility interval for Y if*

$$P(Y \in A_x | X = x) \quad = \quad P_x(Y \in A_x) \quad \geq \quad \beta.$$

Here, the probability $P(Y \in A_x|X = x)$ is evaluated using the posterior density $f_{Y|X=x}(y)$ of Y. Two very common forms of credibility intervals in the case of two-sided intervals are the *equal-tail interval* and the *highest probability density (HPD) interval.*

Definition 3.3 (Equal-tail and HPD Credibility Interval)**.** *Let $X = x$ and $\beta \in (0; 1)$. A credibility interval A_x is called an equal-tail $\beta \cdot 100\%$ credibility interval for Y if*

$$A_x \quad = \quad \left[z_F\left(\frac{1-\beta}{2}\right); \quad z_F\left(\frac{1+\beta}{2}\right)\right],$$

where $z_F(\alpha)$ is the $\alpha \cdot 100\%$ quantile of the posterior distribution $F_{Y|X=x}$ of Y with density $f_{Y|X=x}(y)$. The interval

$$A_x \quad = \quad [z_F(\alpha_1); z_F(1-\alpha_2)]$$

is called a highest posterior density (HPD) interval if $0 < \alpha_1, \alpha_2 < 1$ *are chosen such that* $P(z_F(\alpha_1) \leq Y \leq z_F(1-\alpha_2)|X = x)$ *is minimal with* $P(z_F(\alpha_1) \leq Y \leq z_F(1-\alpha_2)|X = x) \geq \beta$*, and for any* $y' \notin A_x$ *and* $y \in A_x$*, we have* $f_{Y|X=x}(y') \leq f_{Y|X=x}(y)$*.*

A credibility interval has to be distinguished from a frequentist confidence interval in interpretation. A confidence interval describes an interval into which the true, but unknown parameter falls in approximately $\gamma \cdot 100\,\%$ of the cases in repeated sampling, where $\gamma \in (0; 1)$ is the confidence level. For one execution of the confidence procedure, the confidence interval either contains the true parameter value or not, hence the interval is either right or wrong. The parameter is regarded as an unknown fixed value. In contrast, the Bayesian credibility interval contains the true parameter value with approximately $\beta \cdot 100\%$ certainty, by which it delivers the more intuitive interpretation, see Congdon (2006, p. 2). The Bayesian interpretation considers the parameter a random variable.

The Bayesian HPD credibility interval A_x has several appealing properties, see Box & Tiao (1973, p. 123) or Berger (1985, pp. 140–141):

Remark 3.4 (Properties of the HPD Interval)**.**

(a) The HPD interval A_x *for* Y *under observed* $X = x$ *and a given probability content* $\beta \in (0; 1)$ *has smallest possible volume in the parameter space* R_2 *of* Y*.*

(b) If $f_{Y|X=x}(y)$ *is nonuniform over every region* $\subset R_2$*, then the HPD interval* A_x *of content* β *is unique. Two points* y_1 *and* y_2 *with* $f_{Y|X=x}(y_1) = f_{Y|X=x}(y_2)$ *are either simultaneously contained or not contained in a* $\beta \cdot 100\,\%$ *HPD interval* A_x*. Conversely, if* $f_{Y|X=x}(y_1) \neq f_{Y|X=x}(y_2)$*, then there exists a* β *such that* $y_1 \in A_x$ *and* $y_2 \notin A_x$ *or vice versa.*

3.3 Bayesian Measurement and Prediction Spaces

In analogy to frequentist level β measurement and prediction spaces (MPS), we define Bayesian level $\beta \in (0; 1)$ MPSs.

Definition 3.5 (Bayesian Level β MPS)**.** *Let* $\mathcal{A}_1, \mathcal{A}_2$ *be* σ*-fields in* R_1, R_2*, respectively. Let* $f_{X,Y}$ *be the joint density of* X, Y *with respect to a product measure* $\mu_1 \otimes \mu_2$ *on the product field* $\mathcal{A}_1 \otimes \mathcal{A}_2$*, and let* f_X, f_Y *be the respective marginal densities. For sets* $B \in \mathcal{A}_1$ *let*

$$P_y(B) = P(B|Y = y) = \int_B f_{X|Y=y}(x)\ d\mu_1(x)$$

be the conditional probability under $Y = y$. *For sets* $A \in \mathcal{A}_1 \otimes \mathcal{A}_2$ *let* $A_x = \{y|(x,y) \in A\}$, $A_y = \{x|(x,y) \in A\}$ *be the projections for fixed* $x \in R_1$, $y \in R_2$, *respectively. Let* $0 < \beta < 1$. *A set* $A \in \mathcal{A}_1 \otimes \mathcal{A}_2$ *is called a* Bayesian level β measurement and prediction space for $X|Y$ (Bayesian level β MPS for $X|Y$) *if the projection* A_x *constitutes a credibility region for the unknown value of* Y, *i. e.*

$$\beta \;\leq\; P_x(Y \in A_x) \;=\; P(Y \in A_x|X = x) \quad \textit{for all } x \in R_1.$$

The weighted volume $V(A)$ *of a Bayesian level* β *MPS is defined as*

$$V(A) \quad = \quad \int_{R_1} \int_{A_x} d\nu(y) f_X(x) d\mu_1(x),$$

where ν *is the Borel measure, i. e.* $\int_{A_x} d\nu(y) = \nu(A_x)$ *is the geometric volume of the credibility region* A_x.

Bayesian credibility intervals are usually not judged by the frequentist measure of coverage probability, but by their posterior probability properties. Yet, as Bayarri & Berger (2004) pointed out, observing a satisfying number of successes in the sense that the true parameter value actually lies in the Bayesian interval is also a desirable quality of a Bayesian credibility interval. For this reason, we formulate the frequentist coverage probability of a Bayesian MPS.

Remark 3.6. *Let* A *be a Bayesian level* β *MPS. For fixed* $y \in R_2$ *consider the projection* $A_y = \{x|(x,y) \in A\}$. *Then the frequentist coverage probability of the prediction region* A_y *under* $Y = y$ *is given by*

$$C(y) \quad := \quad P_y(X \in A_y) \quad = \quad P(X \in A_y|Y = y).$$

The probability P_y is calculated using the conditional density function $f_{X|Y=y}(x)$ of X under $Y = y$.

Likewise, a frequentist level γ MPS as introduced in Section 2.2 can be evaluated in a Bayesian manner.

Remark 3.7. *Let* A *be a frequentist level* γ *MPS. For fixed* $x \in R_1$ *consider the projection* $A_x = \{y|(x,y) \in A\}$. *Then the probability content of the confidence region* A_x *under* $X = x$ *is given by*

$$P_x(Y \in A_x) \quad = \quad P(Y \in A_x|X = x).$$

The probability P_x is calculated using the posterior density function $f_{Y|X=x}(x)$ of Y under $X = x$.

3.4 HPD Credibility Intervals for a Probability

We use Bayes' theorem to determine credibility intervals for the parameter p of a binomial distribution. We express prior knowledge on p in terms of the so-called *conjugate prior* of the binomial distribution: the beta distribution. In general, conjugate priors are supposed to deliver easily calculable posterior distributions, see Berger (1985, p. 130). We provide a formal definition of a family of conjugate priors following Lindley (1972):

Definition 3.8 (Conjugate Family)**.** *Let $\mathcal{F}$ be a family of distributions over R_2. Let $f_Y(y) \in \mathcal{F}$ be a prior distribution with $y \in R_2$ and let $f_{X|Y=y}(x)$ be the likelihood with $x \in R_1$. Then $\mathcal{F}$ is* closed under sampling *with respect to the distribution with density $f_{X|Y=y}(x)$ if $f_{Y|X=x}(y) \in \mathcal{F}$ for every $x \in R_1$, where $f_{Y|X=x}(y)$ is the posterior density function. The family $\mathcal{F}$ is* conjugate *with respect to $f_{X|Y=y}(x)$.*

Berger (1985, p. 142) remarked that natural conjugate priors are usually unimodal and result in unimodal posterior densities. In that case, the HPD credibility sets are never disconnected credibility sets, but always intervals. While this prevents the credibility sets from looking unusual, it also prevents from detecting possibly conflicting information obtained by the prior and the data, something which according to Berger (1985, p. 142) advises caution. In the present case of inference about a binomial proportion, the beta prior fulfils the requirement to be easily calculable and results in a beta posterior distribution, as becomes obvious from the following proposition. We only consider beta priors on the support $[0; 1]$.

Proposition 3.9 (Posterior Density Function for a Probability)**.** *Consider the binomial density $f_{X|Y=y}(x) = \binom{n}{x} y^x (1-y)^{n-x}$ of X under the probability $Y = y$ and a Beta(a, b) distribution as prior distribution for Y. Then the marginal density $f_X(x)$ of X is given by*

$$f_X(x) \quad = \quad \frac{\binom{n}{x}}{B(a,b)} \int_0^1 y^{x+a-1}(1-y)^{n-x+b-1}\, dy \quad =: \quad w_{a,b}(x).$$

The posterior distribution of Y is the beta distribution Beta$(x + a, n - x + b)$.

Proof. The assertion about the posterior distribution follows from elementary application of Bayes' theorem, see Appendix 3.A, Section 3.A.1. □

From Proposition 3.9, we can directly infer on credibility intervals for a probability $y = p$.

Proposition 3.10 (Credibility Interval for a Probability)**.** *Consider the binomial density $f_{X|Y=y}(x) = \binom{n}{x} y^x (1-y)^{n-x}$ of X under the probability $Y = y$ and a Beta(a, b)*

distribution as the prior for Y. Let $0 < \beta < 1$. Then a $\beta \cdot 100\%$ credibility interval A_x for Y under $X = x$ is given by

$$A_x \quad = \quad \left[z_{Beta(x+a,n-x+b)}(\alpha_1); \quad z_{Beta(x+a,n-x+b)}(1-\alpha_2)\right],$$

where $\alpha_1 + \alpha_2 = 1 - \beta$, $\alpha_1, \alpha_2 \geq 0$.

Proof. From Proposition 3.9 it follows

$$\begin{aligned} \mathrm{P}_x(Y \in A_x) &= \mathrm{P}_x\left(z_{\mathrm{Beta}(x+a,n-x+b)}(\alpha_1) \leq Y \leq z_{\mathrm{Beta}(x+a,n-x+b)}(1-\alpha_2)\right) \\ &= \mathrm{P}_x\left(Y \leq z_{\mathrm{Beta}(x+a,n-x+b)}(1-\alpha_2)\right) - \mathrm{P}_x\left(Y < z_{\mathrm{Beta}(x+a,n-x+b)}(\alpha_1)\right) \\ &= 1 - \alpha_2 - \alpha_1 \quad = \quad \beta. \end{aligned}$$

□

The equal-tail credibility interval for the parameter $y = p$ of a binomial distribution is obtained if we set $\alpha_1 = \alpha_2 = \frac{1-\gamma}{2}$ in Proposition 3.10. The HPD credibility interval for a probability requires more calculation effort. The following propositions are helpful for the calculation of the HPD intervals.

Proposition 3.11 (Monotonicity of the Beta Density)**.** *For the density function of the beta distribution $Beta(a,b)$ with shape parameters $a, b > 0$, that is defined by $f(x) = \frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1}$ for $x \in (0;1)$ and $f(x) = 0$, otherwise, the following assertions hold:*

(a) $f(x)$ is strictly increasing on $(0;1)$ if $b \leq 1 < a$ or $b < 1 = a$.

(b) $f(x)$ is constant on $[0;1]$ if $a = b = 1$.

(c) $f(x)$ is strictly decreasing if $a \leq 1 < b$ or $a < 1 = b$.

(d) In the case $b > 1$, $a > 1$, $f(x)$ is strictly increasing on $(0;x^*)$ and strictly decreasing on $(x^*;1)$, where $x^* := \frac{a-1}{a+b-2}$.

(e) In the case $a < 1, b < 1$, $f(x)$ is strictly decreasing on $(0;x^*)$ and strictly increasing on $(x^*;1)$, where $x^* := \frac{a-1}{a+b-2}$.

Proof. See Appendix 3.A, Section 3.A.2. □

A direct consequence of Proposition 3.11 is the following corollary, which is helpful for the calculation of HPD credibility intervals for a probability.

Corollary 3.12 (Monotonicity of the Posterior Density)**.** *Let the random variable Y that denotes the binomial proportion have a beta prior distribution, i. e. $Y \sim Beta(a,b)$, $a, b > 0$. Let $f_{X|Y=y}(x) = \binom{n}{x}y^x(1-y)^{n-x}$ be the likelihood function with $n \in \mathbb{N}, x = \{0, 1, \ldots, n\}$. Then the following assertions hold:*

(a) $f_{Y|X=x}(y)$ *is strictly increasing on* $(0;1)$ *if*

(i) $n-x+b \leq 1 < a+x$*,*

(ii) $n-x+b < 1 = a+x$*.*

Then necessarily $b \leq 1$ *and* $x = n$*.*

(b) $f_{Y|X=x}(y)$ *is strictly decreasing on* $(0;1)$ *if*

(i) $a+x \leq 1 < n-x+b$

(ii) $a+x < 1 = n-x+b$

Then necessarily $a \leq 1$ *and* $x = 0$*.*

(c) In the case $n-x+b > 1, a+x > 1$*,* $f_{Y|X=x}(y)$ *is strictly increasing on* $(0;y^*)$ *and strictly decreasing on* $(y^*;1)$*, where* $y^* := \frac{a+x-1}{n+a+b-2}$*.*

(d) The case that $f_{Y|X=x}(y)$ *is strictly decreasing on a interval* $(0;y^*)$ *and strictly increasing on* $(y^*;1)$ *with* $y^* \in (0;1)$ *cannot occur.*

(e) The case that $f_{Y|X=x}(y)$ *is constant cannot occur.*

Proof. See Appendix 3.A, Section 3.A.3. □

The preceding corollary allows to infer on HPD credibility intervals for a binomial proportion. We explicitly describe them in Corollary 3.13.

Corollary 3.13 (HPD Credibility Interval for a Binomial Probability)**.** *Let* $0 < \beta < 1$ *and* $f_Y(y) = \frac{1}{B(a,b)} y^{a-1}(1-y)^{b-1}$ *be the prior density of the binomial probability* Y*. Let* $f_{X|Y=y}(x) = \binom{n}{x} y^x (1-y)^{n-x}$ *be the likelihood function with* $n \in \mathbb{N}, x = \{0,1,\ldots,n\}$*, and* $f_{Y|X=x}(y)$ *be the posterior density of* Y *under* $X = x$*. The level* β *HPD credibility interval* A_x *for* Y *under* $X = x$ *is*

(a) $A_x = [l_x; u_x] = [z_{Beta(a+n,b)}(1-\beta); 1]$ *if*

(i) $n-x+b \leq 1 < a+x$ *or*

(ii) $n-x+b < 1 = a+x$*,*

(b) $A_x = [l_x; u_x] = [0; z_{Beta(a,n+b)}(\beta)]$ *if*

(i) $a+x \leq 1 < n-x+b$ *or*

(ii) $a+x < 1 = n-x+b$*,*

(c) $A_x = [l_x; u_x] = [z_{Beta(x+a,n-x+b)}(\alpha_1); z_{Beta(x+a,n-x+b)}(1-\alpha_2)]$ *if* $n-x+b > 1$ *and* $a+x > 1$*, where* α_1, α_2 *are chosen such that* $\alpha_1 + \alpha_2 = 1-\beta$ *and* $f_{Y|X=x}(z_{Beta(x+a,n-x+b)}(\alpha_1)) = f_{Y|X=x}(z_{Beta(x+a,n-x+b)}(1-\alpha_2))$*. The maximum point* y^* *of the posterior density function is* $y^* = \frac{a+x-1}{n+a+b-2}$ *with* $y^* \in A_x$*.*

Proof. See Appendix 3.A, Section 3.A.4. □

Figure 3.1 shows instances of beta prior distributions and their corresponding level 95 % HPD credibility intervals and coverage probability graphs for a binomial proportion $p = y$ if the sample size is $n = 30$. The investigated prior distributions are the uniform distribution $\text{Beta}(1, 1)$, the bathtub shaped distribution $\text{Beta}(0.5, 0.5)$, the right-skewed distribution $\text{Beta}(0.2, 1)$ and the left-skewed distribution $\text{Beta}(7, 3)$. The HPD credibility intervals under $X = x \in \{0, \dots, 30\}$ vary in length depending on the prior. For example, the HPD credibility intervals for small x are comparably wide under the prior $\text{Beta}(7, 3)$. This distribution represents the belief that low values of x are unlikely to occur. The coverage probability graphs take very different shapes depending on the prior. A common feature of all analysed priors is the fact that the coverage probability often lies below the specified credibility level, in this case 0.95. In the case of the $\text{Beta}(0.5, 0.5)$ prior – which deems probabilities close to 0 and 1 more likely than midpoints – the coverage probability is especially high for values of p near the boundaries. In contrast, the coverage probability of the right-skewed prior $\text{Beta}(0.2, 1)$ falls below 0.95 for a wide range of small values of p and has satisfactory coverage probabilities for large values of p. The left-skewed prior $\text{Beta}(7, 3)$, which takes its mode for $p = 0.75$, especially fails for small values of p and for p close to 1, while showing fairly good coverage properties between approximately 0.5 and 0.9.

Figure 3.2 shows posterior distributions for p if $X = x \in \{0, \dots, 7\}$ binomial successes are observed in a sample of size $n = 7$ under different forms of beta priors. The bounds of the corresponding level 80 % credibility intervals are displayed by dashed vertical lines. In the case of the priors $\text{Beta}(1, 1)$, $\text{Beta}(0.5, 0.5)$ and $\text{Beta}(0.2, 1)$, i. e. both shape parameters a and b of the beta prior $\text{Beta}(a, b)$ are ≤ 1, the posterior distributions are decreasing for $x = 0$ and increasing for $x = n$, a behaviour following from Corollary 3.12. The modes of the posterior density functions are increasing in x, a characteristic which can be easily seen by taking the derivative in x of the maximum function $g_{n,a,b}(x) := \frac{a+x-1}{n+a+b-2}$. Under the prior $\text{Beta}(7, 3)$, the posterior distributions are each of inverted bathtub shape with a maximum within the interval $(0; 1)$ for each $x \in \{0, 1, \dots, n\}$. Consequently, there are $p \in [0; 1]$ which are contained in none of the corresponding HPD credibility intervals $[l_x; u_x]$. The frequentist coverage probability at these p results to 0 (compare Fig. 3.1), a property that will be resumed in Theorem 3.16.

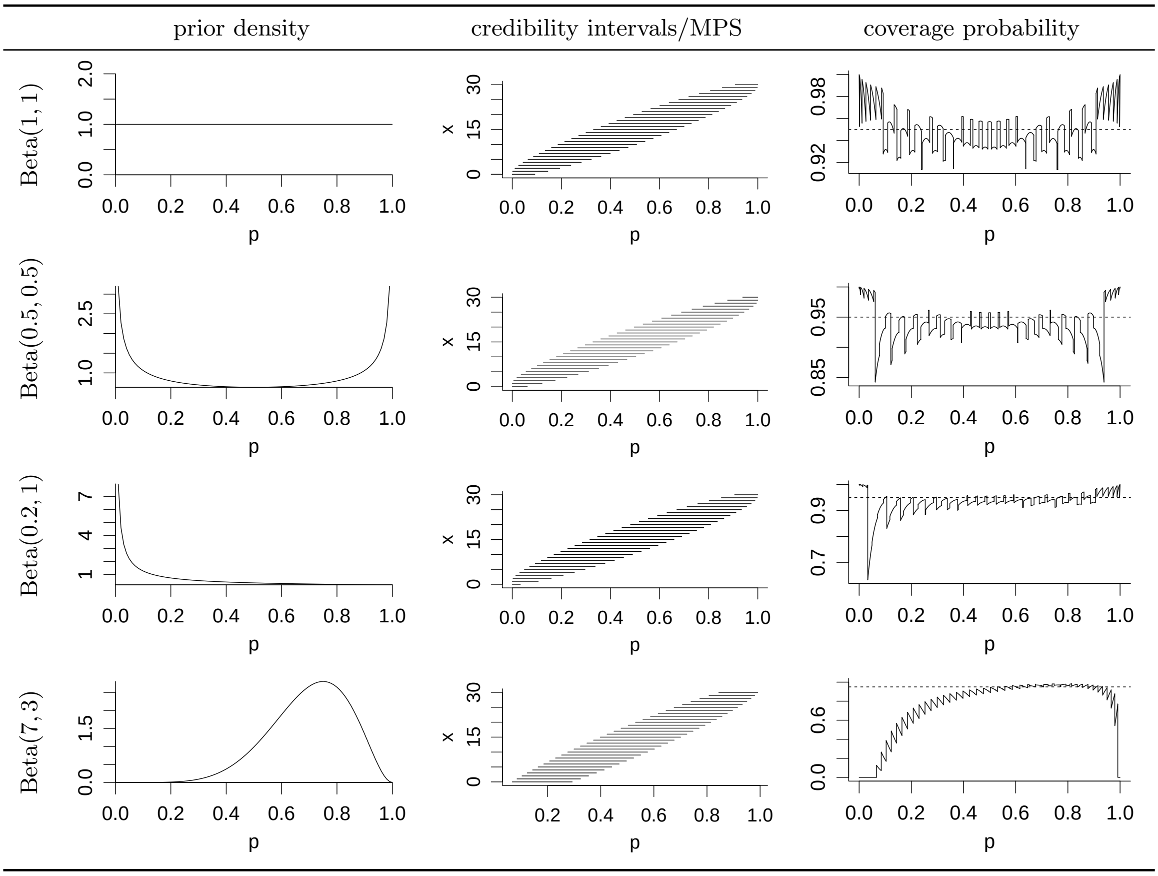


Figure 3.1: Level 95 % HPD credibility intervals for a binomial proportion p and coverage probability for a selection of beta prior distributions. Sample size $n = 30$.

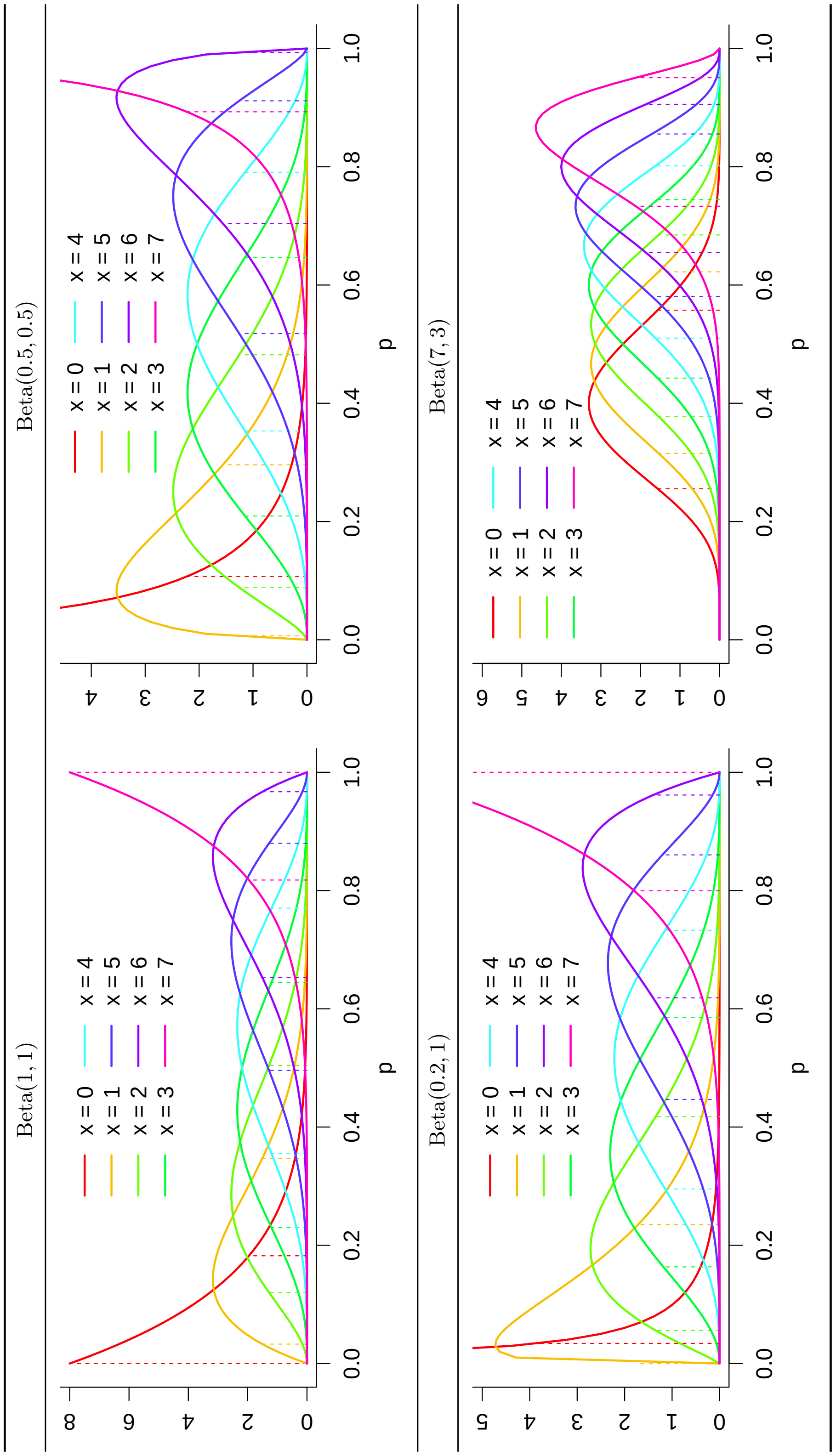


Figure 3.2: Posterior density functions for a binomial proportion p under observed $X = x$ with level 80 % HPD credibility intervals (dashed) for different beta prior distributions. Sample size $n = 7$.

3.5 Relation between the Probability Content of Bayesian HPD Intervals and the Coverage Probability

Using Definition 3.5 and Proposition 3.10, we can describe the volume of a Bayesian MPS for a binomial probability p.

Remark 3.14. *The volume of a level β Bayesian MPS with credibility intervals $A_x = \left[z_{Beta(x+a,n-x+b)}(\alpha_1); \quad z_{Beta(x+a,n-x+b)}(1-\alpha_2)\right]$ for the parameter p of a binomial distribution $Bi(n,p)$ is given by*

$$V(A) \quad = \quad \sum_{x=0}^{n} \nu(A_x) w_{a,b}(x),$$

where $\nu(A_x) = u_x - l_x$ is the length of the credibility interval $A_x = [l_x; u_x]$ obtained from Corollary 3.13, and the weights $w_{a,b}(x)$ are given by Proposition 3.9.

If the objective is to find a *minimum volume* Bayesian MPS for a probability p, the weighted volume $V(A)$ from Remark 3.14 has to be minimised. Since the weights $w_{a,b}(x)$ from Proposition 3.9 are constants under a given x and fixed prior distribution $\text{Beta}(a,b)$, minimising $V(A)$ results to minimising the lengths $\nu(A_x)$ of the Bayesian credibility intervals for $x \in \{0, 1, \ldots, n\}$. This is equivalent to finding the HPD credibility intervals A_x for $x \in \{0, 1, \ldots, n\}$, cf. Box & Tiao (1973).

From Fig. 3.1 it has become obvious that the coverage probability of Bayesian HPD credibility intervals can drop substantially to values far below a certain threshold and therefore can turn out to be unsatisfactory from a frequentist point of view. To get a better idea about the coverage properties of both types of intervals, we analyse the relationship between the confidence level γ of a frequentist confidence interval and the credibility β of a Bayesian HPD credibility interval for a binomial proportion.

The statement of the subsequent proposition about the monotonicity of the quantiles of the posterior distribution will be needed for the proof of Theorem 3.16.

Proposition 3.15 (Quantiles of the Posterior Distribution)**.** *Let $a, b > 0, n \in \mathbb{N}, x \in \{0, 1, \ldots, n\}, 0 < \gamma < 1$. Then the $\gamma \cdot 100\%$ quantiles $z_{Beta(a+x,n-x+b)}(\gamma)$ of the beta distribution $Beta(a+x, n-x+b)$ are increasing in x.*

Proof. See Appendix 3.A, Section 3.A.5. □

The following theorem describes conditions for the $\text{Beta}(a,b)$ prior distribution, under which the Bayesian HPD credibility intervals for a binomial proportion have a minimum coverage probability given the probability content β of the credibility intervals is chosen appropriately.

Theorem 3.16. *Let $\{A_x|x \in \{0,1,\dots,n\}, n \in \mathbb{N}\}$ denote the set of level $\beta \in (0;1)$ credibility intervals for a binomial proportion $Y \in [0;1]$ under the prior $f_Y(y) = \frac{1}{B(a,b)} y^{a-1}(1-y)^{b-1}$, $a,b > 0$ and let $f_{X|Y=y}(x) = \binom{n}{x} y^x (1-y)^{n-x}$ be the likelihood function. Let $A = \{(x,y)|y \in A_x, x = \{0,1,\dots,n\}\}$ be the Bayesian level β MPS and $A_y = \{x|(x,y) \in A\}$ be the projection on $R_1 = \{0,1,\dots,n\}$ for fixed $y \in [0;1]$. Let $C(y) = P_y(X \in A_y)$, $y \in [0;1]$, denote the frequentist coverage probability function. Then for any given $0 < \gamma < 1$, the following two assertions are equivalent:*

i) There exists a credibility level $0 < \beta < 1$ such that for the projection A_y of the corresponding level β HPD intervals we have $C(y) \geq \gamma$ for all $y \in [0;1]$.

ii) The shape parameters a and b of the beta prior Beta(a,b) fulfil the conditions $a \leq 1, b \leq 1$.

Proof. See Appendix 3.A, Section 3.A.6. □

Theorem 3.16 proves the existence of a credibility level β under the described conditions. The credibility level β derived in the proof is not necessarily the lowest possible level fulfiling the condition described in the theorem. In fact, it will in most cases be far from it. For a comparison of the Bayesian HPD credibility intervals with the frequentist confidence intervals for a binomial probability a fair approach would be to take the smallest level β such that the frequentist coverage is greater or equal to a prescribed confidence level γ for all $y \in [0;1]$.

It follows from Theorem 3.16 that if a minimum coverage level on the whole interval $[0;1]$ is intended, a comparison of the two interval types for the probability Y only makes sense if the prior distribution Beta(a,b) with $a,b \leq 1$ is considered. This requirement includes priors of very different shape. The uniform prior Beta$(1,1)$ as well as the bathtub shaped prior Beta$(0.5,0.5)$ fulfil the requirement. By appropriate choices of a and b, a variety of other forms and skewnesses can be modelled, e. g. a distribution with decreasing density function such as Beta$(0.02,1)$, increasing density function such as Beta$(1,0.3)$, symmetric and unsymmetric bathtub shapes such as Beta$(0.3,0.3)$ or Beta$(0.8,0.5)$.

The following proposition allows the description of an algorithm for finding the minimum probability content $0 < \beta < 1$ such that under the premises of Theorem 3.16, the frequentist coverage probability of the credibility intervals is at least a certain prescribed level $0 < \gamma < 1$.

Proposition 3.17 (Monotonicity of the Coverage Probability in β)**.** *The coverage probability $C(y)$ of a Bayesian level β MPS A, where the projection A_X constitutes a level β HPD credibility interval for the binomial probability Y, is an increasing function in β with $C(y) \overset{\beta \to 1}{\nearrow} 1$.*

Proof. See Appendix 3.A, Section 3.A.7. □

3.6 Numerical Comparison of Bayesian and Frequentist MPS

We empirically compare the frequentist confidence interval for a proportion under prior information from Chapter 2 with the Bayesian HPD credibility interval from the preceding sections in the spirit of Theorem 3.16. In our analysis we consider several beta distributions with right-skewed density functions $\text{Beta}(a, 1)$ with $a < 1$, as well as the uniform prior $\text{Beta}(1, 1)$ and the Jeffreys prior $\text{Beta}(0.5, 0.5)$. The investigated distributions all fulfil the crucial requirement of Theorem 3.16 that both shape parameters $a, b > 0$ of the beta prior distribution $\text{Beta}(a, b)$ are smaller than or equal to 1. We compare the frequentist confidence interval under prior information $\text{Beta}(a, b)$ with the HPD intervals under the prior $\text{Beta}(a, b)$ both in terms of their coverage probabilities as well as the volumes of the corresponding MPSs.

The following simple calculation shows that a confidence interval of level γ and a Bayesian HPD credibility interval of level $\beta = \gamma$ cannot simply be compared without further consideration. Let $n = 50$ and consider the prior distribution $\text{Beta}(0.02, 1)$ for the binomial probability Y. Let the number of observed binomial successes X in a sample of size $n = 50$ be 0. (In the application of Chapter 5 this is not unusual.) In this case, the frequentist level 80 % confidence interval for $p = y$ results to $(0.00000; 0.03205)$, whereas the level 80 % HPD interval for Y under $X = 0$ results to $(0.00000; 0.00001)$. The latter is thus extremely narrow and its usability in practice has to be questioned. In contrast, the level 99.83 % credibility interval under $X = 0$ is $(0.00000; 0.03175)$ and is hence considerably broader. For $x = 1$, we obtain the level 80 % confidence interval $(0.00445; 0.06099)$ and the credibility interval $(0.00000; 0.03225)$ under the level 80 % and the interval $(0.00000; 0.12063)$ under the level 99.83 %. This example shows that a confidence level γ of a frequentist confidence interval and a credibility level β of a Bayesian HPD interval are completely different input parameters for two different estimation procedures for a probability.

To better investigate the relation between the confidence level γ and the credibility level β, we consider the Bayesian level β MPS under different levels β. Figure 3.3 shows the coverage probabilities of the HPD credibility intervals under the uniform prior Beta$(1, 1)$ (left-hand side) and the right-skewed prior Beta$(0.02, 1)$ (right-hand side) for three different values of β each. The plots in the first row show the frequentist coverage probabilities of the HPD intervals of level $\beta = 0.9$ (under sample size $n = 10$) and $\beta = 0.8$ (under sample size $n = 50$), respectively. In both cases, probabilities $y = p \in [0; 1]$ exist for which the coverage probability falls below the credibility levels $\beta = 0.9$ and $\beta = 0.8$, respectively. In the case of the uniform prior, the coverage probability is close to or exceeds 0.9 for values of p close to 0 or 1 and mostly falls below the level for values of p in the middle of the unit interval. In the case of the right-skewed prior, the coverage probability clearly fails to meet the level 0.8 for small values of p. The extremely narrow level 80 % credibility interval $(0.00000; 0.00001)$ under $X = 0$ might be one of the reasons for this unsatisfying behaviour of the HPD interval for small p.

According to Theorem 3.16, it is possible to find for both prior distributions Beta$(1, 1)$ and Beta$(0.02, 1)$ credibility levels β such that the pointwise coverage probability does not drop below a certain prescribed level γ. The second row of Fig. 3.3 shows the coverage probabilities for credibility levels which only just do not fulfil pointwise coverage probability of at least γ, whereas in the third row the credibility levels are the minimum credibility levels such that the coverage probabilities exceed the levels $\gamma = 0.9$ and $\gamma = 0.8$, respectively. While in the presented case of the right-skewed Beta$(0.02, 1)$ prior the level $\gamma = 0.8$ can be achieved exactly for $\beta = 0.99830$ and $p = 0.03174$, the obtained minimum coverage probability for the Beta$(1, 1)$ prior undergoes a jump when slightly changing the credibility level β from 0.96456 to 0.96457. While in the first case, the minimum coverage probability of 0.899554 for $p = 0.39138$ and $p = 0.60862$ still lies below $\gamma = 0.9$, it jumps to 0.92452 for $p = 0.29965$ and $p = 0.70035$ if the credibility level is risen to 0.96457. This behaviour demonstrates once more the discontinuities that are present in the context of interval estimation of the parameter p due to the discreteness of the binomial distribution, even under common priors like the uniform prior Beta$(1, 1)$.

The values 0.96457 (bottom left) and 0.99830 (bottom right) in Fig. 3.3 are consequently values for the credibility β such that a prescribed coverage probability of at least 90 % or 80 %, respectively, is achieved pointwise in $[0; 1]$. A more extensive overview over the minimally necessary credibility levels β such that the HPD credibility intervals have a coverage probability exceeding a certain prescribed level γ is provided in Table 3.1 for prior distributions Beta$(1, 1)$, Beta$(0.5, 0.5)$ and right-skewed priors Beta$(a, 1)$ with $a \in \{0.5, 0.1, 0.05, 0.02\}$. The credibility level β which is necessary to achieve pointwise coverage of at least γ exceeds the value γ considerably in a majority of the presented cases. For example, to obtain a minimum pointwise coverage probability of at least $\gamma = 0.8$, a level $\beta = 0.91525$ HPD interval is required under the uniform prior and sample size $n = 10$. There is a tendency of lower minimum β values for increasing sample size, but this is not a general rule. For $n = 150$ in the case of a uniform prior only the credibility level 0.84985 is needed to ensure a coverage of 0.8 in contrast to 0.91525 for $n = 10$, 0.86967 for $n = 50$ or 0.85256 for $n = 100$. For example, the minimum β level of 0.99779 is necessary for $\beta = 0.99$, $n = 100$ under the prior $\mathrm{Beta}(0.5, 0.5)$, and the even larger value 0.99783 for $n = 150$. In general, it has to be mentioned that the differences in the credibility levels for the different priors or sample sizes are often a matter of the 4th or 5th decimal place in β, a precision that is irrelevant in practice. Nevertheless, to demonstrate that also for levels of γ very close to 1, as e. g. $\gamma = 0.99$, the minimum level β required to ensure pointwise coverage of the HPD intervals of at least γ is smaller than 1, we report the results of Table 3.1 up to the 5th digit behind the decimal separator.

In Table 3.1, we also record the volumes of the minimum volume frequentist MPS of level γ and the volumes of the HPD credibility intervals of level β ensuring a frequentist pointwise coverage probability of at least γ. The volume of the frequentist MPS is throughout smaller than the volume of the Bayesian MPS. This is not surprising because the frequentist interval is constructed in such a way that the coverage criterion is fulfiled pointwise and the volume is minimal. For the uniform prior $\mathrm{Beta}(1, 1)$, the volume of the Bayesian MPS for the presented sample sizes is between 0.7 % (for $n = 10$, $\gamma = 0.99$) and 11.6 % (for $n = 10$, $\gamma = 0.8$) higher, which is harmless in comparison to the $\mathrm{Beta}(0.5, 0.5)$ prior, where the volume increases are between 6.0 % ($n = 10$, $\gamma = 0.99$) and 34.8 % ($n = 150$, $\gamma = 0.8$). The volume increase is even worse for the right-skewed priors $\mathrm{Beta}(a, 1)$ with a close to 0. While for certain constellations of sample size n and minimum coverage level γ the volume of the Bayesian MPS can be small, e. g. in the case of the prior $\mathrm{Beta}(0.02, 1)$ for $n = 10$, $\gamma = 0.95$ it is only 4.4 % larger than the frequentist MPS, it can almost double for different parameter combinations, as from a frequentist volume of 0.010 to a Bayesian volume of 0.019 in the case $n = 100$, $\gamma = 0.8$. Consequently, constellations of sample size n and level γ can be found such that there is little difference in the volume of the frequentist and the Bayesian MPS, but there also occur cases where the Bayesian MPS shows clearly broader intervals if a certain pointwise coverage probability is demanded.

The problem in the presented investigation of the coverage properties of the Bayesian level β MPS under different levels β is that often the requirement of pointwise exceedance

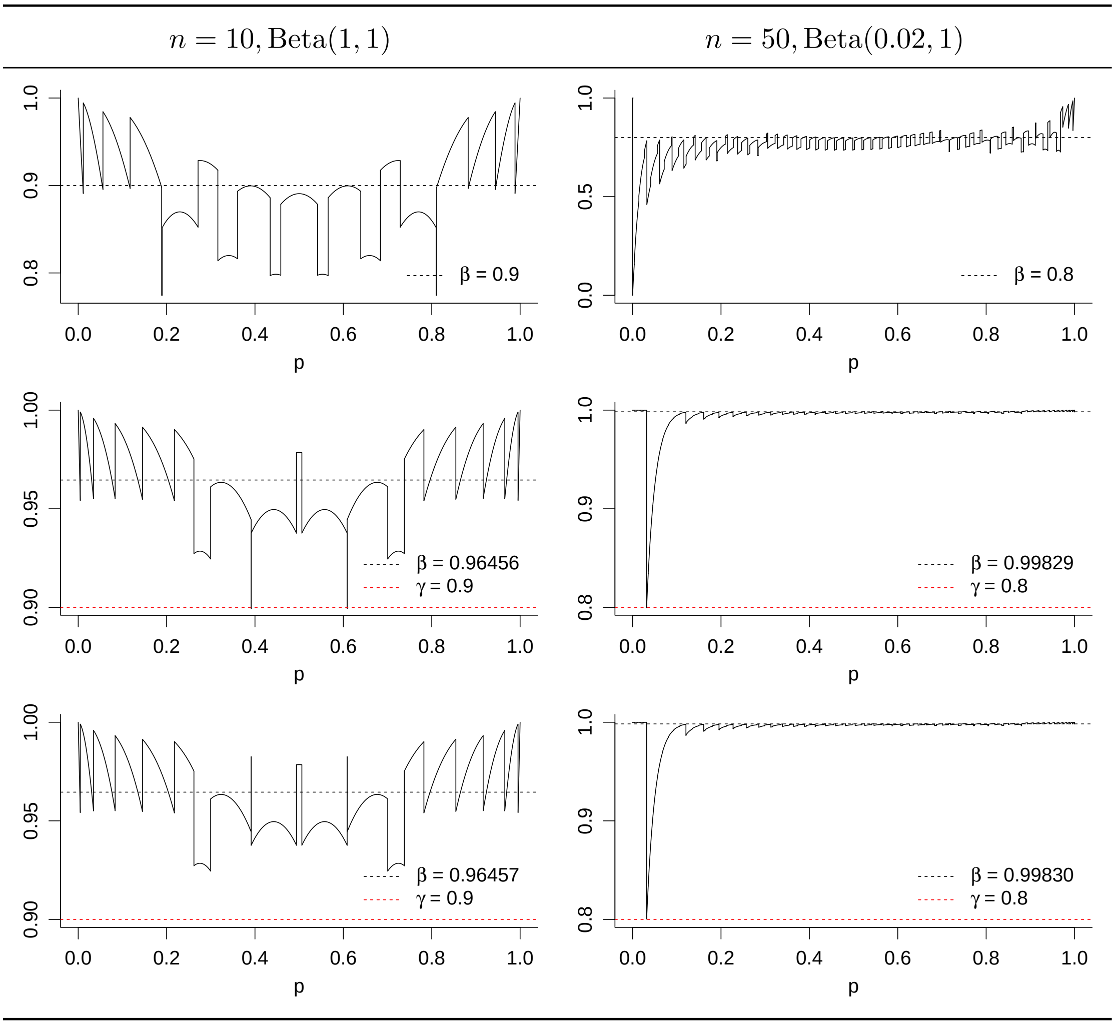


Figure 3.3: Coverage probabilities of the HPD credibility intervals for two different priors and sample sizes under various credibility levels β. In the first two rows, the coverage probabilities do not always exceed a level γ, while in the third row they do.

of a certain prescribed coverage probability fails only for a few single points of p. However, the property that the coverage probability function behaves in a discontinuous way and meets the actual prescribed level exactly only for a finite number of points between $[0;1]$ and clearly exceeds it for the majority of points, is just as well a characteristic of the frequentist minimum volume confidence interval for a binomial proportion, see e. g. Fig. 2.3.

Since it is somewhat difficult to imagine what a certain magnitude of the weighted volume actually means, we explore the length of the confidence and credibility intervals exemplarily under $X = x$ with $x = 0, 1, 2$. These are values of X that should be considered rather likely if right-skewed prior distributions $\mathrm{Beta}(a, 1)$ with $a < 1$ are

Table 3.1: Minimum credibility level β such that the level β HPD credibility interval has coverage probability $\geq \gamma$ with corresponding volume of the level β Bayesian MPS (left in brackets) and the volume of the minimum volume frequentist level γ MPS (right in brackets).

γ	$n = 10$	$n = 50$	$n = 100$	$n = 150$
Beta(1, 1)				
0.8	0.91525 (0.375; 0.336)	0.86967 (0.163; 0.149)	0.85256 (0.112; 0.104)	0.84985 (0.091; 0.085)
0.9	0.96457 (0.451; 0.417)	0.93643 (0.199; 0.188)	0.93082 (0.140; 0.133)	0.92891 (0.114; 0.108)
0.95	0.98453 (0.512; 0.476)	0.96894 (0.231; 0.221)	0.96976 (0.167; 0.157)	0.96859 (0.136; 0.128)
0.99	0.99531 (0.584; 0.580)	0.99437 (0.294; 0.286)	0.99432 (0.212; 0.204)	0.99395 (0.173; 0.167)
Beta(0.5, 0.5)				
0.8	0.93544 (0.331; 0.295)	0.93749 (0.162; 0.124)	0.93341 (0.114; 0.086)	0.93205 (0.093; 0.069)
0.9	0.97432 (0.404; 0.368)	0.97156 (0.191; 0.156)	0.97205 (0.137; 0.109)	0.97225 (0.113; 0.088)
0.95	0.98950 (0.464; 0.427)	0.98665 (0.216; 0.185)	0.98683 (0.155; 0.129)	0.98691 (0.128; 0.105)
0.99	0.99813 (0.562; 0.530)	0.99784 (0.270; 0.241)	0.99779 (0.192; 0.169)	0.99783 (0.158; 0.137)
Beta(0.5, 1)				
0.8	0.94162 (0.350; 0.299)	0.93741 (0.169; 0.128)	0.93339 (0.120; 0.089)	0.93204 (0.098; 0.072)
0.9	0.97679 (0.421; 0.373)	0.97234 (0.200; 0.162)	0.97244 (0.144; 0.114)	0.97241 (0.118; 0.092)
0.95	0.99028 (0.478; 0.432)	0.98710 (0.227; 0.192)	0.98706 (0.163; 0.135)	0.98706 (0.134; 0.109)
0.99	0.99855 (0.581; 0.539)	0.99785 (0.280; 0.250)	0.99785 (0.201; 0.176)	0.99786 (0.165; 0.143)
Beta(0.1, 1)				
0.8	0.99153 (0.245; 0.200)	0.99060 (0.100; 0.066)	0.99050 (0.069; 0.042)	0.99046 (0.056; 0.033)
0.9	0.99684 (0.302; 0.260)	0.99635 (0.118; 0.085)	0.99629 (0.080; 0.055)	0.99627 (0.064; 0.043)
0.95	0.99875 (0.353; 0.315)	0.99850 (0.135; 0.104)	0.99847 (0.091; 0.066)	0.99846 (0.072; 0.051)
0.99	0.99984 (0.457; 0.426)	0.99979 (0.171; 0.143)	0.99979 (0.113; 0.090)	0.99978 (0.089; 0.070)
Beta(0.05, 1)				
0.8	0.99602 (0.206; 0.176)	0.99557 (0.057; 0.034)	0.99551 (0.040; 0.022)	0.99550 (0.032; 0.017)
0.9	0.99853 (0.262; 0.235)	0.99831 (0.066; 0.044)	0.99828 (0.046; 0.028)	0.99827 (0.037; 0.022)
0.95	0.99943 (0.315; 0.290)	0.99932 (0.074; 0.053)	0.99930 (0.051; 0.034)	0.99930 (0.041; 0.027)
0.99	0.99993 (0.422; 0.400)	0.99991 (0.092; 0.073)	0.99991 (0.062; 0.047)	0.99990 (0.049; 0.036)
Beta(0.02, 1)				
0.8	0.99846 (0.175; 0.160)	0.99830 (0.027; 0.015)	0.99828 (0.019; 0.010)	0.99827 (0.015; 0.008)
0.9	0.99944 (0.232; 0.218)	0.99936 (0.031; 0.020)	0.99935 (0.021; 0.013)	0.99934 (0.017; 0.010)
0.95	0.99978 (0.284; 0.272)	0.99974 (0.035; 0.024)	0.99974 (0.024; 0.015)	0.99974 (0.019; 0.012)
0.99	0.99998 (0.409; 0.382)	0.99997 (0.044; 0.033)	0.99997 (0.029; 0.021)	0.99997 (0.023; 0.016)

imposed for the binomial probability. In Fig. 3.4, the lengths of the minimum volume level γ confidence intervals as functions of the first shape parameter a of the prior distribution Beta$(a, 1)$ under $X = 0, 1, 2$ are compared with the lengths of the level β credibility intervals. β has been chosen such that the Bayesian intervals have pointwise coverage of at least γ on $[0; 1]$. From Fig. 3.4 we can see that the level β as a function of the first shape parameter a is rather smooth. Under $X = 0$, the lengths of the minimum volume confidence intervals and the HPD interval are almost indiscernible. For $X = 1, 2$, the HPD interval is throughout wider than the minimum volume confidence interval for the investigated range of values for a. In general, the lengths of the intervals grow with $x = 0, 1, 2$. Apart from the intuitively plausible fact that under the higher sample size $n = 50$ the intervals are considerably shorter than under $n = 10$, it can be observed that the general relative behaviour of minimum volume confidence intervals and HPD intervals towards each other in the case $n = 50$ are very similar to the case $n = 10$.

3.7 Eliciting Prior Information

Precedingly we have described ways to construct frequentist confidence intervals (Chapter 2) and Bayesian (HPD) credibility intervals (Section 3.4) for a binomial proportion. Both approaches require the definition of a distribution that expresses prior information for p. The process of selecting an appropriate prior distribution is called *elicitation*. Elicitation means the acquisition of expert knowledge about the parameter of interest that is then translated to a statistical distribution that reflects the prior belief. See Kadane & Wolfson (1998) and Jenkinson (2005) for reviews on the topic.

Appropriately eliciting a prior distribution is frequently considered as one of the difficulties of Bayesian analysis. Berger (1985, p. 82) remarks that the determination of a prior distribution is naturally dependent on the ability of people. Since they often overestimate their prior knowledge, untrained elicitors can do quite poorly. To avoid misspecification or because information on the parameter of interest is lacking, *non-informative* prior distributions are frequently used. These distributions carry little to no information about the parameter in contrast to informative priors. The first therefore appear in the context of objective determination of priors, whereas the latter usually are outcomes of a subjective choice of prior distributions. To the most common non-informative priors belongs the uniform prior, which assigns equal probability mass to every point from the parameter space. For unbounded parameter spaces, this frequently produces so-called *improper priors* who's density integrated over the parameter space

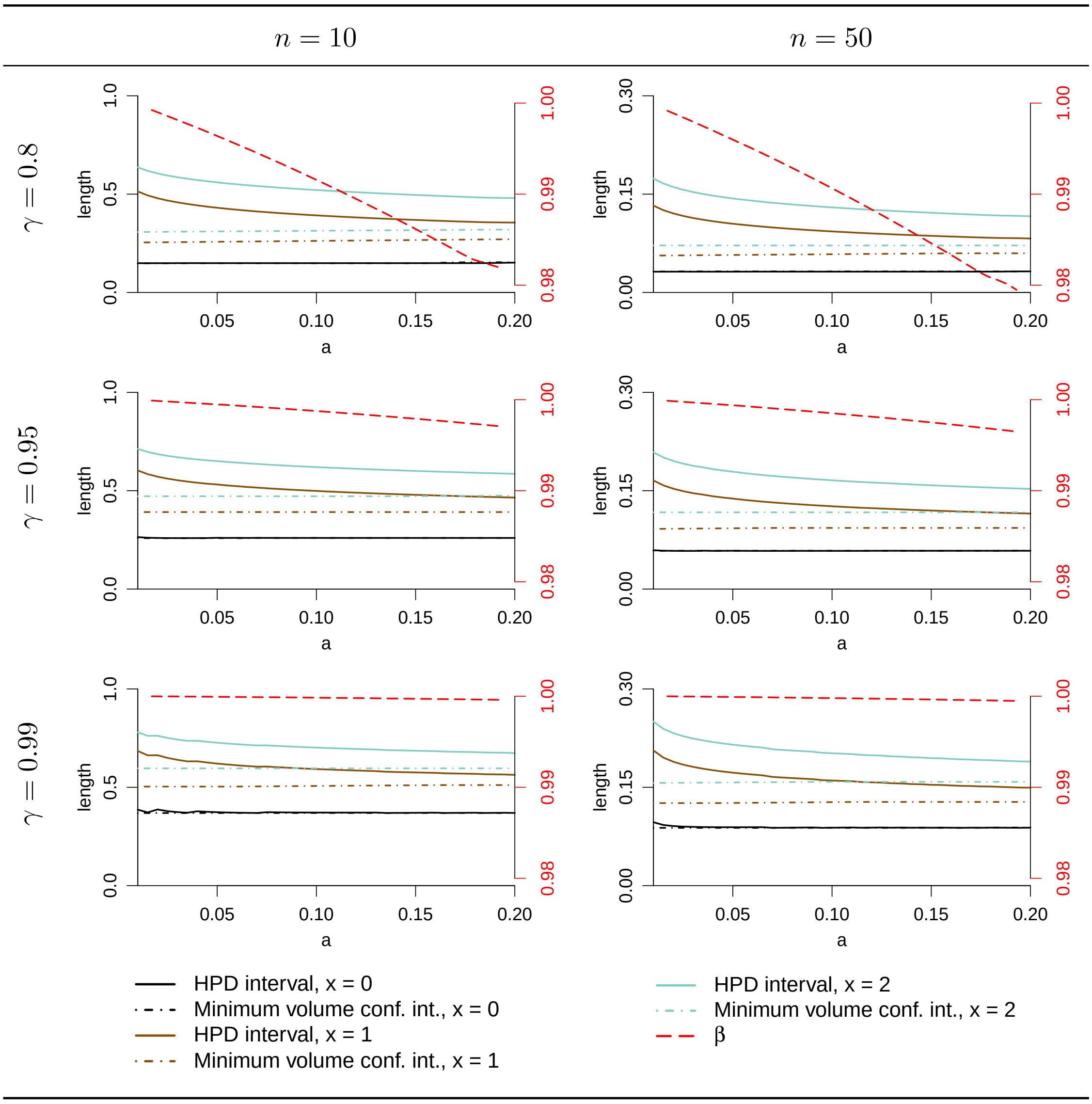


Figure 3.4: Length $u_x - l_x$ of the confidence/credibility interval $[l_x; u_x]$, i. e. under $X = x$ with $x = 0, 1, 2$, as a function of the first shape parameter a of the beta prior $\text{Beta}(a, 1)$. The credibility level β is chosen such that a pointwise coverage probability of $\geq \gamma$ is ensured; $\gamma \in \{0.8, 0.95, 0.99\}$.

integrates to infinity. Less this behaviour than the drawback that the uniform prior is not invariant under transformation led Jeffreys (1946) to work on priors that fulfil this invariance requirement and produce a prior that is defined by the root of the Fisher information matrix. However, non-informative priors involve not only these two kinds of priors. Often, many of them exist, which is one issue that raises criticism with respect to their use, see Berger (1985).

Since, according to Berger (1985, p. 90), many Bayesians believe only in proper priors, real elicitation of prior distributions does not so much mean selecting a non-informative, but an *informative* prior. Berger (1985) describes various general ways of how to elicit an informative prior distribution. We revise the most important ones.

In the *histogram approach*, the parameter space is devided into a number of intervals and the probabilities that the parameter of interest falls into these intervals are seeked for. A smooth density function is then matched to the resulting histogram.

The *relative likelihood approach* is, beneath the histogram approach, the one that is favoured by Berger (1985). Here, the likelihoods of several points are obtained by comparing them relatively to each other, for example, one point could be considered twice as likely as some other point. A smooth density function is then determined from the likelihood of a couple of points (Berger 1985, pp. 77–78).

A way to elicit a prior distribution that is often considered problematic by Berger (1985) is the *determination via the density* function. It involves prescribing a certain functional form of the prior density function, e. g. a beta distribution Beta(a, b) with parameters $a, b > 0$, and choosing the distribution parameters appropriately. The parameters are determined by means of an estimation of moments or quantiles. Berger's (1985) criticism is that the tails of the distribution, although they might be carrying little probability mass, can have a lot of effect on the moments. He therefore finds the moment approach suspect, but somewhat reasonable for bounded parameter spaces, and considers the way through estimating quantiles as more attractive (Berger 1985, p. 79).

The quantile technique can also be applied to the cumulative distribution function, where a smooth distribution function is searched for that matches several pre-specified quantiles as good as possible. This approach is referred to by Berger (1985, pp. 81–82) as the *determination via the cumulative distribution function* approach.

With respect to the choice of a prior distribution for a binomial parameter p, we consider in Chapter 2 and Section 3.4 only beta priors. We are therefore in the situation of eliciting a distribution for a univariate parameter on a bounded parameter space.

In Chapter 2 we express prior information on a probability p in terms of the four-parametric beta distribution $\mathrm{Beta}(p_0, p_1, a, b)$ on a subset $[p_0; p_1]$ of the unit interval, while in Section 3.4 we concentrate on the most important special case of a beta distribution $\mathrm{Beta}(0, 1, a, b) = \mathrm{Beta}(a, b)$ on the support $[0; 1]$. To restrict attention to the interval $[p_0; p_1]$ only is reasonable if probabilities from $[0; 1] \backslash [p_0; p_1]$ can be excluded with certainty. The disadvantage of the clear cut-off at the boundaries of $[p_0; p_1]$ has been warned against in Section 2.3 already. Setting $[p_0; p_1] = [0; 1]$ would at least provide against the risk of a complete misspecification of the prior information by allowing positive probabilities on the interval $(0; 1)$. After the choice of the support of the beta prior distribution, the shape parameters $a, b > 0$ need to be decided upon. In the following, we focus on the beta distribution $\mathrm{Beta}(a, b)$ and the elicitation of its two parameters a and b defined on the whole unit interval $[0; 1]$.

Several non-informative priors exist for the binomial parameter, see Berger (1985, p. 89) or Geisser (1984). To those that are instances of beta distributions belong the uniform prior $\mathrm{Unif}(0, 1)$ and the Jeffreys prior $\mathrm{Beta}(0.5, 0.5)$.

To elicit an informative beta distribution as a prior for the binomial parameter p, the following interpretation taken from Kerman (2011) can be helpful. It regards the posterior distribution $\mathrm{Beta}(x + a, n - x + b)$ under x observed binomial successes as an encouragement of the frequent interpretation of the parameters a and b of the beta prior $\mathrm{Beta}(a, b)$ as the prior number of successes and failures: While believing prior to drawing the sample that about a successes and b failures are most likely to be observed from a total of $a + b$ instances, the observed number of successes x in the actual sample of size n updates this prior belief to $a + x$ successes with $n - x + b$ failures out of a total of $a + b + n$ instances.

Ignoring the popular interpretation of the parameters a and b, we describe ways how to elicit a beta prior distribution for the binomial parameter p as instances of elicitation via the density function. In the first step, after having decided on the support, the shape of the beta density function needs to be determined. The density function of the beta distribution can be constant – in this case it equals the uniform distribution –, strictly increasing, strictly decreasing, bathtub shaped, or reverse bathtub shaped. Which shape the beta distribution takes in association with the parameter set (a, b) can be taken from Proposition 3.11.

The following Proposition 3.18 describes how to choose the parameters a, b of a beta distribution $\mathrm{Beta}(a, b)$ by means of specifying a quantile and deciding on the monotonicity of the density function.

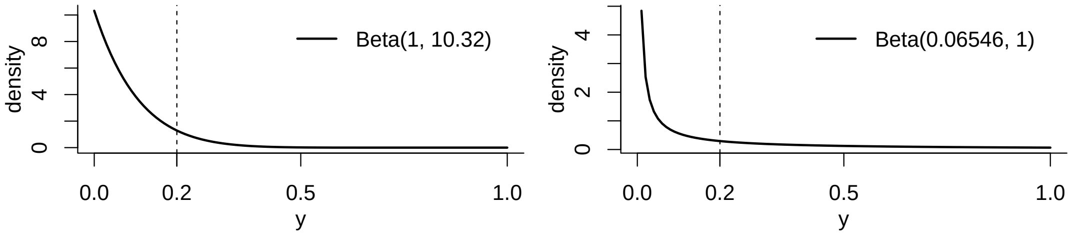


Figure 3.5: Two beta distributions with strictly decreasing density functions on $[0;1]$ with 0.2 as the 90 % quantiles.

Proposition 3.18. *Let* $p_\rho \in (0;1), \rho \in (0;1)$*. If*

(i) $a = 1, b = \frac{\ln(1-\rho)}{\ln(1-p_\rho)}$, *or*

(ii) $b = 1, a = \frac{\ln(\rho)}{\ln(p_\rho)}$,

then $p_\rho \in (0;1)$ *is the* $\rho \cdot 100\,\%$*-quantile of the beta distribution Beta*(a,b)*.*

If $p_\rho = \rho$*, the beta distribution is the uniform distribution on* $[0;1]$*. If* $0 < p_\rho < \rho < 1$*, the beta distribution is strictly decreasing. If* $0 < \rho < p_\rho < 1$*, the beta distribution is strictly increasing.*

Proof. See Appendix 3.A, Section 3.A.8. □

We provide an instance of the elicitation approach described in Proposition 3.18.

Example 3.19 (Specifying a Quantile)**.** *Assume, the prior distribution is expected to be strictly decreasing on* $[0;1]$*. The probability that the unknown parameter p is at most 0.2 is estimated to 90 %, which leaves the probability that p exceeds 0.2 at 10 %. By setting either* $a = 1$ *(see (i) in Proposition 3.18) or* $b = 1$ *(see (ii) in Proposition 3.18), we obtain the beta distributions Beta*$(1, 10.32)$ *and Beta*$(0.06546, 1)$*, respectively, as possible prior distributions.*

Example 3.19 is illustrated by Fig. 3.5, where both elicited beta densities are displayed. Besides specifying a quantile and deciding on a monotonously increasing/decreasing or constant density function, we can take one of the following two elicitation approaches.

(a) Specify two quantiles $\rho_1 < \rho_2 \in (0;1)$ with corresponding probabilities $p_{\rho_1} < p_{\rho_2}$, i. e. $\rho_1 = z_{\text{Beta}(a,b)}(p_{\rho_1})$ and $\rho_2 = z_{\text{Beta}(a,b)}(p_{\rho_2})$.

(b) Specify a quantile $\rho \in (0;1)$ with corresponding probability p_ρ, i. e. $\rho = z_{\text{Beta}(a,b)}(p_\rho)$, and the mean $\frac{a}{a+b}$ of the beta distribution Beta(a,b).

In both cases, we do not need to assume monotonicity properties for they will become clear from the assumptions on the quantiles/the mean. The parameters a, b of the

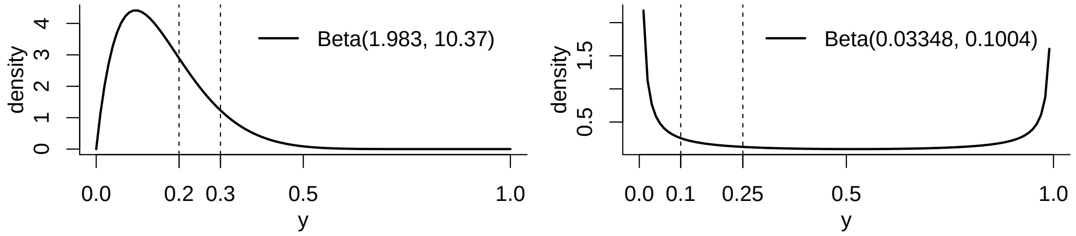


Figure 3.6: Beta distribution with 70 % quantile 0.2 and 90 % quantile 0.3 (left-hand side) and beta distribution with 70 % quantile 0.1 and mean 0.25 (right-hand side).

resulting beta prior distribution can be solved for numerically. We provide examples for both approaches.

Example 3.20 (Specifying two Quantiles)**.** *Assume, the probability that the unknown parameter p is smaller or equal to 0.2 is estimated to 70 % and the probability that p is smaller or equal to 0.3 is estimated to 90 %. The beta distribution fulfiling these conditions is given by* $Beta(1.983, 10.37)$*.*

The unimodal beta prior distribution obtained from Example 3.20 is depicted in the left-hand side of Fig. 3.6.

Example 3.21 (Specifying one Quantile and the Mean)**.** *Assume, the probability that the unknown parameter p is smaller or equal to 0.1 is estimated to 70 % and 0.25 is estimated to be the mean of the beta distribution. The beta distribution fulfiling these conditions is given by* $Beta(0.03348, 0.1004)$*.*

The prior distribution arising from the assumptions in Example 3.21 has an asymmetrical bathtub shape. The distribution is illustrated in the right-hand side of Fig. 3.6.

3.8 Conclusion and Outlook

We have reviewed Bayesian credibility regions and derived HPD credibility intervals for a binomial probability p. The Bayesian approach is the natural approach to make use of prior information, which is why we have conducted a comparison with the confidence intervals for a binomial probability presented in Chapter 2, which also make use of prior information, but are of frequentist type. Both approaches apply the conjugate prior of the binomial – the beta distribution – as prior information distribution on the unknown parameter p.

The concept of measurement and prediction spaces (MPSs) presented in Chapter 2 that simultaneously looks at confidence and prediction regions, can be formulated for the

Bayesian approach as well. Consequently, the credibility intervals can be evaluated by means of their coverage probability. The Bayesian credibility intervals do not in general fulfil the common frequentist requirement that under repeated sampling the parameter of interest is covered about a prescribed number of times by the intervals (*Repeated Sampling Principle*). In this sense, the Bayesian intervals frequently do not feature exactness: Under a given level $\beta \in (0;1)$, the Bayesian credibility intervals for a binomial proportion show a coverage probability below β for many $y = p$ from $[0;1]$. However, the graphs of the coverage probability function can take very different shapes under different prior distributions. We have demonstrated that for prior distributions $\text{Beta}(a,b)$ with parameters $0 < a, b \leq 1$, a level $\beta \in (0;1)$ can be found such that the credibility intervals are exact with respect to a prescribed minimum coverage level $\gamma \in (0;1)$. A comparison of the weighted volume of the Bayesian MPS with the frequentist minimum volume MPS under the same prior information have revealed that the latter exhibit in general smaller weighted volumes. This finding is not surprising since the frequentist intervals are constructed in such a way that their weighted volumes under a given prior information distribution are minimal.

We are aware that Bayesian credibility intervals are usually not judged by the concept of coverage probability, which is a frequentist measure of evaluating confidence intervals. Bayesian statistics follows the likelihood principle. In this sense, the comparison we presented is not completely fair and the credibility level β and the confidence level γ are not the same either. Scricciolo (1999), for example, raises the concern in the context of matching priors that designing priors such that final answers satisfy frequentist properties come with the risk of putting emphasis mainly on frequentist requirements and by that are in danger of violating the likelihood principle of the Bayesian approach. The findings from Theorem 3.16 are certainly in danger of falling victim to this criticism.

However, there is also the other point of view that these two conceptually different approaches – Bayesian and frequentist – "... should not be yielding fundamentally different answers in practice" (Berger et al. 1997): Something would be seriously wrong if a Bayesian credibility interval of level 90 % contained the true parameter in only 70 % of the cases (Bayarri & Berger 2004). On the other hand, Bayarri & Berger (2004) refer to Neyman (1977) when they point out that the motivation of the frequentist principle means the repeated application on different real problems and not on one problem with a fixed unknown parameter. This refers to the concept of the Bayesian approach that in contrast to the frequentist one does not treat the true parameter as an unknown constant, but as a random variable.

As to which of the two approaches to trust, we doubt a definite answer can ever be given. With respect to the Bayesian approach, an object of criticism is the fact that it requires the choice of a prior distribution and therefore can appear arbitrary (Scricciolo 1999). Particular difficulty arises in the Bayesian approach when prior information is not available or vague (Thatcher 1964). Here, the frequentist approach presented in Chapter 2 has the advantage that a misspecification of the prior distribution does not totally lead to catastrophic results: A minimal coverage is ensured under any prior information. The weighing induced by the prior can at most lead to intervals that are wider for certain x than they had been if the prior information had been chosen more aptly.

There are obvious disadvantages of the frequentist exact confidence interval approach. While the Bayesian credibility intervals can be easily calculated based on a closed formula, closed formulas do not exist for the frequentist minimum volume confidence intervals. Their calculation requires the effort of a numerical algorithm. Furthermore, a lot of emphasis is put on single points to ensure that a sufficient coverage is provided at any point in the parameter space $[0; 1]$. We can see from the numerical results in Section 3.6 that if it were not for a small number of isolated points, a credibility level β that leads to credibility intervals fulfiling the coverage criterion could more easily be found. This is certainly a particular property of the discreteness of the binomial distribution, which probably also encouraged Thatcher (1964) to say: “From the mathematical point of view, predictions about binomial samples are not the easiest of cases to consider” (Thatcher 1964, p. 192).

Taking it from there, one suggestion for further research would be to loosen this restriction of pointwise coverage and allow for a “smooth” coverage probability function in the comparison. Currently, the performance in this respect is judged by the behaviour of the function $C(y)$ at the point of the infimum $\inf_{y \in R_2} C(y)$. A less conservative approach would be to act in the spirit of Woodroofe (1986) and Wang (2009) and consider an average coverage probability. This involves averaging the coverage probability over the whole parameter space R_2 of Y or a subset thereof, for example in an environment of the parameter of interest. Comparisons between the Bayesian and the frequentist coverage probability would then have to be based on the average coverage probability, which, according to Woodroofe (1986), may better assess frequentist properties than the strictly pointwise evaluated measure.

The empirical results from Section 3.6 have been calculated numerically. It should be explored how by identifying the points of minimum coverage probability, the minimum probability level β ensuring a pointwise coverage probability of $\geq \gamma$ can be determined analytically.

Regardless of the question concerning the coverage, it would be of interest what the difference in information content of both approaches would be and how exactly prior information impacts credibility and confidence intervals.

3.A Appendix

3.A.1 Proof of Proposition 3.9

The prior distribution of Y – the beta distribution $\text{Beta}(a, b)$ with parameters $a, b > 0$ from Eq. (2.11) – has the density function $f_Y(y) = \frac{1}{B(a,b)} y^{a-1}(1-y)^{b-1}$.

By $f_{X|Y=y}(x) f_Y(y) = f_X(x)$ with $f_{X|Y=y}(x) = \binom{n}{x} y^x (1-y)^{n-x}$ we obtain the marginal density $f_X(x)$ as

$$f_X(x) \quad = \quad \frac{\binom{n}{x}}{B(a,b)} \int_0^1 y^{x+a-1}(1-y)^{n-x+b-1} \,\mathrm{d}y.$$

Using Theorem 3.1, we obtain the posterior density

$$\begin{aligned} f_{Y|X=x}(y) \quad &= \quad \frac{\frac{1}{B(a,b)} y^{a-1}(1-y)^{b-1} \binom{n}{x} y^x (1-y)^{n-x}}{\frac{\binom{n}{x}}{B(a,b)} \int_0^1 z^{x+a-1}(1-z)^{n-x+b-1} \,\mathrm{d}z} \\ &= \quad \frac{1}{\int_0^1 z^{x+a-1}(1-z)^{n-x+b-1} \,\mathrm{d}z} y^{a+x-1}(1-y)^{n-x+b-1} \\ &= \quad \frac{1}{B(x+a, n-x+b)} y^{a+x-1}(1-y)^{n-x+b-1}, \end{aligned}$$

i. e. the density of the beta distribution $\text{Beta}(x + a, n - x + b)$.

3.A.2 Proof of Proposition 3.11

Consider $x \in (0; 1)$. We derive the monotonicity properties of the beta density function. The derivative of $f(x)$ with respect to x is

$$f'(x) \quad = \quad \underbrace{\frac{1}{B(a,b)} x^{a-2}(1-x)^{b-2}}_{>0} \Big((a-1)(1-x) - (b-1)x \Big).$$

Hence, to investigate the monotonicity properties of $f(x)$, we analyse the sign of $(a - 1)(1 - x) - (b - 1)x$. In the case $b \leq 1, a > 1$ we have

$$\begin{aligned} (a-1)(1-x) - (b-1)x \quad &= \quad a - 1 - ax + x - \underbrace{b}_{\leq 1} x + x \\ &\geq \quad a - 1 - ax + x - x + x \\ &= \quad (a-1) - (a-1)x \quad = \quad \underbrace{(a-1)}_{>0} \underbrace{(1-x)}_{>0} > 0. \end{aligned}$$

In the case $b < 1, a = 1$ we have

$$\begin{aligned}(a-1)(1-x)-(b-1)x &= a-1-ax+x-bx+x\\ &= 1-1-x+x-bx+x\\ &= -bx+x \quad = \quad \underbrace{(1-b)}_{>0}\underbrace{x}_{>0} \quad > \quad 0.\end{aligned}$$

Consequently, $f(x)$ is strictly increasing on $(0;1)$ and assertion (a) follows.

In the case $a = b = 1$, we have $(a-1)(1-x)-(b-1)x = 0$ and hence $f'(x) \equiv 0$ on $(0;1)$. Therefore $f(x)$ is constant on $(0;1)$. Assertion (b) follows.

In the case $a < 1 < b$ we have

$$\begin{aligned}(a-1)(1-x)-(b-1)x &= a-1-ax+x-bx+x\\ &< a-1-ax+x-x+x\\ &= \underbrace{(a-1)}_{<0}\underbrace{(1-x)}_{>0} \quad < \quad 0,\end{aligned}$$

i. e. $f(x)$ is strictly decreasing. In the case $a = 1, b > 1$ we have

$$(a-1)(1-x)-(b-1)x = -\underbrace{(b-1)}_{>0}\underbrace{x}_{>0} < 0.$$

In the case $a < 1, b = 1$ we have

$$(a-1)(1-x)-(b-1)x \quad = \quad \underbrace{(a-1)}_{<0}\underbrace{(1-x)}_{>0} < 0.$$

This proves assertion (c).

In the case $b > 1, a > 1$ we have

$$\begin{aligned} & (a-1)(1-x)-(b-1)x && \substack{<\\=\\>} \quad 0\\ \Leftrightarrow \quad & (a-1)(1-x) && \substack{<\\=\\>} \quad (b-1)x\\ \Leftrightarrow \quad & \frac{1-x}{x} && \substack{<\\=\\>} \quad \frac{b-1}{a-1}\\ \Leftrightarrow \quad & \frac{1}{x} && \substack{<\\=\\>} \quad \frac{b-1}{a-1}+1\\ \Leftrightarrow \quad & x && \substack{>\\=\\<} \quad \frac{a-1}{b+a-2}\begin{cases} < \dfrac{a-1}{1+a-2} \;=\; 1,\\ > 0.\end{cases}\end{aligned}$$

Hence $f(x)$ is strictly increasing for $x < x^*$ and strictly decreasing for $x > x^*$, where $x^* = \frac{a-1}{a+b-2}$. Assertion (d) follows.

In the case $a < 1, b < 1$ we have

$$\begin{aligned}
& (a-1)(1-x)-(b-1)x && \substack{<\\=\\>} && 0 \\
\Leftrightarrow\quad & (a-1)(1-x) && \substack{<\\=\\>} && (b-1)x \\
\overset{a-1<0}{\Leftrightarrow}\quad & \frac{1-x}{x} && \substack{>\\=\\<} && \frac{b-1}{a-1} \\
\Leftrightarrow\quad & \frac{1}{x} && \substack{>\\=\\<} && \frac{b-1}{a-1}+1 \\
\Leftrightarrow\quad & x && \substack{<\\=\\>} && \frac{1}{\underbrace{\frac{b-1}{a-1}}_{>0}+1} \begin{cases} < 1, \\ > 0. \end{cases}
\end{aligned}$$

Assertion (e) follows.

3.A.3 Proof of Corollary 3.12

The assertions of Corollary 3.12 follow directly from Proposition 3.9 about the form of the posterior density $f_{Y|X=x}(y)$ under a beta prior and Proposition 3.11 about the monotonicity of the beta density function. To have $f_{Y|X=x}(y)$ strictly decreasing on a interval $(0; x^*)$ and strictly increasing on $(x^*; 1)$ requires according to Proposition 3.11 $\underbrace{a}_{>0} + \underbrace{x}_{\in\mathbb{N}_0} < 1 \Rightarrow x = 0$ and $n - x + b = n + b < 1$, a contradiction to $n \geq 1$ and $b > 0$. This proves assertion (d).

Assume, $f_{Y|X=x}(y)$ is constant on $[0; 1]$. Then from Proposition 3.9 together with assertion (b) from Proposition 3.11 it follows that $n - x + b = a + x = 1$. With the right-hand side of the equality and $a > 0$ follows $x = 0$. With the left-hand side of the equality follows $n - x + b = \underbrace{n}_{\in\mathbb{N}} + b \overset{!}{=} 1$, and therefore $b = 0$, a contradiction to the definition of the shape parameter b. This proves assertion (e).

3.A.4 Proof of Corollary 3.13

The assertions (a) and (b) follow directly from Corollary 3.12. In assertion (c), $y^* \in A_x$ follows from $f_{Y|X=x}(y^*) \geq f_{Y|X=x}(y)$ for every $y \in [0; 1]$. The existence of a value $f_{Y|X=x}(z_{\text{Beta}(x+a,n-x+b)}(\alpha_1)) = f_{Y|X=x}(z_{\text{Beta}(x+a,n-x+b)}(1-\alpha_2)) \in (0; y^*)$ with $\text{P}(z_{\text{Beta}(x+a,n-x+b)}(\alpha_1) \leq Y \leq z_{\text{Beta}(x+a,n-x+b)}(1-\alpha_2)) = \beta \in (0; 1)$ follows from the continuity of the density of the posterior distribution $\text{Beta}(x + a, n - x + b)$ on $[0; 1]$ and $f_{Y|X=x}(0) = f_{Y|X=x}(1) = 0$.

3.A.5 Proof of Proposition 3.15

We prove the monotonicity property of the quantiles of the beta distribution $\mathrm{Beta}(a+x, n-x+b)$. For an arbitrary $x \in \{0,1,\dots,n\}$ let $z = z_{\mathrm{Beta}(a+x,n-x+b)}(\gamma) \in (0;1)$ be the $\gamma \cdot 100\%$ quantile of the beta distribution $\mathrm{Beta}(a+x, n-x+b)$. Then we have by definition of a quantile

$$\gamma \quad = \quad \frac{1}{B(a+x,n-x+b)} \int_0^z y^{a+x-1}(1-y)^{n-x+b-1}\,\mathrm{d}y \quad = \quad I_z(a+x,n-x+b),$$

where

$$I_z(v,w) \quad = \quad \frac{1}{B(v,w)} \int_0^z t^{v-1}(1-t)^{w-1}\,\mathrm{d}t, \quad 0 \le z \le 1$$

is the *incomplete beta function*, see Abramowitz & Stegun (1972, Formula 26.5.1). From Abramowitz & Stegun (1972, Formula 26.5.15) we obtain the recurrence relation

$$I_z(v,w) \quad = \quad \frac{\Gamma(v+w)}{\Gamma(v+1)\Gamma(w)} z^v (1-z)^{w-1} + I_z(v+1,w-1).$$

Consequently, we obtain with $v = a+x, w = n-x+b$

$$\begin{aligned}\gamma &= I_z(a+x,n-x+b)\\ &= \underbrace{\frac{\Gamma(a+n+b)}{\Gamma(a+x+1)\Gamma(n-x+b)}}_{>0}\underbrace{z^{a+x}}_{>0}\underbrace{(1-z)^{n-x+b-1}}_{>0} + I_z(a+x+1,n-x+b-1)\end{aligned}$$

$$\begin{aligned}\Leftrightarrow \quad \gamma &= \frac{1}{B(a+x,n-x+b)} \int_0^z y^{a+x-1}(1-y)^{n-x+b-1}\,\mathrm{d}y\\ &> \frac{1}{B(a+x+1,n-x+b-1)} \int_0^z y^{a+x}(1-y)^{n-x+b-2}\,\mathrm{d}y.\end{aligned}$$

Therefore necessarily $z_{\mathrm{Beta}(a+(x+1),n-(x+1)+b)}(\gamma) > z_{\mathrm{Beta}(a+x,n-x+b)}(\gamma) = z$ and the assertion follows.

3.A.6 Proof of Theorem 3.16

1. $a \le 1, b \le 1$ is a necessary condition. Let $b > 1$ and $0 < \beta < 1$ be arbitrary. The level β HPD credibility intervals A_x are intervals of the form $A_x = [l_x; u_x]$. Let $u_{max} := \max\{u_x | x \in \{0,1,\dots,n\}\}$ and $x_{max} := \max\{x | u_x = u_{max}\}$.

From Corollary 3.12 (a) we obtain $n - x_{max} + b \ge b > 1$, and hence the posterior density function $f_{Y|X=x_{max}}(y)$ is not strictly increasing on $[0;1]$. Consequently, $f_{Y|X=x_{max}}(y)$ is either strictly decreasing on $[0;1]$ or of inverted bathtub shape, i. e. strictly increasing on

$\left(0; \frac{a+x_{max}-1}{n+a+b-2}\right)$ and strictly decreasing on $\left(\frac{a+x_{max}-1}{n+a+b-2}; 1\right)$, see Corollary 3.12. The corresponding HPD credibility interval is the interval $A_{x_{max}} := [l_{x_{max}}; u_{x_{max}}]$. Hereby $u_{x_{max}} < 1$, since in the case $u_{x_{max}} = 1$ we would have $f_{Y|X=x}(u_{x_{max}}) = 0 = f_{Y|X=x}(l_{x_{max}})$ and $[l_{x_{max}}; u_{x_{max}}] = [0; 1]$, which would mean $\beta = 1$, a contradiction to the assumption. Hence $u_{x_{max}} < 1$ and $f_{Y|X=x_{max}}(l_{x_{max}}) > 0$, $f_{Y|X=x_{max}}(u_{x_{max}}) > 0$. Let $1 > p > u_{x_{max}}$ (e. g. $p = \frac{u_{x_{max}}+1}{2}$). Then due to the definition of an HPD credibility interval we have $f_{Y|X=x_{max}}(p) < \min\left\{f_{Y|X=x_{max}}(l_{x_{max}}), f_{Y|X=x_{max}}(u_{x_{max}})\right\}$ and $p \notin [l_{x_{max}}; u_{x_{max}}]$ and since $u_{x_{max}}$ is the maximum of all upper bounds u_x, we also find $p \notin [l_x; u_x]$ for all $x \in \{0, 1, \ldots, n\}$. Consequently, $A_p = \{x|(x, p) \in A\} = \emptyset$ and $C(p) = 0$. Hence, for every $0 < \beta < 1$ we find a $p \in [0; 1]$ such that $C(p) = 0$. Therefore we have the necessary condition $b \leq 1$.

In the same manner we prove $a \leq 1$:

Let $a > 1$ and $0 < \beta < 1$ be arbitrary. The level β HPD credibility intervals A_x are intervals with $A_x = [l_x; u_x]$. Let $l_{min} := \min\{l_x|x \in \{0, 1, \ldots, n\}\}$ and $x_{min} := \max\{x|l_x = l_{min}\}$.

From Corollary 3.12 (a) we obtain $a + x \geq a > 1$, and hence the posterior density function $f_{Y|X=x_{min}}(y)$ is not strictly decreasing on $[0; 1]$. Consequently, $f_{Y|X=x_{min}}(y)$ is either strictly increasing on $[0; 1]$ or of inverted bathtub shape, i. e. strictly increasing on $\left(0; \frac{a+x_{min}-1}{n+a+b-2}\right)$ and strictly decreasing on $\left(\frac{a+x_{min}-1}{n+a+b-2}; 1\right)$, see Corollary 3.12. The corresponding HPD credibility interval is the interval $[l_{x_{min}}; u_{x_{min}}]$. Hereby $l_{x_{min}} > 0$, since in the case $l_{x_{min}} = 0$ we would have $f_{Y|X=x}(l_{x_{min}}) = 0 = u_{x_{min}}$ and $[l_{x_{max}}; u_{x_{max}}] = [0; 1]$, which would mean $\beta = 1$, a contradiction to the assumption. Hence $l_{x_{min}} > 0$ and $f_{Y|X=x_{min}}(l_{x_{min}}) > 0$, $f_{Y|X=x_{min}}(u_{x_{min}}) > 0$. Let $0 < p < l_{x_{min}}$ (e. g. $p = \frac{l_{x_{min}}}{2}$). Then due to the definition of a credibility interval we have $f_{Y|X=x_{min}}(p) < \min\{f_{Y|X=x_{min}}(l_{x_{min}}), f_{Y|X=x_{min}}(u_{x_{min}})\}$ and $p \notin [l_{x_{min}}; u_{x_{min}}]$ and since $l_{x_{min}}$ is the minimum of all lower bounds l_x, we also find $p \notin [l_x; u_x]$ for all $x \in \{0, 1, \ldots, n\}$. Consequently, $A_p = \{x|(x, p) \in A\} = \emptyset$ and $C(p) = 0$. Hence, for every $0 < \beta < 1$ we find a $p \in [0; 1]$ such that $C(p) = 0$. Therefore we have the necessary condition $a \leq 1$.

2. $a \leq 1, b \leq 1$ is a sufficient condition. Let $a \leq 1, b \leq 1$. Then we know from Corollary 3.12 that under $X = 0$ the posterior density function $f_{Y|X=0}(y)$ is strictly decreasing on $[0; 1]$, since $a \leq 1 < \underbrace{n}_{\geq 1} + b = n - x + b$. Similarly, the posterior density function $f_{Y|X=n}(y)$ is strictly increasing on $[0; 1]$ for $X = n$, since the condition $n - x + b \leq 1 < a + x = \underbrace{a}_{>0} + \underbrace{n}_{\geq 1}$ is fulfiled.

Define $q_1 \in [0;1]$ such that $\binom{n}{0} q_1^0 (1-q_1)^n = (1-q_1)^n = \gamma \quad \Leftrightarrow \quad q_1 = 1-\gamma^{1/n} \in (0;1)$. Then due to the monotonicity of the function $(1-y)^n$ in y on $[0;1]$, we have $C(y) = \mathrm{P}_y(X \in A_y) \geq \mathrm{P}_y(0 \in A_y) \geq \gamma$ for $y \in [0;q_1]$.

Define $q_2 \in [0;1]$ such that $\binom{n}{n} q_2^n (1-q_2)^0 = q_2^n = \gamma \quad \Leftrightarrow \quad q_2 = \gamma^{1/n} \in (0;1)$. Then due to the monotonicity of the function y^n in y on $[0;1]$, we have $C(y) = \mathrm{P}_y(X \in A_y) \geq \mathrm{P}_y(n \in A_y) \geq \gamma$ for $y \in [q_2;1]$.

Define $0 < \beta < 1$ such that $z_2 := z_{\mathrm{Beta}(a+n,b)}(1-\beta) < \min\{q_1, q_2\}$ and $z_1 := z_{\mathrm{Beta}(a,n+b)}(\beta) > \max\{q_1, q_2\}$. We show that β is an appropriate level fulfiling the assertion from the theorem.

Since $[0;z_2] \subset [0;q_1]$ and $C(y) \geq \gamma$ for $y \in [0;q_1]$, we have $C(y) \geq \gamma$ for $y \in [0;z_2]$. Furthermore, since $[z_1;1] \subset [q_2;1]$ and $C(y) \geq \gamma$ for $y \in [q_2;1]$, we have $C(y) \geq \gamma$ for $y \in [z_1;1]$.

For $y \in (z_2;z_1)$, due to the definition of β, we have $0 \in A_y$ and $n \in A_y$. If $n = 1$, $C(y) = 1$ for $y \in (z_2;z_1)$ and the assertion follows. We show that for $n \geq 2$ under level β we have also $\{1, 2, \ldots, n-1\} \in A_y$ for $y \in (z_2;z_1)$.

Let $[l_x;u_x]$ denote the level β credibility region if $X = x$, $x \in \{0, 1, \ldots, n\}$. Then by the definition of β we have $l_0 = 0, u_0 = z_1$ and $l_n = z_2, u_n = 1$. For every $x \in \{1, \ldots, n-1\}$ we have

$$\int_{l_n}^{1} f_{Y|X=x}(y)\,\mathrm{d}y \overset{\text{Prop. 3.15}}{<} \int_{l_n}^{1} f_{Y|X=n}(y)\,\mathrm{d}y \quad = \quad \beta,$$

and therefore necessarily $l_x < l_n$. Furthermore,

$$\int_{0}^{u_0} f_{Y|X=x}(y)\,\mathrm{d}y \overset{\text{Prop. 3.15}}{<} \int_{0}^{u_0} f_{Y|X=0}(y)\,\mathrm{d}y \quad = \quad \beta,$$

and therefore necessarily $u_x > u_0$ for every $x \in \{1, \ldots, n-1\}$. So, for every $x \in \{1, \ldots, n-1\}$, the interval $(z_2;z_1)$ is contained in the credibility region $[l_x;u_x]$, i.e. $(z_2;z_1) \subset [l_x;u_x]$. Consequently, for every $y \in (z_2;z_1)$ and $x \in \{1, \ldots, n-1\}$ we have $x \in A_y$. Together with $0 \in A_y$ and $n \in A_y$ it follows $C(y) = \mathrm{P}_y(X \in A_y) = \mathrm{P}_y\left(X \in \{0, 1, \ldots, n\}\right) = 1 > \gamma$ for $y \in (z_2;z_1)$.

Consequently, $C(y) \geq \gamma$ for every $y \in [0;1]$ and β is an appropriate credibility level fulfiling the assertion.

This completes the proof of Theorem 3.16.

3.A.7 Proof of Proposition 3.17

Let $A_x^{(\beta)} = [l_x^{(\beta)}; u_x^{(\beta)}]$ denote the level β HPD interval for Y under $X = x$. Let $\beta_1, \beta_2 \in (0;1)$ be arbitrarily chosen with $\beta_2 > \beta_1$. For each $x \in \{0, 1, \ldots, n\}$, we have $A_x^{(\beta_1)} \subset A_x^{(\beta_2)}$ due to the properties of the HPD interval. Since the density functions of the posterior distribution $f_{Y|X=x}(y)$ for the binomial probability under a beta prior are never constant on any non-empty subinterval of $[0;1]$, we have $A_x^{(\beta_1)} \subsetneq A_x^{(\beta_2)}$ if $\beta_1 \lneq \beta_2$. Let $A^{(\beta)} = \{(x,y)|y \in A_x^{(\beta)}, x = \{0, 1, \ldots, n\}\}$. For the projection $A_y^{(\beta)} = \{x|(x,y) \in A^{(\beta)}\}$ we have $A_y^{(\beta_1)} \subset A_y^{(\beta_2)}$ for $\beta_1 < \beta_2$. Therefore, we can conclude for the coverage probability function $C(y)^{(\beta)} = \mathrm{P}(X \in A_y^{(\beta)})$:

$$C(y)^{(\beta_1)} \;=\; \mathrm{P}(X \in A_y^{(\beta_1)}) \;\leq\; \mathrm{P}(X \in A_y^{(\beta_2)}) \;=\; C(y)^{(\beta_2)},$$

and $C(y)^{(\beta)}$ is increasing in β.

The limiting characteristic $C(y)^\beta \overset{\beta\to 1}{\nearrow} 1$ follows from $l_x^{(\beta)} \searrow 0$ and $u_x^{(\beta)} \nearrow 1$ for $\beta \to 1$.

3.A.8 Proof of Proposition 3.18

Let X be a random variable distributed according to the beta distribution $\mathrm{Beta}(a, b)$.

(i) Let $a = 1$. Then

$$\begin{aligned}
& & \rho &= \mathrm{P}(X \leq p_\rho) \\
\Leftrightarrow & & \rho &= \frac{\Gamma(1+b)}{\Gamma(1)\Gamma(b)} \int_0^{p_\rho} y^{1-1}(1-y)^{b-1}\,\mathrm{d}y \\
\Leftrightarrow & & \rho &= \frac{b\Gamma(b)}{1 \cdot \Gamma(b)} \int_0^{p_\rho} y^{0}(1-y)^{b-1}\,\mathrm{d}y \\
\Leftrightarrow & & \rho &= b\int_0^{p_\rho} (1-y)^{b-1}\,\mathrm{d}y \quad = \quad b\left[-\frac{1}{b}(1-y)^b\right]_0^{p_\rho} \\
\Leftrightarrow & & \rho &= -\Big((1-p_\rho)^b - (1-0)^b\Big) \\
\overset{b>0}{\Leftrightarrow} & & (1-p_\rho)^b &= 1-\rho \\
\Leftrightarrow & & b\ln(1-p_\rho) &= \ln(1-\rho) \\
\Leftrightarrow & & b &= \frac{\ln(1-\rho)}{\ln(1-p_\rho)}.
\end{aligned}$$

(ii) Let $b = 1$. Then

$$\begin{aligned}
& \rho &=& \ \mathrm{P}(X \leq p_\rho) \\
\Leftrightarrow\quad & \rho &=& \ \frac{\Gamma(a+1)}{\Gamma(a)\Gamma(1)} \int_0^{p_\rho} y^{a-1}(1-y)^{1-1}\,\mathrm{d}y \\
\Leftrightarrow\quad & \rho &=& \ \frac{a\Gamma(a)}{\Gamma(a)\cdot 1} \int_0^{p_\rho} y^{a-1}\,\mathrm{d}y \\
\Leftrightarrow\quad & \rho &=& \ a\int_0^{p_\rho} y^{a-1}\,\mathrm{d}y \quad = \quad a\left[\frac{1}{a}y^a\right]_0^{p_\rho} \\
\Leftrightarrow\quad & \rho &=& \ p_\rho^a - 0^a \\
\overset{a>0}{\Leftrightarrow}\quad & \ln(\rho) &=& \ a\ln(p_\rho) \\
\Leftrightarrow\quad & a &=& \ \frac{\ln(\rho)}{\ln(p_\rho)}.
\end{aligned}$$

If $p_\rho = \rho$ we obtain with both (i) and (ii) $a = 1, b = 1$, and hence the beta distribution Beta$(1,1)$, i. e. the uniform distribution on $[0;1]$. If $0 < p_\rho < \rho < 1$, we obtain with (i) and $1-\rho < 1-p_\rho < 1$ that $a = 1$, $b = \dfrac{\ln(\overbrace{1-\rho}^{<1-p_\rho})}{\ln(1-p_\rho)} > 1$ and with (ii) $b = 1$, $a = \dfrac{\ln(\rho)}{\ln(p_\rho)} < 1$. In both cases, Beta$(a,b)$ is strictly decreasing according to Proposition 3.11. If $0 < \rho < p_\rho < 1$, we obtain with (i) and $1-p_\rho < 1-\rho < 1$ that $a = 1$, $b = \dfrac{\ln(\overbrace{1-\rho}^{>1-p_\rho})}{\ln(1-p_\rho)} < 1$ and with (ii) $b = 1$, $a = \dfrac{\ln(\rho)}{\ln(p_\rho)} > 1$. In both cases, Beta$(a,b)$ is strictly increasing according to Proposition 3.11.

4 Minimum Volume Confidence Intervals for the Poisson Parameter under Prior Information

4.1 Introduction

The Poisson distribution is a discrete distribution taking all positive integers as well as zero as possible outcomes. A random variable X that is distributed according to the Poisson distribution $Po(\lambda)$ with $\lambda > 0$, has the probability mass function

$$f_X(x) \;=\; \mathrm{P}(X = x) \;=\; \frac{\lambda^x}{x!}\exp(-\lambda) \text{ for } x = 0, 1, \ldots \tag{4.1}$$

and the distribution function $F_X(c)$ or operating characteristic (OC) function $L_c(\lambda)$ with

$$F_X(c) \quad = \quad \mathrm{P}(X \leq c) \;=\; \sum_{x=0}^{c} \frac{\lambda^x}{x!}\exp(-\lambda) \;=\; L_c(\lambda), \qquad \text{for } c = 0, 1, \ldots \tag{4.2}$$

See Johnson et al. (1992) for a definition and properties of the Poisson distribution.

The Poisson distribution is often used for modelling the occurrences of rare events by means of assuming an average of λ occurrences in a period of time. Since the Poisson distribution is the limiting distribution of the binomial distribution if $n \to \infty, p \to 0, np \to \lambda$, it is frequently used as an approximation of binomial probabilities, especially for large n and small p, see Eqs. (2.1) and (2.2) for the definition of the parameters of the binomial distribution. An important characteristic of a Poisson distributed random variable X is that its variance $V[X]$ and its expected value $E[X]$ take the same value λ (Johnson et al. 1992), i. e. *equidispersion* holds. Lewis (2004) describes case scenarios where the Poisson distribution is appropriate to model loss data. Several more examples where the Poisson distribution is frequently used as a model are given by Sahai & Khurshid (1993): Applications of the Poisson model are e. g. the number of radioactive counts per time unit, the number of birth defects or the number of victims suffering from specific diseases.

Lewis (2004) gives a rule of thumb to decide which of the three most important discrete distributions binomial, Poisson and negative binomial should be chosen given a certain application. According to this rule, the binomial distribution might be appropriate if the variance of the analysed phenomenon is lower than the arithmetic mean (i. e. underdispersion holds), the Poisson distribution is to be applied if the variance equals the arithmetic mean (i. e. equidispersion holds) and the negative binomial distribution might be a good choice if the variance clearly exceeds the arithmetic mean (i. e. overdispersion holds).

Confidence limits for the parameter λ of a Poisson distribution have been in practice since the 1930s. Przyborowski & Wileński (1935) provided upper confidence limits for λ for outcomes x of up to 50 for selected confidence levels. Garwood (1936) seems to have been the first to calculate lower confidence limits, too.

When considering two-sided confidence intervals, authors at the early stages of Poisson confidence intervals, as Garwood (1936) and Ricker (1937), focused on central intervals. These are intervals with equal tail probabilities, i. e. the probability of exceeding the upper or falling below the lower limit are both bounded by $(1-\gamma)/2$ for a confidence level $\gamma \in (0;1)$. We review the simple equal-tail interval B considered by Garwood (1936), see also Johnson et al. (1992), Sahai & Khurshid (1993) or Casella & Berger (2001). The interval fulfils the exactness criterion of a pointwise minimum coverage probability of at least γ, i. e. $\mathrm{P}_\lambda(\lambda \in B) \geq \gamma$ for arbitrary $\lambda \in (0;+\infty)$. Its calculation requires quantiles of the chi-square distribution.

Let $X \sim Po(\lambda)$ and $X = x$ be the observed value of X. An exact equal-tail confidence interval of level $\gamma \in (0;1)$ for the parameter λ is given by

$$(\lambda_L;\ \lambda_U) \quad = \quad \left(\frac{1}{2} z_{\chi^2_{2x}}\left(\frac{1-\gamma}{2}\right);\ \ \frac{1}{2} z_{\chi^2_{2(x+1)}}\left(\frac{1+\gamma}{2}\right)\right), \tag{4.3}$$

where λ_L is the solution in λ of the equation

$$\mathrm{P}(X \geq x) = \sum_{k=0}^{x-1} \frac{\lambda^k}{k!} \exp(-\lambda) = \frac{1+\gamma}{2}$$

and λ_U is the solution in λ of the equation

$$\mathrm{P}(X \leq x) = \sum_{k=0}^{x} \frac{\lambda^k}{k!} \exp(-\lambda) = \frac{1-\gamma}{2},$$

respectively. $z_{\chi^2_k}(\alpha)$ is the $\alpha{\cdot}100\,\%$-quantile of the chi-square distribution $\chi^2(k)$.

One-sided versions of a confidence interval of level $\gamma \in (0;1)$ for the parameter of a Poisson distribution under $X = x$ are given by

$$\begin{aligned}
(0;\ \lambda_U) &= \left(0;\ \frac{1}{2} z_{\chi^2_{2(x+1)}}(\gamma)\right) && (4.4)\\
&\quad \text{(one-sided interval with upper bound)}, && (4.5)\\
(\lambda_L;\ +\infty) &= \left(\frac{1}{2} z_{\chi^2_{2x}}(1-\gamma);\ +\infty\right) && (4.6)\\
&\quad \text{(one-sided interval with lower bound)}, && (4.7)
\end{aligned}$$

where λ_U, λ_L are solutions in λ of the equations

$$\begin{aligned}
\sum_{k=0}^{x} \frac{\lambda^k}{k!} \exp(-\lambda) &= 1-\gamma, && (4.8)\\
\sum_{k=0}^{x-1} \frac{\lambda^k}{k!} \exp(-\lambda) &= \gamma, && (4.9)
\end{aligned}$$

respectively. The exploited relation between the Poisson distribution and the chi-square distribution is derived e. g. by Johnson et al. (1992).

In analogy to Crow (1956), who first considered the total geometric volume of confidence regions for a binomial proportion, Crow & Gardner (1959) refrain from central intervals and focus on confidence intervals for a Poisson parameter λ that are optimal in a geometrical sense. They exploit the connection between acceptance regions as subsets of $\mathbb{N}_0$ and confidence regions from $(0; +\infty)$. In their paper, Crow & Gardner (1959) describe the construction of these intervals and provide tables of confidence limits under common confidence levels.

Crow & Gardner's (1959) confidence intervals lack the appealing characteristic of strictly increasing lower and upper confidence bounds. The intervals of Casella & Robert (1988) fulfil this property, but Kabaila & Byrne (2001) make aware of the disadvantages of Casella & Robert's (1988) intervals: The algorithm to compute the intervals is complicated and relies on normal approximation if the number x of occurrences is large. To overcome these disadvantages, Kabaila & Byrne (2001) provide a simpler algorithm for computing confidence intervals for the Poisson parameter that are as short as possible while ensuring the exactness criterion and strictly increasing bounds.

The above described intervals of Garwood (1936), Casella & Robert (1988) and Crow & Gardner (1959) are all exact in the sense that their coverage probability functions equal or exceed a prescribed confidence level $\gamma \in (0;1)$ pointwise for each $\lambda \in (0; +\infty)$. However, due to the fact that the Poisson distribution is a discrete distribution, equality rarely holds. For most $\lambda \in (0; +\infty)$, the actual coverage probability exceeds γ. Intervals, which are not constructed to fulfil this criterion are intervals based on normal approximation. Sahai & Khurshid (1993) give a continuity corrected normal approximation confidence interval $(\lambda_L;\ \lambda_U)$ of level $\gamma \in (0; 1)$ for λ with bounds

$$\lambda_L = x - \frac{1}{2} + \frac{1}{2}z^2 - z\sqrt{x - \frac{1}{2} + \frac{1}{4}z^2}, \tag{4.10}$$

$$\lambda_U = x + \frac{1}{2} + \frac{1}{2}z^2 + z\sqrt{x + \frac{1}{2} + \frac{1}{4}z^2}, \tag{4.11}$$

that can be used unless the value of λ is small. Here, $z = z_{N(0,1)}\left((1+\gamma)/2\right)$ is the $(1+\gamma)/2{\cdot}100\,\%$-quantile of the standard normal distribution. Sahai & Khurshid (1993) argue that for large values of x the continuity correction may be omitted. This results in the approximative interval

$$x\left(1 + \frac{1}{2x}z^2\left(1 \mp \sqrt{1 + \frac{4x}{z^2}}\right)\right) = \left(x + \frac{1}{2}z^2 \mp z\sqrt{x + \frac{1}{4}z^2}\right),$$

where again $z = z_{N(0,1)}\left((1+\gamma)/2\right)$ is the $(1+\gamma)/2{\cdot}100\,\%$-quantile of the standard normal distribution. This interval can also be found in Johnson et al. (1992), where they argue that it can be used if the parameter λ is expected to be fairly large, which in their view would be larger than 15.

Another approximation referred to by Sahai & Khurshid (1993) to be used for very large values of x – which they quantify as larger than 100 – is obtained by the interval

$$x \mp z\sqrt{x}. \tag{4.12}$$

Here, $z = z_{N(0,1)}\left((1+\gamma)/2\right)$ is the $(1+\gamma)/2{\cdot}100\,\%$-quantile of the standard normal distribution.

An improved approximate confidence interval appearing in Sahai & Khurshid (1993) is based on an approximation given by Molenaar (1970): Let $X \sim Po(\lambda)$ and $X = x$ be the observed value of X and $\gamma \in (0; 1)$ the confidence level. An approximate confidence interval for λ is given by

$$x + \left(2z^2 + 1\right)/6 \mp \left(\frac{1}{2} + \sqrt{z^2\left(x \mp \frac{1}{2} + \frac{z^2+2}{18}\right)}\right), \tag{4.13}$$

where $z = z_{N(0,1)}\left((1-\gamma)/2\right)$ is the $(1-\gamma)/2 \cdot 100\,\%$-quantile of the standard normal distribution.

Several other confidence intervals for the parameter of a Poisson distribution based on normal approximation to the chi-square distribution or using square root transformations are reviewed by Sahai & Khurshid (1993).

Sahai & Khurshid (1993) draw the attention to the fact that all their presented derivations of confidence intervals can be applied not only to the parameter λ, but also to $n\lambda$ in case the sum $X = \sum_{i=1}^{n} X_i$ of Poisson distributed random variables $X_i \sim Po(\lambda)$ is of interest. Confidence limits for $n\lambda$ would have to be divided by n to obtain confidence limits for λ. The difference between both cases is the (time) unit in which the number of successes are counted: If a confidence interval for $n\lambda$ is of interest, the number of occurrences in n times the time unit is considered and λ is the average of occurrences per time unit.

All of the so far presented confidence intervals do not take prior information into account. The natural environment in which prior information plays a role is the Bayesian framework. Here, regions in which the parameter of interest lies with a high probability, so-called *credibility intervals*, contain information about the posterior distribution, see Chapter 3 for an introduction. The confidence intervals which we present in the present chapter are not of Bayesian type – although we briefly consider the Bayesian approach for reasons of comparison –, but of frequentist's. That is, we present confidence intervals which under repeated sampling contain the parameter of the Poisson distribution in at least $\gamma \cdot 100\,\%$ of the cases, where $\gamma \in (0; 1)$. Prior information is used by means of stipulating a gamma distribution on the unknown parameter λ. Methods to calculate frequentist confidence intervals for the Poisson parameter under the use of prior information, as introduced in Chapter 2 for the binomial distribution, do not seem to be available up to this point.

The remainder of the chapter is structured as follows: Section 4.2 presents the model for the parameter λ of a Poisson distribution using a gamma distribution to express prior information on λ. Sections 4.3 and 4.4 deal with important properties of the functions related to the computation of the confidence intervals. The confidence intervals themselves are presented in Section 4.5. Bayesian credibility intervals for a Poisson parameter are briefly considered in Section 4.6. The minimum volume confidence intervals are compared with existing confidence intervals for the Poisson parameter in Section 4.7 and the effect of prior information on the minimum volume confidence intervals is investigated in Section 4.8.

4.2 The Gamma Prior Model for Inference on the Poisson Parameter

In Section 2.2 of Chapter 2 we established the general theory of minimum volume confidence intervals for a distribution parameter and applied it subsequently to the binomial probability parameter p. In this section, another instance of the model is derived in the form of confidence intervals for the parameter $y = \lambda$ of a Poisson distribution. Prior information on λ is employed by means of a gamma distribution. We impose the following two assumptions on the model:

1) The univariate random variable Y with values in $R_2 = (0; +\infty)$ represents a random parameter which varies according to the gamma distribution $\text{Gamma}(\kappa, \vartheta)$ on the support $(0; +\infty)$. The parameter $\kappa > 0$ is a shape parameter, and $\vartheta > 0$ is a scale parameter. The distribution $\text{Gamma}(\kappa, \vartheta)$ has the density function

$$f_{\vartheta,\kappa}(y) = \begin{cases} \frac{y^{\kappa-1}}{\vartheta^{\kappa}\Gamma(\kappa)} \exp\left(\frac{-y}{\vartheta}\right) & \text{for } y > 0, \\ 0, & \text{elsewhere.} \end{cases} \tag{4.14}$$

   Here, the measure μ_2 introduced in Section 2.2 is the Lebesgue measure on R_2. The function $\Gamma(z) = \int_0^{+\infty} t^{z-1} \exp(-t)\, \mathrm{d}t$ denotes the gamma function, see e. g. Abramowitz & Stegun (1972, Section 6.1).

   The gamma distribution as defined in Eq. (4.14) has support $(0; +\infty)$. There is a generalisation of the gamma distribution that includes a threshold parameter $\tau \in \mathbb{R}$, such that the gamma distribution has support $(\tau; +\infty)$, see e. g. Bowman & Shenton (1988). In the following we consider only the most important case of a gamma distribution with threshold $\tau = 0$.

2) Given a value $Y = y$, the random count X of occurrences within a given time interval is conditionally distributed by $Po(y)$. The range of X is $R_1 = \{0, 1, 2, \ldots\}$. The measure μ_1 introduced in Section 2.2 is the counting measure on R_1.

From assumptions 1) and 2) we obtain the unconditional density $f_X(x)$ of X or the *volume weights* $w_{\vartheta,\kappa}(x)$ as

$$\begin{aligned} f_X(x) \quad = \quad w_{\vartheta,\kappa}(x) \quad &= \frac{1}{\Gamma(x+1)\Gamma(\kappa)\vartheta^{\kappa}} \int_0^{+\infty} y^{x+\kappa-1} \exp\left(-y - \frac{y}{\vartheta}\right) \mathrm{d}y \\ &= \frac{1}{x!\Gamma(\kappa)\vartheta^{\kappa}} \int_0^{+\infty} y^{x+\kappa-1} \exp\left(-y - \frac{y}{\vartheta}\right) \mathrm{d}y \end{aligned} \tag{4.15}$$

for $x \in \{0, 1, 2, \ldots\}$. The random variable Y has expectation and variance

$$\mu_Y = E[Y] = \kappa\vartheta, \qquad \sigma_Y^2 = V[Y] = \kappa\vartheta^2, \tag{4.16}$$

see e. g. Johnson et al. (1994).

From the general formula (2.8) we obtain the weighted volume of a measurement and prediction space (MPS) A as

$$V(A) \quad = \quad \sum_{x=0}^{+\infty} \int_{A_x} \mathrm{d}\nu(y) w_{\vartheta,\kappa}(x) \quad = \quad \sum_{x=0}^{+\infty} \nu(A_x) w_{\vartheta,\kappa}(x), \tag{4.17}$$

where ν is the Borel measure, and A_x is the confidence region for y formed under the observation x. For the definition of an MPS, see Section 2.2. Mind that, in contrast to the binomial case, the total weighted volume of an MPS in the Poisson case is the result of an infinite sum.

The gamma distribution is an important example of a distribution with bounded support to the left and is therefore often used for modelling life lengths. It is a favourable model as a prior distribution for the Poisson parameter λ arising from Bayes theory: The gamma distribution is the conjugate prior for the Poisson distribution, see e. g. George et al. (1993). Although we do not pursue the Bayesian approach here, we make use of the favourable relation between the Poisson and the gamma distribution for our frequentist confidence intervals.

According to Jenkinson (2005), little information exists about eliciting the Poisson parameter with a gamma prior. In the Bayesian approach, non-informative gamma prior distributions in the Poisson model are not necessarily flat or uniform priors, in contrast to the binomial-beta model. They are distributions that rely primarily on the likelihood while creating the posterior distribution in a Bayesian approach, see Kerman (2011). An instance of a diffuse or non-informative gamma prior is obtained by setting the parameters $\kappa = 1$ and ϑ very large in Eq. (4.14). The density then becomes

$$f_{\vartheta,\kappa}(y) = \begin{cases} \frac{1}{\vartheta} \exp\left(\frac{-y}{\vartheta}\right) & \text{for } y > 0, \\ 0, & \text{elsewhere} \end{cases}$$

and hence it is the density of an exponential distribution, see Ross (2003).

4.3 Prediction Likelihood Maximisation and the Interval Property of Confidence Regions for the Poisson Parameter

We study confidence regions for the parameter of a Poisson distribution that are of minimum volume. According to Theorem 2.2, this requires taking into consideration prediction regions which are subsets of the areas $D_{\geq s}(y)$ of largest prediction likelihood ratio, see Eq. (2.10). Under the model from Section 4.2 we obtain with Eq. (2.9) the prediction likelihood ratio

$$\begin{aligned} Q_{\vartheta,\kappa,y}(x) &= \frac{\frac{y^x}{x!}\exp(-y)}{w_{\vartheta,\kappa}(x)} = \frac{y^x \exp(-y)}{v_{\vartheta,\kappa}(x)} \\ &= \frac{y^x \exp(-y)\Gamma(\kappa)\vartheta^\kappa}{\int_0^{+\infty} t^{x+\kappa-1}\exp\left(-t-\frac{t}{\vartheta}\right)\,\mathrm{d}t}, \end{aligned} \tag{4.18}$$

where the *relative volume weights* are defined by

$$v_{\vartheta,\kappa}(x) = x! \cdot w_{\vartheta,\kappa}(x) = \frac{1}{\Gamma(\kappa)\vartheta^\kappa}\int_0^{+\infty} y^{x+\kappa-1}\exp\left(-y-\frac{y}{\vartheta}\right)\,\mathrm{d}y. \tag{4.19}$$

Proposition 4.2 in the subsequent section shows that the prediction likelihood ratio $Q_{\vartheta,\kappa,\lambda}(x)$ as a function of x is either decreasing or of inverted bathtub shape. Hence the areas of largest values of $Q_{\vartheta,\kappa,\lambda}(x)$ are always intervals. Conditions under which the confidence regions for λ are intervals can be taken from Proposition 2.3.

4.4 Properties of Weights and Prediction Likelihood Ratios

We investigate the essential quantities of the gamma prior model introduced by Sections 4.2 and 4.3 as functions of the number x of occurrences under a Poisson distribution. Proposition 4.1 considers the volume weights $w_{\vartheta,\kappa}(x)$ presented in Section 4.2. The weights $w_{\vartheta,\kappa}(x)$ determine the influence of the confidence region A_x formed under the total volume $V(A)$ of the MPS, see Eq. (4.17).

Proposition 4.1 (Properties of Volume Weights). *Consider the weights $w_{\vartheta,\kappa}(x), x \in \mathbb{N}_0, \vartheta, \kappa > 0$, defined by Eq. (4.15).*

a) The explicit formula

$$w_{\vartheta,\kappa}(x) = \frac{\vartheta^x \Gamma(\kappa+x)}{(\vartheta+1)^{\kappa+x}\Gamma(x+1)\Gamma(\kappa)} = \begin{cases} \frac{\vartheta^x(\kappa+x-1)\cdot\ldots\cdot\kappa}{(\vartheta+1)^{\kappa+x}\Gamma(x+1)}, & \text{if } x \geq 1, \\ \frac{1}{(1+\vartheta)^\kappa} & \text{if } x = 0, \end{cases}$$

and the recursive formula

$$w_{\vartheta,\kappa}(x+1) = \frac{\vartheta(\kappa+x)}{(\vartheta+1)(x+1)} \cdot w_{\vartheta,\kappa}(x)$$

hold.

b) Let $x_0 := \vartheta(\kappa - 1) - 1$*. Then we have*

$$w_{\vartheta,\kappa}(x+1) \begin{cases} > w_{\vartheta,\kappa}(x) \text{ if } x < x_0, \\ = w_{\vartheta,\kappa}(x) \text{ if } x = x_0, \\ < w_{\vartheta,\kappa}(x) \text{ if } x > x_0. \end{cases}$$

Consequently, the following monotonicity properties hold:

$$w_{\vartheta,\kappa} \text{ is } \begin{cases} \text{strictly decreasing for } x = \{0,1,\ldots\} \text{ if } \kappa \leq 1, \\ \text{strictly decreasing for } x = \{0,1,\ldots\} \text{ if } \kappa > 1 \text{ and } \vartheta < \frac{1}{\kappa-1}, \\ \text{strictly increasing for } x = \{0,\ldots,x_0\} \text{ and strictly decreasing for } x = \{x_0+1, \\ \quad x_0+2,\ldots\} \text{ with } w_{\vartheta,\kappa}(x_0) = w_{\vartheta,\kappa}(x_0+1) \text{ if } \kappa > 1 \text{ and } \vartheta \geq \frac{1}{\kappa-1} \text{ and } x_0 \in \mathbb{N}_0, \\ \text{strictly increasing for } x = \{0,\ldots,\lceil x_0 \rceil\} \text{ and strictly decreasing for } x = \{\lceil x_0 \rceil, \\ \quad \lceil x_0 \rceil + 1,\ldots\} \text{ if } \kappa > 1 \text{ and } \vartheta \geq \frac{1}{\kappa-1} \text{ and } x_0 \notin \mathbb{N}_0. \end{cases}$$

Proof. See Appendix 4.A, Section 4.A.1. □

Proposition 4.2 gives an alternative presentation of the prediction likelihood ratio and investigates its monotonicity properties.

Proposition 4.2 (Likelihood Ratio)**.** *For* $\vartheta, \kappa > 0, x \in \mathbb{N}_0, \lambda > 0$*, consider the likelihood ratios* $Q_{\vartheta,\kappa,\lambda}(x) = \frac{\lambda^x \exp(-\lambda)\Gamma(\kappa)\vartheta^\kappa}{\int_0^{+\infty} y^{x+\kappa-1} \exp\left(-y-\frac{y}{\vartheta}\right) dy}$ *defined for* $x \in \{0,1,2,\ldots\}$*, see Eq. (4.18).*

a) Then the explicit formula

$$Q_{\vartheta,\kappa,\lambda}(x) = \frac{\lambda^x \exp(-\lambda)(\vartheta+1)^{\kappa+x}\Gamma(\kappa)}{\vartheta^x \Gamma(\kappa+x)} \tag{4.20}$$

and the recursive formula

$$Q_{\vartheta,\kappa,\lambda}(x+1) = \frac{\lambda(\vartheta+1)}{(\kappa+x)\vartheta} \cdot Q_{\vartheta,\kappa,\lambda}(x) \tag{4.21}$$

hold.

b) Let $\widetilde{x} := \lambda\frac{\vartheta+1}{\vartheta} - \kappa$*. Then we have*

$$Q_{\vartheta,\kappa,\lambda}(x+1) \begin{cases} > Q_{\vartheta,\kappa,\lambda}(x) \text{ if } x < \widetilde{x}, \\ = Q_{\vartheta,\kappa,\lambda}(x) \text{ if } x = \widetilde{x}, \\ < Q_{\vartheta,\kappa,\lambda}(x) \text{ if } x > \widetilde{x}. \end{cases}$$

Consequently, the following monotonicity properties hold:

$$Q_{\vartheta,\kappa,\lambda} \text{ is } \begin{cases} \textit{strictly decreasing for } x = \{0, 1, \ldots\} \textit{ if } \widetilde{x} < 0, \\ \textit{strictly increasing for } x = \{0, \ldots, \widetilde{x}\} \textit{ and strictly decreasing for } x = \{\widetilde{x}+1, \\ \quad \widetilde{x}+2, \ldots\} \textit{ with } Q_{\vartheta,\kappa,\lambda}(\widetilde{x}) = Q_{\vartheta,\kappa,\lambda}(\widetilde{x}+1) \textit{ if } \widetilde{x} \in \mathbb{N}_0, \\ \textit{strictly increasing for } x = \{0, \ldots, \lceil\widetilde{x}\rceil\} \textit{ and strictly decreasing for } x = \{\lceil\widetilde{x}\rceil, \\ \quad \lceil\widetilde{x}\rceil + 1, \ldots\} \textit{ if } \widetilde{x} \notin \mathbb{N}_0. \end{cases}$$

c) *We have* $Q_{\vartheta,\kappa,\lambda}(0) = \exp(-\lambda)(\vartheta+1)^{\kappa} > 0$ *and* $\lim_{x\to\infty} Q_{\vartheta,\kappa,\lambda}(x) = 0$.

d) *Consider* $Q_{\vartheta,\kappa,\lambda}(x)$ *as a function of* $\lambda \in (0; +\infty)$ *for fixed* $x \in \mathbb{N}_0$. *Then* $Q_{\vartheta,\kappa,\lambda}(x)$ *is strictly increasing on* $(0; x]$ *and strictly decreasing on* $[x; +\infty)$.

Proof. See Appendix 4.A, Section 4.A.2. □

4.5 Prediction Intervals and Confidence Intervals for the Poisson Parameter

As in the case of confidence intervals for a binomial proportion, see Section 2.6, prediction regions of largest prediction likelihood ratio in the case of the Poisson distribution are always intervals, as follows from Proposition 4.2 in the preceding section. However, from Proposition 2.3 it is known that the interval property of confidence regions is not implied by the interval property of the prediction region alone. Since we would like to have both prediction regions and confidence regions to be intervals for reasons of better interpretability, we restrict attention to MPSs with the following characteristics:

Q1) The prediction regions are nonempty and of the form

$$A_y \quad = \quad \Big\{c_L(y), c_L(y)+1, \ldots, c_U(y)\Big\} \quad \text{for each } y \in (0; +\infty),$$

where the *prediction limits* $c_L(y)$, $c_U(y)$ are increasing in $y \in (0; +\infty)$.

Q2) The confidence intervals are nonempty open intervals of the form

$$A_x \quad = \quad (\lambda_L(x); \lambda_U(x)) \subset (0; +\infty) \quad \text{for each } x \in \{0, 1, 2, \ldots\},$$

where the *confidence limits* $\lambda_L(x)$, $\lambda_U(x)$ are increasing in $x \in \{0, 1, 2, \ldots\}$.

From Q2) it follows in particular that A_x cannot be a singleton.

Among the MPSs satisfying the properties Q1) and Q2), we search for an MPS $A^{\star}$ of minimum volume $V(A^{\star})$, see Eq. (4.17) for the formula of the volume.

The subsequent proposition shows how appropriately prescribed prediction limits or appropriately prescribed confidence limits can determine the entire MPS with properties Q1) and Q2).

Proposition 4.3 (Confidence Limits and Prediction Limits)**.** *Let A be an MPS with nonempty projections A_x and A_y for all $x \in \{0, 1, 2, \ldots\}$ and all $y \in (0; +\infty)$.*

a) Let A satisfy the property Q1), and for the projections A_x let the lower prediction limit $c_L \colon (0; +\infty) \to \{0, 1, 2, \ldots\}$ be right-continuous in $\sup A_x$ and let the upper prediction limit $c_U \colon (0; +\infty) \to \{0, 1, 2, \ldots\}$ be left-continuous in $\inf A_x$. Then A satisfies the property Q2).

b) If A satisfies the property Q2), then A satisfies the property Q1).

Proof. See Appendix 4.A, Section 4.A.3. □

For the definition of the projections A_x and A_y in Proposition 4.3 we refer to Section 2.2. Section 2.2 establishes the duality between level γ prediction and level γ confidence regions. For prediction intervals of type Q1) and confidence intervals of type Q2), the defining characteristic (2.7) amounts to

$$\begin{aligned} \gamma \;\; &\le \;\; \mathrm{P}_y\Big(\lambda_L(X) < y < \lambda_U(X)\Big) \;\; = \;\; \mathrm{P}_y\Big(c_L(y) \le X \le c_U(y)\Big) \\ &= \;\; L_{c_U(y)}(y) - L_{c_L(y)-1}(y) \quad \text{for each } y \in (0; +\infty), \end{aligned} \tag{4.22}$$

where the *Poisson operating characteristic (OC) function* $L_c(y)$ is defined in Eq. (4.2).

The following proposition describes the minimum content of a level γ prediction interval of type Q1) and corresponding confidence interval of type Q2).

Proposition 4.4 (Minimum Content of Level γ Prediction Interval)**.** *Let $0 < \gamma < 1$ and $L_x(\lambda)$ be the Poisson OC function. For $x = 1, 2, \ldots$, let $\lambda_{x,\gamma}$ be the unique solution of the equation $L_{x-1}(\lambda) \overset{!}{=} \gamma$, and let $\lambda_{0,\gamma} = 0$. For $x = 0, 1, 2, \ldots$, let $\widetilde{\lambda}_{x,\gamma}$ be the unique solution of the equation $L_x(\lambda) \overset{!}{=} 1 - \gamma$. Let c_L, c_U be level γ prediction limits with corresponding confidence limits λ_L, λ_U as characterised by (4.22). Then the following assertions hold:*

a) The sequences $(\lambda_{x,\gamma})_{x \in \mathbb{N}_0}$ and $(\widetilde{\lambda}_{x,\gamma})_{x \in \mathbb{N}_0}$ are strictly increasing.

b) In the case of $\gamma \ge 0.5$ we have $\lambda_{x,\gamma} < \widetilde{\lambda}_{x,\gamma}$ for $x \in \{0, 1, 2, \ldots\}$.

c) For $x \in \{0, 1, 2, \ldots\}$ and $y \in (\lambda_{x,\gamma}; \widetilde{\lambda}_{x,\gamma})$ we have $c_L(y) \le x \le c_U(y)$.

Proof. See Appendix 4.A, Section 4.A.4. □

For the efficient computation of shortest Poisson confidence intervals, Proposition 4.4 turns out to be very useful. The minimum intervals $(\lambda_{x,\gamma};\widetilde{\lambda}_{x,\gamma})$ are contained in any level γ confidence interval A_x, but they are in general too narrow and not yet of level γ. However, they can be used as starting intervals to be extended to both sides to the proper shortest intervals $(\lambda_L(x);\lambda_U(x))$.

For the computation of the intervals, several other results play a role, see Propositions 4.5 and 4.6. Proposition 4.5 describes the monotonicity of the prediction region coverage. The result goes back to Crow & Gardner (1959). The proof is obtained by elementary differential calculus.

Proposition 4.5 (Monotonicity of Prediction Region Coverage)**.** *For* $0 \leq x_1 < x_2$, $\lambda > 0$, *let* $\Delta_{x_1,x_2}(y) = L_{x_2}(y) - L_{x_1-1}(y)$, *where* $L_x(y)$ *is the OC function from Eq.* (4.2). *In the case* $x_1 > 0$ *let*

$$\rho_{x_1,x_2} \quad := \quad (x_1 \cdot \ldots \cdot x_2)^{\frac{1}{x_2-x_1+1}}\,.$$

Then we have:

a) In the case $x_1 = 0$, $\Delta_{x_1,x_2} = L_{x_2}$ *is strictly decreasing on* $(0;+\infty)$.

b) In the case $x_1 > 0$, Δ_{x_1,x_2} *is strictly increasing on* $(0;\rho_{x_1,x_2}]$ *and strictly decreasing on* $[\rho_{x_1,x_2};+\infty)$.

Proof. See Appendix 4.A, Section 4.A.5. □

The following proposition deals with the comparison of the prediction likelihood ratios of two prediction points $x_1 \neq x_2$.

Proposition 4.6 (Comparison of Two Likelihood Ratios)**.** *Consider the likelihood ratios* $Q_{\vartheta,\kappa,y}(x)$ *defined for* $x \in \{0,1,2,\ldots\}$, *see* (4.18). *Let* $0 \leq x_1 < x_2, y \in (0;+\infty)$. *Let*

$$\lambda_{x_1,x_2} \quad := \quad \frac{\vartheta}{\vartheta+1}\Big((\kappa+x_1)\cdot\ldots\cdot(\kappa+x_2-1)\Big)^{\frac{1}{x_2-x_1}}\,.$$

Then we have

$$Q_{\vartheta,\kappa,y}(x_1) \begin{cases} > Q_{\vartheta,\kappa,y}(x_2) & \text{if } y < \lambda_{x_1,x_2}, \\ = Q_{\vartheta,\kappa,y}(x_2) & \text{if } y = \lambda_{x_1,x_2}, \\ < Q_{\vartheta,\kappa,y}(x_2) & \text{if } y > \lambda_{x_1,x_2}. \end{cases} \tag{4.23}$$

Proof. See Appendix 4.A, Section 4.A.6. □

In the case of the binomial distribution, the calculation of the prediction regions requires to compute and compare a limited number of $n+1$ values of the binomial prediction likelihood ratio $Q_{p_0,p_1,a,b,y}(x)$ due to the limited support $\{0,1,\ldots,n\}$ of the binomial distribution, see Chapter 2. The Poisson distribution, in contrast, has the unlimited support $\{0,1,2,\ldots\}$ and a priori requires to take into account an unlimited number of x with their respective prediction likelihood ratios. The following proposition shows that for a given y, values of x above a certain threshold can be neglected.

Proposition 4.7. *Let $G_y(t) := P_y\left(Q_{\vartheta,\kappa,y}(X) > t\right)$, where $Q_{\vartheta,\kappa,y}(x)$ is the prediction likelihood ratio from Eq. (4.18). Let $0 < \gamma < 1$, and s_y be chosen such that $s_y = \inf\{t|G_y(t) \le \gamma\}$. Let $D_{\ge t}(y) := \left\{x|Q_{\vartheta,\kappa,y}(x) \ge t\right\}$ for $y \in (0;+\infty)$.*
Then there exists a $c \in \mathbb{N}$ such that $D_{\ge s_y} \subset \{0,1,\ldots,c\}$.

Proof. See Appendix 4.A, Section 4.A.7. □

The general definition and properties of the function $G_y(t)$ can be found in Proposition 2.1. The segments of largest prediction likelihood ratio $D_{\ge t}(y)$ are defined in Eq. (2.10).

The proof of Proposition 4.7 suggests to choose $c \in \mathbb{N}$ such that

i) $Q_{\vartheta,\kappa,\lambda}(c) < Q_{\vartheta,\kappa,\lambda}(0)$ and simultaneously

ii) $F_X(c) > \gamma$.

Any $\mathbb{N} \ni x > c$ also fulfils the requirements and c is not necessarily the smallest number doing so. With Proposition 4.2 it can be shown that condition i) is equivalent to the condition

$$\left(\frac{1}{\lambda}\frac{\vartheta}{\vartheta+1}\right)^c \frac{\Gamma(\kappa+c)}{\Gamma(\kappa)} \quad > \quad 1.$$

It follows from Proposition 4.7 that it is sufficient to calculate and compare the prediction likelihood ratios $Q_{\vartheta,\kappa,\lambda}(x)$ for $x = 0,1,\ldots,c$ in the calculation of the prediction regions for the Poisson distribution under a fixed $\lambda \in (0;+\infty)$.

Proposition 4.7 is important in constructing prediction regions for a given $\lambda \in (0;+\infty)$. The problem that is reversed to the one addressed in Proposition 4.7 of finding confidence regions for a given x is not covered by Proposition 4.7. Given an x, confidence intervals for λ are bounded by the infimum and supremum in $\lambda \in (0;+\infty)$ such that $x \in A_\lambda$. However, due to the unlimitedness of the parameter space $R_2 = (0;+\infty)$ of $y = \lambda$, the problem is not as easily solved as in the case of the binomial distribution where both x and y are bounded from below and above. The question to be asked and answered to solve the problem for the Poisson distribution is to find for a given $x \in \{0,1,\ldots\}$ a $\lambda_1 \in (0;+\infty)$ such that for all $y \ge \lambda_1$ the condition $x \notin D_{\ge s_y}(y)$ holds. Although we eventually would like to restrict attention to prediction regions that are increasing in $\lambda = y$ in their lower and upper limits, monotonicity cannot be expected a priori for the bounds of $D_{\geq s_y}(y)$. This makes the problem a non-trivial one. If monotonicity was present, a threshold λ_1 would be given by the infimum in y such that $x < \min D_{\geq s_y}(y)$. The reason for this unfavourable behaviour is the discreteness of the Poisson distribution.

When following the procedure of constructing a confidence interval for a Poisson parameter λ as presented in the preceding sections, the result has to be interpreted as follows: Let x be the number of observed successes in the time interval T. Then the confidence interval $A_x = (\lambda_L(x); \lambda_U(x))$ provides a confidence interval for the number of successes in time interval T. In particular, only one time unit T is investigated, i. e. the sample size is 1. However, the result can be interpreted differently if looking at a different time unit: Let $T = nt$ be n times a time unit of length t. Let the parameter λ of interest be the mean in time unit t. Then $A_x =: A_{x,1} = (\lambda_L(x); \lambda_U(x))$ provides a confidence interval for the total number $n\lambda$ of successes in time $T = nt$. A confidence interval for the parameter λ is then given by $A_{x,n} = (\lambda_L(x)/n; \lambda_U(x)/n)$ and the sample size is n. The justification for this relation is the fact that the sum of n independent Poisson distributed random variables $X_i \sim Po(\lambda)$ is distributed according to $\sum_{i=1}^{n} X_i \sim Po(n\lambda)$, see e. g. Sahai & Khurshid (1993), Johnson et al. (1992) and references therein.

Mind that since the construction of the minimum volume confidence intervals for a Poisson mean λ in time unit t is done by considering the mean number of successes $n\lambda$ in a time interval $T = nt$, the corresponding prediction regions consequently reflect the possible number of successes in time T. Confidence intervals for $n\lambda$ are derived by exploiting the relation between prediction and confidence regions for $n\lambda$. Confidence intervals for λ are calculated subsequently. Prediction regions for $\lambda = y/n$ are not directly considered. The coverage probability under λ has to be derived from the coverage probability under $n\lambda$.

We illustrate the use and interpretation of the prediction and confidence intervals by means of examples in Section 4.8.

4.6 Bayesian Credibility Intervals for the Poisson Parameter

We briefly cover the Bayesian approach of interval estimation of a Poisson parameter to emphasise the difference between the frequentist approach of confidence intervals under prior information and the Bayesian approach yielding credibility intervals. An introduction into credibility intervals, in particular highest posterior density (HPD) intervals, is provided by Chapter 3 with applications to a binomial probability. Credibility intervals for the Poisson parameter can be obtained by applying Bayes' theorem, see Theorem 3.1, to the Poisson distribution. Using a gamma prior, which is the conjugate prior for the Poisson distribution, we obtain HPD credibility intervals with the help of the posterior distribution. The result of the following proposition can be taken from George et al. (1993) with different parametrisations of the Poisson and gamma distributions.

Proposition 4.8 (Posterior Distribution in the Poisson-gamma Model)**.** *Consider the Poisson density $f_{X|Y=y} = \frac{y^x}{x!}\exp(-y)$ of X under $Y = y$ and a Gamma(κ, ϑ) distribution as the prior for Y, where the gamma distribution is parametrised as in Eq.* (4.14)*. The posterior distribution of Y is the gamma distribution Gamma$(\kappa + x, \frac{\vartheta}{\vartheta+1})$.*

Proof. The assertion follows by elementary application of Bayes' theorem, see Appendix 4.A, Section 4.A.8. □

From Proposition 4.8 we obtain for $\kappa = 1$ and $\vartheta \to \infty$ the limit of the posterior density as

$$\lim_{\vartheta\to\infty} \frac{1}{\left(\frac{\vartheta}{\vartheta+1}\Gamma(\kappa+x)\right)} y^{\kappa+x-1} \exp\left(-\frac{y}{\frac{\vartheta}{\vartheta+1}}\right) = \frac{1}{\Gamma(1+x)} y^x \exp(-y) = \frac{1}{x!} y^x \exp(-y).$$

It is the limiting posterior density in the case of the non-informative gamma prior Gamma$(1, \vartheta)$ for $\vartheta \to \infty$, see also Ross (2003) and the last paragraph of Section 4.2.

The subsequent proposition explains the monotonicity properties of the gamma distribution.

Proposition 4.9 (Monotonicity of the Gamma Distribution)**.** *The density function $f_{\vartheta,\kappa}(x)$ of the gamma distribution as defined in Eq.* (4.14) *has the following monotonicity properties:*

(a) $f_{\vartheta,\kappa}(x)$ is strictly decreasing on $(0; +\infty)$ if $\kappa \leq 1$.

(b) $f_{\vartheta,\kappa}(x)$ is strictly increasing on $(0; x^*)$ and strictly decreasing on $(x^*; +\infty)$ if $\kappa > 1$, where $x^* = \vartheta(\kappa - 1)$.

Proof. See Appendix 4.A, Section 4.A.9. □

With Proposition 4.9, we can derive monotonicity properties of the posterior distribution in the Poisson-gamma model from Proposition 4.8.

Corollary 4.10 (Monotonicity of the Posterior Density)**.** *Let $0 < \beta < 1$ and $f_Y(y) = f_{\vartheta,\kappa}(y) = \frac{y^{\kappa-1}}{\vartheta^\kappa\Gamma(\kappa)}\exp\left(\frac{-y}{\vartheta}\right)$ the prior density of the Poisson parameter $\lambda > 0$. Let*

$f_{X|Y=y}(x) = \frac{y^x}{x!}\exp(-y)$ be the likelihood function, where $x = 0, 1, \ldots$ Then the following assertions hold:

(a) $f_{Y|X=x}(y)$ is strictly decreasing on $(0; +\infty)$ if $\kappa \leq 1$ and $x = 0$.

(b) $f_{Y|X=x}(y)$ is strictly increasing on $(0; y^*)$ and strictly decreasing on $(y^*; +\infty)$ if $\kappa > 1$ or $x \geq 1$, where $y^* := \frac{\vartheta}{\vartheta+1}(\kappa + x - 1)$.

Corollary 4.10 allows the derivation of HPD intervals for the parameter λ of a Poisson distribution. They are explicitly stated in the following corollary.

Corollary 4.11 (HPD Credibility Interval for a Poisson Parameter)**.** *Let $0 < \beta < 1$ and Gamma(κ, ϑ), as defined in Eq. (4.14), be the prior distribution for the Poisson parameter $y = \lambda$. Let $f_{X|Y=y}(x) = \frac{y^x}{x!}\exp(-y)$ be the likelihood function and $f_{Y|X=x}(y)$ the posterior density of Y under $X = x$. The level β HPD credibility interval A_x for Y under $X = x$ is given by*

(a) $A_x = (l_x; u_x) = \left(0; z_{Gamma\left(\kappa+x, \frac{\vartheta}{\vartheta+1}\right)}(\beta)\right)$ if $\kappa \leq 1$ and $x = 0$,

(b) $A_x = \left(z_{Gamma\left(\kappa+x, \frac{\vartheta}{\vartheta+1}\right)}(\alpha_1), z_{Gamma\left(\kappa+x, \frac{\vartheta}{\vartheta+1}\right)}(\alpha_2)\right)$, where $f_{Y|X=x}\left(z_{Gamma\left(\kappa+x, \frac{\vartheta}{\vartheta+1}\right)}(\alpha_1)\right) = f_{Y|X=x}\left(z_{Gamma\left(\kappa+x, \frac{\vartheta}{\vartheta+1}\right)}(1-\alpha_2)\right)$ with $\alpha_1 + \alpha_2 = 1 - \beta$ if $\kappa > 1$ or $x \geq 1$. The maximum point y^* of the posterior density function is $y^* = \frac{\vartheta}{\vartheta+1}(\kappa + x - 1)$ with $y^* \in A_x$.

Proof. See Appendix 4.A, Section 4.A.10. □

Figure 4.1 shows posterior distributions for λ under $X = x \in \{0, 1, \ldots, 7\}$ for four different gamma priors Gamma(κ, ϑ) and sample size $n = 1$. The bounds of the corresponding level 80 % credibility intervals are displayed by dashed vertical lines. In the case of the priors Gamma$(1, 1)$ and Gamma$(1, 0.2)$, where the shape parameter κ is equal to 1, the posterior distributions under $x = 0$ are strictly decreasing. For $x > 0$ and $\kappa > 1$, all posterior gamma density functions are unimodal with mode $y^* > 0$, cf. Corollary 4.10. In those cases, as e. g. for all x under the priors Gamma$(2, 1)$ and Gamma$(15, 1)$, there are $\lambda > 0$ such that $\lambda \notin (l_x; u_x)$ for all $x \in \{0, 1, 2, \ldots\}$. The frequentist coverage probability at these λ results to 0, i. e. $\mathrm{P}_\lambda(\lambda \in (l_X; u_X)) = 0$.

Theorem 3.16 states under which conditions on the beta prior distribution, the HPD interval for the binomial proportion has a frequentist coverage of at least a prescribed level $\gamma \in (0; 1)$. With respect to the equivalent statement in the Poisson case, $\kappa \leq 1$ is a necessary condition because it is required that there is at least one possible outcome $x \in \{0, 1, 2, \ldots\}$ which leads to a decreasing posterior density. Besides of at least one decreasing posterior density function, the proof of Theorem 3.16 relies on the fact that under certain conditions the posterior density is increasing. Consequently, the idea of the proof of Theorem 3.16 will not work in the Poisson-gamma case: From Corollary 4.10 we know that the posterior density is either decreasing or unimodal and therefore never increasing on the whole parameter space $(0; +\infty)$. Whether a similar assertion as in Theorem 3.16 can be derived for the Poisson-gamma case remains to be investigated. Based on the current findings, we conjecture that the Poisson credibility intervals can exceed a prescribed frequentist coverage probability at most asymptotically.

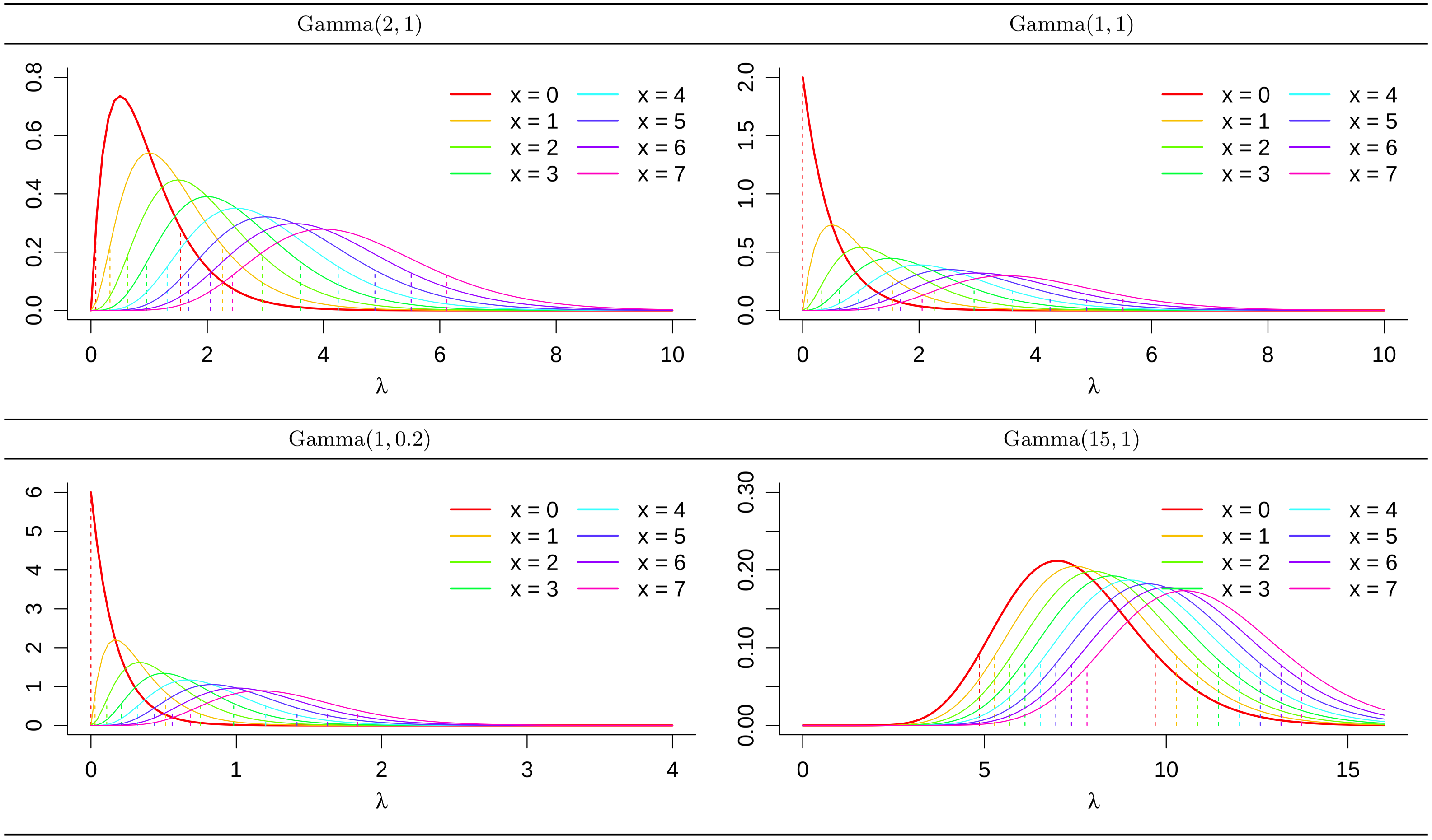


Figure 4.1: Posterior density functions for the Poisson parameter λ under $X = x \in \{0, 1, \ldots, 7\}$ with level 80 % HPD credibility intervals (dashed) for different gamma priors. Sample size $n = 1$.

4.7 Comparison with other Confidence Intervals for the Poisson Parameter

In this section, we compare various types of confidence intervals for the parameter of a Poisson distribution in terms of their coverage probabilities. We investigate the following six confidence interval approaches:

i) the minimum volume confidence interval proposed in the previous sections under the exemplary prior information $\text{Gamma}(2, 1)$,

ii) the classical exact confidence interval based on quantiles of the chi-square distribution from Eq. (4.3),

iii) a simple normal approximation based confidence interval as in Eq. (4.13),

iv) the continuity corrected confidence interval by Sahai & Khurshid (1993) as in Eqs. (4.10)–(4.11),

v) the improved approximate confidence interval by Molenaar (1970) as in Eq. (4.13),

vi) the HPD credibility interval from Section 4.6 under the prior $\text{Gamma}(2, 1)$.

The coverage probability functions of the five confidence interval approaches are displayed in Fig. 4.2 for a nominal confidence/credibility level of 95 % and sample size 1. The simple normal approximation based interval clearly violates the nominal confidence level in actual coverage probability for small λ and shows satisfying coverage probability close to the nominal level from approximately $\lambda = 15$ on. The improved approximate confidence interval by Molenaar (1970) is very similar to the classical exact confidence interval unless λ is very close to 0. In fact, the coverage probabilities of the interval by Molenaar (1970) as well as the continuity corrected interval by Sahai & Khurshid (1993) exceed the nominal confidence level nearly everywhere in the investigated area. The exception are values of λ very close to 0. The interval by Sahai & Khurshid (1993) seems to be slightly less conservative than the exact interval for λ smaller than 4. The Bayesian HPD interval with prior $\text{Gamma}(2, 1)$ shows a coverage probability of 0 for λ very close to 0 and is over-conservative for λ up to 3. The coverage probability decreases drastically for increasing λ and drops to nearly 0 for $\lambda \geq 20$. The pattern of the minimum volume confidence interval under prior information is different from the other displayed patterns. Since the interval is an exact interval, the actual coverage probability exceeds the nominal confidence level for all λ in the displayed area. Other than the five other coverage probability functions, the coverage probability of the minimum volume confidence interval reaches exactly 0.95, the nominal confidence level, many times. It therefore seems at least in these points and for $\lambda < 4$ less conservative than the other intervals.

4.8 Numerical Analysis of the Minimum Volume Confidence Intervals

In this section, we investigate the shortest confidence intervals for the parameter of a Poisson distribution on a numerical basis.

Figure 4.3 shows the lower ends (prediction points $0, 1, \ldots, 15$) of the measurement and prediction spaces (MPSs) of the minimum volume confidence intervals for a Poisson parameter as well as the corresponding coverage probability functions. A variety of prior information distributions is considered. The sample size is $n = 1$ and the confidence level $\gamma = 0.95$. Prior information is generally expressed by means of a gamma distribution, whose density function can only take right-skewed shapes. The investigated prior distributions are the gamma distributions $\text{Gamma}(1, 1)$ and $\text{Gamma}(1, 0.2)$ with strictly decreasing density functions as well as several gamma distributions with unimodal densities. The coverage probability functions all look very similar in the investigated parameter space of λ. In terms of the MPS, a slightly differing behaviour is visible for different prior information distributions. They therefore produce confidence intervals of different lengths depending on which of the λ and x get more weight by the prior information.

Since we currently cannot empirically compare the total volume of the MPS from Eq. (4.17) due to the unboundedness of the prediction space $\{0, 1, 2, \ldots\}$, we have a look at the widths of the confidence intervals if the number of successes is between 0 and 15. In Fig. 4.4, the lengths of the minimum volume confidence intervals obtained under three different prior distributions are compared with the lengths of the classical exact confidence interval for a sample size of 1 and a confidence level of $\gamma = 0.95$. The prior distributions $\text{Gamma}(1, 1)$ and $\text{Gamma}(2, 1)$ lead to confidence intervals which are narrower than the classical exact confidence interval for $x = 0, 1, \ldots, 9$. They are decreasing or with a mode at $\lambda = 1$, respectively, and therefore put considerable weight to values of λ close to 0. Only from $x = 10$ onwards they start losing their advantages in length. The prior information $\text{Gamma}(7, 0.5)$, while hardly being inferior to the priors $\text{Gamma}(1, 1)$ and $\text{Gamma}(2, 1)$ for $x = 0, 1, \ldots, 6$, partly delivers even tighter confidence bounds for $x = 7, \ldots, 11$ than the other two. For $x = 13, 14, 15$, the classical exact confidence

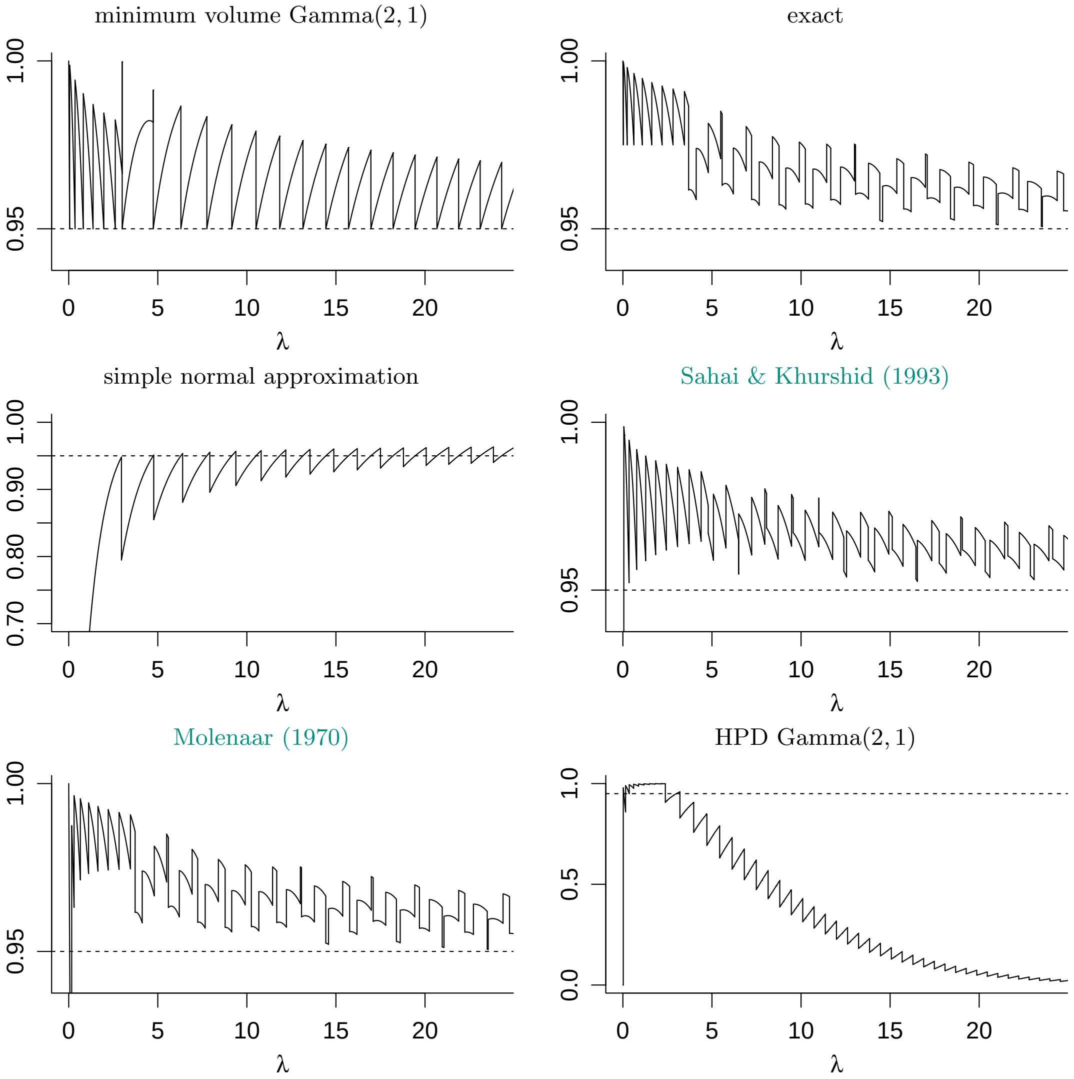


Figure 4.2: Coverage probability of several Poisson confidence intervals as a function of λ for sample size 1 under the nominal confidence level 95 %.

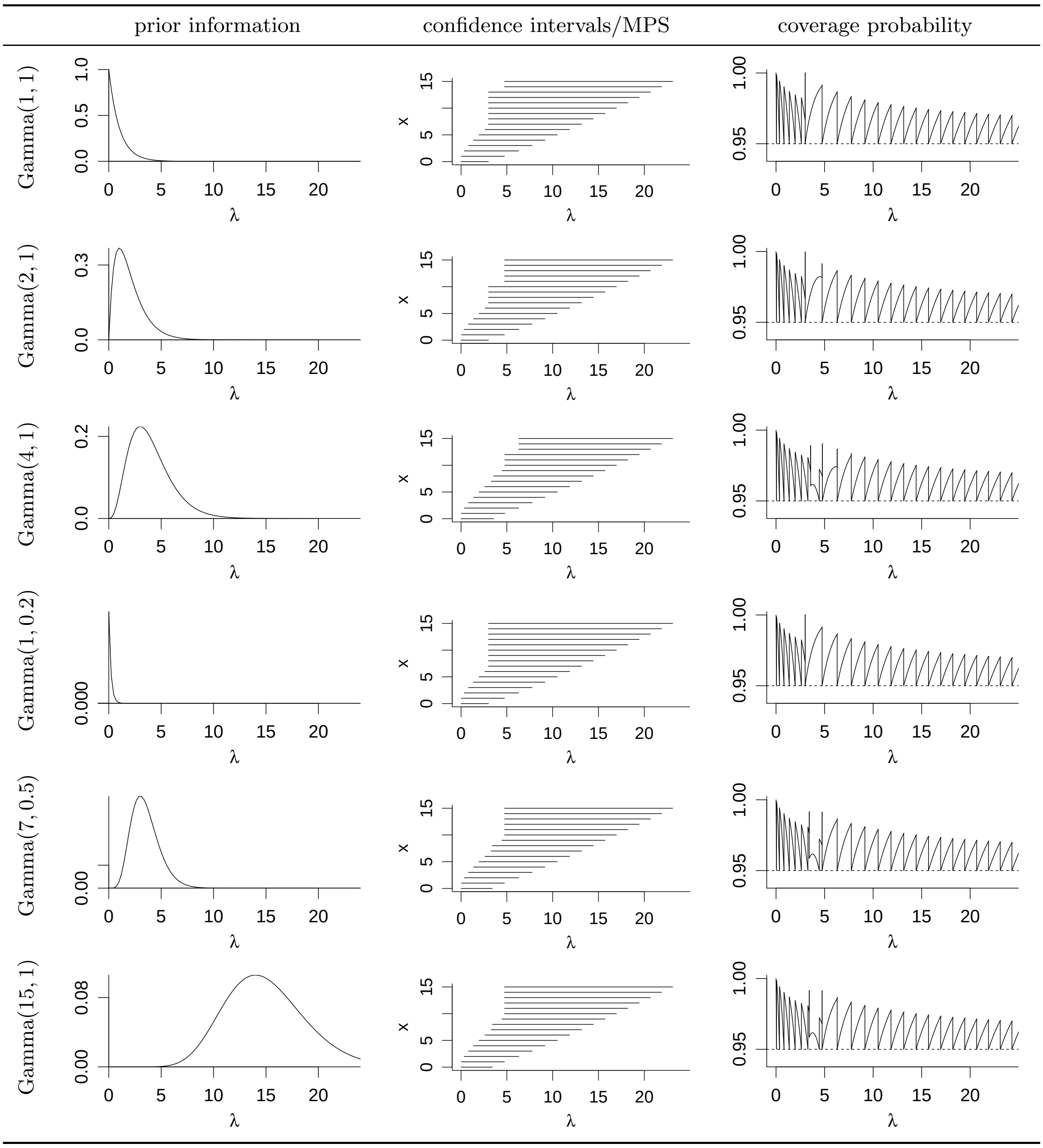


Figure 4.3: Level 95 % minimum volume confidence intervals for a Poisson parameter λ and coverage probability for a selection of gamma prior distributions. Sample size $n = 1$.

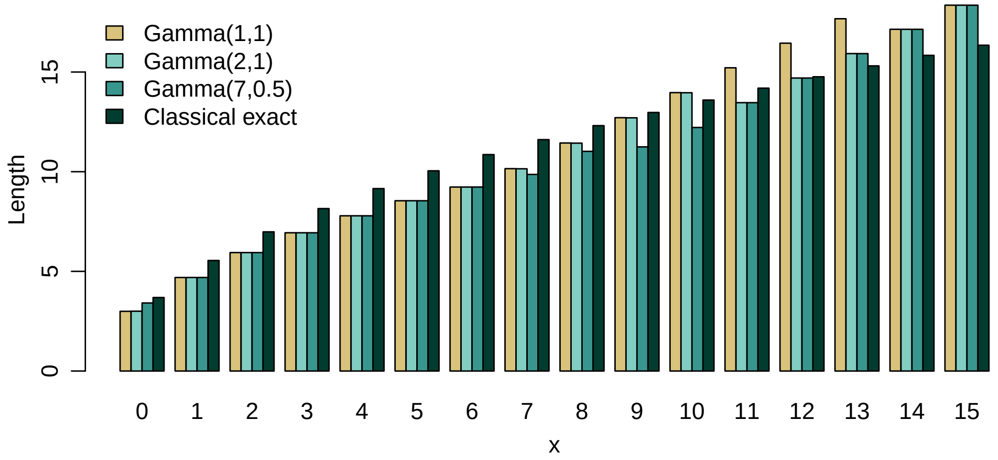


Figure 4.4: Lengths of the minimum volume confidence interval for a Poisson parameter λ for a selection of gamma prior distributions in comparison to the classical exact Poisson confidence interval. Sample size $n = 1$. Confidence level 95 %.

Table 4.1: Minimum volume confidence intervals for a Poisson parameter under prior information distribution Gamma$(2, 1)$ for sample sizes $n = 1, 5$ and confidence level $\gamma = 0.95$.

$\sum_{i=1}^{n} x_i$	$n = 1$			$n = 5$		
	$\widehat{\lambda}$	λ_L	λ_U	$\widehat{\lambda}$	λ_L	λ_U
0	0	0.000	3.002	0.0	0.000	0.600
1	1	0.051	4.744	0.2	0.010	0.949
2	2	0.355	6.296	0.4	0.071	1.259
3	3	0.818	7.754	0.6	0.164	1.551
4	4	1.366	9.154	0.8	0.273	1.831
5	5	1.970	10.513	1.0	0.394	2.103
6	6	2.613	11.842	1.2	0.523	2.368
7	7	3.002	13.148	1.4	0.600	2.630
8	8	3.002	14.435	1.6	0.600	2.887
9	9	3.002	15.705	1.8	0.600	3.141
10	10	3.002	16.962	2.0	0.600	3.392

interval provides smallest lengths in comparison to the investigated minimum volume intervals under prior information.

All empirical results above hold in the case of a sample size of 1, i. e. the number of occurrences X in only one time interval of length T is considered with $X \sim Po(\lambda)$. Simultaneously, this can be interpreted as $X = \sum_{i=1}^{n} X_i$ occurrences in n time intervals with $X_i \sim Po(\lambda/n)$, where X_i is the number of occurrences in time interval i of length $t = T/n$. Minimum volume confidence intervals for λ/n if the sample size is $n > 1$ can be obtained from the intervals under sample size 1 by deviding their lower and upper bounds by n. We provide an example to illustrate the relationship in Table 4.1.

Table 4.1 provides lower and upper bounds of the minimum volume confidence intervals

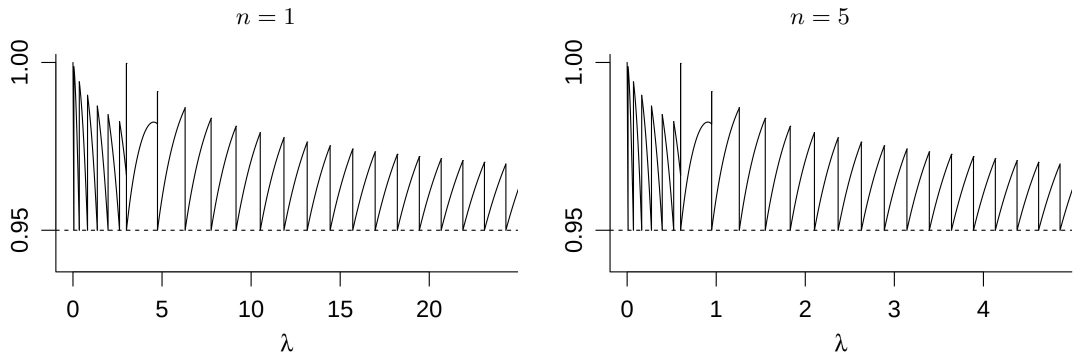


Figure 4.5: Coverage probability of the minimum volume confidence interval for a Poisson parameter under prior information distribution Gamma$(2,1)$ for sample sizes $n = 1, 5$ and confidence level $\gamma = 0.95$.

$(\lambda_L; \lambda_U)$ for the Poisson parameter λ under prior information distribution Gamma$(2,1)$ and sample sizes $n = 1, 5$. The estimates $\widehat{\lambda}$ are estimates for λ in one time interval of length t, where the number of successes $\sum_{i=1}^{n} x_i$ have been observed in time interval $T = nt$. The lower and upper bounds λ_L, λ_U under sample size $n = 5$ have been obtained by deviding the corresponding lower and upper bounds under sample size $n = 1$ by 5.

The left-hand side of Fig. 4.5 shows the coverage probability under sample size $n = 1$. The right-hand side under sample size $n = 5$ shows exactly the same picture, the only difference is the axis scale: The parameter of interest on the right-hand side is the mean λ of the Poisson distribution in a time interval whose length is one fifth of the length of the time interval on the left-hand side. The calculation of the coverage probability function exploits the relation between the confidence intervals and the prediction points, where the prediction points for sample size $n = 5$ reflect predictions for the number of occurrences $\sum_{i=1}^{5} X_i$ in all five time intervals.

Figure 4.6 shows the coverage probability function for the shortest confidence interval for a Poisson parameter under the prior information distribution Gamma$(2,1)$ for sample size 1 and confidence level $\gamma = 0.95$ if the values $x \in \{0, 1, \ldots, 60\}$ are taken into account for the construction of the prediction intervals. The coverage probability exceeds the prescribed nominal confidence level of 95 % for $\lambda \in (0; 31)$. It only rarely drops below 95 % in the interval $[31; 45)$ and many times so for $\lambda > 45$. That the coverage probability drops below the prescribed threshold is due to the fact that only prediction points $\{0, 1, \ldots, 60\}$ have been taken into account here. The problem could be overcome by considering sufficiently many prediction points larger than 60. Proposition 4.7 explains which prediction points are worth considering under a fixed $\lambda = y$, but not which λ

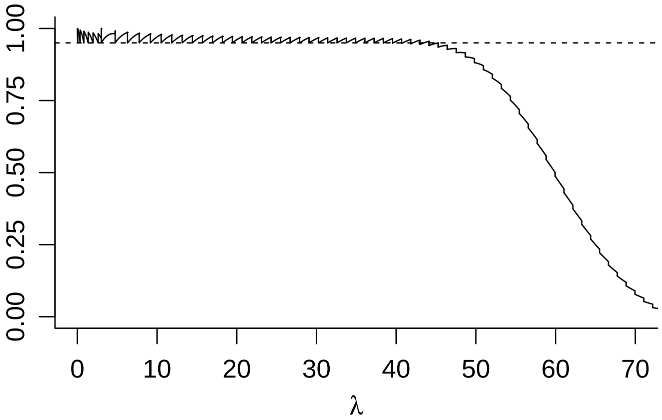


Figure 4.6: Coverage probability of the minimum volume confidence interval for a Poisson parameter λ if for the prediction regions $x \in \{0, 1, \dots, 60\}$ are taken into account. Prior information for λ: Gamma$(2, 1)$. Sample size $n = 1$. Confidence level 95 %.

should be covered in this respect. At this stage it is unclear which $\lambda \in (0; +\infty)$ are worth considering under a given prediction point x. Mind that this difficulty of the Poisson confidence interval is not an issue in the context of the minimum volume confidence interval for a binomial probability as presented in Chapter 2 because the prediction space is bounded from above by 1.

To currently overcome that we have not yet completely solved the problem of a threshold in λ from when on to stop considering values above, the empirical results in this chapter have been obtained by taking small to considerable large values for x into account and stopping to consider prediction points when the area of interest in λ shows sufficient coverage. In case the described problem cannot be solved analytically, a numerical algorithm can be thought of accordingly.

4.9 Conclusion and Outlook

We have applied the theory of minimum volume confidence intervals under prior information that have been suggested in Section 2.2 to the expectation λ of a Poisson distribution. Prior information has been used by imposing a gamma prior distribution on λ. The approach is a frequentist approach and the outcomes are exact intervals, i. e. a prescribed coverage probability of at least γ is maintained for all $\lambda \in (0; +\infty)$. The purpose of the prior distribution is to differently weigh the confidence intervals for the different possible numbers of occurrences $x = \{0, 1, 2, \dots\}$. The theory necessary to construct these intervals for the Poisson parameter has been developed and presented nearly completely in this chapter. It has become obvious that many results are obtained more easily than in the case of the binomial distribution. Several difficulties that arise in the context of the Poisson distribution trace back to the unboundedness of the parameter region $(0; +\infty)$ and the prediction region $\{0, 1, 2, \ldots\}$, which is not an issue in the binomial case. What remains to be investigated is the problem stated subsequent to Proposition 4.7: For a given x, find a threshold $\lambda_1 \in (0; +\infty)$ such that the point x is not contained in the region $D_{\geq s_y}(y)$ of greatest prediction likelihood ratio filling up the level γ for any $y \geq \lambda_1$.

A field for further investigation is the comparison between confidence intervals for the binomial distribution and confidence intervals for the Poisson distribution from the point of view that the Poisson distribution is the limiting distribution of the binomial distribution if $n \to \infty, p \to 0, np \to \lambda$. It is expected from this analysis that insight can be gained into the limiting characteristics of the minimum volume confidence intervals for a binomial probability from Chapter 2.

A more detailed comparison with Bayesian credibility intervals should be performed, among it the investigation of an equivalent to Theorem 3.16 that tries to map the coverage probabilities of frequentist and Bayesian confidence intervals for the Poisson parameter under certain conditions on the prior distribution.

Possibilities of how to practically and appropriately elicit the parameters of the gamma prior information distribution need to be investigated.

With the Poisson distribution, we have applied the general theory of minimum volume confidence intervals under prior information to a discrete distribution for which expectation and variance coincide. The theory has been developed for and applied to the binomial distribution, see Chapter 2, for which underdispersion holds. As a third important instance of a discrete distribution, this time showing overdispersion, we consider the negative binomial distribution to be worth investigating.

4.A Appendix

4.A.1 Proof of Proposition 4.1

We prove the properties of the volume weights.

For the proof of a) let $\kappa_1 := \kappa + x \geq \kappa > 0, \vartheta_1 := \frac{\vartheta}{\vartheta+1} \overset{\vartheta>0}{>} 0$. Then we have

$$\begin{aligned}
w_{\vartheta,\kappa}(x) &\overset{(4.15)}{=} \frac{1}{\Gamma(x+1)\Gamma(\kappa)\vartheta^\kappa} \int_0^{+\infty} y^{x+\kappa-1} \exp\left(-y - \frac{y}{\vartheta}\right) \mathrm{d}y \\
&= \frac{\vartheta_1^{\kappa_1}\Gamma(\kappa_1)}{\Gamma(x+1)\Gamma(\kappa)\vartheta^\kappa} \underbrace{\frac{1}{\vartheta_1^{\kappa_1}\Gamma(\kappa_1)} \int_0^{+\infty} y^{\kappa_1-1} \exp\left(-\frac{y}{\vartheta_1}\right) \mathrm{d}y}_{=1} \\
&= \frac{\vartheta_1^{\kappa_1}\Gamma(\kappa_1)}{\Gamma(x+1)\Gamma(\kappa)\vartheta^\kappa} = \frac{\left(\frac{\vartheta}{\vartheta+1}\right)^{\kappa+x}\Gamma(\kappa+x)}{\Gamma(x+1)\Gamma(\kappa)\vartheta^\kappa} \\
&= \frac{\vartheta^x\Gamma(\kappa+x)}{(\vartheta+1)^{\kappa+x}\Gamma(x+1)\Gamma(\kappa)} = \begin{cases} \frac{\vartheta^x(\kappa+x-1)\cdot\ldots\cdot\kappa}{(\vartheta+1)^{\kappa+x}\Gamma(x+1)} & \text{if } x \geq 1, \\ \frac{1}{(1+\vartheta)^\kappa} & \text{if } x = 0. \end{cases} \\
w_{\vartheta,\kappa}(x+1) &= \frac{\vartheta^{x+1}\Gamma(\kappa+x+1)}{(\vartheta+1)^{\kappa+x+1}\Gamma(x+2)\Gamma(\kappa)} \\
&= \frac{\vartheta(\kappa+x)}{(\vartheta+1)(x+1)} \cdot \underbrace{\frac{\vartheta^x\Gamma(\kappa+x)}{(\vartheta+1)^{\kappa+x}\Gamma(x+1)\Gamma(\kappa)}}_{=w_{\vartheta,\kappa}(x)}.
\end{aligned}$$

For the proof of b) we have

$$\begin{aligned}
& & w_{\vartheta,\kappa}(x+1) &\gtreqqless w_{\vartheta,\kappa}(x) \\
&\Leftrightarrow & \frac{\vartheta(\kappa+x)}{(\vartheta+1)(x+1)} &\gtreqqless 1 \\
&\overset{(\vartheta+1)(x+1)>0}{\Leftrightarrow} & \vartheta(\kappa+x) &\gtreqqless (\vartheta+1)(x+1) \\
&\Leftrightarrow & \vartheta\kappa + \vartheta x &\gtreqqless \vartheta x + x + \vartheta + 1 \\
&\Leftrightarrow & \vartheta\kappa - \vartheta - 1 &\gtreqqless x \\
&\Leftrightarrow & \underbrace{\vartheta(\kappa-1)-1}_{=x_0} &\gtreqqless x.
\end{aligned}$$

If $\kappa \leq 1$, we have $x_0 = \vartheta\underbrace{(\kappa-1)}_{<0}-1 \leq -1 < 0$ since $\vartheta > 0$ and hence $w_{\vartheta,\kappa}$ is strictly decreasing in $x = \{0, 1, \ldots\}$.

Let $\kappa > 1$ and $\vartheta < \frac{1}{\kappa-1}$. Then $x_0 = \vartheta(\kappa-1) - 1 < 0$ and $w_{\vartheta,\kappa}$ is strictly decreasing in $x = \{0, 1, \ldots\}$.

Let $\kappa > 1$ and $\vartheta \geq \frac{1}{\kappa-1}$. Then $w_{\vartheta,\kappa}(x)$ is strictly increasing for $x \in \{0, \dots, x_0\}$ and strictly decreasing for $x \in \{x_0+1, x_0+2, \dots\}$ with $w_{\vartheta,\kappa}(x_0) = w_{\vartheta,\kappa}(x_0+1)$ if $x_0 = \vartheta(\kappa-1) - 1 \in \mathbb{N}_0$. If $x_0 = \vartheta(\kappa-1) - 1 \notin \mathbb{N}_0$, then $w_{\vartheta,\kappa}(x)$ is strictly increasing for $x \in \{0, \dots, \lceil x_0 \rceil\}$ and strictly decreasing for $x \in \{\lceil x_0 \rceil, \lceil x_0 + 1 \rceil, \dots\}$.

4.A.2 Proof of Proposition 4.2

We prove the assertions about the prediction likelihood ratio $Q_{\vartheta,\kappa,\lambda}$.

We derive the explicit formula (4.20) and recursive formula (4.21) for the prediction likelihood ratio.

$$\begin{aligned}
Q_{\vartheta,\kappa,\lambda}(x) &= \frac{\lambda^x \exp(-\lambda)\Gamma(\kappa)\vartheta^\kappa}{\int_0^{+\infty} y^{x+\kappa-1} \exp\left(-y - \frac{y}{\vartheta}\right)\, \mathrm{d}y} \\
&= \frac{\lambda^x \exp(-\lambda)}{\Gamma(x+1) \cdot \frac{1}{\Gamma(x+1)\Gamma(\kappa)\vartheta^\kappa} \int_0^{+\infty} y^{x+\kappa-1} \exp\left(-y - \frac{y}{\vartheta}\right)\, \mathrm{d}y} \\
&= \frac{\lambda^x \exp(-\lambda)}{\Gamma(x+1) w_{\vartheta,\kappa}(x)} \\
&\overset{\text{Prop. 4.1}}{=} \frac{\lambda^x \exp(-\lambda)(\vartheta+1)^{\kappa+x}\Gamma(x+1)\Gamma(\kappa)}{\Gamma(x+1)\vartheta^x\Gamma(\kappa+x)} \\
&= \frac{\lambda^x \exp(-\lambda)(\vartheta+1)^{\kappa+x}\Gamma(\kappa)}{\vartheta^x\Gamma(\kappa+x)}.
\end{aligned}$$

$$\begin{aligned}
Q_{\vartheta,\kappa,\lambda}(x+1) &= \frac{\lambda^{x+1} \exp(-\lambda)(\vartheta+1)^{\kappa+x+1}\Gamma(\kappa)}{\vartheta^{x+1}\Gamma(\kappa+x+1)} \\
&= \frac{\lambda(\vartheta+1)}{(\kappa+x)\vartheta} \underbrace{\frac{\lambda^x \exp(-\lambda)(\vartheta+1)^{\kappa+x}\Gamma(\kappa)}{\vartheta^x\Gamma(\kappa+x)}}_{=Q_{\vartheta,\kappa,\lambda}(x)} \\
&= \frac{\lambda(\vartheta+1)}{(\kappa+x)\vartheta} \cdot Q_{\vartheta,\kappa,\lambda}(x).
\end{aligned}$$

This proves assertion a).

For assertion b), we obtain from a) with $Q_{\vartheta,\kappa,\lambda} > 0$ that the sign of $Q_{\vartheta,\kappa,\lambda}(x+1) - Q_{\vartheta,\kappa,\lambda}(x)$ is the sign of $\frac{\lambda(\vartheta+1)}{(\kappa+x)\vartheta} - 1$. We have

$$\begin{aligned}
& \frac{\lambda(\vartheta+1)}{(\kappa+x)\vartheta} - 1 \quad \substack{>\\=\\<} \quad 0 \\
\overset{\kappa+x>0}{\Leftrightarrow} \quad & \frac{\lambda(\vartheta+1)}{\vartheta} \quad \substack{>\\=\\<} \quad \kappa + x \\
\Leftrightarrow \quad & \frac{\lambda(\vartheta+1)}{\vartheta} - \kappa \quad \substack{>\\=\\<} \quad x.
\end{aligned}$$

The assertion follows.

Assertion c) is easily seen by $Q_{\vartheta,\kappa,\lambda}(0) = \frac{\lambda^0 \exp(-\lambda)(\vartheta+1)^\kappa \Gamma(\kappa)}{\vartheta^0 \Gamma(\kappa+0)} = \exp(-\lambda)(\vartheta+1)^\kappa \overset{\vartheta>0}{>} 0$ and

$$\begin{aligned}
\lim_{x\to\infty} Q_{\vartheta,\kappa,\lambda}(x) &= \lim_{x\to\infty} \frac{\lambda^x \exp(-\lambda)(\vartheta+1)^{\kappa+x}\Gamma(\kappa)}{\vartheta^x \Gamma(\kappa+x)} \\
&= \exp(-\lambda)\Gamma(\kappa)(\vartheta+1)^\kappa \underbrace{\lim_{x\to\infty} \left(\frac{\lambda(\vartheta+1)}{\vartheta}\right)^x \cdot \frac{1}{\Gamma(\kappa+x)}}_{=0} = 0.
\end{aligned}$$

For the proof of d) consider the derivation $\frac{\mathrm{d}}{\mathrm{d}\lambda} Q_{\vartheta,\kappa,\lambda}(x)$:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\lambda} Q_{\vartheta,\kappa,\lambda}(x) &= \frac{\mathrm{d}}{\mathrm{d}\lambda} \lambda^x \exp(-\lambda) \frac{(\vartheta+1)^{\kappa+x}\Gamma(\kappa)}{\vartheta^\kappa \Gamma(\kappa+x)} \\
&= \frac{(\vartheta+1)^{\kappa+x}\Gamma(\kappa)}{\vartheta^\kappa \Gamma(\kappa+x)} \left(x\lambda^{x-1}\exp(-\lambda) - \lambda^x \exp(-\lambda)\right) \\
&= \underbrace{\frac{(\vartheta+1)^{\kappa+x}\Gamma(\kappa)}{\vartheta^\kappa \Gamma(\kappa+x)}}_{>0} \underbrace{\lambda^{x-1}}_{>0} \underbrace{\exp(-\lambda)}_{>0} (x-\lambda)\,.
\end{aligned}$$

Then

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\lambda} Q_{\vartheta,\kappa,\lambda}(x) > 0 &\quad\Leftrightarrow\quad \lambda < x, \\
\frac{\mathrm{d}}{\mathrm{d}\lambda} Q_{\vartheta,\kappa,\lambda}(x) = 0 &\quad\Leftrightarrow\quad \lambda = x, \\
\frac{\mathrm{d}}{\mathrm{d}\lambda} Q_{\vartheta,\kappa,\lambda}(x) < 0 &\quad\Leftrightarrow\quad \lambda > x.
\end{aligned}$$

Hence $Q_{\vartheta,\kappa,\lambda}(x)$ as a function of λ is strictly increasing on $(0;x]$ and strictly decreasing on $[x;+\infty)$.

4.A.3 Proof of Proposition 4.3

Consider prediction regions with characteristics Q1). By Proposition 2.3, $A_x = \{y \in (0;+\infty) | c_L(y) \leq x \leq c_U(y)\}$ is an interval for all $x \in \{0,1,\ldots\}$ with endpoints $\inf A_x$, $\sup A_x$ increasing in x. We have to prove that the endpoints $\inf A_x$, $\sup A_x$ are not elements of A_x for $x \in \{0,1,\ldots\}$. Let $x \in \{0,1,\ldots\}$. Let (y_l) be a sequence in A_x with $y_1 > y_2 > \ldots$, $\lim_l y_l = \inf A_x$. Then $c_L(y_l) \leq x \leq c_U(y_l)$ for all l. Since c_U is left-continuous in $\inf A_x$, but not continuous in $\inf A_x$, c_U is not right-continuous in $\inf A_x$. Consequently, we have $\lim_l c_U(y_l) \neq c_U(\lim_l y_l) = c_U(\inf A_x)$. In particular, since c_U is increasing in $y \in (0;+\infty)$, we have $\inf c_U(y_l) = \lim_l c_U(y_l) > c_U(\inf A_x)$. Consequently, $x > c_U(\inf A_x)$ and $\inf A_x \notin A_x$. The proof of $\sup A_x \notin A_x$ follows analogously.

Assertion b) is an application of Proposition 2.3.

4.A.4 Proof of Proposition 4.4

To prove Proposition 4.4, we need the following proposition on the Poisson OC $L_c(\lambda)$:

Proposition 4.12 (Poisson OC)**.** *Let $c \in \mathbb{N}_0$ and $L_c(\lambda)$ the OC function of the Poisson distribution from Eq. (4.2). For $\lambda > 0$ we have*

$$\frac{d}{d\lambda} L_c(\lambda) = -\frac{\lambda^c}{c!} \exp(-\lambda).$$

Then $L_c(\lambda)$ is strictly decreasing on $(0; +\infty)$ with $\lim\limits_{\lambda \to 0} L_c(\lambda) = 1$, $\lim\limits_{\lambda \to +\infty} L_c(\lambda) = 0$.

Proof. For the derivative of $L_c(\lambda) = \sum\limits_{x=0}^{c} \frac{\lambda^x}{x!} e^{-\lambda} = \exp(-\lambda) + \sum\limits_{x=1}^{c} \frac{\lambda^x}{x!} e^{-\lambda}$ we have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\lambda} L_c(\lambda) &= -\exp(-\lambda) + \sum_{x=1}^{c} \left\{ \frac{x}{x!} \lambda^{x-1} \exp(-\lambda) - \frac{\lambda^x}{x!} \exp(-\lambda) \right\} \\
&= -\exp(-\lambda) + \sum_{x=1}^{c} \left\{ \frac{\lambda^{x-1}}{(x-1)!} \exp(-\lambda) - \frac{\lambda^x}{x!} \exp(-\lambda) \right\} \\
&= -\exp(-\lambda) + \sum_{x=0}^{c-1} \frac{\lambda^x}{x!} \exp(-\lambda) - \sum_{x=1}^{c} \frac{\lambda^x}{x!} \exp(-\lambda) \\
&= -\exp(-\lambda) + \frac{\lambda^0}{0!} \exp(-\lambda) + \sum_{x=1}^{c-1} \frac{\lambda^x}{x!} \exp(-\lambda) - \sum_{x=1}^{c-1} \frac{\lambda^x}{x!} \exp(-\lambda) - \frac{\lambda^c}{c!} \exp(-\lambda) \\
&= -\underbrace{\frac{\lambda^c}{c!}}_{>0} \underbrace{\exp(-\lambda)}_{>0} \quad < \quad 0.
\end{aligned}$$

Hence $L_c(\lambda)$ is strictly decreasing on $(0; +\infty)$. We have

$$\begin{aligned}
\lim_{\lambda \to 0} L_c(\lambda) \quad &= \quad \lim_{\lambda \to 0} \sum_{x=0}^{c} \frac{\lambda^x}{x!} e^{-\lambda} \quad = \quad \lim_{\lambda \to 0} \frac{\lambda^0}{0!} \exp(-\lambda) + \sum_{x=1}^{c} \frac{1}{x!} \underbrace{\lim_{\lambda \to 0} \frac{\lambda^x}{e^\lambda}}_{=0} \\
&= \quad \underbrace{\lim_{\lambda \to 0} \lambda^0}_{=1} \cdot \underbrace{\lim_{\lambda \to 0} \exp(-\lambda)}_{=1} \quad = \quad 1, \\
\lim_{\lambda \to +\infty} L_c(\lambda) \quad &= \quad \lim_{\lambda \to +\infty} \sum_{x=0}^{c} \frac{\lambda^x}{x!} e^{-\lambda} \quad = \quad \sum_{x=0}^{c} \frac{1}{x!} \underbrace{\lim_{\lambda \to +\infty} \frac{\lambda^x}{e^\lambda}}_{=0} \quad = \quad 0.
\end{aligned}$$

□

We can now prove Proposition 4.4:

Assertion a) of Proposition 4.4 follows directly from the definition of $\lambda_{x,\gamma}$ and $\widetilde{\lambda}_{x,\gamma}$, and monotonicity properties of the Poisson OC function from Proposition 4.12.

For the proof of assertion b), let $\gamma \geq 0.5, x > 0$. Then

$$L_x(\widetilde{\lambda}_{x,\gamma}) = 1 - \gamma \overset{\gamma\geq 0.5}{\leq} \gamma = L_{x-1}(\lambda_{x,\gamma}) < L_x(\lambda_{x,\gamma}),$$

hence $\widetilde{\lambda}_{x,\gamma} > \lambda_{x,\gamma}$ since L_x is strictly decreasing on $(0;+\infty)$, see Proposition 4.12. In the case $x = 0$, we have by definition $\lambda_{x,\gamma} = 0 < \widetilde{\lambda}_{x,\gamma}$.

For the proof of assertion c), let $x \in \mathbb{N}_0, y \in (\lambda_{x,\gamma}; \widetilde{\lambda}_{x,\gamma})$. Then $L_x(\widetilde{\lambda}_{x,\gamma}) < L_x(y) < L_x(\lambda_{x,\gamma})$. Assume $x \leq c_L(y) - 1$. Then $L_x(y) \leq L_{c_L(y)-1}(y)$ and

$$L_{c_U(y)}(y) - L_{c_L(y)-1}(y) \leq 1 - L_{c_L(y)-1}(y) \leq 1 - L_x(y) < 1 - \underbrace{L_x(\widetilde{\lambda}_{x,\gamma})}_{=1-\gamma} = \gamma,$$

in contradiction to property (4.22). Now assume $x \geq c_U(y) + 1$. Then $c_U(y) \leq x - 1$ and $L_{c_U(y)}(y) \leq L_{x-1}(y)$ and

$$L_{c_U(y)}(y) - L_{c_L(y)-1}(y) \leq L_{c_U(y)}(y) \leq L_{x-1}(y) < L_{x-1}(\lambda_{x,\gamma}) = \gamma,$$

in contradiction to property (4.22). This proves $c_L(y) \leq x \leq c_U(y)$.

4.A.5 Proof of Proposition 4.5

We derive the monotonicity properties of the prediction region coverage function.

Assertion a) is obvious from Proposition 4.12.

For assertion b), let $x_1 > 0$. By Proposition 4.12 we have for $y > 0$

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}y}\Delta_{x_1,x_2}(y) &= \frac{\mathrm{d}}{\mathrm{d}y}L_{x_2}(y) - \frac{\mathrm{d}}{\mathrm{d}y}L_{x_1-1}(y) = \frac{y^{x_1-1}}{(x_1-1)!}\exp(-y) - \frac{y^{x_2}}{x_2!}\exp(-y) \\ &= \underbrace{\exp(-y)}_{>0}\left(\frac{y^{x_1-1}}{(x_1-1)!} - \frac{y^{x_2}}{x_2!}\right). \end{aligned}$$

Hence

$$\begin{aligned} & 0 &\lesseqgtr\;& \frac{\mathrm{d}}{\mathrm{d}y}\Delta_{x_1,x_2}(y) \\ \Leftrightarrow\quad & \frac{y^{x_2}}{x_2!} &\lesseqgtr\;& \frac{y^{x_1-1}}{(x_1-1)!} \\ \Leftrightarrow\quad & \frac{y^{x_2}}{y^{x_1-1}} &\lesseqgtr\;& \frac{x_2!}{(x_1-1)!} \\ \Leftrightarrow\quad & y^{x_2-x_1+1} &\lesseqgtr\;& x_1 \cdot \ldots \cdot x_2 \\ \Leftrightarrow\quad & y &\lesseqgtr\;& (x_1 \cdot \ldots \cdot x_2)^{\frac{1}{x_2-x_1+1}}. \end{aligned}$$

This proves Proposition 4.5.

4.A.6 Proof of Proposition 4.6

We prove the properties about the comparison of two likelihood ratios.

With Proposition 4.2 we get

$$\begin{aligned}\frac{Q_{\vartheta,\kappa,y}(x_1)}{Q_{\vartheta,\kappa,y}(x_2)} &= \frac{\frac{y^{x_1}\exp(-y)(\vartheta+1)^{\kappa+x_1}\Gamma(\kappa)}{\vartheta^{x_1}\Gamma(\kappa+x_1)}}{\frac{y^{x_2}\exp(-y)(\vartheta+1)^{\kappa+x_2}\Gamma(\kappa)}{\vartheta^{x_2}\Gamma(\kappa+x_2)}} = \frac{y^{x_1}\vartheta^{x_2}(\vartheta+1)^{\kappa+x_1}\Gamma(\kappa+x_2)}{y^{x_2}\vartheta^{x_1}(\vartheta+1)^{\kappa+x_2}\Gamma(\kappa+x_1)} \\ &= y^{x_1-x_2}\left(\frac{\vartheta}{\vartheta+1}\right)^{x_2-x_1}\Big((\kappa+x_1)\cdot\ldots\cdot(\kappa+x_2-1)\Big),\end{aligned}$$

and hence

$$\begin{aligned} && Q_{\vartheta,\kappa,y}(x_1) &\overset{<}{\underset{>}{=}} Q_{\vartheta,\kappa,y}(x_2) \\ \Leftrightarrow && \frac{Q_{\vartheta,\kappa,y}(x_1)}{Q_{\vartheta,\kappa,y}(x_2)} &\overset{<}{\underset{>}{=}} 1 \\ \Leftrightarrow && y^{-(x_2-x_1)}\left(\frac{\vartheta}{\vartheta+1}\right)^{x_2-x_1}\Big((\kappa+x_1)\cdot\ldots\cdot(\kappa+x_2-1)\Big) &\overset{<}{\underset{>}{=}} 1 \\ \Leftrightarrow && \left(\frac{\vartheta}{\vartheta+1}\right)^{x_2-x_1}\Big((\kappa+x_1)\cdot\ldots\cdot(\kappa+x_2-1)\Big) &\overset{<}{\underset{>}{=}} y^{x_2-x_1} \\ \Leftrightarrow && \left(\left(\frac{\vartheta}{\vartheta+1}\right)^{x_2-x_1}\Big((\kappa+x_1)\cdot\ldots\cdot(\kappa+x_2-1)\Big)\right)^{\frac{1}{(x_2-x_1)}} &\overset{<}{\underset{>}{=}} y \\ \Leftrightarrow && \frac{\vartheta}{\vartheta+1}\Big((\kappa+x_1)\cdot\ldots\cdot(\kappa+x_2-1)\Big)^{\frac{1}{(x_2-x_1)}} &\overset{<}{\underset{>}{=}} y.\end{aligned}$$

4.A.7 Proof of Proposition 4.7

Let $F_X(c)$ denote the distribution function of the Poisson distribution from Eq. (4.2). From Proposition 4.2, assertion b), we know that the likelihood ratio $Q_{\vartheta,\kappa,y}(x)$ as a function of x is decreasing or of inverted bathtub shape. From Proposition 4.2, assertion c), we know $Q_{\vartheta,\kappa,y}(0) > 0$ and $\lim_{x\to\infty} Q_{\vartheta,\kappa,y}(x) = 0$. Consequently, there exists an $x_1 \in \{1,2,\ldots\}$ such that $Q_{\vartheta,\kappa,y}(0) > Q_{\vartheta,\kappa,y}(x_1)$. Let x_2 be a natural number with $x_2 \geq x_1$ and $\mathrm{P}_y\left(X \in \{0,1,\ldots,x_2\}\right) = F_X(x_2) > \gamma$. The existence of such an x_2 is clear since $F_X(x)$ is a distribution function and hence increasing in x with $\lim_{x\to\infty} F_X(x) = 1 > \gamma$. Furthermore, x_2 is not unique and any $x > x_2$ also fulfils the requirements. Let $\widetilde{s_y} := Q_{\vartheta,\kappa,y}(x_2)$. Then $D_{\geq\widetilde{s_y}} = \{0,1,\ldots,x_2\}$.

We prove $\widetilde{s_y} \leq s_y$ by contradiction:

Assume $\widetilde{s_y} > s_y$. From the definition of s_y we have

$$s_y = \inf\{t|G_y(t) \leq \gamma\} = \inf\{t|\mathrm{P}_y(Q_{\vartheta,\kappa,y}(X) > t) \leq \gamma\}.$$

From Proposition 2.1 we know that $G_y(t) = \mathrm{P}_y(Q_{\vartheta,\kappa,y}(X) > t)$ is decreasing in t. Therefore, we have for $\widetilde{s_y}$ necessarily $G_y(\widetilde{s_y}) = \mathrm{P}_y(Q_{\vartheta,\kappa,y}(X) > \widetilde{s_y}) \leq \mathrm{P}_y(Q_{\vartheta,\kappa,y}(X) \geq \widetilde{s_y}) = \mathrm{P}_y(X \in D_{\geq \widetilde{s_y}}) = \mathrm{P}_y(X \in \{0, 1, \ldots, x_2\}) \leq \gamma$, a contraction against the choice of x_2. This proves $\widetilde{s_y} \leq s_y$. From $Q_{\vartheta,\kappa,y}(x) \geq s_y \geq \widetilde{s_y}$ then follows $Q_{\vartheta,\kappa,y}(x) \geq \widetilde{s_y}$ for an $x \in \mathbb{N}_0$ and therefore $\{x | Q_{\vartheta,\kappa,y}(x) \geq s_y\} \subset \{x | Q_{\vartheta,\kappa,y}(x) \geq \widetilde{s_y}\}$. Hence we can conclude $D_{\geq s_y} \subset D_{\geq \widetilde{s_y}} = \{0, 1, \ldots, x_2\}$, which proves the assertion of the proposition. By letting $c := x_2$, we obtain a natural number fulfiling the condition of Proposition 4.7.

4.A.8 Proof of Proposition 4.8

The prior density of Y is given by $f_Y(y) = \dfrac{y^{\kappa-1}}{\vartheta^\kappa \Gamma(\kappa)} \exp\left(-\dfrac{y}{\vartheta}\right)$. Hence by Theorem 3.1 we have

$$\begin{aligned} f_{Y|X=x}(y) \propto f_Y(y) f_{X|Y=y}(x) &= \frac{y^{\kappa-1}}{\vartheta^\kappa \Gamma(\kappa)} \exp\left(-\frac{y}{\vartheta}\right) \frac{y^x}{x!} \exp(-y) \\ &\propto y^{\kappa+x-1} \exp\left(-y\left(\frac{\vartheta+1}{\vartheta}\right)\right), \end{aligned}$$

i. e. the posterior density $f_{Y|X=x}(y)$ is proportional to the density of the gamma distribution Gamma $\left(\kappa + x, \frac{\vartheta}{\vartheta+1}\right)$ and therefore results to

$$f_{Y|X=x}(y) = \frac{1}{\left(\frac{\vartheta}{\vartheta+1}\Gamma(\kappa+x)\right)} y^{\kappa+x-1} \exp\left(-\frac{y}{\frac{\vartheta}{\vartheta+1}}\right).$$

4.A.9 Proof of Proposition 4.9

We derive monotonicity properties of the density function of the gamma distribution.

Let $\kappa = 1$. The derivative of

$$f_{\vartheta,1}(x) = \begin{cases} \frac{1}{\vartheta} \exp\left(\frac{-x}{\vartheta}\right) & \text{for } x \geq 0, \\ 0, & \text{elsewhere}, \end{cases}$$

for $x > 0$ is $\dfrac{\mathrm{d}}{\mathrm{d}x} f_{\vartheta,1}(x) = \dfrac{1}{\vartheta} \exp\left(\dfrac{-x}{\vartheta}\right)\left(-\dfrac{1}{\vartheta}\right) < 0$, hence $f_{\vartheta,\kappa}(x)$ is strictly decreasing.
Let $\kappa \neq 1$. The derivative of

$$f_{\vartheta,\kappa}(x) = \begin{cases} \frac{x^{\kappa-1}}{\vartheta^\kappa \Gamma(\kappa)} \exp\left(\frac{-x}{\vartheta}\right) & \text{for } x \geq 0, \\ 0, & \text{elsewhere}, \end{cases}$$

for $x > 0$ is

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x} f_{\vartheta,\kappa}(x) &= \frac{1}{\vartheta^\kappa \Gamma(\kappa)} \left((\kappa-1)x^{\kappa-2} \exp\left(-\frac{x}{\vartheta}\right) + x^{\kappa-1} \exp\left(-\frac{x}{\vartheta}\right) \left(-\frac{1}{\vartheta}\right) \right) \\ &= \underbrace{\frac{1}{\vartheta^\kappa \Gamma(\kappa)}}_{>0} \underbrace{x^{\kappa-2}}_{\overset{x>0}{>}0} \underbrace{\exp\left(-\frac{x}{\vartheta}\right)}_{>0} \left((\kappa-1) - \frac{x}{\vartheta} \right).\end{aligned}$$

Therefore, the sign of $\frac{\mathrm{d}}{\mathrm{d}x} f_{\vartheta,\kappa}(x)$ is the sign of $(\kappa-1) - \frac{x}{\vartheta}$. We have $(\kappa-1) - \frac{x}{\vartheta} < 0$ if $\kappa < 1$. For $\kappa > 1$ we have

$$\frac{\mathrm{d}}{\mathrm{d}x} f_{\vartheta,\kappa}(x) \begin{cases} > 0 \text{ if } 0 < x < \vartheta(\kappa-1), \\ = 0 \text{ if } x = \vartheta(\kappa-1), \\ < 0 \text{ if } x > \vartheta(\kappa-1), \end{cases}$$

and the assertion follows.

4.A.10 Proof of Corollary 4.11

Assertion (a) follows directly from Corollary 4.10. In assertion (b), $y^* \in A_x$ follows from $f_{Y|X=x}(y^*) \geq f_{Y|X=x}(y)$ for every $y \in (0; +\infty)$. The existence of a value

$$f_{Y|X=x}\left(z_{\mathrm{Gamma}\left(\kappa+x, \frac{\vartheta}{\vartheta+1}\right)}(\alpha_1)\right) = f_{Y|X=x}\left(z_{\mathrm{Gamma}\left(\kappa+x, \frac{\vartheta}{\vartheta+1}\right)}(1-\alpha_2)\right) \in (0; y^*)$$

with

$$\mathrm{P}\left(z_{\mathrm{Gamma}\left(\kappa+x, \frac{\vartheta}{\vartheta+1}\right)}(\alpha_1) \leq Y \leq z_{\mathrm{Gamma}\left(\kappa+x, \frac{\vartheta}{\vartheta+1}\right)}(1-\alpha_2)\right) = \beta \in (0; 1)$$

follows from the continuity of the density of the posterior distribution Gamma $\left(\kappa + x, \frac{\vartheta}{\vartheta+1}\right)$ on $(0; +\infty)$ and $f_{Y|X=x}(0) = 0 = \lim\limits_{y \to \infty} f_{Y|X=x}(y)$.

5 Confidence Intervals in Audit Populations

5.1 Introduction

Auditing is concerned with the independent verification of the fairness of financial statements, which summarise and report a company's economic activities, see CPSMR (1988). According to the American Accounting Association (as cited in Gillet (2000)), auditing can be defined as "a systematic process of objectively obtaining and evaluating evidence regarding assertions about economic actions and events to ascertain the degree of correspondence between those assertions and established criteria and communicating the results to interested users". An audit procedure involves a variety of audit tests. Among them are the *compliance test*, which investigates whether the accounting procedure is in accordance with prescribed procedures for internal control. In a *substantive test of details*, the total monetary error in the balance is of interest, see CPSMR (1988).

The conformance of monetary book values U kept in lists and databases on items like accounts, articles in an inventory or transactions is compared with the corresponding de facto values or audit values W of the items in reality. The degree of misstatement of U on W can be measured by the tainting ratio $Y = (U - W)/U$, i. e. the deviation of the book value U from the de facto value W relative to the stipulated book value. Misstatements occur when the amount stated in the balance is not identical to the amount in reality. In many contexts, misstatements tend to be overstatements where $0 \leq W < U$. Misstatement by overstatement is the dominant error mode particularly in asset accounts, accounts receivable and revenue accounts, see the empirical studies by Ramage et al. (1979), Johnson et al. (1981), Ham et al. (1985) and Icerman & Hillison (1990). Under the overstatement error mode – which is what the statistical methodology presented in this chapter focuses on –, the tainting ratio ranges from 0 to 1, where $Y = 0$ represents a correct statement, and $Y = 1$ represents the case that a de facto value $W = 0$ is erroneously stated by a positive account entry $U > 0$.

In regular accounting practice, most account entries are correct, and small deviations are more frequent than large deviations. Hence the regular distribution of the tainting ratio has three properties: i) highly right-skewed, ii) zero inflation, i. e. large probability point mass at 0, iii) probability density function decreasing on $[0;1]$. This pattern is often addressed as "reverse J-shaped", see e. g. Neter et al. (1985).

CPSMR (1988) stress the necessity of acquiring reliable audit information at low cost. Statistical sampling can contribute to this requirement by examining only a sample of audit evidence instead of the whole population. In consequence, there is a need for inferential procedures that judge an audit population on the basis of a sample. Stringer (1963) and many subsequent authors warned against using methods based on normal distribution approximations for the statistical inference on tainting distributions. In particular, because of the large probability mass at $Y = 0$, confidence intervals for the mean as known from survey sampling are evidently questionable in the analysis of tainting distributions. Various authors, for example Kaplan (1973), Teitlebaum & Robinson (1975) and Neter & Loebbecke (1977), have demonstrated that normal confidence limits cannot guarantee the prescribed confidence level when sampling from tainting distributions. Not rarely, only few or no errors are observed and a reliable estimation of a variance is impossible.

The basic heuristic idea of a specific upper confidence bound μ_U for the population mean μ in sampling inference on audit populations was launched by Stringer (1963). The formal development of the bound nowadays associated with Stringer's name (*Stringer bound*) is mainly due to studies of Meikle (1972), Anderson & Teitlebaum (1973) and Goodfellow et al. (1974). The Stringer bound μ_Y was intended as a robust alternative, conservative in the sense that a prescribed nominal confidence level γ is exceeded by the actual confidence level $\mathrm{P}(\mu \leq \mu_Y)$. Stringer's (1963) approach was motivated by financial auditing. However, the problem of analysing populations with zero inflation arises in other contexts, too. Examples are accident costs in insurance, contamination and earthquake measurement. In addition, zero inflation can be a consequence of measurement imprecision where small signals are cumulated on zero.

Not strict model-based theory, but intuitive heuristic reasoning drove the initiation of the Stringer bound. It was several decades after Stringer's (1963) paper that Bickel (1992) presented a motivation for the Stringer bound. The intuitive basis in a multinomial error model is nicely described by Pap & van Zuijlen (1996). Many authors have investigated the actual confidence level of the Stringer confidence bound, both by simulation and analytical studies. Throughout, simulation studies have corroborated the conservatism of the Stringer bound, see Burdick & Reneau (1978), Reneau (1978), Leitch et al. (1982), Plante et al. (1985) and Tsui et al. (1985). Bickel (1992) initiated the study of the asymptotic behaviour of the bound for large sample size n. Pap & van

Zuijlen (1995) provided finite sampling results for uniformly distributed taintings, and a finite sampling example for the anticonservatism in the case of $0 < \gamma < 0.5$. Pap & van Zuijlen (1996) showed that the bound is asymptotically conservative for confidence level $0.5 \leq \gamma < 1$, and suggested a modified bound that has asymptotically the nominal confidence level. Further asymptotic results and modifications of the bound so as to remove the anticonservatism in the case of $0 < \gamma < 0.5$ were developed by de Jager et al. (1997). Godfrey & Neter's (1984) idea was to use a Bayesian approach to exploit prior knowledge. Fienberg et al. (1977) and Neter et al. (1978) introduced the multinomial bound which essentially recurs to the heuristic basis underlying the Stringer bound, see Pap & van Zuijlen (1996).

A small total amount of misstatements is usually tolerated in an audit. A misstatement beyond a tolerance level is called a *material misstatement*. IAASB (2013) define materiality with respect to financial misstatements as follows: "Misstatements, including omissions, are considered to be material if they, individually or in the aggregate, could reasonably be expected to influence the economic decisions of users taken on the basis of the financial statements" (IAASB 2013, ISA 320). In view of financial, administrative, and legal consequences, the auditor's primary interest is to restrict the type 1 risk of incorrectly not detecting an excessive mean tainting in the accounts. This risk of incorrect acceptance of an account is what is frequently referred to as *beta risk* and is related to audit effectiveness (Guy et al. 2002). This interest has lead research to focus on upper confidence bounds for the mean tainting. However, the type 2 risk of incorrectly assuming an excessive mean tainting also has considerable impact on the auditor and the auditee. This risk, frequently referred to as *alpha risk*, is concerned with audit efficiency (Guy et al. 2002). In particular, the supposition of misstatements will increase subsequent audit efforts, to the disadvantage of both the auditor and the auditee. A lower confidence bound enables the rejection of the population if too many misstatements are detected in the sample. Accordingly, the Commission on Physical Sciences, Mathematics, and Applications (CPSMR 1988) considered lower confidence bounds as "an area of considerable importance where research is needed". In spite of this encouraging suggestion, lower confidence limits have not received much attention. A lower Stringer bound can simply be obtained by applying the Stringer method to the observations $1 - Y$, but these intervals use to be wide. Plante et al. (1984) developed a lower bound by the multinomial method. Tsui et al. (1985) take a Bayesian approach by assuming a Dirichlet prior for the multinomial probability parameters that can be used to construct either an upper or a lower bound. The simulation study by Matsumura et al.

(1991) showed that for both methods the actual confidence level of the lower bounds is reasonably close to the prespecified nominal level.

The study in the present chapter combines two objectives: 1) Find tight two-sided confidence intervals of Stringer type. 2) Establish a simple way of using prior information on the target population. This is achieved by using the shortest binomial confidence intervals introduced in Chapter 2. In this scheme, prior information is specified by postulating a beta distribution $\text{Beta}(p_0, p_1, a, b)$ with shape parameters $a, b > 0$ on the support $[p_0; p_1] \subset [0; 1]$ on the target probability p to be estimated. The case $p_0 = 0$, $p_1 = 1$, $a = b = 1$ represents the absence of prior information. By an appropriate choice of p_0, p_1, a, b, a high degree of prior knowledge can be expressed.

In our case, the target probability on which prior knowledge is applied is $p = \text{P}(Y > 0)$, in the auditing context the probability of an overstatement of the book value U over the de facto value W. From past audits and from the ongoing audit process, the auditor acquires considerable insight into a company's conduct, particularly by auditing the internal control system. Often, this information is sufficient to specify at least some upper bound p_1 for the overstatement rate p or a value that p is most likely not to exceed. In general, empirical studies show that extremely high misstatement rates close to 1 occur rarely, see Johnson et al. (1981) and Ham et al. (1985). Reliable prior knowledge on $p = \text{P}(Y > 0)$, particularly on upper bounds or some quantile, will mostly also be available in the analysis of other zero inflation phenomena like accident costs, contamination or earthquake measurements.

The subsequent sections are organised as follows: Section 5.2 reviews several sampling techniques that play a role in auditing. Section 5.3 describes some characteristics of audit populations that determine the sampling and evaluation method. The upper Stringer bound for the mean tainting as proposed by Stringer (1963) is reviewed in Section 5.4. In Section 5.5, Stringer's method is extended to the two-sided case. A way is proposed how to make the interval more precise by the use of prior information on the error probability p. Section 5.6 evaluates the performance of the two-sided Stringer bound by means of a simulation study. An indication how under- and overstatement errors could be simultaneously dealt with is provided in Section 5.7.

5.2 Sampling Techniques in Auditing

The necessity to consider sampling in auditing arises from the fact that audit populations are frequently very large and the inspection of the whole lot is unfeasible or extremely costly. Consequently, it is common practice to select only a sample for inspection. Statistical as well as non-statistical sampling concepts can be encountered. In non-statistical sampling, the probabilities of selecting items from the population are uncontrolled. *Judgmental selection* (also *purposive selection*), where the auditor selects specific items from a population that are deemed close to the population average (Guy et al. 2002), is an instance of non-statistical sampling. *Haphazard selection* or *block selection*, as the selection of adjacent or similar items or subsequent items in a time period, are non-statistical sampling techniques as well. This contrasts with statistical sampling, where all items have a positive, controlled probability of being in the sample. Instances of statistical sampling are simple random sampling, where all items have identical probabilities of being in the sample (Cochran 1963), structured random sampling, where the probabilities might be non-identical, stratified random sampling, systematic sampling with random start (e. g. every 10th item) or probability proportional to size (PPS) sampling.

The nature of audit sampling is thought of differently among the different auditing standards. According to AICPA (1981), "Audit sampling is the application of an audit procedure to less than 100 percent of the items within an account balance or class of transactions for the purpose of evaluating some characteristic of the balance or class." In particular, judgmental sampling can be considered audit sampling according to AICPA (1981). IAASB (2013, ISA 500) refrain from considering judgmental sampling as audit sampling: "While selective examination of specific items from a class of transactions or account balance will often be an efficient means of obtaining audit evidence, it does not constitute audit sampling. The results of audit procedures applied to items selected in this way cannot be projected to the entire population; accordingly, selective examination of specific items does not provide audit evidence concerning the remainder of the population" (IAASB 2013, ISA 500, A55). Their definition of audit sampling involves the additional requirement of the sampling units to have a chance of selection, a requirement which is violated by judgmental sampling: "Audit sampling (sampling) – The application of audit procedures to less than 100 % of items within a population of audit relevance such that all sampling units have a chance of selection in order to provide the auditor with a reasonable basis on which to draw conclusions about the entire population" (IAASB 2013, ISA 530).

The advantages of statistical sampling are obvious. Not only does it allow an objective statement about the population by means of mathematical calculations, but it also enables a quantification of the reliability on the sample as well as precise sample size calculations that avoid under- or overauditing, see Guy et al. (2002). Guy et al. (2002) explain that the mathematical evaluation, besides the random selection of sample items, is an integral part of statistical sampling. After the publication of the first literature on statistical sampling by the American Institute of Certified Public Accountants (AICPA) in 1962 (as cited in Guy et al. 2002), it had not been widely accepted for a long time because some practitioners believed statistical sampling would interfer with their professional judgment. However, Guy et al. (2002) stress that statistical sampling in auditing "sharpens the professional judgment of auditors and enhances understanding of the audit process". Statistical sampling became gradually more accepted in auditing procedures from about 1981 on, when in AICPA (1981) the section about statistical sampling moved to a more prominent place in the standard, see Guy et al. (2002).

Apart from the technique to select items from an audit population, sampling techniques can be distinguished by the nature of the target variable. The aim is to detect potential misstatements in a population. The items under investigation can be judged either by being misstated or not, or by their degree of misstatement. This determines the classification of sampling techniques into *attribute sampling* and *variable sampling*. The sampling technique *monetary unit sampling* has been developed for the specific auditing requirements. It is mostly applied in combination with an evaluation technique combining attribute sampling with variable sampling ideas, which coins the name *combined attribute and variable sampling*. Both terms are many times not distinguished because they often appear together. However, in theory, monetary unit sampling could be evaluated by means of other evaluation techniques, e. g. together with variable sampling. We deem it helpful to devote each of the two schemes their own section (Sections 5.2.3 and 5.2.4) to help avoid the confusion of both terms.

5.2.1 Attribute Sampling

Attribute sampling reports about sampled items only whether or not they share a certain characteristic or attribute. The purpose of this sampling technique is usually to estimate the proportion or percentage of items carrying these attributes. In Cochran (1963), this sampling technique can be found under the name *sampling for proportions and percentages*. If a certain item carries the characteristic of interest, it is commonly coded by 1, whereas an item not carrying the characteristic is coded by 0. In the auditing context, attribute sampling with its respective evaluation technique is primarily used in tests of controls, which are concerned with assessing the quality of internal control procedures. It finds application, for example, when the auditor is interested in estimating the frequency of inventory items which are not properly priced (as in inventory tests) or invoice quantities not conforming with shipping data (as in sales tests), see Guy et al. (2002).

Evaluation of samples by means of attribute sampling can be conducted, for instance, by means of confidence intervals for a proportion. If the population is of finite size, the hypergeometric distribution is the theoretically correct model for most applications, as in the example of auditing populations. However, the binomial or Poisson distribution are frequently used instead due to simpler calculations and tables, resulting in conservative approximations, see Gillet (2000). Using the binomial distribution as a model, it is assumed that irrespective of the number of items already drawn from a finite population, the probability of drawing an item having the characteristic of interest in n trials is constantly p. This assumption is justified if the population is large in relation to the sample, see Cochran (1963). The theoretical justification of using the binomial or Poisson distribution arises from the limiting characteristic of the binomial distribution approximating the hypergeometric if the population is large, and the Poisson approximating the binomial distribution, both limitations holding under certain parameter limiting characteristics.

Ways in which attribute sampling plans are applied in auditing are described in Guy et al. (2002).

5.2.2 Variable Sampling

In *variable sampling*, less the fact whether an item carries a certain feature or not is important, but rather its magnitude. It is applicable if the interest is on the mean or some other function of the sample values, see Loebbecke & Neter (1975). Variable sampling is useful for accounting applications rather than auditing applications (Guy et al. 2002), a typical field of application being substantive tests of details. Evaluation techniques following variable sampling are usually a point estimate and confidence interval for the total audit population value, see Reneau (1978). In this context, the feature of interest could be the total monetary value or the amount of misstatement of an account balance. Guy et al. (2002) describe several techniques (unstratified mean-per-unit, stratified mean-per-unit, difference estimation, ratio estimation) how to apply variable sampling to accounting estimation. The evaluation techniques frequently use normal distribution quantiles and hence vitally rely on the applicability of the central limit theorem. However, this is often not appropriate in auditing. The problem is the typical skewness in accounting populations, which have a large number of correctly stated items and a small number of misstated items, see Kaplan (1973), Neter & Loebbecke (1977), Johnson et al. (1981) and the brief review in Section 5.3 for a characterisation of audit populations. In the case that no errors are found in the sample, Stringer (1963) hints to the problem that this leads to an estimated standard error of zero in a ratio estimation procedure when in fact the estimate is not perfect. According to Kaplan (1973), a strong dependency between the estimated mean and the standard deviation lead to too narrow confidence intervals. Variable sampling and the associated evaluation techniques in auditing can be used when the population is not too heavily skewed and shows frequent misstatements of rather small magnitude in both directions. In other words, both under- and overstatement errors are present. These characteristics apply, for example, to inventory populations.

5.2.3 Monetary Unit Sampling

Monetary unit sampling (MUS) is a special case of *probability proportional to size sampling (PPS)*, the latter dating back to Hansen & Hurwitz (1943). In PPS, units are selected from the population with probabilities proportional to their sizes. Sampling is done with replacement to have the advantage of simpler formulas for the variances of the estimates. If the sample size is small in comparison to the population size, this is similar to sampling without replacement, see Cochran (1963). In accounting with the particular nature of the involved populations, MUS can be particularly effective and is recommended, for example, by Teitlebaum & Robinson (1975).

In MUS, a variant of PPS is applied with the size of each item being identical to the monetary value of the item. The idea to sample for individual dollars had already been suggested by Deming (1960). Among the first to apply the technique, which was frequently known under the name *dollar unit sampling (DUS)*, in auditing were Stringer (1963) and Teitlebaum (1973). Later, the technique was often more generally called *monetary unit sampling*. Instead of the number of items or accounts in the population, the number of individual monetary units is considered. Consequently, instead of population items having certain monetary values, individual monetary units are sampled. By that, monetary units corresponding to items having a larger monetary value are more likely to be in the sample than monetary units associated with units of smaller amount.

In accounting it is of interest to check for possible errors in the book values. If a monetary unit is sampled and a misstatement detected in the item to which the monetary unit belongs, the misstatement is assigned in equal parts to each of the monetary units building the item. For example, if the sampled monetary unit belongs to an item of the book value 100, which has de facto value 90 only and is hence overstated, each of the 100 monetary units of the item is 10 % overstated and has a de facto value of 0.90. Two or more monetary units from the same item can be part of the sample, all having equal tainting values, which would be 0.1 in the example. MUS employs the highest degree of stratification by book amount possible, such that advantages of stratification are exploited without really using stratification, see Neter et al. (1978). Sampling by stratification is in more detail described in Cochran (1963).

With respect to the question how to actually sample the individual monetary units, various solutions are available, among them *simple random sampling*, *systematic sampling* or *cell-selection*, see Teitlebaum (1973), Anderson & Teitlebaum (1973) or Leslie et al. (1979). However, with the availability of computers these days, suitable solutions for sampling monetary units can easily be implemented.

In the context of auditing, there are authors who under the term MUS or PPS understand the sampling technique described above in combination with the technique to evaluate the sample, as e. g. Guy et al. (2002). The evaluation technique is mostly a combination of attribute and variable sampling, delivering an upper bound for the total monetary misstatement in the accounting population. The reason for the confusion is that these combined attribute and variable sampling plans mostly make use of MUS to create the sample. However, this is not a necessity (although it proved to be advantageous in the auditing context due to Anderson & Teitlebaum 1973), which is why we treat what Guy et al. (2002) understands by PPS in the subsequent section.

5.2.4 Combined Attribute and Variable Sampling

Combined attribute and variable sampling (CAV) has emerged in statistical auditing through the endeavours to overcome deficiencies in both attribute and variable sampling plans (CPSMR 1988). CAV is more precisely described as an evaluation technique of a sample making use of both attribute and variable sampling evaluation techniques. In auditing, they are commonly applied in overstatements contexts, delivering an upper bound for the monetary amount of misstatement in the population, see Guy et al. (2002). The result of a CAV approach is usually a point estimate and confidence interval for the quantity under investigation. A simulation study comparing five different CAV approaches was conducted by Reneau (1978).

The most famous type of a CAV technique is the Stringer bound, proposed by Stringer (1963). It is commonly applied by making use of MUS as the corresponding sampling technique (Anderson & Teitlebaum 1973). We present the Stringer bound in more detail in Section 5.4.

5.3 Audit Populations

Neter & Loebbecke (1975, 1977) were the first to conduct an empirical investigation about the error characteristics and statistical procedures to evaluate them based on the errors discovered in four audits. Several other empirical studies revealed the error characteristics in audit population: While Neter & Loebbecke (1975), Ramage et al. (1979) and Johnson et al. (1981) examined inventory and accounts receivable populations only, Ham et al. (1985) additionally had access to accounts payable, purchases and sales.

The studies of Johnson et al. (1981), Ramage et al. (1979) and Ham et al. (1985) commonly found accounts receivable errors to be mostly overstatement errors while inventory populations are rather balanced in terms of overstatement and understatement errors. Accounts payable and purchases mostly come with understatement errors and sales with overstatement errors, as was found out by Ham et al. (1985). The findings of Icerman & Hillison (1990) are in accordance with this. Accounts receivable populations can have understatement in errors as well (Johnson et al. 1981; Ham et al. 1985). However, since they occur rather rarely, Johnson et al. (1981, p. 288) suggest that it might be useful to focus on the distribution of overstatement taintings when dealing with accounts receivable.

Regarding the error amount distributions, both Johnson et al. (1981) and Ham et al. (1985) found that the mean error detected for accounts receivable is larger, and in particular positive, than the mean error in inventory audits, which can turn out to be negative. Accounts payable and purchases have more negative error means, whereas sales have slightly more positive error means, see Ham et al. (1985). In terms of error rates, however, Johnson et al. (1981) found the error rates for inventory to be in tendency higher than those for accounts receivable with median errors rates of 0.154 (inventories) and 0.024 (accounts receivable), respectively.

The distributions of the different accounting classes clearly differ. Apart from possibly inventories, accounting error populations are not normally distributed (Ham et al. 1985; Icerman & Hillison 1990). Johnson et al. (1981) and Ham et al. (1985) both found the error distributions to be often heavily skewed. Accounts receivable show rather positive skewness, while accounts payable and inventories are frequently negatively skewed.

In many auditing procedures, not the absolute auditing error is considered, but the tainting Y_i, the relative deviation of the de facto value from the book value of an item i. Procedures making use of the tainting cannot deal with book values of 0 or less. These items, which do not occur very frequently, are often subject to separate auditing, see Johnson et al. (1981, p. 286). Just as well as they detected a balance between overstatement and understatement in the error amounts of inventory audits, Johnson et al. (1981) found negative and positive taintings with approximate equal relative frequency. While for inventory audits, the error tainting fell below -100 % in 2–5 % of the cases, it never did in the case of accounts receivable. For both types of account classes, the maximum taintings observed were not more than 125 %, whereby a considerable percentage of especially accounts receivable items have a tainting of 100 %, causing a discontinuity at that point. Regarding the location of the tainting distributions, inventory audits show a smaller mean tainting than receivable audits. This finding was supported by Ham et al. (1985). The shape of the taintings is of skewed kind. Inventory audits mostly show negative skewness, which sometimes even occurs in accounts receivable audits (Johnson et al. 1981).

5.4 The Stringer Bound

The Stringer bound, which was proposed by Stringer (1963), is an instance of a CAV technique (see Section 5.2.4) for an audit population in which overstatement is expected. The bound constitutes a nonparametric upper bound for the mean tainting Y in an audit population. Due to the assumption of overstatement and that bookvalues and de-facto-values are non-negative, the tainting is a quantity between 0 and 1. The taintings $Y_1, \ldots, Y_n$ of n sampled items are assumed to be realisations of an independent and identically distributed random variable Y. Since it relies on the definition of a tainting, the Stringer bound is applicable only to populations with book values $U_i > 0$. For a population in which bookvalues of 0 occur, the Stringer bound can be applied only to those items with bookvalue $U_i > 0$. Items with bookvalues of 0 are subject to a total inspection. An overview over the quantities related to the notation of the Stringer bound is given in Table 5.1. The sampling procedure to use with the Stringer bound was suggested to be DUS by Anderson & Teitlebaum (1973).

We present the upper Stringer bound for the mean of a population that consists of items with taintings between 0 and 1 with probability 1.

Table 5.1: Notation of important quantities

quantity	description
μ	average misstatement in the population
N	number of items in the population
$U_1, \dots, U_N$	book values
$W_1, \dots, W_N$	de facto values
$Y_i = \frac{U_i - W_i}{U_i}$	tainting of item i
n	sample size
$0 \leq Y_{(1,n)} \leq \dots \leq Y_{(n,n)} \leq 1$	ordered Y_i in the sample
$\gamma \in (0; 1)$	confidence level

Definition 5.1 (Stringer Bound). *Let $Y_1, \dots, Y_n$ be an i.i.d. sample of variables with $P(0 \leq Y \leq 1) = 1$. Let $0 \leq Y_{(1,n)} \leq \dots \leq Y_{(n,n)}$ be the corresponding ordered sample and $Y_{(n+1,n)} = 1$. Let $0 < \gamma < 1$ be the confidence level. For $k = -1, \dots, n+1$ let $0 \leq p_U(k) \leq 1$ be the upper confidence limit for the success probability p in a series of Bernoulli trials of size n if k successes are observed, where $p_U(-1) = 0$, $p_U(n+1) = 1$. The one-sided upper Stringer bound for $E[Y] = \mu_Y$ is defined by*

$$\mu_{ST} \quad := \quad p_U(0) + \sum_{x=1}^{n} \Big(p_U(x) - p_U(x-1) \Big) Y_{(n-x+1,n)}.$$

In Bickel (1992), $p_U(x)$ is defined to be the unique solution in p of the equation

$$\sum_{k=x+1}^{n} \binom{n}{k} p^k (1-p)^{n-k} \quad = \quad \gamma,$$

which corresponds to the upper bound of the one-sided exact confidence interval for a binomial probability from Eq. (2.4). According to Neter et al. (1978), $p_U(x)$ was almost always calculated by approximation through limits of the Poisson distribution when it came to applying the Stringer bound as an estimate of the upper mean tainting in an accounting population, i. e. Eq. (4.8) was used as an approximation for Eq. (2.4). The use of Poisson limits was introduced by Goodfellow et al. (1974). Their use is most certainly due to an easier handling of Poisson confidence limits in tables in comparison to binomial limits. With the possibilities to calculate confidence limits in real time, this argument now no longer holds. Nevertheless, the approach by the Poisson distribution is still advocated many times, as, for example, in Guy et al. (2002), where the method can be found hidden under the name *PPS sampling plan*. Regarding the confusion of this term as a sampling or evaluation method, compare Section 5.2.

The theoretical analysis of the Stringer bound is difficult for it combines attribute and variables principles (Neter et al. 1978). A theoretical derivation of the Stringer bound as an upper confidence bound is missing. Its popularity in accountancy is justified rather by having found it to be reliable in practice and conservative in a number of simulation studies. Several important arguments speak for its use: 1) The bound is greater than zero even if no misstatements are found in the sample. 2) The bound is independent of the sample drawn from the population if the population is free of errors.

However, there are limitations with respect to the application of the Stringer bound. It is mainly applicable when overstatement is expected in the population. Due to the use of DUS/MUS as the sampling technique applied with the Stringer bound, it is not recommended in tests for understatement, see Guy et al. (2002). Furthermore, the Stringer bound cannot deal with negative values, which sometimes occur in accounting populations. Therefore, they need to be removed from the population before the Stringer bound can be applied (Guy et al. 2002).

5.5 Two-sided Stringer Bounds

The previous Section 5.4 has established the concept of an upper confidence bound of Stringer type that is frequently used in the auditing context when overstatement is present. The decisions that can be made based on an upper bound for the mean tainting are either to accept the population if the bound is below the maximally acceptable misstatement rate or to be undecided. The latter case usually encourages the auditor to take another sample and continue the auditing process, which effects the efficiency of the audit. The rejection of an audit population is statistically impossible by means of an upper bound only. To have a basis for rejection and an idea about the minimum mean error that is most likely present in the accounting population together with the possibility to accept the population, requires a lower bound as well as an upper bound.

In this section we present a two-sided version of the one-sided Stringer interval presented in Section 5.4. To achieve tight bounds, we make use of prior information to obtain a two-sided confidence interval of the probability of erroneous items p that is used in the formula of the Stringer confidence interval.

5.5.1 Definition

We formulate a two-sided version of the Stringer confidence interval.

Definition 5.2 (Two-sided Stringer Bounds)**.** *Let $Y_1, \ldots, Y_n$ be an i.i.d. sample of variables with $P(0 \leq Y \leq 1) = 1$. Let $0 \leq Y_{(1,n)} \leq \ldots \leq Y_{(n,n)}$ be the corresponding ordered sample and $Y_{(n+1,n)} = 1$. Let $0 < \gamma < 1$ be the confidence level. For $k = -1, \ldots, n+1$ let $0 \leq p_L(k) < p_U(k) \leq 1$ be the two-sided level γ confidence limits for a success probability p in a series of Bernoulli trials of size n if k successes are observed, where $p_L(-1) = 0 = p_U(-1)$, $p_L(n+1) = 1 = p_U(n+1)$. Then the two-sided confidence interval of Stringer type for $E[Y] = \mu_Y$ is defined by $[\mu_{Y,L}; \mu_{Y,U}]$, where*

$$\begin{aligned}\mu_{Y,L} &= p_L(0) + \sum_{x=1}^{n} \Big(p_L(x) - p_L(x-1)\Big) Y_{(n-x+1,n)},\\ \mu_{Y,U} &= p_U(0) + \sum_{x=1}^{n} \Big(p_U(x) - p_U(x-1)\Big) Y_{(n-x+1,n)}.\end{aligned}$$

The formal definitions do not require an interpretation of the probability p both in the two-sided as well as in the one-sided version of the Stringer interval. In the one-sided version of the Stringer bound from Section 5.4, if the values $p_U(x)$ are supposed to be exact confidence bounds for a probability constructed by means of the binomial distribution, the Stringer bound under the use of the limits from Eq. (2.4) are the tightest possible. In the two-sided version, in contrast, inserting the Clopper & Pearson bounds $p_L(x), p_U(x)$ from Eq. (2.3) does not necessarily result in the shortest possible set of confidence intervals. The Clopper & Pearson confidence interval is constructed to be an equal-tail confidence interval, i. e. it restricts both the lower and upper tail probability by $(1-\gamma)/2$ if $\gamma \in (0;1)$ is the confidence level. Varying the tail probabilities appropriately results in two-sided confidence intervals for a probability that might be narrower in a certain sense, but are still exact. Chapter 2 presents a method to construct alternative two-sided intervals for a binomial probability that allow for prior information, where the intervals are still exact and of minimum weighted total volume. The weights are determined by the prior information, which is present in the form of a beta distribution $\text{Beta}(p_0, p_1, a, b)$ with shape parameters $a, b > 0$ on the support $[p_0; p_1]$, where $0 \leq p_0 < p_1 \leq 1$. In the following, we consider the two-sided Stringer bounds under a variety of prior information, including the case $p_0 = 0, p_1 = a = b = 1$, that constitutes the case of no prior information on the probability p. If an interpretation was intended, an obvious interpretation would be to consider p to be the probability of detecting an erroneous item in the population. Prior information in that case expresses the expected error proportion.

Our interest is to use the two-sided Stringer bounds for inference on the mean of a zero-inflated variable Y concentrated on $[0;1]$ in the model

$$Y = ZX, \quad Z, X, \text{ independent}, \quad Z \sim Bi(1,p), \quad \mathrm{P}(0 \leq X \leq 1) = 1. \tag{5.1}$$

In the audit context, this model signifies that the probability of a misstated item is p. Given a misstatement exists, its tainting distribution ranges between 0 and 1 with probability 1. The situation of a random variable $\widetilde{Y}$ concentrated on the compact interval $[s,t]$ can be reduced to the model (5.1) by considering $Y = \left(\widetilde{Y} - s\right)/(t-s)$.

For a tainting variable $Y = (U - W)/U$ in auditing, the model (5.1) follows from the error model

$$\begin{aligned} &W = (1-Z)U + ZQ, \qquad Z \text{ independent from } Q/U, \\ &Z \sim Bi(1,p), \qquad \mathrm{P}(0 \leq Q \leq U) = 1, \end{aligned}$$

i. e. the book value U overstates the de facto value W by the amount $U - Q$ with probability p. Then we have $Y = Z(U-Q)/U = ZX$ with $X = 1 - Q/U$ in the terminology of model (5.1). Q is the random variable representing the de facto value W of an audited item.

The problem is whether the two-sided Stringer bounds are conservative and provide a two-sided level γ confidence interval for μ_Y, i. e. whether $\mathrm{P}\left(\mu_{Y,L} \leq \mu_Y \leq \mu_{Y,U}\right) \geq \gamma$ holds.

5.5.2 Finite Sample Results

In the one-sided case, it can be shown that the Stringer bound is conservative if Y has a two-point distribution (de Jager et al. 1997). The same general result cannot be obtained for the two-sided bounds from Definition 5.2. However, the subsequent Proposition 5.3 shows that the required additional assumptions are not very restrictive. In particular, in the important special case of a prior distribution Beta(p_0, p_1, a, b) on p on the support $[p_0; p_1] = [0;1]$, which includes the case $p_0 = 0, p_1 = a = b = 1$ of no prior information, the two-sided bounds are conservative if Y has a two-point distribution.

Proposition 5.3 (Conservativeness for Two-point Distributions). *Let A be a measurement and prediction space (MPS) for the binomial probability y with projections A_x and A_y for $x \in \{0, \ldots, n\}$ and $y \in [p_0; p_1] \subset [0;1]$, where $A_x = [p_L(x); p_U(x)]$ is an exact confidence interval for y of level $\gamma \in (0;1)$, see Chapter 2. Let $\mu_{Y,L}$, $\mu_{Y,U}$ be the lower* *and upper bounds of the two-sided Stringer interval built under the use of the exact binomial confidence limits. Let* $0 \leq s < t \leq 1, q \in [p_0; p_1]$*, and let* $P(Y = s) = 1 - q$, $P(Y = t) = q$*. In the case of* $(1 - p_1)s = p_0(1 - t)$*, we have* $P(\mu_{Y,L} \leq \mu_Y \leq \mu_{Y,U}) \geq \gamma$*. In particular, the latter inequality holds in each of the subsequent cases: i)* $s = 0, t = 1$*, ii)* $s = 0, p_0 = 0$*, iii)* $p_0 = 0, p_1 = 1$*, iv)* $p_1 = 1, t = 1$*.*

Proof. See Appendix 5.A, Section 5.A.1. □

For the definition of an MPS see Section 2.2. An important special case of Proposition 5.3 is given for Y concentrated on the two points $s = 0$ and $t = 1$ with $\mathrm{P}(Y = 0) = 1 - q$, $\mathrm{P}(Y = 1) = q$. If $K = \{1, \ldots, n \mid Y_i = 1\}$ for $Y_1, \ldots, Y_n$, then

$$\begin{aligned}
\mu_{Y,L} &= p_L(0) + \sum_{x=1}^{K} (p_L(x) - p_L(x-1)) = p_L(K) \\
\mu_{Y,L} &= p_U(0) + \sum_{x=1}^{K} (p_U(x) - p_U(x-1)) = p_U(K)
\end{aligned}$$

by evaluating the telescoping sum, and $[\mu_{Y,L}; \mu_{Y,U}]$ coincides with the exact confidence interval $[p_L(K); p_U(K)]$ for a binomial probability q, for which conservativeness holds by definition.

5.5.3 The Two-sided Stringer Bound in Different Scenarios

We investigate the two-sided Stringer bounds $[\mu_{Y,L}; \mu_{Y,U}]$ under different definitions of a confidence interval $[p_L(x); p_U(x)]$ for a binomial probability p. Especially, we would like to exploit that in the case of an audit, prior information on the misstatement probability is frequently available and investigate the performance of the two-sided Stringer bounds under the use of shortest confidence intervals for a binomial probability under prior information from Chapter 2. In particular, we look at two important special cases: the case that only taintings of zero are observed and the case that exactly one misstatement is detected with the maximum possible tainting of 1.

5.5.3.1 The Case of Zero Non-zero Values

In the case of the two-sided Clopper & Pearson interval for the error probability p, the two-sided Stringer bound if only values of 0 are observed is

$$[\mu_{Y,L}; \mu_{Y,U}] = [p_L(0); p_U(0)] \,. \tag{5.2}$$

The conservativeness of Eq. (5.2) in the case of using an exact confidence interval for a probability is obvious from Proposition 5.3. In the case of using the confidence interval by Clopper & Pearson (1934), the interval becomes

$$[\mu_{Y,L}; \mu_{Y,U}] = \left[0; z_{\mathrm{Beta}(1,n)}\left(\frac{1+\gamma}{2}\right)\right],$$

where $z_{\mathrm{Beta}(1,n)}\left((1+\gamma)/2\right)$ is the $(1+\gamma)/2 \cdot 100\,\%$-quantile of the beta distribution $\mathrm{Beta}(1,n)$. A plot of the upper bound $z_{\mathrm{Beta}(1,n)}\left((1+\gamma)/2\right)$ under the Clopper & Pearson interval as a function of the sample size n is shown on the left-hand side of Fig. 5.1, where $\gamma = 0.9$. Since the lower bound of the two-sided Stringer interval in this case amounts to 0, the upper bound is equal to the length of the interval. The bound is decreasing, approaches 0 as the sample size n decreases, drops quickly for n approaching 20 and is already quite close to 0 for a sample size of more than $n = 100$. In the case of using the shortest exact confidence interval for the error probability p, the lower bound of the two-sided Stringer interval takes the value 0 as well. The right-hand side of Fig. 5.1 shows the difference between the upper bound of the two-sided Stringer interval under the use of the Clopper & Pearson interval for p and the shortest confidence interval for p for various prior information if 0 misstatements are observed. In the case of a uniform prior $\mathrm{Unif}(0,1)$, the upper bound, which for 0 errors equals the length of the two-sided Stringer interval, is close to the one emanating from the Stringer interval under the use of the Clopper & Pearson bounds. While it is the closest to the Clopper & Pearson bound from all investigated prior information types, it is smaller for sample sizes $1,\ldots,6$ and $8,\ldots,17$ and more conservative otherwise. All other prior information distributions displayed in the right-hand side of Fig. 5.1 deliver smaller upper Stringer bounds than the Clopper & Pearson bound. The gain is considerable for small sample sizes and less so for larger n. A huge gain in length can be obtained by considering the prior information $\mathrm{Unif}(0,0.1)$ and even more $\mathrm{Unif}(0,0.05)$, the latter being the best performing of the considered prior information for sample sizes up to about $n = 40$.

Other heavily right-skewed beta prior information distributions $\mathrm{Beta}(a,1)$ with $a = 0.04576, 0.03517, 0.02693$ are not shown in the plot, for they are hardly distinguishable from the curve produced by $\mathrm{Beta}(0.06546,1)$.

5.5.3.2 Minimum Sample Sizes if only Values of Zero are Expected

Guidelines for audit sampling often give recommendations about the required minimum sample size based on the assumption of observing no misstatements in the sample. Table 5.2 shows the minimum sample sizes n in dependence of the magnitude of the tolerable

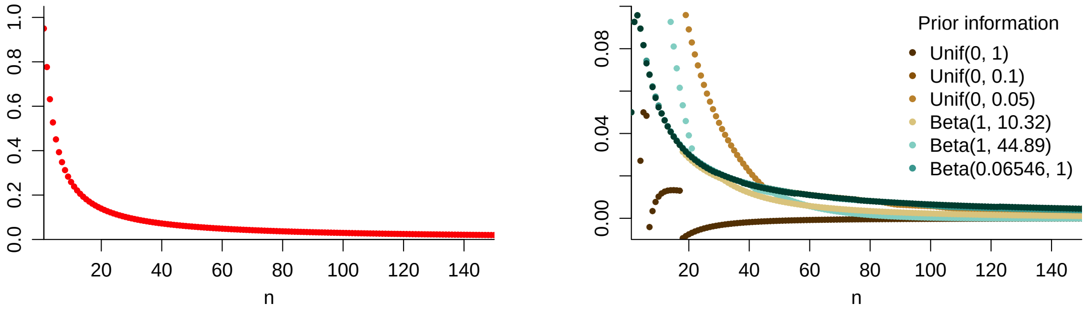


Figure 5.1: Upper bound of two-sided Stringer interval of level $\gamma = 0.9$ under 0 successes (left-hand side) and difference between the upper Clopper & Pearson bound and the upper two-sided confidence limit under prior information (right-hand side).

misstatement if two-sided Stringer bounds of level 90 % are used for inference. The given sample sizes are calculated such that the upper bound $\mu_{U,Y}$ of the two-sided Stringer interval does not exceed the tolerable misstatement. In this case, the investigated audit population would be accepted if 0 errors are detected in the sample. As soon as at least one error is found, the population cannot be accepted based on the given sample. The levels for the tolerable misstatement listed in Table 5.2 are 0.1, 0.05, 0.03, 0.02. For example, the minimum sample size that would yield an upper bound below 3 % under the use of the Clopper & Pearson interval for p is 99. For the minimum volume confidence interval with prior Unif$(0, 1)$, the sample sizes required are one higher than for the Clopper & Pearson interval for all of the investigated tolerable misstatement levels. Evidently, the confidence interval applying Unif$(0, 1)$, though producing the minimum volume by weighing over all possible number of successes $x = 0, 1, \ldots, n$, does not produce mimimum length of the confidence interval if $x = 0$ for most sample sizes. The required sample sizes are considerably smaller if certain other confidence intervals for p with prior information are used. The uniform prior on $[0; 0.05]$ means a misspecification in prior information if the true misstatement is higher than 0.05, which is why the minimum sample sizes of 1 for the misstatements 0.05 and 0.1 should be handled with care. For a tolerable misstatement of 0.03 though, the minimum volume confidence interval under the uniform prior $[0; 0.05]$ yields a considerably low minimum sample size of $n = 76$. The prior distributions Beta$(1, 10.32)$ and Beta$(0.06546, 1)$ both reflect beta distributions with 0.2 as the 90 %-quantile, the distributions Beta$(1, 21.85)$ and Beta$(0.04576, 1)$ beta distributions with 90 %-quantile 0.1, the distributions Beta$(1, 44.89)$ and Beta$(0.03517, 1)$ beta distributions with 90 %-quantile 0.05, the distributions Beta$(1, 114)$ and Beta$(0.02693, 1)$ beta distributions with 90 %-quantile 0.02, where all have a density which is decreasing on $[0; 1]$. Hence, beta distributions of the form Beta$(1, b)$ with high b as well as of

Table 5.2: Minimum sample sizes of the two-sided Stringer interval under a variety of prior information in the minimum volume confidence interval for a probability for confidence level $\gamma = 0.9$.

prior information on p	tolerable misstatement			
	0.1	0.05	0.03	0.02
none (Clopper & Pearson two-sided)	29	59	99	149
$\text{Unif}(0, 1)$	30	60	100	150
$\text{Unif}(0, 0.1)$	1	47	98	150
$\text{Unif}(0, 0.05)$	1	1	76	134
$\text{Beta}(1, 10.32)$	23	51	90	140
$\text{Beta}(1, 21.85)$	22	45	77	129
$\text{Beta}(1, 44.89)$	22	45	77	120
$\text{Beta}(1, 114)$	22	45	76	114
$\text{Beta}(0.06546, 1)$	22	46	77	116
$\text{Beta}(0.04576, 1)$	22	46	77	115
$\text{Beta}(0.03517, 1)$	22	46	77	116
$\text{Beta}(0.02693, 1)$	22	46	77	116
none (Clopper & Pearson one-sided)	22	45	76	114

the form $\text{Beta}(a, 1)$ with low a represent extreme prior distributions in terms of skewness. Yet, between the priors $\text{Beta}(0.06546, 1)$, $\text{Beta}(0.04576, 1)$, $\text{Beta}(0.03517, 1)$, and $\text{Beta}(0.02693, 1)$, there is hardly a difference in terms of the minimum sample size. That for a tolerable misstatement of 0.02 the prior distribution $\text{Beta}(0.04576, 1)$ yields a minimum sample size of 115, which is one smaller than for the more extreme distributions $\text{Beta}(0.03517, 1)$ and $\text{Beta}(0.02693, 1)$, must be one more consequence of the discreteness of the problem. The smallest minimum sample sizes among the observed scenarios are obtained under the $\text{Beta}(1, 114)$ distribution. This setting represents a prior distribution which is so extremely right-skewed that the resulting upper bound of the two-sided confidence interval is close to the one-sided upper Clopper & Pearson bound, which is the smallest exact upper bound for a binomial probability. The latter is reported for reasons of comparison in Table 5.2 and we can see that it produces the same minimum sample sizes as the $\text{Beta}(1, 114)$ prior distribution.

5.5.3.3 The Case of one Non-zero Observation

We investigate another important special case, when exactly one observation > 0 is observed in the sample. In this case, the two-sided Stringer interval becomes

$$[\mu_{Y,L}; \mu_{Y,U}] = [p_L(0) + (p_L(1) - p_L(0))Y_{(n,n)}; p_U(0) + (p_U(1) - p_U(0))Y_{(n,n)}] \quad (5.3)$$

and we can see that the upper and lower bounds are dependent on the value $Y_{(n,n)} > 0$ of the one observation which is greater than zero in the sample. From Eq. (5.3) we can easily see that the upper bound $\mu_{Y,U}$ and lower bound $\mu_{Y,L}$ are monotonous in the value $Y_{(n,n)}$. For a conservative assessment, we consider for $Y_{(n,n)}$ the most extreme value, which is 1, since $Y_{(n,n)}$ is assumed to be a random variable ranging between 0 and 1 with probability 1, see the assumption in Definition 5.2. The Stringer interval coincides with the confidence interval $[p_L(1); p_U(1)]$ for a binomial probability if one success is observed, i. e. the same interval which an evaluation method for attribute sampling would have yielded. The conservativeness of (5.3) is obvious from the fact that the interval is equal to an exact confidence interval for a probability as well as from Proposition 5.3.

In the case that the two-sided Clopper & Pearson (1934) confidence interval for the probability $p = \mathrm{P}(Y > 0)$ is used, we obtain the interval

$$[p_L(1); p_U(1)] = \left[z_{\text{Beta}(1,n)} \left(\frac{1-\gamma}{2} \right); z_{\text{Beta}(2,n-1)} \left(\frac{1+\gamma}{2} \right) \right],$$

see Eq. (2.3).

A comparison between the lower (left-hand side of the figure) and upper bounds (right-hand side of the figure) of the two-sided level 90 % confidence intervals under the use of the Clopper & Pearson interval and several minimum volume confidence intervals for p is shown in Fig. 5.2. The lower bounds of the intervals using the other investigated prior information distributions (among them $\text{Unif}(0, 0.1)$, $\text{Unif}(0, 0.05)$, $\text{Beta}(1, 10.32)$, $\text{Beta}(0.06546, 1)$) are omitted from the plot, for they are equal to the lower bound under the prior information $\text{Unif}(0, 1)$. The upper bounds of the Clopper & Pearson confidence interval and the one using the uniform prior on $[0; 1]$ are hardly distinguishable. The right-skewed distributions $\text{Beta}(1, 10.32)$, and $\text{Beta}(0.06546, 1)$ achieve visibly smaller upper bounds than the uniform prior on $[0; 1]$. The bounds under the distributions $\text{Unif}(0, 0.1)$ and $\text{Unif}(0, 0.05)$ are hardly distinguishable from the right-skewed priors from a sample size of about 37 on and 74 on, respectively. However, they clearly display the cut-offs at 0.1 and 0.05 for small sample sizes.

In contrast to the one-sided intervals, two-sided intervals provide the chance of a rejection of the population. If by using the minimum volume confidence intervals for a probability in the two-sided Stringer bounds one tainting equal to 1 would have been discovered and otherwise zero values in a sample of size $n \leq 5$, the population would have been rejected if the tolerable misstatement were ≤ 0.02. In contrast, the two-sided Stringer interval under the use of the Clopper & Pearson confidence interval for p would have left the

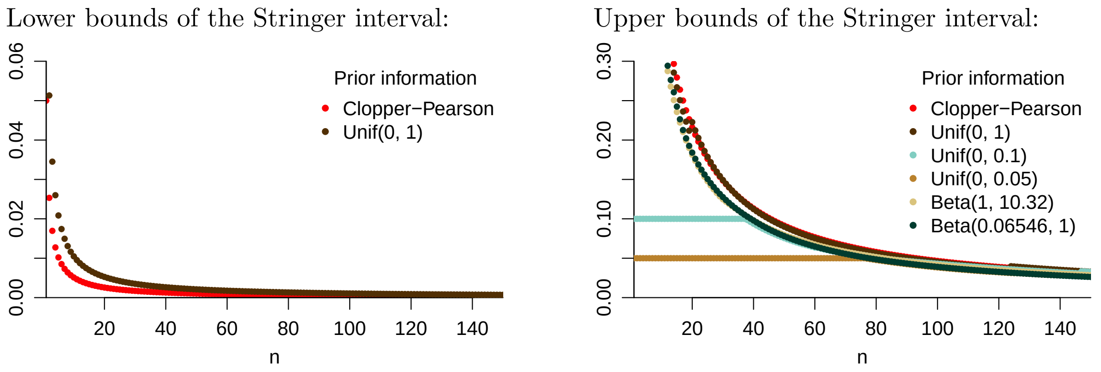


Figure 5.2: Lower and upper bound of two-sided Stringer interval of level $\gamma = 0.9$ if one observation of value 1 is discovered and values of 0 otherwise.

decision maker indifferent already from a sample size of $n = 3$ onwards.

5.5.3.4 Minimum Sample Sizes if one Non-zero Observation is Expected

We investigate the required sample sizes that are minimally necessary such that the auditor has the chance of accepting the population even if one observation > 0 is observed. This is possible when the upper bound of the two-sided Stringer bound does not exceed the tolerable misstatement. As in Section 5.5.3.3, we assume the worst case of a value of 1 for the non-zero observation. The results of the analysis are summarised in Table 5.3 for a variety of prior information distributions under a confidence level of $\gamma = 0.9$. Naturally, the required minimum sample sizes if one observation greater than 0 is observed are larger than in the case of observing only zeros. The upper bound under the same sample size n is $p_U(1) - p_U(0) > 0$ larger if one value of 1 is detected and only values of 0 otherwise. Thus it is more difficult to achieve that the upper bound does not exceed a prescribed threshold. For example, if the tolerable misstatement is given as 5 % and if it should still be possible to find one non-zero observation of a magnitude up to 1, the sample should have at least a size of 91 if no prior information on p is employed, whereas it can drop to 77 under the use of some highly right-skewed prior information distributions. 77 is the sample size that the one-sided exact confidence interval for p would also demand to be able to lead to an acceptance of the population, which is the best possible case, given an exact confidence interval for p is used.

While for the tolerable misstatements 0.1 and 0.05, the Stringer bound under the use of the minimum volume confidence interval for p with prior information Unif$(0, 1)$ delivers slightly smaller sample sizes (45 and 91, respectively) than under the use of the two-sided

Table 5.3: Minimum sample sizes of the two-sided Stringer interval under a variety of prior information in the minimum volume confidence interval for a probability for confidence level $\gamma = 0.9$ if one observation of 1 is observed in the sample and 0 otherwise.

prior information on p	tolerable misstatement			
	0.1	0.05	0.03	0.02
none (Clopper & Pearson two-sided)	46	93	157	236
Unif$(0, 1)$	45	91	165	248
Unif$(0, 0.1)$	1	83	165	248
Unif$(0, 0.05)$	1	1	131	219
Beta$(1, 10.32)$	40	83	142	218
Beta$(1, 21.85)$	38	78	142	216
Beta$(1, 44.89)$	38	77	131	204
Beta$(1, 114)$	38	77	129	194
Beta$(0.06546, 1)$	40	78	131	198
Beta$(0.04576, 1)$	39	78	131	198
Beta$(0.03517, 1)$	39	78	131	198
Beta$(0.02693, 1)$	39	78	131	198
none (Clopper & Pearson one-sided)	38	77	129	194

Clopper & Pearson confidence interval (46 and 93, respectively), the Clopper & Pearson confidence intervals gets slightly advantageous for tolerable misstatements of 0.03 and 0.02.

5.6 Simulation Study

We conduct a simulation study to investigate the performance of the two-sided Stringer interval for inference on the mean of a zero-inflated variable Y on $[0; 1]$ in the model (5.1). With the variable X between 0 and 1 with probability 0 and Z being either 0 or 1, Y also takes values between 0 and 1 with probability 1, i. e. $\mathrm{P}(0 \leq Y \leq 1) = 1$.

For X, we use the beta distribution Beta(a, b) on the support $[0; 1]$ as a model. By choosing the shape parameters of the beta distribution as $a = 0.5$ and $b = 2$, we obtain a decreasing and right-skewed beta distribution, see Proposition 3.11. This choice is supposed to model the positive skewness of non-zero taintings in an audit population. With the beta distribution Beta(a, b) having expected value $a/(a + b)$, we obtain

$$E[X] = E[Y|Y > 0] = \frac{a}{a+b} = \frac{0.5}{0.5+2} = 0.2$$

as the conditional mean of Y, given $Y > 0$. With the independence of Z and X in model (5.1) we obtain the expected value of Y as

$$E[Y] = E[Z]E[X] = p\frac{a}{a+b}. \tag{5.4}$$

We investigate the Stringer bound under $Z \sim Bi(1, p)$ in model (5.1), where p is chosen to be 5 % and 10 %. In the first case, Y takes the value 0 with probability 95 %, and a value > 0 with probability 5 % and its expected value is $0.05 \cdot 0.2 = 0.01$ by formula (5.4). In the second case, Y has mean $0.1 \cdot 0.2 = 0.02$.

Figure 5.3 displays the average length of the two-sided level 90 % Stringer interval using the minimum volume confidence interval for a probability relative to the average length of the Stringer interval using the Clopper & Pearson confidence interval for a probability as a function of the sample size n. Sample sizes between 5 and 150 in intervals of 5 are considered. 10 000 simulation runs per sample size n were executed. The simulation was performed in R. We consider model (5.1) with $p = 0.05, 0.1$ in $Z \sim Bi(1, p)$ and $Y|Y > 0 \sim \text{Beta}(0.5, 2)$. Under the prior information Unif(0, 1), the Stringer interval is slightly wider than the Stringer interval using the Clopper & Pearson bounds for several sample sizes, but mostly comparable. The largest decrease in terms of length can be achieved when using the uniform prior on $[0; 0.05]$ for p. However, with p being 0.05 in average by definition, the analysis of this prior distribution is of little practical use. The uniform prior on $[0; 0.1]$ also produces rather narrow Stringer bounds for sample sizes up to about 40. The prior information Beta(1, 10.32) generates Stringer bounds that are in average between 20 % and 9 % smaller than the Stringer bound under the Clopper & Pearson interval, and Beta(1, 21.85) brings an improvement between 21 % and 13 %. The distributions Beta(1, 44.89) and Beta(0.06546, 1) show enhancements between 20 % and 18 % (Beta(1, 44.89)) and 20 % and 16 % (Beta(0.06546, 1)) for the investigated sample sizes.

We examine the coverage probability of the two-sided Stringer interval in model (5.1) with $Y|Y > 0 \sim \text{Beta}(0.5, 2)$ and $Z \sim Bi(1, p)$. The actual coverage probability under a nominal confidence level of $\gamma = 0.9$ for $p = 0.05, 0.1$ is shown in Fig. 5.4 under various prior information on p. In all investigated cases, the two-sided Stringer interval is very conservative, producing coverage probabilities larger than 96 % for a nominal confidence level of 90 %. In the case $p = 0.05$, the most conservative Stringer interval for sample sizes up to about 120 is obtained by making use of the Clopper & Pearson confidence interval for a probability. The Stringer bounds applying the minimum volume confidence interval for a probability show similar behaviour in terms of coverage probability for sample sizes n up to about 60. For larger sample sizes, they drift apart with prior distributions Unif(0, 1) and Unif(0, 0.1) showing the least conservative behaviour and the coverage probability of the Beta(0.06546, 1) prior even exceeding the one under the Clopper & Pearson interval for sample sizes larger than about 120. In the case

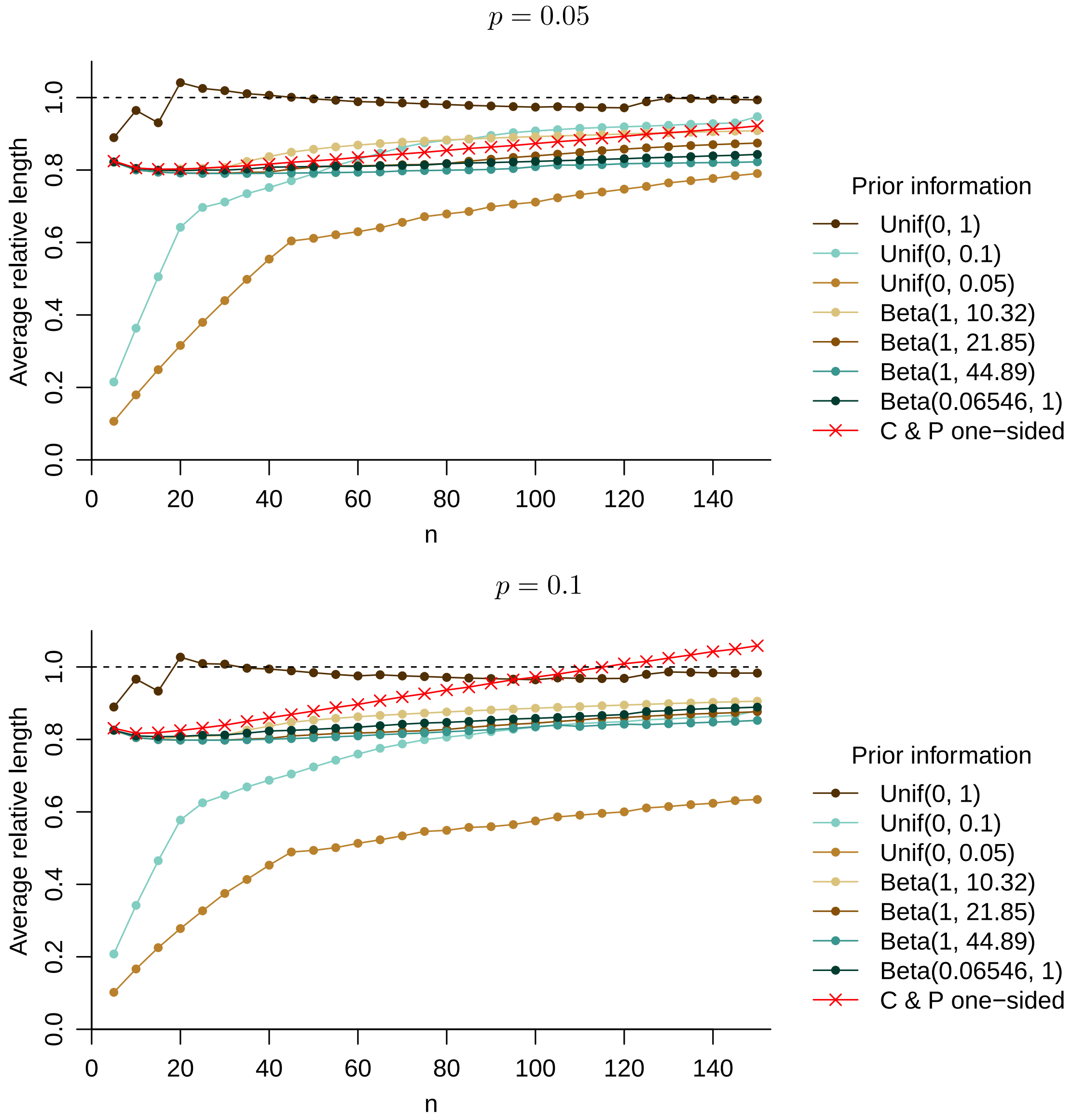


Figure 5.3: Average length of the Stringer interval using the minimum volume confidence interval for a probability relative to the average length of the Stringer interval using the Clopper & Pearson (C & P) confidence interval for a probability as a function of the sample size n for confidence level $\gamma = 0.9$. The stipulated model is (5.1) with $p = 0.05, 0.1$ in $Z \sim Bi(1, p)$ and $Y|Y > 0 \sim \text{Beta}(0.5, 2)$.

of $p = 0.1$, drifting apart starts earlier at sample sizes around 30. Again, the prior Unif$(0, 1)$ produces the least conservative results, whereas Unif$(0, 0.05)$ and the priors Beta$(1, 44.89)$ and Beta$(0.06546, 1)$ are extremely conservative for larger sample sizes with coverage probabilities frequently larger than 99 %.

Figure 5.5 shows a plot of the actual coverage probability of the two-sided Stringer interval for samples of sizes $n = 50$ and $n = 100$ as a function of the error proportion p in model (5.1) with $Z \sim Bi(1, p)$ and $Y|Y > 0 \sim$ Beta$(0.5, 2)$ for different prior information. The Clopper & Pearson confidence interval in the Stringer bound as well as the uniform prior on $[0; 1]$ and the right-skewed priors all show coverage probabilities of more than 0.96 for a prescribed nominal level of 0.9 and are hence overly conservative. The coverage probability of Unif$(0, 0.1)$ and Unif$(0, 0.05)$ drops below 0.9 for p larger than about 0.35 and 0.22, respectively, for sample size $n = 50$. For $n = 100$, their coverage probability drops below 0.9 for p larger than about 0.3 and 0.18, respectively. However, both prior distributions express prior believes of p not exceeding 0.1 and 0.05, respectively, which is why larger values than these are not deemed likely anyway. It has to be remarked, though, that even for p falling outside the range of $[0; 0.1]$ or $[0; 0.05]$, the Stringer bound remains conservative for a while. The general behaviour of the coverage probability of the two-sided Stringer bound under the investigated prior information distributions is not remarkedly different between the sample sizes $n = 50$ and $n = 100$. It can be concluded that the two-sided Stringer bound is conservative in all analysed cases unless the assumption of $p \sim$ Unif$(0, u)$, $0 < u < 1$, is far from being valid.

We investigate the probabilities for the different decisions that can be based on the Stringer confidence intervals. Indifference occurs when the confidence interval contains the tolerable misstatement; acceptance when the upper bound of the confidence interval does not exceed the tolerable misstatement; rejection when the lower bound of the confidence interval is larger than the tolerable misstatement. Figure 5.6 illustrates the indifference (first row), acceptance (second row) and rejection probabilities (third row) of the Stringer interval for the tolerable misstatements 0.05 and 0.1 for confidence level $\gamma = 0.9$ and sample size $n = 100$ as a function of the true misstatement probability $p = \mathrm{P}(Y > 0)$. The underlying error model follows (5.1) with $Z \sim Bi(1, p)$ and $Y|Y > 0 \sim$ Beta$(0.5, 2)$. We consider the two-sided Stringer interval under the two-sided Clopper & Pearson interval as well as under the minimum volume confidence interval for a probability with prior information Unif$(0, 1)$, Unif$(0, 0.1)$, Unif$(0, 0.05)$, Beta$(1, 10.32)$, Beta$(1, 21.85)$, Beta$(1, 44.89)$, Beta$(0.06546, 1)$ as well as the one-sided Stringer interval using the one-sided exact confidence interval for a probability from Eq. (2.4). The indif-

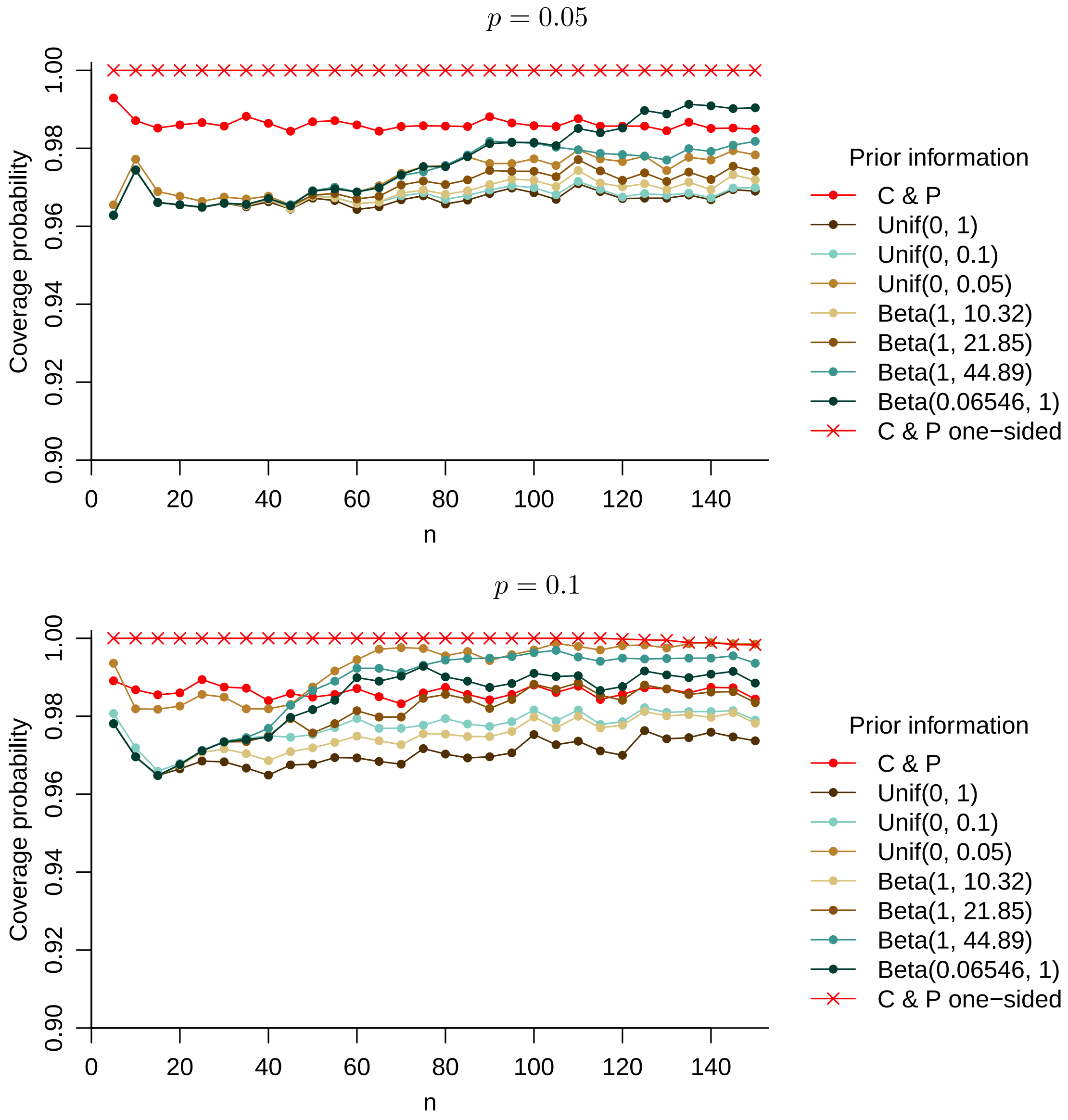


Figure 5.4: Coverage probability of two-sided Stringer interval using the minimum volume confidence interval for a probability and the Clopper & Pearson interval as a function of the sample size n for confidence level $\gamma = 0.9$ under model (5.1) with $Z \sim Bi(1, p)$ and $Y|Y > 0 \sim \text{Beta}(0.5, 2)$.

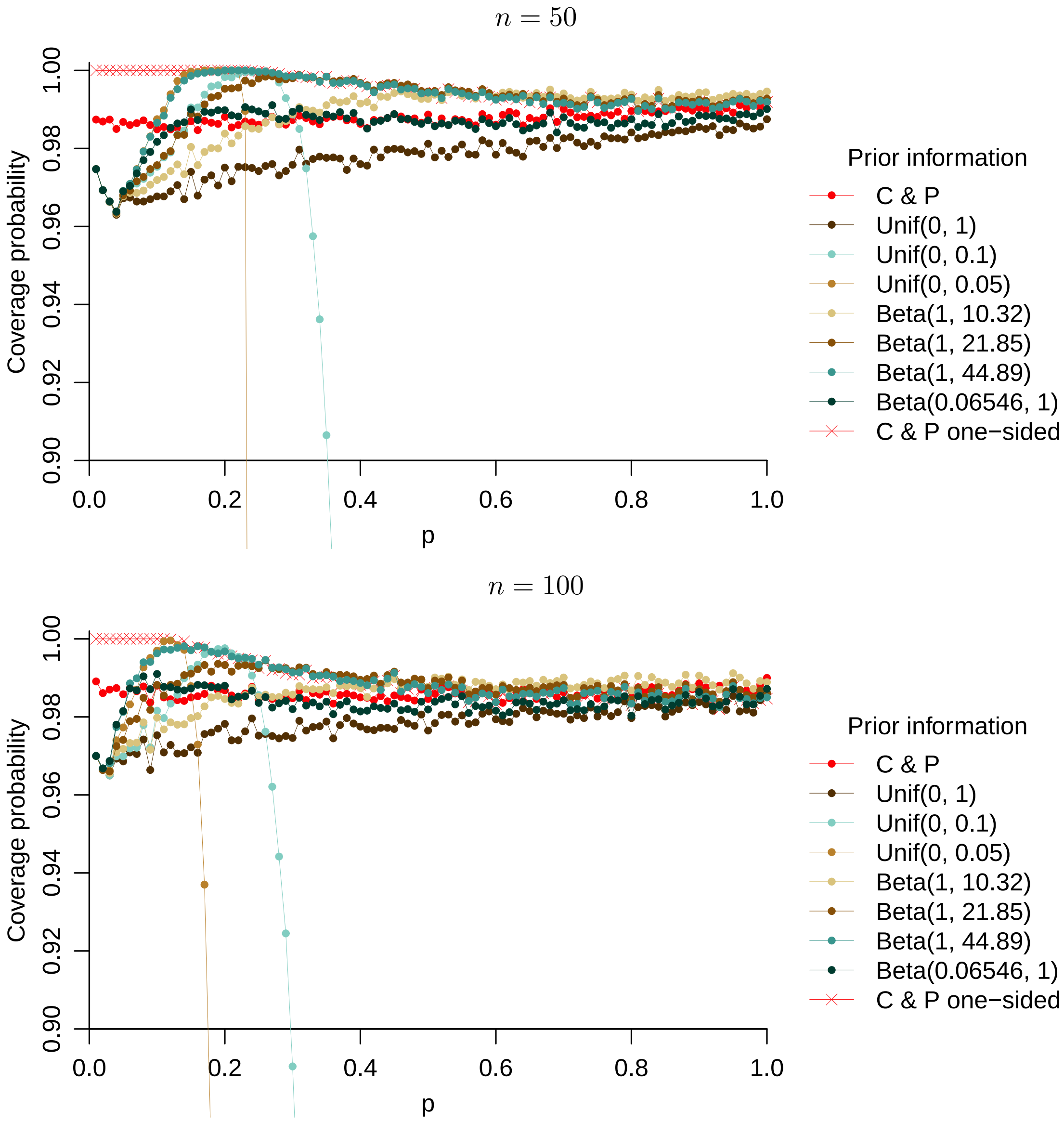


Figure 5.5: Coverage probability of Stringer interval using the minimum volume confidence interval for a probability and the Clopper & Pearson (C & P) interval as a function of the error proportion p for confidence level $\gamma = 0.9$ under model (5.1) with $p = 0.05, 0.1$ in $Z \sim Bi(1, p)$ and $Y|Y > 0 \sim \text{Beta}(0.5, 2)$.

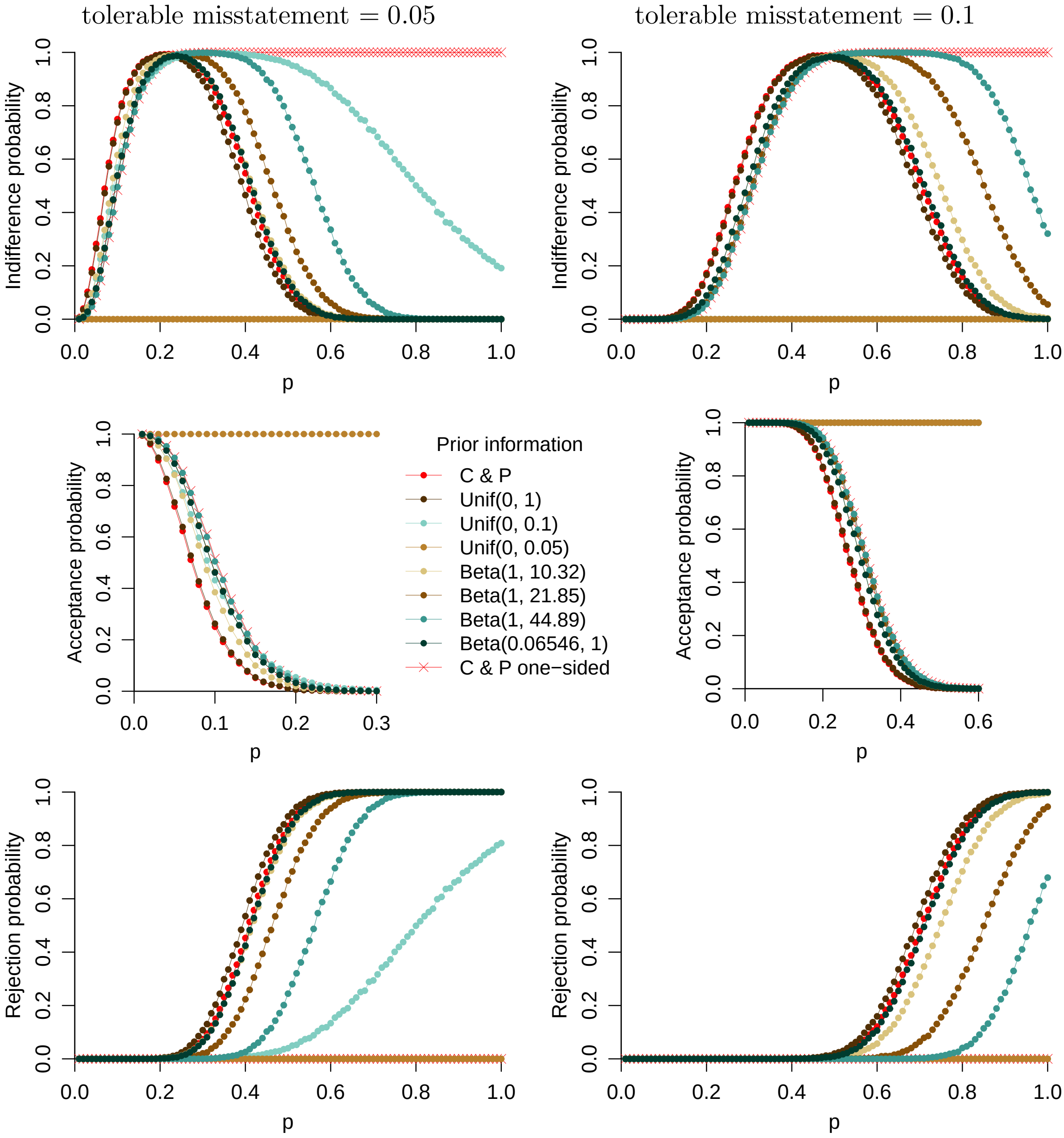


Figure 5.6: Decision probabilities of Stringer interval using the minimum volume confidence interval for a probability and the Clopper & Pearson (C & P) interval as a function of the error proportion p for confidence level $\gamma = 0.9$ and sample size $n = 100$ under model (5.1) with $Z \sim Bi(1, p)$ and $Y|Y > 0 \sim \text{Beta}(0.5, 2)$ for tolerable misstatements 0.05 and 0.1.

ference probability reaches its maximum, which is ≥ 0.9, at the misstatement probability under which the average misstatement is the tolerable misstatement. Since they increase early when approaching their maximum value in p, the two-sided Clopper & Pearson and the minimum volume confidence interval under the uniform prior $\text{Unif}(0, 1)$ show a worse behaviour in terms of indifference probability than all other Stringer intervals. These show lower indifference probabilities for values of p smaller than 0.05 and 0.01, respectively, and are close to the one-sided Stringer interval using the one-sided exact confidence interval for a probability. Under the investigated prior information distributions, the two-sided Stringer interval under $\text{Unif}(0, 1)$ and the two-sided Clopper & Pearson interval are, however, advantageous for larger misstatement rates. The most unfavourable behaviour in terms of indifference probability for larger p is shown by the prior information distributions $\text{Unif}(0, 0.1)$ and $\text{Beta}(1, 44.89)$ for a tolerable misstatement of 0.05 and $\text{Beta}(1, 44.89)$ and $\text{Beta}(1, 21.85)$ for a tolerable misstatement of 0.1. The one-sided Stringer interval clearly always leaves the decision maker indifferent if the true misstatement rate exceeds a certain threshold.

In terms of acceptance probability, the two-sided Stringer interval under the uniform prior information on $[0; 1]$ as well as under the two-sided Clopper & Pearson interval show the least favourable behaviour. The prior information distributions $\text{Unif}(0, 0.5)$ and $\text{Unif}(0, 0.1)$ lead to an acceptance of the population for tolerable misstatements of 0.05 and 0.1, respectively, per definitionem. All other Stringer interval types processing more meaningful prior information distributions come with acceptance probabilities that are very close to the one of the one-sided Stringer interval. From about $p = 0.3$ on in the case of the tolerable misstatement 0.05, and $p = 0.6$ on in the case of the tolerable misstatement 0.1, the acceptance probabilities are virtually 0.

The rejection probabilities start being larger than 0 for misstatement rates p of about 0.2 in the case of the tolerable misstatement 0.05 and about 0.5 in the case of the tolerable misstatement 0.1. To the confidence intervals for a probability that lead to a Stringer interval with the largest rejection probabilities belong the two-sided Clopper & Pearson interval and the minimum volume confidence interval under prior information $\text{Unif}(0, 1)$, $\text{Beta}(0.06546, 1)$ and – at least for a tolerable misstatement of 0.05 – also $\text{Beta}(1, 10.32)$. The Stringer intervals under the prior information $\text{Beta}(1, 21.85)$, $\text{Beta}(1, 44.89)$ and in the case of a tolerable misstatement of 0.05 also $\text{Unif}(0, 1)$, more rarely lead to the rejection of the population. The one-sided Stringer interval with an upper bound naturally never leads to a rejection of the population.

While most other Stringer interval types that come with advantages in one of the decision types have disadvantages in another, the Stringer interval under the minimum volume confidence interval with prior information Beta$(0.06546, 1)$ shows a good overall performance.

5.7 Dealing with both Under- and Overstatement

The previous sections dealt with an overstatement model only, that is, the book value was assumed to be larger than or equal to the de facto value. In this section, we formulate an approach that could be useful when the purpose is to simultaneously treat both overstatement and understatement errors.

Let $W \geq 0$ denote the de facto value and $U > 0$ the book value. Let $W = Q \cdot U$, i. e. $Q = W/U = 1 - (U - W)/U$. Since $W \geq 0$ and $U > 0$, it follows $Q \geq 0$. We frequently encounter an equality of book and de facto values, which is why Q is one-inflated, i. e. takes a considerable amount of values of 1. The case $Q < 1$ reflects the case of overstatement errors. The case $Q > 1$ reflects understatement errors and is not covered by the model (5.1) that was investigated in the previous sections.

The expected value of Q is given by

$$\begin{aligned} \mu_Q \quad = \quad E[Q] \quad &= \quad E[Q \cdot \mathbb{1}_{Q>1}] + 1 \cdot \mathrm{P}(Q = 1) + E[Q \cdot \mathbb{1}_{Q<1}] \\ &= \quad E[Q|Q > 1] \cdot \mathrm{P}(Q > 1) + \mathrm{P}(Q = 1) + E[Q|Q < 1] \cdot \mathrm{P}(Q < 1). \end{aligned}$$

Under the assumption that Q and U are independent we have $\mu_W = \mu_Q \cdot \mu_U$. The total sum of de facto values then results to

$$\sum_{i=1}^{N} W_i \quad = \quad N\mu_W \quad = \quad \mu_Q \cdot N\mu_U \quad = \quad \mu_Q \cdot \sum_{i=1}^{N} U_i.$$

The implications of the approach will not be investigated further here. More steps are necessary to fully decide on the usefulness of the approach, among them the formulation of an error model as well as the understanding to which extent the approach relies on monetary unit sampling.

5.8 Conclusion and Outlook

We have proposed a two-sided confidence interval of Stringer type based on the one-sided bound introduced by Stringer (1963) that is supposed to work especially well in zero-inflated populations. In contrast to the one-sided bound, the two-sided confidence interval allows the rejection of a population at the cost of being able to accept less often. The two-sided Stringer confidence interval inherits some important features from its one-sided version: i) Even if only values of zero are observed, the upper bound μ_U is always greater than zero. ii) If applied to an audit population, the same lower and upper confidence bounds are delivered if no errors are found in the sample, irrespective of the sample book values.

As an instance of a combined attribute and variable sampling (CAV) bound, the two-sided Stringer interval combines features of both attribute and variable sampling, where the attribute sampling part is reflected by the confidence limits for a proportion that are used in the formula for the bounds. We have made use of the exact minimum volume confidence interval for a probability under prior information, where prior information is imposed on the misstatement probability p. Misstatements in the audit context are frequently measured in terms of the *tainting* $Y = (U - W)/U$ of an item, which is defined as the relative deviation of the de facto value W from the book value U. The model we have proposed assumes the tainting to range between 0 and 1, where the value 0 is taken by Y with a usually high probability $1 - p$ (zero-inflation) and values greater than 0 with probability p. Only the overstatement case is covered by the model, which means that the de facto value is assumed to be not larger than the book value.

The approach of applying the minimum volume confidence intervals for a probability delivers markedly narrower Stringer bounds than the Clopper & Pearson confidence interval for a probability inserted in the Stringer bounds under many realistic prior information choices. The minimum sample size required to lead to an acceptance of the population can be reduced not rarely more than 20 % in comparison to the two-sided Stringer interval obtained by means of the Clopper & Pearson bounds.

In some important special cases in which the distribution of Y is concentrated on two points, we have theoretically proven in Proposition 5.3 that the two-sided Stringer confidence interval is conservative. A solid theoretical framework supporting the two-sided Stringer bound and its conservativeness in arbitrary circumstances, however, is missing. A simulation study has been conducted that indicates high conservativeness of the two-sided interval. This means, the nominal confidence level is mostly by far exceeded by the actual coverage probability unless the prior information is clearly misspecified. The conditional distribution of $Y|Y > 0$ in the simulation study has been chosen to follow a right-skewed beta distribution. This is considered a plausible model for an auditing population, which has frequently heavily skewed tainting distributions.

The two-sided Stringer bounds under certain right-skewed prior distributions on the misstatement probability reaches properties that come very close to the best one-sided Stringer confidence interval. The heavily right-skewed prior information distribution Beta$(0.06546, 1)$ shows particulary satisfying indifference and acceptance probabilities, but in contrast to the one-sided interval has the advantage that it can lead to a rejection of the population.

The proposed two-sided Stringer interval cannot only be applied to auditing populations. Situations where the random variable of interest ranges between 0 and 1 and the value 0 is taken with a comparably large probability are potential applications of the two-sided Stringer interval, as accident costs in insurance or earthquake measurements. The performance of the two-sided Stringer bounds in real data applications would be interesting to explore.

Several things remain to be studied with respect to the two-sided Stringer confidence interval. Instead of binomial limits, the one-sided Stringer bound is often applied using Poisson limits in Definition 5.1, see, for example, Guy et al. (2002) or Goodfellow et al. (1974). The two-sided Stringer confidence interval should be investigated under the use of two-sided Poisson confidence intervals as a potential alternative to the binomial based Stringer bounds. Prior information can be exploited in a similar way as in the case of the binomial distribution by means of the minimum volume confidence intervals for a Poisson parameter under prior information as described in Chapter 4. In addition, the Poisson distribution based Stringer interval might reveal important characteristics about the asymptotic behaviour of the Stringer bounds under the use of the binomial limits.

Loebbecke & Neter (1975) mention that CAV bounds might be most appropriate, among other arguments, when there is the danger of outliers in the population, by which they mean “a major and isolated aberration from the rest of the data set” (Loebbecke & Neter 1975, p. 40). The robustness and conservativeness of two-sided Stringer bounds should be explored in the case of outliers.

In many cases, the Stringer bound will not deliver a decision in the first place, which is the case when the tolerable error rate falls between the lower and upper confidence limit. This will create the necessity of drawing another sample to increase the chance of achieving a decision in the second round. Prior knowledge obtained in the first step of the procedure could be exploited when evaluating the second part of the sample. The characteristics of this multi-stage procedure as well as the potentials that arise from an appropriate exploitation of knowledge about the first sample part need to be explored.

The one-sided as well as the here proposed two-sided Stringer interval was developed particularly for the case of overstatement errors in auditing, as appears frequently in accounts receivable populations. It is not applicable when also understatement is present. Section 5.7 sketches a way how over- and understatement errors could be treated simultaneously. The idea should be pursued.

5.A Appendix

5.A.1 Proof of Proposition 5.3

Let $K = \{1 \leq i \leq n \mid Y_i = t\}$. Then K has the binomial distribution $Bi(n, q)$. We have $Y_{(1,n)} = \ldots = Y_{(n-K,n)} = s$, $Y_{(n-K+1,n)} = \ldots = Y_{(n,n)} = t$. Hence

$$\begin{aligned}
\mu_{Y,L} &= p_L(0) + \sum_{j=1}^{K} \Big(p_L(j) - p_L(j-1)\Big)t + \sum_{j=K+1}^{n} \Big(p_L(j) - p_L(j-1)\Big)s \\
&= p_L(0) + t(p_L(K) - p_L(0)) + s(p_L(n) - p_L(K)) \\
&= p_L(0)(1-t) + p_L(K)(t-s) + p_L(n)s \\
&= p_0(1-t) + p_L(K)(t-s) + p_L(n)s
\end{aligned}$$

and analogously

$$\begin{aligned}
\mu_{Y,U} &= p_U(0) + \sum_{j=1}^{K} \Big(p_U(j) - p_U(j-1)\Big)t + \sum_{j=K+1}^{n} \Big(p_U(j) - p_U(j-1)\Big)s \\
&= p_U(0) + t(p_U(K) - p_U(0)) + s(p_U(n) - p_U(K)) \\
&= p_U(0)(1-t) + p_U(K)(t-s) + p_U(n)s \\
&= p_U(0)(1-t) + p_U(K)(t-s) + p_1 s.
\end{aligned}$$

We have $\mu_Y = E[Y] = (1-q)s + qt = s + (t-s)q$. Since $p_0 \leq p_L(x) \leq p_U(x) \leq p_1$ for all $x \in \{0, \ldots, n\}$, we obtain

$$\begin{aligned}
&\mathrm{P}\left(\mu_{Y,L} \leq \mu_Y \leq \mu_{Y,U}\right) \\
= \quad & P\Big(p_0(1-t) + p_L(K)(t-s) + p_L(n)s \leq s + (t-s)q \\
&\qquad\qquad \leq p_U(0)(1-t) + p_U(K)(t-s) + p_1 s\Big) \\
\geq \quad & \mathrm{P}\Big(p_0(1-t) + p_L(K)(t-s) + p_1 s \leq s + (t-s)q \\
&\qquad\qquad \leq p_0(1-t) + p_U(K)(t-s) + p_1 s\Big) \\
= \quad & \mathrm{P}\left(p_L(K) \leq \frac{(1-p_1)s - p_0(1-t)}{t-s} + q \leq p_U(K)\right).
\end{aligned}$$

In the case $(1-p_1)s = p_0(1-t)$, the nominator in the above term results to 0, and we obtain the inequality

$$\mathrm{P}\left(\mu_{Y,L} \leq \mu_Y \leq \mu_{Y,U}\right) \quad \geq \quad \mathrm{P}\left(p_L(K) \leq q \leq p_U(K)\right).$$

The assertion of the proposition follows since $\mathrm{P}\left(p_L(K) \leq q \leq p_U(K)\right) \geq \gamma$ by definition of the exact level γ confidence limits $p_L(K)$ and $p_U(K)$ for the parameter q of the binomial distribution $Bi(n, q)$.

6 Exponential Smoothing with Covariates with Application in Electricity Load Forecasting

6.1 Introduction

Exponential smoothing (ES) traces back to the work of Robert G. Brown, who developed the basics of the method in the 1950s for purposes of inventory management (Brown 1959, 1963). Methods for smoothing of trends and seasonal data were independently developed by Charles C. Holt, see Holt (1957). With Winters (1960) testing Holt's method with empirical data, the methods gained publicity as the Holt-Winters forecasting system, see Gardner (2006). This work provided the scheme for further extensions, in particular, the damped additive trend considered by Gardner & McKenzie (1985), the multiplicative or exponential trend introduced by Pegels (1969) and extended to a damped version by Taylor (2003a). Gardner (1985, 2006) provides a detailed survey of the historical evolution and the various facets of ES.

ES was often considered rather as a heuristic forecasting technique without a precise model foundation guaranteeing optimality. However, relatively early several authors presented model foundations for ES, in particular by regression or ARIMA models. In this regard, the linear ES versions all have an equivalent ARIMA model, see e. g. Gardner & McKenzie (1988) and Yar & Chatfield (1990). The essential step towards a solid model framework was achieved with the single source of error (SSOE) state-space scheme presented by Ord et al. (1997). In particular, the state-space formulation allows to demonstrate the optimality of the classical ES predictors by deriving them as conditional expectations in the SSOE model.

Hyndman et al. (2002) elaborated the SSOE approach to a broad taxonomy of ES methods. These include cross combinations of additive/exponential (damped) trend, additive/multiplicative seasonality and additive/multiplicative error. Extensions of this scheme are due to Taylor (2003a), who added a damped version of the multiplicative trend, and Taylor (2003b, 2010), who involved a second and third seasonal component and applied the model to half-hourly electricity load data.

An appealing characteristic of an ES method is its simplicity; the structure can be formulated in a transparent and interpretable way. This characteristic becomes partially lost when considering the single source of error (SSOE) formulation for ES instead, as presented in the subsequent section, but the advantages of having a model surely overweigh this minor disadvantage.

The formulation of ES methods including covariates is relatively new. Wang (2006) coined exponential smoothing with covariates (ESCov) by introducing into the observation equation of the SSOE model an additive term depending linearly on exogenous variables. An innovation state-space models with exogenous variables (practically ESCov with a slightly different treatment of the damping parameter) was applied by Athanasopoulos & Hyndman (2008). In ARIMA methods, the inclusion of covariates dates back longer than for ES. The ARIMAX model, where the X stands for the availability of exogenous variables in the model, was initiated by Box & Tiao (1975).

In the following sections, we extent the model of Wang (2006) to multiple seasonalities and apply it to forecast hourly electricity load. The study is structured as follows: The general SSOE model for ESCov is presented in Section 6.2. Section 6.3 provides specific versions of the ESCov SSOE model. The empirical fitting of ESCov SSOE models and respective empirical forecasting techniques are explained in Sections 6.4 and 6.6. Maximum likelihood estimation is dealt with in Section 6.5. Section 6.7 explains how to model the covariate influence if is expected to be non-constant over time. Section 6.8 is concerned with the topic of renormalising the seasonal component. Section 6.9 applies the fitting and forecasting techniques provided in Sections 6.4 and 6.6 to hourly electricity consumption data of the customers of an energy vendor in some provinces of Emilia-Romagna, an Italian region. We provide some insights into the implementation of the ESCov model in R and associated numerical difficulties in Section 6.10.

6.2 State-space Models for Exponential Smoothing with Covariates

In this section we revise the SSOE state-space model that was found to be the statistical model underpinning ESCov by Wang (2006). The SSOE model for ESCov is an extension of the SSOE model for ES introduced by Ord et al. (1997). The addition of a linear covariate term by Wang (2006) is supposed to enhance the purely history-based forecasting method. ESCov therefore can in short be described as a combination of multiple linear regression with ES where both components are treated simultaneously. The results of the present and the subsequent sections contain the ES framework without covariates as a special case. Consequently, all results apply in the absence of covariates by setting the covariate parameter $\boldsymbol{\beta} = \mathbf{0}$.

The essential characteristic of the SSOE model is that it assumes only one source of error. In comparison to the multiple source of error model, Hyndman et al. (2002) finds the SSOE model for ES advantageous because it allows that the error-correction form of the classical ES smoothing equations finds itself in the state transition equations. Furthermore, both linear and nonlinear cases can be expressed.

Definition 6.1 (SSOE State-space Model for ESCov, Wang (2006))**.** *Let Y_t, Y_{t-1}, Y_{t-2}, $\ldots$ be an observed real-valued time series and $\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \boldsymbol{x}_{t-2}, \ldots \in \mathbb{R}^k$ a series of k-valued covariate vectors. The SSOE state-space model for ESCov is defined by an observation equation*

$$Y_t = f_{\boldsymbol{\alpha}}(\boldsymbol{u}_{t-1}) + \boldsymbol{\beta}^\top \boldsymbol{x}_t + \xi_t \tag{6.1}$$

and a state transition equation

$$\boldsymbol{u}_t = g_{\boldsymbol{\alpha}}(\boldsymbol{u}_{t-1}) + w_{\boldsymbol{\alpha}}(\boldsymbol{u}_{t-1})\xi_t, \tag{6.2}$$

where $\boldsymbol{u}_{t-1} \in \mathbb{R}^p$ is a state vector, $\boldsymbol{\alpha} \in \mathbb{R}^q$ a parameter vector, $f_{\boldsymbol{\alpha}} : \mathbb{R}^p \to \mathbb{R}$ is a continuously differentiable function and $g_{\boldsymbol{\alpha}}, w_{\boldsymbol{\alpha}} : \mathbb{R}^p \to \mathbb{R}^p$ are continuously differentiable functions. The error or residual ξ_t acts as a noise with $E[\xi_t] = 0$ in the observation and in the state transition equation.

The state vector $\boldsymbol{u}_{t-1}$ substantially influences Y_t via the function $f_{\boldsymbol{\alpha}}(\boldsymbol{u}_{t-1})$ and the subsequent state $\boldsymbol{u}_t$ via the state transition function $g_{\boldsymbol{\alpha}}(\boldsymbol{u}_{t-1})$. The covariate vector $\boldsymbol{x}_t = (x_{t1}, \ldots, x_{tk})^\top$ substantially influences Y_t via the linear term $\boldsymbol{\beta}^\top \boldsymbol{x}_t$. In an empirical context, Y_t and $\boldsymbol{x}_t$ are observable. All other quantities are unobservable or unknown and have to be estimated from observations Y_t.

The SSOE model requires no assumptions on the residuals beyond the basic assumption $E[\xi_t] = 0$, in particular no assumptions on independence or stationarity. Frequently, ξ_t is described as a function

$$\xi_t = k_{\boldsymbol{\alpha}}(\boldsymbol{u}_{t-1})\varepsilon_t \tag{6.3}$$

of ε_t with $E[\varepsilon_t] = 0$ and $V[\varepsilon_t] = \sigma^2$. Then the type of residual is determined by the function $k_{\boldsymbol{\alpha}}(\boldsymbol{u}_{t-1})$. In this study, we consider three types of residuals (ξ_t):

Additive (homoscedastic) residuals: Let $k_{\boldsymbol{\alpha}} \equiv 1$ in Eq. (6.3). The residuals (ξ_t) are independent and homoscedastic with $E[\xi_t] = 0$, $V[\xi_t] = \sigma^2 = V[\varepsilon_t]$.

Multiplicative (heteroscedastic) residuals: The residuals (ξ_t) are independent and heteroscedastic. They are obtained by a transformation of the baseline residuals ε_t with $k_{\boldsymbol{\alpha}}(\boldsymbol{u}_{t-1}) = f_{\boldsymbol{\alpha}}(\boldsymbol{u}_{t-1})$, and hence $\xi_t = f_{\boldsymbol{\alpha}}(\boldsymbol{u}_{t-1})\varepsilon_t$. The observation equation becomes

$$Y_t = f_{\boldsymbol{\alpha}}(\boldsymbol{u}_{t-1})(1 + \varepsilon_t) + \boldsymbol{\beta}^\top \boldsymbol{x}_t \tag{6.4}$$

with $E[1 + \varepsilon_t] = 1$.

Autoregressive residuals of order 1: The residuals (ξ_t) are dependent and subject to an AR(1) process, that is, they fulfil

$$\xi_t \quad = \quad \lambda \xi_{t-1} + \varepsilon_t, \qquad \text{where } -1 \leq \lambda < 1. \tag{6.5}$$

Hereby, (ε_t) are independent and satisfy $E[\varepsilon_t] = 0$ and $V[\varepsilon_t] = \sigma_\varepsilon^2$.

Ord et al. (1997) give an example where the function $k_{\boldsymbol{\alpha}}$ is a transformation of the function $f_{\boldsymbol{\alpha}}$ in the form $k_{\boldsymbol{\alpha}}(\boldsymbol{u}_{t-1}) = f_{\boldsymbol{\alpha}}(\boldsymbol{u}_{t-1})^\gamma$. The parameter $\gamma \in [0; 1]$ determines the magnitude of the heteroscedasticity. The additive residuals are a special case of this formula with $\gamma = 0$. Multiplicative residuals are obtained by setting $\gamma = 1$. Choosing $\gamma \in (0; 1)$ covers the area inbetween, which we do not consider in this study.

An essential virtue of SSOE models is to provide a basis for the derivation of the minimum mean square error (MMSE) h-step-ahead forecasts $Y_{t+h|t}$, i. e. the conditional expectations

$$Y_{t+h|t} \quad = \quad E[Y_{t+h}|\mathcal{H}_t] \tag{6.6}$$

under the process history $\mathcal{H}_t = \{(\xi_s)_{s \leq t}, (\boldsymbol{u}_s)_{s \leq t}\}$ up to time t. The derivation of the expressions for the conditional expectation $E[Y_{t+h}|\mathcal{H}_t]$ in Eqs. (6.11), (6.13) and (6.18) rests on the assumption that the residuals (ξ_s) are the only random drivers of the process, and that there is a causal relationship between the state vectors and the residuals; that means, each $\boldsymbol{u}_t$ can be expressed as a function $\boldsymbol{u}_t = \boldsymbol{u}_t(t, \xi_t, \xi_{t-1}, \ldots)$ of the errors up to time t. This property holds under mild regularity conditions for the linear SSOE model introduced in Definition 6.2, see the discussion by Ord et al. (1997). Although unproven for the nonlinear case, the high plausibility of the assumption warrants its use for inference on the conditional expectation.

It is then obvious that the one-step-ahead MMSE forecast for the SSOE model from Definition 6.1 under independent residuals is

$$Y_{t+1|t} \;=\; E[Y_{t+1}|\mathcal{H}_t] \;=\; f_{\boldsymbol{\alpha}}(\boldsymbol{u}_t) + \boldsymbol{\beta}^\top \boldsymbol{x}_{t+1}, \tag{6.7}$$

and that $Y_{t+1} - Y_{t+1|t} = \xi_{t+1}$. Therefore, the residual ξ_{t+1} is the forecast error of the one-step-ahead MMSE forecast made at time t. MMSE forecasts $Y_{t+h|t} = E[Y_{t+h}|\mathcal{H}_t]$ for specific model classes are provided in Sections 6.3.1 and 6.3.2.

6.3 SSOE Models for ESCov under Multiple Seasonality

Taylor (2003b) considers ES with additive (linear) trend and double multiplicative seasonality and Taylor (2010) with triple seasonality. We generalise this approach in the following respects: admitting covariates, accounting for a damped trend, analogising for additive seasonality and considering an arbitrary number of seasonalities. The resulting schemes are expressed as special cases of the SSOE model for ESCov described in Section 6.2. The models NT-NS, AT-NS, ADT-NS, NT-AS, AT-AS and ADT-AS from Table 6.1 can be expressed in form of a linear SSOE state-space model for ESCov, see Section 6.3.1. The models NT-MS, AT-MS and ADT-MS from Table 6.2 are instances of a partially linear SSOE state-space model for ESCov, see Section 6.3.2. The exponential trend models from Table 6.3 neither fit into the scheme described in Section 6.3.1 nor into the one from Section 6.3.2, and are therefore not considered any further here.

6.3.1 The Linear SSOE Model

The models NT-NS, AT-NS, ADT-NS, NT-AS, AT-AS and ADT-AS in Table 6.1 are all instances of the general formulation of a linear SSOE state-space model for ESCov presented in the subsequent definition. In the trend-free models NT, the state vector $\boldsymbol{u}_t$ contains only the level μ_t in the case of NS. In the case of AS or MS, it additionally contains for each seasonal component $i = 1, \ldots, m$ with seasonal lag d_i the relevant seasonal parameters $e_{i,t}, \ldots, e_{i,t-d_i+1}$. The parameter vector $\boldsymbol{\alpha}$ consists only of the level smoothing coefficient α_1 for NS and additionally of the m seasonal smoothing coefficients $\alpha_{1,3}, \ldots, \alpha_{m,3}$ for AS or MS. In the linear trend models AT, the state vector contains additionally the trend increment Δ_t, and the parameter vector $\boldsymbol{\alpha}$ contains additionally the smoothing coefficient α_2 of the trend increment. The damped trend models ADT have the same state vector as the undamped models AT, the parameter vector $\boldsymbol{\alpha}$ contains

Table 6.1: Linear SSOE models for exponential smoothing with covariates

model	observation equation, forecast	state transition equations
NT-NS	$Y_t = \mu_{t-1} + \boldsymbol{\beta}^\top \boldsymbol{x}_t + \xi_t$ $Y_{t+h\|t} = \mu_t + \boldsymbol{\beta}^\top \boldsymbol{x}_{t+h}$	$\mu_t = \mu_{t-1} + \alpha_1 \xi_t$
AT-NS	$Y_t = \mu_{t-1} + \Delta_{t-1} + \boldsymbol{\beta}^\top \boldsymbol{x}_t + \xi_t$ $Y_{t+h\|t} = \mu_t + h\Delta_t + \boldsymbol{\beta}^\top \boldsymbol{x}_{t+h}$	$\mu_t = \mu_{t-1} + \Delta_{t-1} + \alpha_1 \xi_t$ $\Delta_t = \Delta_{t-1} + \alpha_1 \alpha_2 \xi_t$
ADT-NS	$Y_t = \mu_{t-1} + \phi\Delta_{t-1} + \boldsymbol{\beta}^\top \boldsymbol{x}_t + \xi_t$ $Y_{t+h\|t} = \mu_t + \sum_{j=1}^h \phi^j \Delta_t + \boldsymbol{\beta}^\top \boldsymbol{x}_{t+h}$	$\mu_t = \mu_{t-1} + \phi\Delta_{t-1} + \alpha_1 \xi_t$ $\Delta_t = \phi\Delta_{t-1} + \alpha_1 \alpha_2 \xi_t$
NT-AS	$Y_t = \mu_{t-1} + \sum_{i=1}^m e_{i,t-d_i} + \boldsymbol{\beta}^\top \boldsymbol{x}_t + \xi_t$ $Y_{t+h\|t} = \mu_t + \sum_{i=1}^m e_{i,t+h-\lceil \frac{h}{d_i} \rceil d_i} + \boldsymbol{\beta}^\top \boldsymbol{x}_{t+h}$	$\mu_t = \mu_{t-1} + \alpha_1 \xi_t$ $e_{i,t} = e_{i,t-d_i} + (1-\alpha_1)\alpha_{i,3}\xi_t$
AT-AS	$Y_t = \mu_{t-1} + \Delta_{t-1} + \sum_{i=1}^m e_{i,t-d_i}$ $+\boldsymbol{\beta}^\top \boldsymbol{x}_t + \xi_t$ $Y_{t+h\|t} = \mu_t + h\Delta_t + \sum_{i=1}^m e_{i,t+h-\lceil \frac{h}{d_i} \rceil d_i}$ $+\boldsymbol{\beta}^\top \boldsymbol{x}_{t+h}$	$\mu_t = \mu_{t-1} + \Delta_{t-1} + \alpha_1 \xi_t$ $\Delta_t = \Delta_{t-1} + \alpha_1 \alpha_2 \xi_t$ $e_{i,t} = e_{i,t-d_i} + (1-\alpha_1)\alpha_{i,3}\xi_t$
ADT-AS	$Y_t = \mu_{t-1} + \phi\Delta_{t-1} + \sum_{i=1}^m e_{i,t-d_i}$ $+\boldsymbol{\beta}^\top \boldsymbol{x}_t + \xi_t$ $Y_{t+h\|t} = \mu_t + \sum_{j=1}^h \phi^j \Delta_t + \sum_{i=1}^m e_{i,t+h-\lceil \frac{h}{d_i} \rceil d_i}$ $+\boldsymbol{\beta}^\top \boldsymbol{x}_{t+h}$	$\mu_t = \mu_{t-1} + \phi\Delta_{t-1} + \alpha_1 \xi_t$ $\Delta_t = \phi\Delta_{t-1} + \alpha_1 \alpha_2 \xi_t$ $e_{i,t} = e_{i,t-d_i} + (1-\alpha_1)\alpha_{i,3}\xi_t$

Table 6.2: Partially linear SSOE models for exponential smoothing with covariates

model	observation equation, forecast	state transition equations
NT-MS	$Y_t = \mu_{t-1} \prod_{i=1}^m e_{i,t-d_i} + \boldsymbol{\beta}^\top \boldsymbol{x}_t + \xi_t$ $Y_{t+h\|t} = \mu_t \prod_{i=1}^m e_{i,t+h-\lceil \frac{h}{d_i} \rceil d_i} + \boldsymbol{\beta}^\top \boldsymbol{x}_{t+h}$	$\mu_t = \mu_{t-1} + \frac{\alpha_1 \xi_t}{\prod_{i=1}^m e_{i,t-d_i}}$ $e_{i,t} = e_{i,t-d_i} + \frac{(1-\alpha_1)\alpha_{i,3}\xi_t}{\mu_{t-1}+\phi\Delta_{t-1}}$
AT-MS	$Y_t = (\mu_{t-1} + \Delta_{t-1}) \prod_{i=1}^m e_{i,t-d_i}$ $+\boldsymbol{\beta}^\top \boldsymbol{x}_t + \xi_t$ $Y_{t+h\|t} = (\mu_t + h\Delta_t) \prod_{i=1}^m e_{i,t+h-\lceil \frac{h}{d_i} \rceil d_i}$ $+\boldsymbol{\beta}^\top \boldsymbol{x}_{t+h}$	$\mu_t = \mu_{t-1} + \Delta_{t-1} + \frac{\alpha_1 \xi_t}{\prod_{i=1}^m e_{i,t-d_i}}$ $\Delta_t = \Delta_{t-1} + \frac{\alpha_1 \alpha_2 \xi_t}{\prod_{i=1}^m e_{i,t-d_i}}$ $e_{i,t} = e_{i,t-d_i} + \frac{(1-\alpha_1)\alpha_{i,3}\xi_t}{\mu_{t-1}+\phi\Delta_{t-1}}$
ADT-MS	$Y_t = (\mu_{t-1} + \phi\Delta_{t-1}) \prod_{i=1}^m e_{i,t-d_i}$ $+\boldsymbol{\beta}^\top \boldsymbol{x}_t + \xi_t$ $Y_{t+h\|t} = (\mu_t + \sum_{j=1}^h \phi^j \Delta_t) \prod_{i=1}^m e_{i,t+h-\lceil \frac{h}{d_i} \rceil d_i}$ $+\boldsymbol{\beta}^\top \boldsymbol{x}_{t+h}$	$\mu_t = \mu_{t-1} + \phi\Delta_{t-1} + \frac{\alpha_1 \xi_t}{\prod_{i=1}^m e_{i,t-d_i}}$ $\Delta_t = \phi\Delta_{t-1} + \frac{\alpha_1 \alpha_2 \xi_t}{\prod_{i=1}^m e_{i,t-d_i}}$ $e_{i,t} = e_{i,t-d_i} + \frac{(1-\alpha_1)\alpha_{i,3}\xi_t}{\mu_{t-1}+\phi\Delta_{t-1}}$

additionally the damping coefficient $0 < \phi \leq 1$. Setting the damping parameter $\phi = 1$ in the damped models ADT leads to the undamped versions AT.

Definition 6.2 (Linear SSOE State-space Model for ESCov). *Let Y_t, Y_{t-1}, $Y_{t-2}, \ldots$ be an observed real-valued time series and $\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \boldsymbol{x}_{t-2}, \ldots \in \mathbb{R}^k$ a series of covariate vectors. The linear SSOE state-space model for ESCov is defined by an observation equation*

$$Y_t = \boldsymbol{\delta}_{\boldsymbol{\alpha}}^\top \boldsymbol{u}_{t-1} + \boldsymbol{\beta}^\top \boldsymbol{x}_t + \xi_t \tag{6.8}$$

and a state transition equation

$$\boldsymbol{u}_t = \mathbf{G}_{\boldsymbol{\alpha}} \boldsymbol{u}_{t-1} + \boldsymbol{w}_{\boldsymbol{\alpha}} \xi_t. \tag{6.9}$$

Here, $\boldsymbol{u}_t \in \mathbb{R}^p$ is a state vector, $\boldsymbol{\alpha} \in \mathbb{R}^q$ a parameter vector, $\boldsymbol{\delta}_{\boldsymbol{\alpha}}$ a vector from $\mathbb{R}^p$, $\mathbf{G}_{\boldsymbol{\alpha}}$ a $p \times p$ transition matrix and $\boldsymbol{w}_{\boldsymbol{\alpha}}$ a continuously differentiable function with $\boldsymbol{w}_{\boldsymbol{\alpha}} : \mathbb{R}^p \to \mathbb{R}^p$. (ξ_t) is a series of errors with $E[\xi_t] = 0$.

In the case of the linear SSOE model for ESCov, the heteroscedastic residual series $(\xi_t) = (f_{\boldsymbol{\alpha}}(\boldsymbol{u}_{t-1})\varepsilon_t)$ satisfies $\xi_t = \boldsymbol{\delta}_{\boldsymbol{\alpha}}^\top \boldsymbol{u}_{t-1}\varepsilon_t$ with conditional variance $V[\xi_t|\mathcal{H}_{t-1}] = \sigma^2_{\xi_t|\mathcal{H}_{t-1}} = \left(\boldsymbol{\delta}_{\boldsymbol{\alpha}}^\top \boldsymbol{u}_{t-1}\right)^2 \sigma^2_\varepsilon$.

The MMSE h-step-ahead forecast in the case of independent errors is provided in the subsequent proposition.

Table 6.3: Exponential trend models for exponential smoothing with covariates

model	observation equation, forecast	state transition equations
ET-NS	$Y_t = \mu_{t-1}\Delta_{t-1} + \boldsymbol{\beta}^\top \boldsymbol{x}_t + \xi_t$ $Y_{t+h\|t} = \mu_t\Delta_t^h + \boldsymbol{\beta}^\top \boldsymbol{x}_{t+h}$	$\mu_t = \mu_{t-1}\Delta_{t-1} + \alpha_1\xi_t$ $\Delta_t = \Delta_{t-1} + \alpha_1\alpha_2\xi_t/\mu_{t-1}$
EDT-NS	$Y_t = \mu_{t-1}\Delta_{t-1}^{\phi} + \boldsymbol{\beta}^\top \boldsymbol{x}_t + \xi_t$ $Y_{t+h\|t} = \mu_t\Delta_t^{\sum_{j=1}^h \phi^j} + \boldsymbol{\beta}^\top \boldsymbol{x}_{t+h}$	$\mu_t = \mu_{t-1}\Delta_{t-1}^{\phi} + \alpha_1\xi_t$ $\Delta_t = \Delta_{t-1}^{\phi} + \alpha_1\alpha_2\xi_t/\mu_{t-1}$
ET-AS	$Y_t = \mu_{t-1}\Delta_{t-1} + \sum_{i=1}^m e_{i,t-d_i}$ $+\boldsymbol{\beta}^\top \boldsymbol{x}_t + \xi_t$ $Y_{t+h\|t} = \mu_t\Delta_t^h + \sum_{i=1}^m e_{i,t+h-\lceil \frac{h}{d_i}\rceil d_i}$ $+\boldsymbol{\beta}^\top \boldsymbol{x}_{t+h}$	$\mu_t = \mu_{t-1}\Delta_{t-1} + \alpha_1\xi_t$ $\Delta_t = \Delta_{t-1} + \alpha_1\alpha_2\xi_t/\mu_{t-1}$ $e_{i,t} = e_{i,t-d_i} + (1-\alpha_1)\alpha_{i,3}\xi_t$
EDT-AS	$Y_t = \mu_{t-1}\Delta_{t-1}^{\phi} + \sum_{i=1}^m e_{i,t-d_i}$ $+\boldsymbol{\beta}^\top \boldsymbol{x}_t + \xi_t$ $Y_{t+h\|t} = \mu_t\Delta_t^{\sum_{j=1}^h \phi^j} + \sum_{i=1}^m e_{i,t+h-\lceil \frac{h}{d_i}\rceil d_i}$ $+\boldsymbol{\beta}^\top \boldsymbol{x}_{t+h}$	$\mu_t = \mu_{t-1}\Delta_{t-1}^{\phi} + \alpha_1\xi_t$ $\Delta_t = \Delta_{t-1}^{\phi} + \alpha_1\alpha_2\xi_t/\mu_{t-1}$ $e_{i,t} = e_{i,t-d_i} + (1-\alpha_1)\alpha_{i,3}\xi_t$
ET-MS	$Y_t = (\mu_{t-1}\Delta_{t-1})\prod_{i=1}^m e_{i,t-d_i}$ $+\boldsymbol{\beta}^\top \boldsymbol{x}_t + \xi_t$ $Y_{t+h\|t} = (\mu_t\Delta_t^h)\prod_{i=1}^m e_{i,t+h-\lceil \frac{h}{d_i}\rceil d_i}$ $+\boldsymbol{\beta}^\top \boldsymbol{x}_{t+h}$	$\mu_t = \mu_{t-1}\Delta_{t-1} + \frac{\alpha_1\xi_t}{\prod_{i=1}^m e_{i,t-d_i}}$ $\Delta_t = \Delta_{t-1} + \frac{\alpha_1\alpha_2\xi_t}{\mu_{t-1}\prod_{i=1}^m e_{i,t-d_i}}$ $e_{i,t} = e_{i,t-d_i} + \frac{(1-\alpha_1)\alpha_{i,3}\xi_t}{\mu_{t-1}\Delta_{t-1}}$
EDT-MS	$Y_t = \mu_{t-1}\Delta_{t-1}^{\phi}\prod_{i=1}^m e_{i,t-d_i}$ $+\boldsymbol{\beta}^\top \boldsymbol{x}_t + \xi_t$ $Y_{t+h\|t} = \mu_t\Delta_t^{\sum_{j=1}^h \phi^j}\prod_{i=1}^m e_{i,t+h-\lceil \frac{h}{d_i}\rceil d_i}$ $+\boldsymbol{\beta}^\top \boldsymbol{x}_{t+h}$	$\mu_t = \mu_{t-1}\Delta_{t-1}^{\phi} + \frac{\alpha_1\xi_t}{\prod_{i=1}^m e_{i,t-d_i}}$ $\Delta_t = \Delta_{t-1}^{\phi} + \frac{\alpha_1\alpha_2\xi_t}{\mu_{t-1}\prod_{i=1}^m e_{i,t-d_i}}$ $e_{i,t} = e_{i,t-d_i} + \frac{(1-\alpha_1)\alpha_{i,3}\xi_t}{\mu_{t-1}\Delta_{t-1}^{\phi}}$

Proposition 6.3 (MMSE Forecast for Linear SSOE under Independent Residuals)**.** *Consider the linear SSOE model for ESCov from Definition 6.2 with independent errors* (ξ_t). *For each* t *and each* $k > 0$ *let*

$$E[\xi_{t+k}|\mathcal{H}_t] \quad = \quad E[\xi_{t+k}|(\xi_s)_{s\leq t}] \quad = \quad E[\xi_{t+k}] \quad = \quad 0. \tag{6.10}$$

Let $h > 0$. *The MMSE* h*-step-ahead forecast is given by*

$$Y_{t+h|t} \quad = \quad E[Y_{t+h}|\mathcal{H}_t] \quad = \quad \boldsymbol{\delta}_{\boldsymbol{\alpha}}^\top \mathbf{G}_{\boldsymbol{\alpha}}^{h-1} \boldsymbol{u}_t + \boldsymbol{\beta}^\top \boldsymbol{x}_{t+h} \tag{6.11}$$

and the conditional forecasting variance is given by

$$\sigma^2_{Y_{t+h}|\mathcal{H}_t} \quad = \quad V[Y_{t+h}|\mathcal{H}_t] \quad = \quad \sum_{\ell=0}^{h-2} \left(\boldsymbol{\delta}_{\boldsymbol{\alpha}}^\top \mathbf{G}_{\boldsymbol{\alpha}}^{\ell} \boldsymbol{w}_{\boldsymbol{\alpha}} \right)^2 \sigma^2_{\xi_{t+h-\ell-1}} + \sigma^2_{\xi_{t+h}}. \tag{6.12}$$

Proof. See Appendix 6.A, Section 6.A.1. □

From Proposition 6.3 we can see that the forecast under the SSOE model for ESCov relies on the availability of the covariate vector $\boldsymbol{x}_{t+h}$ at the future time point $t + h$. Under the covariate-free model, knowledge about future values of certain components is naturally not a requirement. Proposition 6.3 furthermore shows that the model with additive and multiplicative residuals have the same point forecast $Y_{t+h|t} = E[Y_{t+h}|\mathcal{H}_t]$. The models differ only in the conditional forecasting variance (6.12).

The particular components $\boldsymbol{u}_{t-1}$, $\boldsymbol{\delta}_{\boldsymbol{\alpha}}$, $\mathbf{G}_{\boldsymbol{\alpha}}$ and $\boldsymbol{w}_{\boldsymbol{\alpha}}$ for a certain ESCov model from Table 6.1 can be taken from Table 6.4. Hereby, the $d_i \times d_i$-matrix $\mathbf{E}_i$ and the vector $\boldsymbol{e}_{i,t-d_i,t-1}$ are defined by

$$\mathbf{E}_i := \begin{pmatrix} \mathbf{0}_{d_i-1}^\top & 1 \\ \mathbf{I}_{d_i-1} & \mathbf{0}_{d_i-1} \end{pmatrix}, \quad \boldsymbol{e}_{i,t-d_i,t-1} := \begin{pmatrix} e_{i,t-1} \\ \vdots \\ e_{i,t-d_i} \end{pmatrix},$$

where $\mathbf{I}_{d_i-1}$ is the $(d_i - 1) \times (d_i - 1)$ identity matrix and $\mathbf{0}_{d_i-1}$ a vector of zeros of length $d_i - 1$. By choosing the parameters according to Table 6.4, the models NT-NS, AT-NS, ADT-NS, NT-AS, AT-AS and ADT-AS can be denoted as instances of the scheme from Definition 6.2.

Another instance of a linear SSOE model for ESCov is the model which considers the residual series (ξ_t) to be subject to a causal AR(1) process, i. e. Eq. (6.5) holds. The MMSE h-step-ahead forecast in the case of AR(1) errors is provided in the subsequent proposition.

Table 6.4: Components of linear SSOE models for ESCov

model	$\boldsymbol{u}_{t-1}$	$\boldsymbol{\delta}_{\boldsymbol{\alpha}}$	$\mathbf{G}_{\boldsymbol{\alpha}}$	$\boldsymbol{w}_{\boldsymbol{\alpha}}$
NT-NS	$\begin{pmatrix}\mu_{t-1}\end{pmatrix}$	$\begin{pmatrix}1\end{pmatrix}$	$\begin{pmatrix}1\end{pmatrix}$	$\begin{pmatrix}\alpha_1\end{pmatrix}$
AT-NS	$\begin{pmatrix}\mu_{t-1}\\ \Delta_{t-1}\end{pmatrix}$	$\begin{pmatrix}1\\ 1\end{pmatrix}$	$\begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix}$	$\begin{pmatrix}\alpha_1\\ \alpha_1\alpha_2\end{pmatrix}$
ADT-NS	$\begin{pmatrix}\mu_{t-1}\\ \Delta_{t-1}\end{pmatrix}$	$\begin{pmatrix}1\\ \phi\end{pmatrix}$	$\begin{pmatrix}1 & \phi\\ 0 & \phi\end{pmatrix}$	$\begin{pmatrix}\alpha_1\\ \alpha_1\alpha_2\end{pmatrix}$
NT-AS	$\begin{pmatrix}\mu_{t-1}\\ \boldsymbol{e}_{1,t-d_1,t-1}\\ \boldsymbol{e}_{2,t-d_2,t-1}\\ \vdots\\ \boldsymbol{e}_{m,t-d_m,t-1}\end{pmatrix}$	$\begin{pmatrix}1\\ \mathbf{0}_{d_1-1}\\ 1\\ \mathbf{0}_{d_2-1}\\ 1\\ \vdots\\ \mathbf{0}_{d_m-1}\\ 1\end{pmatrix}$	$\begin{pmatrix}1 & \mathbf{0}_{d_1}^\top & \mathbf{0}_{d_2}^\top & \cdots & \mathbf{0}_{d_m}^\top\\ \mathbf{0}_{d_1} & \mathbf{E}_1 & \mathbf{0}_{d_1\times d_2} & \cdots & \mathbf{0}_{d_1\times d_m}\\ \mathbf{0}_{d_2} & \mathbf{0}_{d_1\times d_2} & \mathbf{E}_2 & \cdots & \mathbf{0}_{d_2\times d_m}\\ \vdots & \vdots & \ddots & \ddots & \vdots\\ \mathbf{0}_{d_m} & \mathbf{0}_{d_m\times d_1} & \mathbf{0}_{d_m\times d_2} & \cdots & \mathbf{E}_m\end{pmatrix}$	$\begin{pmatrix}\alpha_1\\ (1-\alpha_1)\alpha_{1,3}\\ \mathbf{0}_{d_1-1}\\ (1-\alpha_1)\alpha_{2,3}\\ \mathbf{0}_{d_2-1}\\ \vdots\\ (1-\alpha_1)\alpha_{m,3}\\ \mathbf{0}_{d_m-1}\end{pmatrix}$
AT-AS	$\begin{pmatrix}\mu_{t-1}\\ \Delta_{t-1}\\ \boldsymbol{e}_{1,t-d_1,t-1}\\ \boldsymbol{e}_{2,t-d_2,t-1}\\ \vdots\\ \boldsymbol{e}_{m,t-d_m,t-1}\end{pmatrix}$	$\begin{pmatrix}1\\ 1\\ \mathbf{0}_{d_1-1}\\ 1\\ \mathbf{0}_{d_2-1}\\ 1\\ \vdots\\ \mathbf{0}_{d_m-1}\\ 1\end{pmatrix}$	$\begin{pmatrix}1 & 1 & \mathbf{0}_{d_1}^\top & \mathbf{0}_{d_2}^\top & \cdots & \mathbf{0}_{d_m}^\top\\ 0 & 1 & \mathbf{0}_{d_1}^\top & \mathbf{0}_{d_2}^\top & \cdots & \mathbf{0}_{d_m}^\top\\ \mathbf{0}_{d_1} & \mathbf{0}_{d_1} & \mathbf{E}_1 & \mathbf{0}_{d_1\times d_2} & \cdots & \mathbf{0}_{d_1\times d_m}\\ \mathbf{0}_{d_2} & \mathbf{0}_{d_2} & \mathbf{0}_{d_1\times d_2} & \mathbf{E}_2 & \cdots & \mathbf{0}_{d_2\times d_m}\\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots\\ \mathbf{0}_{d_m} & \mathbf{0}_{d_m} & \mathbf{0}_{d_m\times d_1} & \mathbf{0}_{d_m\times d_2} & \cdots & \mathbf{E}_m\end{pmatrix}$	$\begin{pmatrix}\alpha_1\\ \alpha_1\alpha_2\\ (1-\alpha_1)\alpha_{1,3}\\ \mathbf{0}_{d_1-1}\\ (1-\alpha_1)\alpha_{2,3}\\ \mathbf{0}_{d_2-1}\\ \vdots\\ (1-\alpha_1)\alpha_{m,3}\\ \mathbf{0}_{d_m-1}\end{pmatrix}$
ADT-AS	$\begin{pmatrix}\mu_{t-1}\\ \Delta_{t-1}\\ \boldsymbol{e}_{1,t-d_1,t-1}\\ \boldsymbol{e}_{2,t-d_2,t-1}\\ \vdots\\ \boldsymbol{e}_{m,t-d_m,t-1}\end{pmatrix}$	$\begin{pmatrix}1\\ \phi\\ \mathbf{0}_{d_1-1}\\ 1\\ \mathbf{0}_{d_2-1}\\ 1\\ \vdots\\ \mathbf{0}_{d_m-1}\\ 1\end{pmatrix}$	$\begin{pmatrix}1 & \phi & \mathbf{0}_{d_1}^\top & \mathbf{0}_{d_2}^\top & \cdots & \mathbf{0}_{d_m}^\top\\ 0 & \phi & \mathbf{0}_{d_1}^\top & \mathbf{0}_{d_2}^\top & \cdots & \mathbf{0}_{d_m}^\top\\ \mathbf{0}_{d_1} & \mathbf{0}_{d_1} & \mathbf{E}_1 & \mathbf{0}_{d_1\times d_2} & \cdots & \mathbf{0}_{d_1\times d_m}\\ \mathbf{0}_{d_2} & \mathbf{0}_{d_2} & \mathbf{0}_{d_1\times d_2} & \mathbf{E}_2 & \cdots & \mathbf{0}_{d_2\times d_m}\\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots\\ \mathbf{0}_{d_m} & \mathbf{0}_{d_m} & \mathbf{0}_{d_m\times d_1} & \mathbf{0}_{d_m\times d_2} & \cdots & \mathbf{E}_m\end{pmatrix}$	$\begin{pmatrix}\alpha_1\\ \alpha_1\alpha_2\\ (1-\alpha_1)\alpha_{1,3}\\ \mathbf{0}_{d_1-1}\\ (1-\alpha_1)\alpha_{2,3}\\ \mathbf{0}_{d_2-1}\\ \vdots\\ (1-\alpha_1)\alpha_{m,3}\\ \mathbf{0}_{d_m-1}\end{pmatrix}$

Proposition 6.4 (MMSE Forecast for Linear SSOE under AR(1) Residuals)**.** *Consider the linear SSOE model for ESCov from Definition 6.2 under AR(1) errors, i. e. Eq. (6.5) holds. For each t and each $k > 0$ let $E[\xi_{t+k}|\mathcal{H}_t] = E[\xi_{t+k}|(\xi_s)_{s\leq t}]$. Let $h > 0$. The MMSE h-step-ahead forecast is given by*

$$Y_{t+h|t} \quad = \quad E[Y_{t+h}|\mathcal{H}_t] = \left(\lambda^h + \boldsymbol{\delta}_{\boldsymbol{\alpha}}^\top \sum_{\ell=0}^{h-2} \mathbf{G}_{\boldsymbol{\alpha}}^\ell \boldsymbol{w}_{\boldsymbol{\alpha}} \lambda^{h-\ell-1}\right) \xi_t + \boldsymbol{\delta}_{\boldsymbol{\alpha}}^\top \mathbf{G}_{\boldsymbol{\alpha}}^{h-1} \boldsymbol{u}_t + \boldsymbol{\beta}^\top \boldsymbol{x}_{t+h} \tag{6.13}$$

and the conditional forecasting variance is given by

$$\sigma^2_{Y_{t+h}|\mathcal{H}_t} \quad = \quad V[Y_{t+h}|\mathcal{H}_t] \tag{6.14}$$

$$= \sigma_\varepsilon^2 \left(2\sum_{\ell=0}^{h-2} \boldsymbol{\delta}_{\boldsymbol{\alpha}}^\top \mathbf{G}_{\boldsymbol{\alpha}}^\ell \boldsymbol{w}_{\boldsymbol{\alpha}} \lambda^{\ell+1} \frac{1-\lambda^{2(h-\ell-1)}}{1-\lambda^2} + \sum_{\ell=0}^{h-2} \left(\boldsymbol{\delta}_{\boldsymbol{\alpha}}^\top \mathbf{G}_{\boldsymbol{\alpha}}^\ell \boldsymbol{w}_{\boldsymbol{\alpha}}\right)^2 \frac{1-\lambda^{2(h-\ell-1)}}{1-\lambda^2} + \frac{1-\lambda^{2h}}{1-\lambda^2} \right).$$

Proof. See Appendix 6.A, Section 6.A.2. □

6.3.2 The Partially Linear SSOE Model

The partially linear SSOE state-space model for ESCov covers the exponential smoothing models NT-MS, AT-MS, ADT-MS with (multiple) multiplicative seasonality as provided in Table 6.2. The observation equation and state transition equation are linear in the level μ_{t-1} and the trend increment Δ_{t-1}, but nonlinear in the multiplicative seasonality.

The quantities $\boldsymbol{u}_{t,ne}$ and $\boldsymbol{u}_{t,e}$ in the subsequent definition of the partially linear SSOE state-space model for ESCov are built as follows: The component $\boldsymbol{u}_{t,ne}$ is of the form $\boldsymbol{u}_{t,ne} = (\mu_t)$ for the model NT without trend and of the form $\boldsymbol{u}_{t,ne} = (\mu_t, \Delta_t)^\top$ for the models AT or ADT with trend. $\boldsymbol{u}_{t,e}$ contains the relevant seasonal components at time t.

Definition 6.5 (Partially Linear SSOE State-space Model for ESCov)**.** *Let Y_t, Y_{t-1}, Y_{t-2}, $\ldots$ be an observed real-valued time series and $\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \boldsymbol{x}_{t-2}, \ldots \in \mathbb{R}^k$ a series of covariate vectors. Let $\boldsymbol{u}_t = (\boldsymbol{u}_{t,ne}, \boldsymbol{u}_{t,e})^\top$ be a decomposition of the state vector $\boldsymbol{u}_t$ into a nonseasonal part $\boldsymbol{u}_{t,ne}$ and a seasonal part $\boldsymbol{u}_{t,e}$.*

Let $1 \leq h \leq d_1, \ldots, d_m$.

The general form of the observation equation for the partially linear SSOE model for ESCov is

$$Y_t = \underbrace{\boldsymbol{\delta}_{\boldsymbol{\alpha},ne}^\top \boldsymbol{u}_{t-1,ne} \prod_{i=1}^m e_{i,t-d_i}}_{f_{\boldsymbol{\alpha}}(\boldsymbol{u}_{t-1})} + \boldsymbol{\beta}^\top \boldsymbol{x}_t + \xi_t. \tag{6.15}$$

Table 6.5: Components of partially linear SSOE models for ESCov

model	$\boldsymbol{u}_{t-1,ne}$	$\boldsymbol{u}_{t-1,e}$	$\boldsymbol{\delta}_{\boldsymbol{\alpha},ne}$	$\mathbf{G}_{\boldsymbol{\alpha},ne}$	$\boldsymbol{w}_{\boldsymbol{\alpha},ne}$
NT-MS	$\begin{pmatrix}\mu_{t-1}\end{pmatrix}$	$\begin{pmatrix}\boldsymbol{e}_{1,t-d_1,t-1}\\ \vdots \\ \boldsymbol{e}_{m,t-d_m,t-1}\end{pmatrix}$	$\begin{pmatrix}1\end{pmatrix}$	$\begin{pmatrix}1\end{pmatrix}$	$\begin{pmatrix}\alpha_1\end{pmatrix}$
AT-MS	$\begin{pmatrix}\mu_{t-1}\\ \Delta_{t-1}\end{pmatrix}$		$\begin{pmatrix}1\\ 1\end{pmatrix}$	$\begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix}$	$\begin{pmatrix}\alpha_1\\ \alpha_1\alpha_2\end{pmatrix}$
ADT-MS	$\begin{pmatrix}\mu_{t-1}\\ \Delta_{t-1}\end{pmatrix}$		$\begin{pmatrix}1\\ \phi\end{pmatrix}$	$\begin{pmatrix}1 & \phi\\ 0 & \phi\end{pmatrix}$	$\begin{pmatrix}\alpha_1\\ \alpha_1\alpha_2\end{pmatrix}$

The state transition equation is split into the part for $\boldsymbol{u}_{t,ne}$ given by

$$\boldsymbol{u}_{t,ne} \quad = \quad \mathbf{G}_{\boldsymbol{\alpha},ne}\boldsymbol{u}_{t-1,ne} + \boldsymbol{w}_{\boldsymbol{\alpha},ne}\frac{\xi_t}{\prod_{i=1}^{m} e_{i,t-d_i}}, \tag{6.16}$$

and into the part for $\boldsymbol{u}_{t,e}$ given by the m recursions

$$\boldsymbol{e}_{i,t-d_i+1,t} \quad = \quad \begin{pmatrix} e_{i,t-d_i} + (1-\alpha_1)\alpha_{i,3}\frac{\xi_t}{\boldsymbol{\delta}^\top_{\boldsymbol{\alpha},ne},\boldsymbol{u}_{t-1,ne}} \\ e_{i,t-1} \\ \vdots \\ e_{i,t-d_i+1} \end{pmatrix}, \quad i = 1,\dots,m. \tag{6.17}$$

For the models NT-MS, AT-MS, ADT-MS from Table 6.2, the vectors $\boldsymbol{u}_{t,ne}, \boldsymbol{u}_{t,e}, \boldsymbol{\delta}_{\boldsymbol{\alpha},ne}$, $\boldsymbol{w}_{\boldsymbol{\alpha},ne}$ and the matrix $\mathbf{G}_{\boldsymbol{\alpha},ne}$ can be read from Table 6.5.

The MMSE h-step-ahead forecast for the partially linear SSOE for ESCov is provided in the subsequent proposition.

Proposition 6.6 (MMSE Forecast for Partially Linear SSOE under Independent Residuals)**.**
Consider the partially linear SSOE model for ESCov from Definition 6.5. Let $0 < h \leq d_1,\dots,d_m$ be the forecasting horizon below the m seasonal lags. For each t and each $k > 0$ let $E[\xi_{t+k}|\mathcal{H}_t] = E[\xi_{t+k}|(\xi_s)_{s\leq t}] = 0$. The MMSE h-step-ahead forecast is given by

$$Y_{t+h|t} = E[Y_{t+h}|\mathcal{H}_t] = \boldsymbol{\delta}^\top_{\boldsymbol{\alpha},ne}\mathbf{G}^{h-1}_{\boldsymbol{\alpha},ne}\boldsymbol{u}_{t,ne}\prod_{i=1}^{m} e_{i,t+h-d_i} + \boldsymbol{\beta}^\top\boldsymbol{x}_{t+h}. \tag{6.18}$$

Let (ξ_t) be independent errors. The conditional forecasting variance for forecasting horizon $0 < h \leq d_1,\dots,d_m$ is given by

$$\sigma^2_{Y_{t+h}|\mathcal{H}_t} = V[Y_{t+h}|\mathcal{H}_t] = \sum_{\ell=0}^{h-2}\left(\boldsymbol{\delta}^\top_{\boldsymbol{\alpha},ne}\mathbf{G}^{\ell}_{\boldsymbol{\alpha},ne}\boldsymbol{w}_{\boldsymbol{\alpha},ne}\right)^2\sigma^2_{\xi_{t+h-\ell-1}}\frac{\prod_{i=1}^{m} e^2_{i,t+h-d_i}}{\prod_{i=1}^{m} e^2_{i,t+h-d_i-\ell-1}} + \sigma^2_{\xi_{t+h}}.$$

$$(6.19)$$

Proof. See Appendix 6.A, Section 6.A.3. □

The equations for the MMSE h-step-ahead forecast and its variance in Proposition 6.6 obviously only hold for the case $h \leq \min\{d_1, \ldots, d_m\}$. For $h \geq d_i + 1$, the formula would use seasonal coefficients $e_{i,s}$ with $s > t$. A sensible approximation of the MMSE h-step-ahead forecast $E[Y_{t+h}|\mathcal{H}_t]$ for cases $h \geq d_i + 1$ is provided by using

$$\boldsymbol{\delta}_{\boldsymbol{\alpha},ne}^{\top} \mathbf{G}_{\boldsymbol{\alpha},ne}^{h-1} \boldsymbol{u}_{t,ne} \prod_{i=1}^{m} e_{i,t+h-\left\lceil \frac{h}{d_i} \right\rceil d_i} + \boldsymbol{\beta}^{\top} \boldsymbol{x}_{t+h} \tag{6.20}$$

in the right-hand side of Eq. (6.18). Equation (6.20) leads to the forecasts provided by Table 6.2. Analogously, use

$$\sum_{\ell=0}^{h-2} \left(\boldsymbol{\delta}_{\boldsymbol{\alpha},ne}^{\top} \mathbf{G}_{\boldsymbol{\alpha},ne}^{\ell} \boldsymbol{w}_{\boldsymbol{\alpha},ne} \right)^2 \sigma_{\xi_{t+h-\ell-1}}^2 \frac{\prod_{i=1}^{m} e^2_{i,t+h-\left\lceil \frac{h}{d_i} \right\rceil d_i}}{\prod_{i=1}^{m} e^2_{i,t+h-\left\lceil \frac{h}{d_i} \right\rceil d_i-\ell-1}} + \sigma_{\xi_{t+h}}^2$$

in the right-hand side of Eq. (6.19) to approximate $\sigma^2_{Y_{t+h}|\mathcal{H}_t}$ in cases $h \geq d_i + 1$.

In the case of the partially linear SSOE model for ESCov, the heteroscedastic residual series $(\xi_t) = (f_{\boldsymbol{\alpha}}(\boldsymbol{u}_{t-1})\varepsilon_t)$ satisfies $\xi_t = \boldsymbol{\delta}_{\boldsymbol{\alpha},ne}^{\top} \boldsymbol{u}_{t-1,ne} \prod_{i=1}^{m} e_{i,t-d_i} \varepsilon_t$, where (ε_s) is considered as homoscedastic. For the conditional variance we have $V[\xi_t|\mathcal{H}_{t-1}] = \sigma^2_{\xi_t|\mathcal{H}_{t-1}} = \left(\boldsymbol{\delta}_{\boldsymbol{\alpha},ne}^{\top} \boldsymbol{u}_{t-1,ne} \prod_{i=1}^{m} e_{i,t-d_i}\right)^2 \sigma_\varepsilon^2$.

The MMSE forecast as well as its variance in the case of AR(1) residuals for the partially linear SSOE for ESCov still has to be explored.

6.4 Empirical Fitting of ESCov SSOE Models

Consider a time series (Y_t) that is assumed to follow an ESCov SSOE model defined by an observation equation and a state transition equation as in Definition 6.1, both specified in Tables 6.1 to 6.3. Let $Y_1, \ldots, Y_T$ be T successive observations from the series. The estimation of the model parameters proceeds in the following four steps (i)–(iv):

(i) Iterative estimation of state vectors and residuals under known transition function components $g_{\boldsymbol{\alpha}}$, $\boldsymbol{w}_{\boldsymbol{\alpha}}$, known observation function $f_{\boldsymbol{\alpha}}$ and known parameter vector $\boldsymbol{\theta} = (\boldsymbol{\alpha}^{\top}, \boldsymbol{\beta}^{\top})^{\top}$: (i.1) Provide a starting value $\boldsymbol{u}_0$ for the state vector. (i.2) Estimate the residual ξ_1 from the forecast error by $\widehat{\xi}_1 = Y_1 - f_{\boldsymbol{\alpha}}(\boldsymbol{u}_0) - \boldsymbol{\beta}^{\top}\boldsymbol{x}_1$, and estimate the state vector $\boldsymbol{u}_1$ from $\widehat{\boldsymbol{u}}_1 = g_{\boldsymbol{\alpha}}(\boldsymbol{u}_0) + w_{\boldsymbol{\alpha}}(\boldsymbol{u}_0)\widehat{\xi}_1$. (i.3) Continue the pattern (i.2) to estimate successively $\xi_1, \ldots, \xi_T$ and $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_T$ as functions of $\boldsymbol{u}_0$.

(ii) Estimation of the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\alpha}^\top, \boldsymbol{\beta}^\top)^\top$ under known structure of the transition function components $g_{\boldsymbol{\alpha}}$, $\boldsymbol{w}_{\boldsymbol{\alpha}}$ and known structure of the observation function $f_{\boldsymbol{\alpha}}$: Estimate $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ by minimising an appropriate objective function. If h-step-ahead forecasts are intended, appropriate fit measures are the empirical mean square error $\text{MSE}_h(\boldsymbol{\theta}) = \frac{1}{T-h+1}\sum_{t=0}^{T-h}(Y_{t+h} - Y_{t+h|t})^2$ or the empirical mean absolute percentage error $\text{MAPE}_h(\boldsymbol{\theta}) = \frac{1}{T-h+1}\sum_{t=0}^{T-h}\frac{|Y_{t+h}-Y_{t+h|t}|}{|Y_{t+h}|}$ of the h-step-ahead forecast, where $Y_{t+h|t} = Y_{t+h|t}(\boldsymbol{\theta})$ is the forecast defined in Tables 6.1 to 6.3. The forecast depends on the parameter vector $\boldsymbol{\theta}$ and the associated states estimated by step (i).

(iii) Model specification, i. e. an empirical selection of a specific transition function $g_{\boldsymbol{\alpha}}$, $\boldsymbol{w}_{\boldsymbol{\alpha}}$ and of the observation function $f_{\boldsymbol{\alpha}}$ from a class of alternatives: Among the model alternatives from Tables 6.1 to 6.3, which are cross combinations of a trend type (no trend NT, additive trend AT, additive damped trend ADT, exponential trend ET, exponential damped trend EDT) with a seasonality type (no seasonality NS, additive seasonality AS, multiplicative seasonality MS), select the one which minimises an objective function, e. g. the MSE, the negative likelihood or the Akaike information criterion (AIC, see Section 6.5) of the h-step-ahead forecast. In this context, there should be also taken into account arguments like whether there is seasonality, whether the seasonality amplitude increases with an increasing level of the series (multiplicative seasonality might be more appropriate than additive) or is rather independent of the level and suchlike.

(iv) The final estimates $\widehat{\xi}_1, \ldots, \widehat{\xi}_T$ are obtained by repeating step (i) with the estimate $\widehat{\boldsymbol{\theta}}$ as input. The residual variance σ_ξ^2 is estimated by the empirical variance

$$\widehat{\sigma}^2_{\xi,\text{ES}} = \widehat{\sigma}^2_\xi = \frac{1}{T-1}\sum_{s=1}^{T}(\widehat{\xi}_s - \bar{\xi})^2, \tag{6.21}$$

where $\bar{\xi} = \frac{1}{T}\sum_{s=1}^{T}\xi_s$. Based on an estimate $\widehat{\boldsymbol{\theta}}$, point predictions $Y_{T+h|T}$ of future instances Y_{T+h} are obtained by replacing $\boldsymbol{\theta}$ with the estimate $\widehat{\boldsymbol{\theta}}$ in the equation for $Y_{T+h|T}$ from Tables 6.1 to 6.3. The forecast error variance $\sigma^2_{T+h|T}$ is estimated by inserting $\widehat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ and $\widehat{\sigma}^2_\xi$ for σ^2_ξ in formulas (6.12), (6.14) and (6.19).

If 1-, …, M-steps-ahead forecasts are intended, the above scheme allows to find a separate parameter set for each forecast step h, $h = 1, \ldots, M$. A different approach would be to decide on one forecast step, for example $h = 1$, and produce forecasts $1, \ldots, M$-steps-ahead with the parameter set obtained under this fixed forecast step.

6.5 The Likelihood under the ESCov SSOE Model

With their formulation of the SSOE model, Ord et al. (1997) laid the foundations for maximum likelihood estimation for ES methods. Wang (2006) extended it to ESCov. We revise the computation of the likelihood under ESCov following Ord et al. (1997) and Wang (2006), where the focus is on the one-step-ahead forecast.

Let $Y_1, \dots, Y_T$ be T successive observations from the series. The observations $Y_t, t = 1, 2, \dots, T$, emerge, depending on the chosen model, from certain recursion formulas using the previous states $\boldsymbol{u_t}$, $t = T-1, T-2, \dots$ The states themselves are unknowns as functions of the unknown parameter vector $\boldsymbol{\theta} = (\boldsymbol{\alpha}^\top, \boldsymbol{\beta}^\top)^\top$ (directly of $\boldsymbol{\alpha}$ by means of the state transition equations and indirectly of $\boldsymbol{\beta}$ by means of the errors ξ_t therein) and the initial state vector $\boldsymbol{u}_0$. The initial state vector $\boldsymbol{u}_0$ summarises the behaviour of the time series previous to time point 1, the time point of the first observation Y_1. With the history previous to time point 1 being unknown, the initial state vector $\boldsymbol{u}_0$ is an unknown as well.

Let $\varepsilon_t \sim \text{iid}(0, \sigma^2_\varepsilon)$ and consider Eq. (6.3). Let

$$Y_t \quad = \quad f_{\boldsymbol{\alpha}}(Y_1, \dots, Y_{t-1}, \boldsymbol{u}_0, \boldsymbol{\theta}) + \boldsymbol{\beta}^\top \boldsymbol{x}_t + k_{\boldsymbol{\alpha}}(Y_1, \dots, Y_{t-1}, \boldsymbol{u}_0, \boldsymbol{\theta})\varepsilon_t$$

be the observation equation (6.1) which represents Y_t as a function of the observations $Y_1, \dots, Y_{t-1}$, the initial states $\boldsymbol{u}_0$, and the parameter vector $\boldsymbol{\theta}$, where $f_{\boldsymbol{\alpha}}$ and $k_{\boldsymbol{\alpha}}$ are time-dependent functions. Then we have

$$E[Y_t|\mathcal{H}_{t-1}] \quad = \quad f_{\boldsymbol{\alpha}}(Y_1, \dots, Y_{t-1}, \boldsymbol{u}_0, \boldsymbol{\theta}) + \boldsymbol{\beta}^\top \boldsymbol{x}_t, \tag{6.22}$$

$$\sqrt{V[Y_t|\mathcal{H}_{t-1}]} \quad = \quad k_{\boldsymbol{\alpha}}(Y_1, \dots, Y_{t-1}, \boldsymbol{u}_0, \boldsymbol{\theta})\sigma_\varepsilon. \tag{6.23}$$

Let $f(\cdot)$ denote a density function. Using the rule $\text{P}(A \cap B) = \text{P}(A|B)\text{P}(B)$, the joint probability of $Y_1, \dots, Y_n$ is given by

$$f(Y_1, \dots, Y_T \mid \boldsymbol{x}_1, \dots, \boldsymbol{x}_T, \boldsymbol{u}_0, \boldsymbol{\theta}, \sigma^2_\varepsilon) = \prod_{t=1}^{T} f(Y_t \mid Y_1, \dots, Y_{t-1}, \boldsymbol{x}_1, \dots, \boldsymbol{x}_{t-1}, \boldsymbol{u}_0, \boldsymbol{\theta}, \sigma^2_\varepsilon).$$

Assuming that the ε_t are normally distributed, seeking the maximum likelihood estimates $\boldsymbol{\theta}, \boldsymbol{u}_0$ for the smoothing parameter vector and initial state vector requires with Ord et al. (1997) the maximisation of the conditional likelihood

$$L(\boldsymbol{\theta}, \boldsymbol{u}_0 \mid Y_1, \dots, Y_T) = \frac{1}{(2\pi s^2)^{T/2}} \frac{1}{\left|\prod_{t=1}^{T} k_{\boldsymbol{\alpha}}(Y_1, \dots, Y_{t-1}, \boldsymbol{u}_0, \boldsymbol{\theta})\right|} \cdot \exp\left(-\frac{1}{2}T\right),$$

where $s^2 = \frac{1}{T}\sum_{t=1}^{T}\varepsilon_t^2$.

Consequently, the log-likelihood is given by

$$\ln L(\boldsymbol{\theta}, \boldsymbol{u}_0 \mid Y_1, \ldots, Y_T) = -\frac{T}{2}\ln(2\pi s^2) - \left(\sum_{t=1}^{T} \ln\left(k_{\boldsymbol{\alpha}}(Y_1, \ldots, Y_{t-1}, \boldsymbol{u}_0, \boldsymbol{\theta})\right)\right) - \frac{1}{2}T.$$

In the case of homoscedastic errors $\xi_t = \varepsilon_t$, we have $k_{\boldsymbol{\alpha}}(Y_1, \ldots, Y_{t-1}, \boldsymbol{u}_0, \boldsymbol{\theta}) \equiv 1$, and the conditional likelihood is

$$L(\boldsymbol{\theta}, \boldsymbol{u}_0 \mid Y_1, \ldots, Y_T) = \frac{1}{(2\pi s^2)^{T/2}} \cdot \exp\left(-\frac{1}{2}T\right)$$

and the log-likelihood

$$\ln L(\boldsymbol{\theta}, \boldsymbol{u}_0 \mid Y_1, \ldots, Y_T) = -\frac{T}{2}\ln(2\pi s^2) - \frac{1}{2}T.$$

In the case of multiplicative errors ξ_t, we have $k_{\boldsymbol{\alpha}}(Y_1, \ldots, Y_{t-1}, \boldsymbol{u}_0, \boldsymbol{\theta}) = f_{\boldsymbol{\alpha}}(\boldsymbol{u}_{t-1}) = \boldsymbol{\delta}_{\boldsymbol{\alpha}}^\top \boldsymbol{u}_{t-1}$ in the case of a linear SSOE for ESCov, and $k_{\boldsymbol{\alpha}}(Y_1, \ldots, Y_{t-1}, \boldsymbol{u}_0, \boldsymbol{\theta}) = f_{\boldsymbol{\alpha}}(\boldsymbol{u}_{t-1})$ $= \boldsymbol{\delta}_{\boldsymbol{\alpha},ne}^\top \boldsymbol{u}_{t-1,ne} \prod_{i=1}^{m} e_{i,t-d_i}$ in the case of a partially linear SSOE for ESCov.

A complete maximum likelihood estimation would require to estimate both the vector $\boldsymbol{\theta}$ of the smoothing and covariate parameters and the initial state vector $\boldsymbol{u}_0$. As Ord et al. (1997) mention, the corresponding computational load can be huge. In the case of the ADT-AS model with one seasonality of length 12, for example, one would have to estimate a 4-dimensional parameter vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_{1,3}, \phi)^\top$, a k-dimensional covariate vector $\boldsymbol{\beta}$ as well as an initial state vector $\boldsymbol{u}_0 = (\mu_0, \Delta_0, e_{1,0}, e_{1,-1}, \ldots, e_{1,-d_1+1})^\top$ of dimension $1+1+12$ (where, possibly, the 12th seasonal component can be chosen by a rule of the kind that the 12 components are supposed to add up to 0). Consequently, the amount of unknown parameters to be estimated by the maximum likelihood methods is considerably large. To reduce computational loads in such cases, Ord et al. (1997) suggest to estimate the initial state vector $\boldsymbol{u}_0$ independently of $\boldsymbol{\alpha}$ and treat $\boldsymbol{u}_0$ as a constant in the estimation. For ESCov, this would require to choose $\boldsymbol{u}_0$ not only independently of $\boldsymbol{\alpha}$, but also of $\boldsymbol{\beta}$. Rules how to do so appropriately without interfering too much with the estimation of $\boldsymbol{\theta} = (\boldsymbol{\alpha}^\top, \boldsymbol{\beta}^\top)^\top$ still need to be explored.

The above formulas provide the likelihood for the case of the one-step-ahead forecast only. Optimising the h-step-ahead forecast with $h > 1$ requires the derivation of formulas that take into account the covariance structure of (Y_t). One would have to replace Eqs. (6.22) and (6.23) by $E[Y_t|\mathcal{H}_{t-h}]$ and $\sqrt{V[Y_t|\mathcal{H}_{t-h}]}$, respectively, i. e. by the expectation and standard deviation of the h-step-ahead forecast Y_t given the history until time $t-h$, and follow along the lines of this section while doing appropriate adaptations. For the linear SSOE for ESCov, the formulas for $E[Y_t|\mathcal{H}_{t-h}]$ and $V[Y_t|\mathcal{H}_{t-h}]$ can be taken from Eqs. (6.11)–(6.12) in the case of independent errors, and from Eqs. (6.13)–(6.14) in the case of AR(1) errors. For the partially linear SSOE for ESCov, the respective formulas can be found in Eqs. (6.18)–(6.19).

Popular measures of fit based on the likelihood are the AIC, AICc and BIC:

The *Akaike information criterion (AIC)* dates back to Akaike (1974) and is defined by

$$AIC \quad := \quad 2k - 2\ln L, \tag{6.24}$$

where k denotes the number of estimated parameters and $\ln L$ the log-likelihood.

The *corrected Akaike information criterion AICc* proposed by Sugiura (1978) is a measure that is recommended as a variant of the AIC in particular for small samples. It is defined by

$$AICc \quad := \quad 2k\frac{T}{T-k-1} - 2\ln L. \tag{6.25}$$

Here, k denotes the number of estimated parameters, $\ln L$ the log-likelihood and T the sample size.

A third popular likelihood-based measure is the *Bayesian information criterion (BIC)* by Schwarz (1978). It is defined by

$$BIC \quad := \quad k\ln T - 2\ln L, \tag{6.26}$$

where k denotes the number of estimated parameters, $\ln L$ the log-likelihood and T the sample size.

AIC, AICc and BIC play important roles when it comes to identifying and selecting predictive models, see e. g. Stone (1977), Shmueli (2010) and Diebold (2012). As reverse monotonous transformations of the likelihood, small values of AIC, AICc and BIC are deemed advantageous to larger values in contrast to the likelihood, which is intended to be maximised.

6.6 Empirical Forecasting under ESCov SSOE Models

Assume that the model type, the model parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, the residuals $\xi_1, \ldots, \xi_T$, and the state vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_T$ have been estimated from a time series segment $Y_1, \ldots, Y_T$. Consider further observations $Y_{T+1}, \ldots, Y_t$. Under the determined model with fixed parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, the estimation of state vectors and residuals can be continued with $\boldsymbol{u}_T$

as a starting value as in the step (i) of Section 6.4 to obtain $\boldsymbol{u}_{T+1}, \ldots, \boldsymbol{u}_t$. Based on $\boldsymbol{u}_t$, h-step-ahead MMSE forecasts $Y_{t+h|t}$ are made with the formulas provided by Tables 6.1 to 6.3.

For the estimation of the forecasting variance $\sigma^2_{Y_{t+h}|\mathcal{H}_t}$ we have to distinguish between nonseasonal or additive seasonal and multiplicative seasonal models. For the models NT-NS, AT-NS, ADT-NS, NT-AS, AT-AS, ADT-AS, $\sigma^2_{Y_{t+h}|\mathcal{H}_t}$ is estimated by applying the formula for the conditional variance from Proposition 6.3 for independent residuals and from Proposition 6.4 for AR(1) residuals with the explanatory Table 6.4. For the multiplicative seasonal models NT-MS, AT-MS, ADT-MS, $\sigma^2_{Y_{t+h}|\mathcal{H}_t}$ is estimated by applying the formula for the conditional variance from Proposition 6.6 with the explanatory Table 6.5. Apart from the model parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, the formulas require estimates of the residual variances $\sigma^2_{\xi_{t+1}}, \ldots, \sigma^2_{\xi_{t+h}}$. The estimation of the residual variances $\sigma^2_{\xi_{t+1}}, \ldots, \sigma^2_{\xi_{t+h}}$ requires further assumptions on the residual process (ξ_s), as expressed by the subsequent models of the type *additive residual*, *multiplicative residual* and *AR(1) residual*.

Additive residual model: The residuals (ξ_s) are independent and variance stationary with constant variance $\sigma^2_{\xi_s} = \sigma^2_\xi$. Then σ^2_ξ is estimated from all historical residuals $\widehat{\xi}_1, \ldots, \widehat{\xi}_t$ by the empirical variance $\widehat{\sigma}^2_\xi = \frac{1}{t-1}\sum(\widehat{\xi}_s - \overline{\xi})^2$. The additive residual model is appropriate when the residual in the observation equation (6.1) can be considered independent of the states and covariates.

Multiplicative residual model: The residuals (ξ_s) are linear transformations $\xi_s = f_{\boldsymbol{\alpha}}(\boldsymbol{u}_{s-1})\varepsilon_s$ of independent and variance stationary variables (ε_s) with constant variance $\sigma^2_{\varepsilon_s} = \sigma^2_\varepsilon$. Then ε_s is estimated by $\widehat{\varepsilon}_s = \widehat{\xi}_s / f_{\widehat{\boldsymbol{\alpha}}}(\widehat{\boldsymbol{u}}_{s-1})$. The variance σ^2_ε is estimated from $\widehat{\varepsilon}_1, \ldots, \widehat{\varepsilon}_t$ by the empirical variance $\widehat{\sigma}^2_\varepsilon = \frac{1}{t-1}\sum(\widehat{\varepsilon}_s - \overline{\varepsilon})^2$. The time dependent variances $\sigma^2_{\xi_{t+1}}, \ldots, \sigma^2_{\xi_{t+h}}$ are estimated by

$$\widehat{\sigma}^2_{\xi_{t+\ell}} \quad = \quad f_{\boldsymbol{\alpha}}(\boldsymbol{u}_{t+\ell-1})^2 \cdot \widehat{\sigma}^2_\varepsilon, \quad \ell = 1, \ldots, h. \tag{6.27}$$

Estimates for the quantities $f_{\boldsymbol{\alpha}}(\boldsymbol{u}_t), \ldots, f_{\boldsymbol{\alpha}}(\boldsymbol{u}_{t+h-1})$ are provided by

$$\widehat{f}_{\boldsymbol{\alpha}}(\boldsymbol{u}_{t+\ell-1}) \quad = \quad Y_{t+\ell|t} - \boldsymbol{\beta}^\top \boldsymbol{x}_{t+\ell}, \quad \ell = 1, \ldots, h. \tag{6.28}$$

The multiplicative residual model is appropriate when the residual in the observation equation (6.1) is multiplicative in the sense of Eq. (6.4), see Section 6.3.

AR(1) residual model: The residuals (ξ_s) are an AR(1) process satisfying $\xi_s = \lambda\xi_{s-1} + \varepsilon_s$ with independent and variance stationary errors (ε_s) with $\sigma^2_{\varepsilon_s} = \sigma^2_\varepsilon$. Then ε_s is estimated by $\widehat{\varepsilon}_s = \widehat{\xi}_s - \widehat{\lambda}\widehat{\xi}_{s-1}$. The variance σ^2_ε is estimated from $\widehat{\varepsilon}_1, \ldots, \widehat{\varepsilon}_t$ by the empirical variance $\widehat{\sigma}_{\varepsilon}^2 = \frac{1}{t-1}\sum(\widehat{\varepsilon}_s - \overline{\varepsilon})^2$. The variance σ_{ξ}^2 of the AR(1) process variables ξ_s is estimated by $\widehat{\sigma}_{\xi}^2 = \widehat{\sigma}_{\varepsilon}^2/(1-\lambda^2)$.

The AR(1) residual model is appropriate when the residual in the observation equation Eq. (6.1) is an AR(1) process, i. e. $\xi_t = \lambda\xi_{t-1} + \varepsilon_t$ holds.

Prediction intervals for ESCov are in more detail considered in Chapter 7. In the empirical electricity demand forecasting study of Section 6.9, we use the method called "plug-in" by Ord et al. (1997):

Let the forecast error be normally distributed with mean $Y_{t+h|t}$ and with variance equal to the estimate of $\sigma^2_{Y_{t+h}|\mathcal{H}_t}$. Let $z_{N(0,1)}$ be the one-sided upper $(1+\gamma)/2 \cdot 100\,\%$ quantile of the standard normal distribution with $\gamma \in (0;1)$. Then a symmetric prediction interval of level γ around the point forecast is given by

$$\left(Y_{t+h|t} - z_{N(0,1)}\sigma_{Y_{t+h}|\mathcal{H}_t};\quad Y_{t+h|t} + z_{N(0,1)}\sigma_{Y_{t+h}|\mathcal{H}_t}\right). \tag{6.29}$$

As remarked by Ord et al. (1997), the plug-in method is a widely used but less-than-ideal method. Chapter 7 considers alternatives to the plug-in method.

6.7 Non-constant Covariate Coefficients

Assume, for now, that $k = 1$ in the Definition 6.1 of the SSOE state-space model for ESCov, that is, the covariate vector $x_t \in \mathbb{R}$ is of dimension 1 and there are T observations $x_1, x_2, \dots, x_T$ available. Let $\boldsymbol{x} := (x_1, x_2, \dots, x_T)^\top$. The scheme from Definition 6.1 makes the assumption that the influence of the covariate on the dependent variable Y_t is constantly β over time, i. e. for time point t we have $Y_t = f_{\boldsymbol{\alpha}}(\boldsymbol{u}_{t-1}) + \beta x_t + \xi_t$. Imagine, this is not the case and consider the following two scenarios:

(i) The covariate parameter is dependent on the magnitude of the covariate, i. e. we have

$$\beta = \begin{cases} \beta_1 & \text{if} \quad x_t \le c_1, \\ \beta_2 & \text{if} \quad c_1 < x_t \le c_2, \\ \dots & \\ \beta_l & \text{if} \quad c_{l-1} < x_t, \end{cases}$$

where $c_1, c_2, \dots, c_{l-1} \in \mathbb{R}$.

(ii) The covariate parameter varies in dependence of the period in time, i. e. we have

$$\beta = \begin{cases} \beta_1 & \text{if} \quad t \leq t_1, \\ \beta_2 & \text{if} \quad t_1 < t \leq t_2, \\ \dots & \\ \beta_l & \text{if} \quad t_{l-1} < t, \end{cases}$$

where $t_1, t_2, \dots, t_{l-1} \in \mathbb{N}$.

Then we can fit these models into the framework of the SSOE state-space model for ESCov of the form (6.1)–(6.2) by an appropriate definition of the covariate vector $\boldsymbol{x}_t$:

(i) Let

$$\begin{pmatrix} \mathbb{1}_{x_t \leq c_1} \boldsymbol{x} & \mathbb{1}_{c_1 < x_t \leq c_2} \boldsymbol{x} & \cdots & \mathbb{1}_{c_{l-1} < x_t} \boldsymbol{x} \end{pmatrix} =: \begin{pmatrix} \boldsymbol{x}_1^\top \\ \boldsymbol{x}_2^\top \\ \vdots \\ \boldsymbol{x}_T^\top \end{pmatrix} \in \mathbb{R}^{T \times l} =: \mathbf{X} \quad (6.30)$$

and the covariate vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_l)^\top \in \mathbb{R}^l$. Then scenario (i) can be covered by choosing $Y_t = f_{\boldsymbol{\alpha}}(\boldsymbol{u}_{t-1}) + \boldsymbol{\beta}^\top \boldsymbol{x}_t + \xi_t$ as the observation equation in Definition 6.1, where $\boldsymbol{x}_t$ is the t-th row of the matrix $\mathbf{X}$ defined in (6.30).

(ii) Let

$$\begin{pmatrix} \mathbb{1}_{t \leq t_1} \boldsymbol{x} & \mathbb{1}_{t_1 < t \leq t_2} \boldsymbol{x} & \cdots & \mathbb{1}_{t_{l-1} < t} \boldsymbol{x} \end{pmatrix} =: \begin{pmatrix} \boldsymbol{x}_1^\top \\ \boldsymbol{x}_2^\top \\ \vdots \\ \boldsymbol{x}_T^\top \end{pmatrix} \in \mathbb{R}^{T \times l} =: \mathbf{X} \quad (6.31)$$

and the covariate vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_l)^\top \in \mathbb{R}^l$. Then scenario (ii) can be covered by choosing $Y_t = f_{\boldsymbol{\alpha}}(\boldsymbol{u}_{t-1}) + \boldsymbol{\beta}^\top \boldsymbol{x}_t + \xi_t$ as the observation equation in Definition 6.1, where $\boldsymbol{x}_t$ is the t-th row of the matrix $\mathbf{X}$ defined in (6.31).

In both scenarios, the one-dimensional model is transferred into an l-dimensional model. This means, in particular, that we have to deal with l covariate coefficients $\beta_1, \dots, \beta_l$ instead of only one covariate coefficient β.

Example 6.7 (Example for Scenario (i))**.** *If a certain ambient temperature threshold in an environment or region is exceeded, air conditioning will be more frequently used and the temperature will drive the electricity consumption more heavily than in the case of* *lower temperatures. In this case it might be reasonable to deal with two separate covariate parameters β_1, β_2, where β_1 indicates the temperature influence below the threshold and β_2 the temperature influence greater or equal to the threshold.*

Example 6.8 (Example for Scenario (ii))**.** *During times of a heavy crisis, the influence of a leading indicator on sales during normal economic periods is overridden by an exaggerated reaction of the sales in dependence of the leading indicator. It might be reasonable to work with separate covariate parameters β_1, β_2, where β_1 is the indicator coefficient during normal periods and β_2 the indicator coefficient during crises times.*

The presented scheme can be easily extended to the case of more than one covariate.

6.8 Renormalisation of Seasonal Patterns

In this section, let us consider a seasonal ESCov model. Commonly, the initial values for the seasonal pattern are normalised such that they have a mean of 0 in the case of additive seasonality and a mean of 1 in the case of multiplicative seasonality, respectively. However, in the course of successive smoothing by means of the state transition equation (6.2), the normalisation gets lost. This is due to the fact that only one seasonal factor is updated at each point in time, see Archibald & Koehler (2003). In fact, Hyndman et al. (2008, Chapter 8) showed that the series of means of the seasonal patterns behaves like a random walk.

The loss of a normalised seasonal pattern is disadvantageous with respect to the appealing characteristic of exponential smoothing to deliver a decomposition of a time series into its components level, season, noise and potentially trend and covariate component. As long as the average of one seasonal component is close to 0 (for additive seasonality) or 1 (for multiplicative seasonality), this might not be a problem. However, if the average of a seasonal component is far away from the desired value 0 or 1, respectively, difficulties might occur in interpreting this seasonal component. There are several examples of real data sets, in which this has been found to be the case, as in Makridakis et al. (1982).

For the additive seasonality models, Lawton (1998) observed a bias both in the level and the seasonal components. The errors in estimating level and season were found to be counter-balancing and hence without impact on the forecasts. Therefore, when adjusting both components appropriately at each point in time, we can achieve a normalised seasonal component while maintaining the point forecasts, which is what is recommended by Lawton (1998). The proposal for the renormalisation of the seasonal pattern in the case of additive seasonality, which we revise in this section, goes back to Roberts (1982) and McKenzie (1986). Roberts (1982), McKenzie (1986), Lawton (1998) and Archibald & Koehler (2003) considered ES without covariates. Nevertheless, the fact that their correction is applied to the state transition equations, which are structurally the same for both ES and ESCov, allows to apply their renormalisation equation also to ESCov.

In this section, we consider the ESCov models NT-AS, AT-AS, ADT-AS, NT-MS, AT-MS, ADT-MS with no trend or a linear trend and either additive or multiplicative seasonality.

6.8.1 Renormalisation for Additive Seasonality

The renormalisation scheme described by Roberts (1982) and revised by Archibald & Koehler (2003) can be applied to ES methods without covariates with one additive seasonality and no trend or linear trend, i. e. NT-AS, AT-AS, ADT-AS. The scheme can be extended to multiple seasonality and ESCov. In this section, we present the equations for renormalisation under ESCov only for the model ADT-AS, which includes the models NT-AS and AT-AS as special cases by setting $\alpha_2 = 0, \Delta_0 = 0$ and $\phi = 1$, respectively.

We follow the notation by Archibald & Koehler (2003) and Roberts (1982), which we find convenient to describe the normalisation scheme, and adapt their idea proposed for a single seasonality to multiple seasonalities. For the notation, consider at each time point t the present $e_{i,t}^{(d_i)}$ and the previous $d_i - 1$ smoothed seasonal components $e_{i,t}^{(d_i-1)}, e_{i,t}^{(d_i-2)}, \dots, e_{i,t}^{(1)}$, where d_i is the length of season $i, i = 1, \dots, m$. Here, $e_{i,t}^{(s)}$ with $s = 1, \dots, d_i$ serves as a prediction of the season factor for the future time points $t + s, t+s+d_i, t+s+2d_i, t+s+3d_i, \dots$ made at time t. In the standard smoothing scheme without renormalisation of the season factors and $d_i \geq 2$, the factors $e_{i,t}^{(s)}$ and $e_{i,t-1}^{(s+1)}$ are identical. However, under renormalisation, the prediction for the season factor for time point $t + s$ is updated at each point in time t and therefore usually changes with t. Consequently, we need to differentiate the smoothed season values depending on the point of time t the prediction is made.

The following definition rephrases the usual smoothing equations for model ADT-AS from Table 6.1 in the new notation:

Definition 6.9. *Let $e_{i,t}^{(s)}$ denote the value of the smoothed component of season i corresponding to time point $t + s$ made at time t. The forecast and state transition equations* *for model ADT-AS with multiple additive seasonality from Table 6.1 can be written as*

$$\begin{array}{lcll}
Y_{t+h|t} & = & \mu_t + \sum_{j=1}^{h} \phi^j \Delta_t + \sum_{i=1}^{m} e_{i,t}^{(h)} + \boldsymbol{\beta}^\top \boldsymbol{x}_{t+h}, & \\
\mu_t & = & \mu_{t-1} + \phi \Delta_{t-1} + \alpha_1 \xi_t, & \\
\Delta_t & = & \phi \Delta_{t-1} + \alpha_1 \alpha_2 \xi_t, & \\
e_{i,t}^{(d_i)} & = & e_{i,t-1}^{(1)} + (1-\alpha_1)\alpha_{i,3}\xi_t, & \\
e_{i,t}^{(s)} & = & e_{i,t-1}^{(s+1)}, \quad s = 1, \ldots, d_i - 1. & i\text{=1,}\ldots,m,
\end{array}$$

Similar to the renormalisation scheme for additive seasonality proposed by Roberts (1982) and McKenzie (1986), we denote the renormalised smoothed components by $\widetilde{\mu}_t, \widetilde{\Delta}_t$ and $\widetilde{e}_{i,t}^{(s)}$, $s = 1, \ldots, d_i, i = 1, \ldots, m$, and the corresponding residual by $\widetilde{\xi}_t$. To achieve the renormalisation, a time-dependent term r_t is subtracted from the seasonal component, which in return is added to the level and trend component:

Definition 6.10. *The forecast and state transition equations in the renormalised scheme for model ADT-AS with multiple additive seasonality are given by*

$$\begin{array}{lcll}
\widetilde{Y}_{t+h|t} & = & \widetilde{\mu}_t + \sum_{j=1}^{h} \phi^j \widetilde{\Delta}_t + \sum_{i=1}^{m} \widetilde{e}_{i,t}^{(h)} + \boldsymbol{\beta}^\top \boldsymbol{x}_{t+h}, & \\
r_{i,t} & = & (1-\alpha_1)\alpha_{i,3}\widetilde{\xi}_t / d_i, & \\
\widetilde{\mu}_t & = & \widetilde{\mu}_{t-1} + \phi \widetilde{\Delta}_{t-1} + \alpha_1 \widetilde{\xi}_t + \sum_{i=1}^{m} r_{i,t}, & \\
\widetilde{\Delta}_t & = & \phi \widetilde{\Delta}_{t-1} + \alpha_1 \alpha_2 \widetilde{\xi}_t, & \\
\widetilde{e}_{i,t}^{(d_i)} & = & \widetilde{e}_{i,t-1}^{(1)} + (1-\alpha_1)\alpha_{i,3}\widetilde{\xi}_t - r_{i,t}, & \\
\widetilde{e}_{i,t}^{(s)} & = & \widetilde{e}_{i,t-1}^{(s+1)} - r_{i,t}, \quad s = 1, \ldots, d_i - 1. & i\text{=1,}\ldots,m,
\end{array}$$

Mind that the above formulation holds for additive or multiplicative errors $\widetilde{\xi}_t$. In the additive case, we have $\widetilde{\xi}_t = \widetilde{\varepsilon}_t$, in the multiplicative case $\widetilde{\xi}_t = f_{\boldsymbol{\alpha}}(\widetilde{\boldsymbol{u}}_{t-1})\widetilde{\varepsilon}_t$.

The following proposition shows that the seasonal patterns under the above renormalisation scheme have indeed the desired property of an average, or – equivalently – a sum of 0.

Proposition 6.11 (Renormalisation under Additive Seasonality)**.** *Let the initial season factors under the ESCov model ADT-AS add up to 0, i. e.* $\sum_{k=1}^{d_i} \widetilde{e}_{i,0}^{(k)} = 0$ *for* $i = 1, \ldots, m$*. Let* $R_{i,t} := \frac{1}{d_i}\sum_{k=1}^{d_i} e_{i,t}^{(k)}$ *be the cumulative renormalisation correction factor. Then, under the renormalisation equations from Definition 6.10, the following assertions hold:*

(a) *The season factors of season i add up to 0 at any point in time $t = 0, 1, 2\dots$, i. e.* $\sum_{k=1}^{d_i} \widetilde{e}_{i,t}^{(k)} = 0$.

(b) *The correction factor can be calculated iteratively by $R_{i,t} = R_{i,t-1} + r_{i,t}$.*

(c) *The renormalisation can be achieved at any point in time by*

$$\begin{aligned} \widetilde{\mu}_t &= \mu_t + \sum_{i=1}^{m} R_{i,t}, \\ \widetilde{\Delta}_t &= \Delta_t, \\ \widetilde{e}_{i,t}^{(k)} &= e_{i,t}^{(k)} - R_{i,t}, \quad k = 1, \dots, d_i. \end{aligned}$$

(d) *The predictions are the same under the renormalised and the classical scheme, i. e.* $\widetilde{Y}_{t+h|t} = Y_{t+h|t}$ *for* $t \geq 0, h \geq 1$.

Proof. The proof for renormalisation in ES with a single additive seasonality goes back to Archibald & Koehler (2003). See Appendix 6.A, Section 6.A.4 for the proof with multiple additive seasonalities in ESCov. □

In particular, it follows from Proposition 6.11 (d) that the one-step-ahead forecast errors ξ_t of the classical and the renormalised scheme coincide, that is, we have $\xi_t = \widetilde{\xi}_t$.

By replacing $\boldsymbol{u}_t$, $\boldsymbol{e}_{i,t}$ and $\boldsymbol{w}_{\boldsymbol{\alpha}}$ in Table 6.4 by the subsequently (Eqs. (6.32) and (6.33)) defined quantities $\widetilde{\boldsymbol{u}}_t$, $\widetilde{\boldsymbol{e}}_{i,t}$ and $\widetilde{\boldsymbol{w}}_{\boldsymbol{\alpha}}$, the smoothing scheme for renormalisation in the case of additive seasonality (see Definition 6.10) can be identified as an instance of a linear SSOE state-space model (see Definition 6.2):

$$\widetilde{\boldsymbol{e}}_{i,t} = \begin{pmatrix} \widetilde{e}_{i,t}^{(d_i)} \\ \widetilde{e}_{i,t}^{(d_i-1)} \\ \vdots \\ \widetilde{e}_{i,t}^{(2)} \\ \widetilde{e}_{i,t}^{(1)} \end{pmatrix} = \begin{pmatrix} \widetilde{e}_{i,t-1}^{(1)} + (1-\alpha_1)\alpha_{i,3}\widetilde{\xi}_t \\ \widetilde{e}_{i,t-1}^{(d_i)} \\ \vdots \\ \widetilde{e}_{i,t-1}^{(3)} \\ \widetilde{e}_{i,t-1}^{(2)} \end{pmatrix} \in \mathbb{R}^{d_i}, \tag{6.32}$$

$$\widetilde{\boldsymbol{u}}_{t-1} = \begin{pmatrix} \widetilde{\mu}_{t-1} \\ \widetilde{\Delta}_{t-1} \\ \widetilde{\boldsymbol{e}}_{1,t-1} \\ \widetilde{\boldsymbol{e}}_{2,t-1} \\ \vdots \\ \widetilde{\boldsymbol{e}}_{m,t-1} \end{pmatrix}, \qquad \widetilde{\boldsymbol{w}}_{\boldsymbol{\alpha}} = \begin{pmatrix} \alpha_1 + \sum\limits_{i=1}^{m} \dfrac{(1-\alpha_1)\alpha_{i,3}}{d_i} \\ \alpha_1\alpha_2 \\ (1-\alpha_1)\alpha_{1,3}\cdot\left(1-\frac{1}{d_1}\right) \\ \left.\begin{matrix} -\frac{(1-\alpha_1)\alpha_{1,3}}{d_1} \\ \vdots \\ -\frac{(1-\alpha_1)\alpha_{1,3}}{d_1} \end{matrix}\right\} d_1 - 1 \\ (1-\alpha_1)\alpha_{2,3}\cdot\left(1-\frac{1}{d_2}\right) \\ \left.\begin{matrix} -\frac{(1-\alpha_1)\alpha_{2,3}}{d_2} \\ \vdots \\ -\frac{(1-\alpha_1)\alpha_{2,3}}{d_2} \end{matrix}\right\} d_2 - 1 \\ \vdots \\ (1-\alpha_1)\alpha_{m,3}\cdot\left(1-\frac{1}{d_m}\right) \\ \left.\begin{matrix} -\frac{(1-\alpha_1)\alpha_{m,3}}{d_m} \\ \vdots \\ -\frac{(1-\alpha_1)\alpha_{m,3}}{d_m} \end{matrix}\right\} d_m - 1 \end{pmatrix}. \tag{6.33}$$

Under the above renormalisation scheme, one seasonal pattern of length d_i averages 0 at any given time point t. However, when repeating the renormalisation at every time point anew, the series of the seasonal component “resulting in the very end” is the series $\widetilde{e}_{i,1}^{(d_i)}, \widetilde{e}_{i,2}^{(d_i)}, \ldots, \widetilde{e}_{i,T}^{(d_i)}$ for observed $Y_1, Y_2, \ldots, Y_T$. It does not necessarily fulfil

$$\frac{1}{d_i}\sum_{j=t+1}^{t+d_i} \widetilde{e}_{i,j}^{(d_i)} \quad = \quad 0, \qquad t = 0, \ldots, T - d_i$$

in the case of additive seasonality or

$$\frac{1}{d_i}\sum_{j=t+1}^{t+d_i} \widetilde{e}_{i,j}^{(d_i)} \quad = \quad 1, \qquad t = 0, \ldots, T - d_i$$

in the case of multiplicative seasonality. If that were the case, the seasonal component $\widetilde{e}_{i,t+d_i}^{(d_i)}$ would be equal to $\widetilde{e}_{i,t}^{(d_i)}$ for all $t = 0, 1, \ldots, T - d_i$, which clearly does not hold in general.

6.8.2 Renormalisation for Multiplicative Seasonality

The following renormalisation scheme adapted to multiple multiplicative seasonality and damped linear trend in the ESCov model was formulated by Archibald & Koehler (2003) for a single multiplicative seasonality and an undamped linear trend model for ES. We rephrase the usual state transition equations for the damped trend and multiplicative seasonality ESCov model ADT-MS from Table 6.2 with the notation by Archibald & Koehler (2003). We present the equations for renormalisation under ESCov only for the model ADT-MS, which includes the models NT-MS and AT-MS as special cases by setting $\alpha_2 = 0, \Delta_0 = 0$ and $\phi = 1$, respectively.

Definition 6.12. *Let $e_{i,t}^{(s)}$ denote the value of the smoothed seasonal component of season i corresponding to time point $t+s$ made at time t. The forecast and state transition equations for model ADT-MS with multiple multiplicative season and damped linear trend can be written as*

$$\begin{aligned}
Y_{t+h|t} &= \left(\mu_t + \sum_{j=1}^{h} \phi^j \Delta_t\right) \prod_{i=1}^{m} e_{i,t}^{(h)} + \boldsymbol{\beta}^\top \boldsymbol{x}_{t+h}, \\
\mu_t &= \mu_{t-1} + \phi\Delta_{t-1} + \frac{\alpha_1 \xi_t}{\prod_{i=1}^m e_{i,t-1}^{(1)}}, \\
\Delta_t &= \phi\Delta_{t-1} + \frac{\alpha_1\alpha_2\xi_t}{\prod_{i=1}^m e_{i,t-1}^{(1)}}, \\
e_{i,t}^{(d_i)} &= e_{i,t-1}^{(1)} + \frac{(1-\alpha_1)\alpha_{i,3}\xi_t}{(\mu_{t-1}+\phi\Delta_{t-1})}, \qquad\qquad \textit{i=1,...,m,} \\
e_{i,t}^{(s)} &= e_{i,t-1}^{(s+1)}, \quad s = 1, \ldots, d_i - 1.
\end{aligned}$$

As in the renormalisation scheme for multiplicative seasonality proposed by Archibald & Koehler (2003), we denote the renormalised smoothed components by $\widetilde{\mu}_t, \widetilde{\Delta}_t, \widetilde{e}_{i,t}^{(s)}$ and $\widetilde{\xi}_t$. To achieve the renormalisation, the seasonal component i is devided by a time-dependent term $r_{i,t}$, which in return is multiplied to both the level and trend component:

Definition 6.13. *The forecast and state transition equations in the renormalised scheme for model ADT-MS with multiple multiplicative season are given by*

$$\begin{aligned}
\widetilde{Y}_{t+h|t} &= \left(\widetilde{\mu}_t + \sum_{j=1}^{h} \phi^j \widetilde{\Delta}_t\right) \prod_{i=1}^{m} \widetilde{e}_{i,t}^{(h)} + \boldsymbol{\beta}^\top \boldsymbol{x}_{t+h}, \\
r_{i,t} &= 1 + (1-\alpha_1)\alpha_{i,3} \frac{\widetilde{\xi}_t}{d_i\left(\widetilde{\mu}_{t-1}+\phi\widetilde{\Delta}_{t-1}\right)}, \\
\widetilde{\mu}_t &= \left(\widetilde{\mu}_{t-1} + \phi\widetilde{\Delta}_{t-1} + \frac{\alpha_1\widetilde{\xi}_t}{\prod_{i=1}^m \widetilde{e}_{i,t-1}^{(1)}}\right) \prod_{i=1}^{m} r_{i,t}, \\
\widetilde{\Delta}_t &= \left(\phi\widetilde{\Delta}_{t-1} + \frac{\alpha_1\alpha_2\widetilde{\xi}_t}{\prod_{i=1}^m \widetilde{e}_{i,t-1}^{(1)}}\right) \prod_{i=1}^{m} r_{i,t},
\end{aligned}$$

$$\begin{aligned}
\widetilde{e}_{i,t}^{(d_i)} &= \left(\widetilde{e}_{i,t-1}^{(1)} + \frac{(1-\alpha_1)\alpha_{i,3}\widetilde{\xi}_t}{(\widetilde{\mu}_{t-1}+\phi\widetilde{\Delta}_{t-1})}\right)/r_{i,t}, \qquad \textit{i=1,}\dots\textit{,m,}\\
\widetilde{e}_{i,t}^{(s)} &= \widetilde{e}_{i,t-1}^{(s+1)}/r_{i,t}, \quad s=1,\dots,d_i-1.
\end{aligned}$$

Proposition 6.14 summarises several important results with respect to the renormalisation of the multiplicative seasonality ESCov model.

Proposition 6.14 (Renormalisation under Multiplicative Seasonality)**.** *Let the initial season factors under the ESCov model ADT-MS have sum d_i, i. e., $\sum_{k=1}^{d_i}\widetilde{e}_{i,0}^{(k)} = d_i$ for $i=1,\dots,m$. Let $R_{i,t} := \frac{1}{d_i}\sum_{k=1}^{d_i} e_{i,t}^{(k)}$ be the cumulative renormalisation correction factor. Then, under the renormalisation equations from Definition 6.13, the following assertions hold:*

(a) The season factors of season i have sum d_i at any point in time $t=0,1,2\dots$, i. e. $\sum_{k=1}^{d_i}\widetilde{e}_{i,t}^{(k)} = d_i$.

(b) In the case of $m=1$, the correction factor can be calculated iteratively by $R_{1,t} = R_{1,t-1}r_{1,t}$.

(c) In the case of $m=1$, the renormalisation can be achieved at any point in time by

$$\begin{aligned}
\widetilde{\mu}_t &= \mu_t R_{1,t}\\
\widetilde{\Delta}_t &= \Delta_t R_{1,t}\\
\widetilde{e}_{1,t}^{(k)} &= e_{1,t}^{(k)}/R_{1,t}, \quad k=1,\dots,d_i.
\end{aligned}$$

(d) In the case of $m=1$, the predictions are the same under the renormalised and the classical scheme, i. e. $\widetilde{Y}_{t+h|t} = Y_{t+h|t}$ for $t\geq 0, h\geq 1$.

Proof. The proof for renormalisation in ES with a single multiplicative seasonality goes back to Archibald & Koehler (2003). See Appendix 6.A, Section 6.A.5 for the proof with multiple multiplicative seasonalities in ESCov. □

Under the correction for the multiple multiplicative seasonal model ADT-MS as proposed in Definition 6.13, assertions (b)–(d) cannot easily be extended to the multiple seasonality case. In particular, if the number of seasonalities m is ≥ 2, the predictions under the renormalised scheme are not equal to the predictions under the classical scheme.

As stated by Archibald & Koehler (2003), the point forecast and hence also the residuals are the same for both the renormalised and standard method. Nevertheless, in order to avoid potentially inadequate interpretation of the components, Archibald & Koehler

(2003) recommend 1) to renormalise the components in every time period, and 2) to examine only normalised components. Due to its efficiency, Gardner (2006) also favours renormalisation, especially in the case of multiplicative seasonality.

6.9 Application in Electricity Load Forecasting in Italy

6.9.1 Introduction to the Load Forecasting Study

The topic of forecasting electricity load is attracting rapidly increasing interest of business management, politics and social discourse. The interest is driven by two factors: the deregulation of the electricity market and the reorganisation of electricity production. Starting with the privatisation of the electricity sector in Chile in the 1980s, many developed countries have deregulated and liberalised their originally monopolistic electricity markets in the 1990s and 2000s, particularly in the European Union and in North America. Concerns on the safety of nuclear power production and the climate change have initiated a move towards decentralised power production from wind, bio or solar energy.

Different spheres of interest require different forecasting horizons and different resolutions (half-hourly, hourly, daily, weekly or annual) of the forecasting target. In that sense, electricity load forecasting is sometimes classified into short-term, medium-term and long-term forecasts, see e. g. Weron (2006) or Fan & Hyndman (2012). Political decision makers are particularly interested in long-term forecasts on the macro level, e. g. a ten year projection of national electricity demand. Power distributors are interested in short-term (e. g. a few hours or a few days ahead) and medium-term forecasts (e. g. a few weeks or months ahead), often with spatially highly resolved prediction targets.

Energy companies use electricity load forecasting for the scheduling of their power systems. The need for reliable load forecasts arises from a) the lack of (cheap) storing capacities and the fact that an overflow in electricity can be discharged only at additional costs, see Cho et al. (2013), and b) that an undersupply with electricity requires last-minute acquisition of electricity at very expensive prices. Cho et al. (2013) point out that already a small improvement in the load forecasting can cause a considerable reduction in costs.

In an increasingly decentralised power production environment it is important to get the appropriate amount of electricity to the right place in the right time. By deregulation and liberalisation, electricity has become a trading commodity. Energy exchanges like the

EEX (European Energy Exchange) in Leipzig (Germany) or the Italian Power Exchange (IPEX) provide markets for trading of long-term futures and short-term spot markets for intra-day and one-day-ahead trading. The IPEX (now operated by the GME) took off on 31 March 2004 and its spot market comprises the day-ahead-market and the intra-day market of hourly energy blocks: The former opens nine days before the day of delivery and closes the day before, whereas the latter opens the day before and closes at 11:45 a.m. on the day of delivery. Therefore, energy vendors participating in this market need short-term forecasts from one hour ahead up to ten days ahead of the energy requested by their customers. In this work we consider the vendor's point of view.

The demand from industry has been stimulating the research on load forecasting methodology. Basically, two approaches can be distinguished: i) methodology from the area of informatics and machine learning, e. g. classification techniques like SVM (support vector machines), neural networks, expert systems; ii) statistical methodology, mainly from statistical time series analysis. A considerable amount of literature has been appearing from both classes, see Weron (2006) for a concise review. The essential methodological challenge for either approach is the ability to relate electricity load to exogenous factors or covariates, like meteorological variables (temperature, humidity, cloud shading), calendar and time data (time of year, month, weekday, hour), electricity prices or customer clusters.

Statistical approaches have been concentrating both on calendar and time, which can be modelled by seasonal components. The preferred modelling approaches are regression, the Box-Jenkins methodology like AR, ARMA or ARIMA models and seasonal derivatives thereof, like SARIMA, and state-space models. Short-term predictions receive more attention than long-term forecasts. Weron (2006) provides a survey of statistical studies until 2006. Some later references in the short-term prediction area without meteorological covariates are Diego J. Pedregal (2010) using state-space modelling, Chakhchoukh et al. (2009, 2011) and Soares & Medeiros (2008) applying SARIMA.

In the field of electricity load forecasting, ES has played an important role especially through various studies of Taylor, see Taylor (2003b, 2010), Taylor et al. (2006) and Taylor & McSharry (2007). An earlier reference of ES applied in the context of short-term load forecasting is Park et al. (1991), where the method is combined with an autoregressive model and a recursive least squares method.

Calendar and time variables are the main influencing factors, especially for a short-term horizon, which call for methods including seasonal or periodic terms. In this context, electricity load forecasting – depending on the nature of the data – often requires to deal with multiple seasonalities: Electricity demand can be subject to an annual seasonality as well as a weekly seasonality. Half-hourly or hourly data additionally involve an intra-day seasonality. For short-term forecasts, intra-day and intra-week seasonality are predominant. To deal with multiple overlaying seasonalities, different approaches are presented in the literature. The intra-day seasonality, which is present when dealing with hourly or half-hourly data, is often considered complicated to model. Consequently, many authors choose to estimate 24 or 48 separate models for each hour or half-hour of the day, as for example Fan & Hyndman (2012), Dordonnat et al. (2012), Hinman & Hickey (2009), Soares & Medeiros (2005) or Ramanathan et al. (1997). Empirical studies like Taylor (2003b, 2010) show that treating multiple seasonalities simultaneously are competitive approaches. Taylor (2003b) applied double seasonal exponential smoothing as well as double seasonal ARIMA. Taylor (2010) accounted also for annual seasonality by triple seasonal methods. As further calendar-related factors, special days and (public) holidays often have an influence on the electricity demand and have to be taken into account.

Studies that have highlighted within the statistical approach the importance of meteorological covariates for short-term forecasting, besides calendar variables, are, for example, Papalexopoulos & Hesterberg (1990), Ramanathan et al. (1997), Dordonnat et al. (2008) and Hinman & Hickey (2009). In particular, Hinman & Hickey (2009) review the literature in this area and extract the information that temperature affects the load in a nonlinear way, a fact also noted by Dordonnat et al. (2008). More recent contributions, such as Fan & Hyndman (2012) and Ba et al. (2012), add semi-parametric additive models to the picture. In the past, the poor precision of meteorological forecasts may have discouraged statisticians from studying electricity load models under meteorological covariates on a broad scale. However, meteorologists have markedly increased the accuracy of their forecasts, particularly for a short-term and mid-term perspective, in the last twenty years. Bunn (1982) considers meteorological data as substantial for mid-term and long-term forecasting, but as dispensable for a short-term horizon, essentially because of two arguments: i) Difficulties in guaranteeing the regular input of weather variables are potential threats for the stability of the forecasting procedure. ii) Adaptive forecasting procedures without covariates may be able to induce by themselves the effect of weather changes, so that the inclusion of covariates provides no extra prediction accuracy. However, argument i) has lost its impact in view of modern data transmission technology which guarantees the rapid and reliable transport of large amounts of data. Argument ii) may have been true 30 years ago, but it has lost its plausibility in view of the impressive progress meteorology has made. In particular, the high variability of weather conditions in Western Europe rather suggests that precise weather forecasts may improve upon load forecasting also in the short-term context.

In general, temperature forecasts provided by meteorological institutes have become remarkably exact. The remaining uncertainty will not affect the accuracy of load forecasts substantially. Khotanzad (2007) study the effect of three temperature forecasts on the economic benefits of 24-hours-ahead load forecasts: (i) the persistence forecast, i. e. use today's temperature as a forecast for tomorrow's temperature; (ii) a forecast obtained from a National Weather Service (NWS) computer model; and (iii) the perfect forecast, that is, use tomorrow's exact temperature. Clearly, the perfect forecast can only be evaluated retrospectively. Khotanzad (2007) conclude that "about 70 % of the total potential benefits of a perfect forecast versus a persistence forecast have already been realized by using the current NWS forecast". There is a huge gain in using a scientifically sound temperature forecast instead of a naïve one, but a relatively small difference between using the meteorologist forecast and the hypothetical perfect forecast.

The subsequent study has the objectives to explore the use of meteorological covariates in short-term load forecasting and to gain empirical experience with the novel time series analysis technique of ESCov introduced in Sections 6.2 and 6.3. In Section 6.9.2, the fitting and forecasting techniques provided by Sections 6.4 and 6.6 are applied to hourly electricity consumption data of the customers of an energy vendor in some provinces of Emilia Romagna, an Italian region. Unlike Dordonnat et al. (2008), Hinman & Hickey (2009) and Fan & Hyndman (2012), who consider a very large number of customers (the national hourly load of France, about 3.8 million customers in Northern Illinois and the whole Victoria, Australia, respectively), we deal with a much smaller aggregate of 103 customers overall. In consequence, an unusual behaviour of a few customers may have a significant effect on the aggregate. Section 6.9.3 applies the theory of renormalising seasonality patterns from Section 6.8 to the electricity load data. The consequences of the load forecasting study for the development of a forecasting system are pointed out in Section 6.9.4.

6.9.2 Model Estimation for Electricity Demand Time Series

We apply the ESCov SSOE models introduced in Section 6.3 to a dataset representing the total amount of electricity sold by a single energy vendor in Italy. The dataset consists of 103 single series of hourly electricity demand, each of which comes from a certain point of delivery (PoD). A PoD is defined as any point where there is a power meter measuring the electricity load of one or more customers. The data set starts in 1 January 2005 and ends in 30 January 2007 and thus covers a total of 760 days and 18240 hourly observations. The customers are consistently either large consumers or small businesses, but no private households, since the latter were allowed to access the free energy market in Italy not earlier than 1 July 2007. In the course of time, the vendor gradually acquired new or lost customers, which leaves many of the series incomplete, especially in the beginning of the time period at hand. The PoDs in the dataset mainly come from Emilia Romagna, only four are located in the region Piedmont, both in Northern Italy. To be able to buy the appropriate amount of electricity from the energy market at a certain point of time, the vendor needs forecasts of the electricity demand of the customers to avoid both an overflow and a shortage on electricity. The vendor is interested in the total demand of his portfolio, which determines the amount of electricity to be bought from the energy market, rather than the individual PoD's electricity demand.

Available covariates for predicting the electricity load are

i) calendar variables indicating bank holidays, regional holidays, bridge days and special holiday seasons (indicator variables) and

ii) the realised meteorological variables temperature (TMP), humidity (HMDTY), cloud cover (CLD), wind speed (WND) (all on an hourly basis).

The meteorological data is available for Emilia Romagna only on the spatial resolution of provinces, whereas there is no meteorological data available for Piedmont. Hence we discard the analysis of the four PoDs in Piedmont and stick to the remaining 99 PoDs in Emilia Romagna, as we would like to investigate, among other things, the value added by an inclusion of weather information. Since an analysis and aggregation of the 99 individual, partly very irregular, PoD series is expected to come with a high variance, we aggregate these 99 PoDs to obtain a more stable load series, especially in view of the vendor's objective of a prediction on an aggregated level. Weather information is not perfectly complete. Missing hourly observations are replaced by interpolation between the respective hourly observations of adjacent days. The PoDs to be investigated are situated in the following provinces of Emilia Romagna, for all of which meteorological data is basically available: Bologna (BO, 24 PoDs), Forlì-Cesena (FC, 15 PoDs), Ferrara (FE, 11 PoDs), Modena (MO, 14 PoDs), Ravenna (RA, 12 PoDs), Reggio Emilia (RE, 3 PoDs) and Rimini (RN, 20 PoDs). To obtain weather information for the whole region Emilia Romagna, we take the average of the weather variables of the seven provinces, so as to forecast the total electricity load in Emilia Romagna of the vendor. A plot of the total hourly electricity demand is provided in Fig. 6.1.

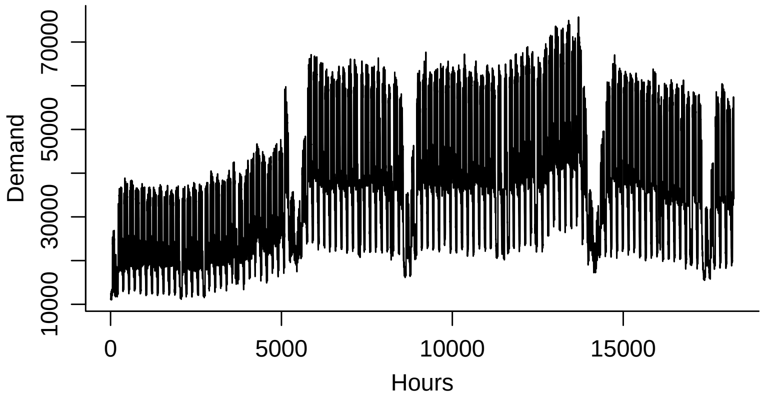


Figure 6.1: Hourly electricity demand (in kWh) from 1 January 2005 to 30 January 2007.

The meteorological variables humidity, cloud cover and wind speed do not show any visible influence on the electricity consumption. Therefore we only consider the temperature as a meteorological covariate to forecast the electricity load. For the estimation we use the actually realised meteorological data, which we assume to be well predictable by weather forecasts. Actual weather forecasts for the analysed time period were not available, which is why we have to concentrate on the analysis using the true values.

Figure 6.2, which shows the hourly temperature (in °C) plotted against the hourly deseasonalised electricity demand, reveals the nonlinear relationship between the two variables. A U-shaped relationship between the electricity demand and the temperature, sometimes observed in some European countries (see e. g. Dordonnat et al. (2008)) where electrical heating causes an increase in the electricity load when colder temperatures occur, is not true for the Italian data at hand. In our case, the relationship between the temperature and the electricity load can be well described as piecewise linear with a cut in the regression line at a temperature around 15°C, similar to the study of Dordonnat et al. (2008). In Dordonnat et al.'s (2008) study, however, a strong decrease in electricity load is observed for temperatures increasing up to 15°C, and a slight increase for temperatures above 15°C, whereas our data show a behaviour of approximately the opposite way: A clear increase in the electricity load, mainly tracing back to the widespread use of air conditioning in Italy during the summer, is visible as soon as a threshold of approximately 15°C in the temperature is exceeded, whereas it is approximately constant below that threshold. The relationship between the temperature and the electricity demand is known to be of nonlinear nature, which was observed, for example, by Fan & Hyndman

(2012) analysing the Australian National Electricity Market. The chosen threshold of 15° C appears to be reasonable in the light of the data, also considering that there are no domestic customers, who would perhaps switch on air conditioning at higher temperatures. Some customers are hospitals, so we can assume that there are cooling appliances other than air conditioning that start operating more often at warm air temperatures.

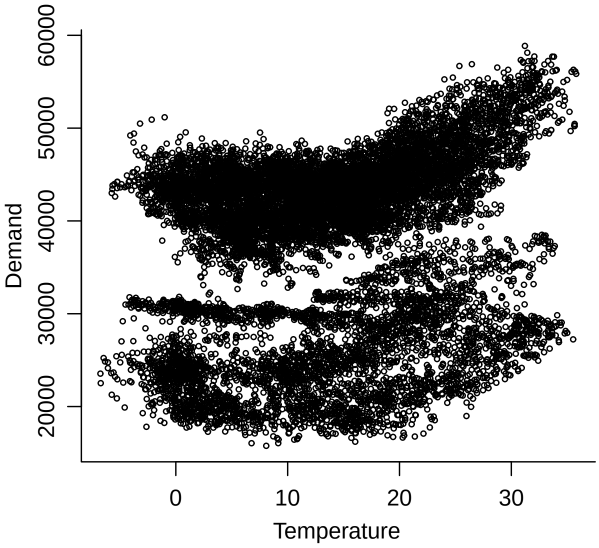


Figure 6.2: Hourly temperature (in °C) plotted against the hourly deseasonalised electricity demand (in kWh) from 1 January 2005 to 30 January 2007.

Because only two years of data are available, estimating the effect of days and periods with unusual behaviour is not possible. Periods showing a behaviour which is remarkably different from "normal" weeks are the holiday period in August as well as the two weeks around Christmas and New Year. To avoid a deteriorating influence of these untypical periods on estimation quality, the respective weeks were deleted from the sample period. Furthermore, special days, like bank holidays and bridge days, destroy the weekly seasonality pattern in the data and are left out for the estimation of the parameters by deleting the whole week in which they occur. The same is done for the week around 25 November 2005, when a national strike took place. Not deleted are the weeks in which a regional holiday in one of the seven provinces occurred. Since a regional holiday occurred in only one province at a time and since the modelled electricity consumption involves the total of all seven provinces, the effect of a regional holiday is minor, also in view of similar disturbances caused by immigration and emigration of customers, which are also present in the data. By this approach of handling bank holidays, the relations between successive days or hours, respectively, are lost. However, the eventual drawbacks from this method are negligible compared to the effect that an inclusion of the exceptional weeks would have on the estimation of the parameters. If we had a longer series of data, special days could be included in the model by introducing appropriate dummy variables. For some holidays, this can go as far as having a term for every hour in that day, as seen in Dordonnat et al. (2008).

For the hourly electricity data displayed in Fig. 6.1, we consider a daily as well as a weekly seasonality under the ESCov model from Tables 6.1 and 6.2. The model is estimated in the subsequent section.

6.9.2.1 Model Estimation

We apply the ESCov model with the temperature covariate to the dataset of hourly electricity load data. To be able to evaluate the performance of the ESCov model under the use of the covariate temperature, we choose July 2006 for a post-sample evaluation of our model. This is a period in which a change in the temperature has an influence on the electricity load and which is not disturbed by holidays. Therefore the parameter estimation period terminates by 30 June 2006. Since the customer portfolio of the energy vendor undergoes considerable structural changes in the first seven months of 2005, which are mainly caused by acquiring new customers, we choose September 2005 as the beginning of the estimation period. This ten months period includes several months in which the temperature has a noticeable effect on the electricity load, so that we are able to estimate its influence. Having omitted the special weeks containing holidays as well as the Christmas period from the dataset, we end up with a period of 233 days and $233 \cdot 24 = 5592$ hours for the estimation of the parameters. A plot of the now comparably regular series from 1 September 2005 to 25 July 2006 is provided in Fig. 6.3.

The relationship between temperature and load of the reduced data set is unchanged, and if we plotted the load against the temperature it would appear very similar to Fig. 6.2.

Using an idea of Dordonnat et al. (2008), the piecewise linear relationship between the temperature and the electricity consumption visible in the plot is transferred to a linear relationship by imposing the transformation

$$\widetilde{x}_t := \begin{cases} 0 & \text{if } x_t \leq 15, \\ x_t - 15 & \text{if } x_t > 15, \end{cases}$$

on the covariate, where 15°C is the threshold below which the electricity consumption is deemed unaffected by the temperature.

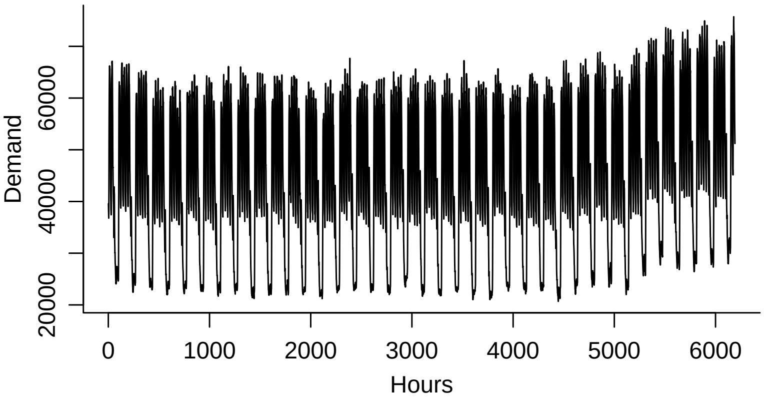


Figure 6.3: Hourly electricity demand (in kWh) from 1 September 2005 to 25 July 2006 with special weeks removed.

The considered vendor's customer portfolio is relatively small. So immigration and emigration of customers cause considerable changes in the electricity load series. Nevertheless, we apply the method of ESCov to these volatile series as they happen in reality, and exploit the ability of exponential smoothing to adapt quickly to changes in the behaviour of a time series.

We evaluate the precision of the forecasts with the help of the mean absolute percentage error (MAPE) and the mean square error (MSE) defined as follows:

Definition 6.15 (Accuracy Measures)**.** *Let $Y_{T+1|T+1-h}, \dots, Y_{T+N|T+N-h}$ be the h-step-ahead forecasts of the electricity load for time points $T+1, \dots, T+N$ and $Y_{T+1}, \dots, Y_{T+N}$ the true observations. The* mean absolute percentage error (MAPE) *of the forecast is defined as*

$$MAPE := \frac{1}{N} \sum_{i=1}^{N} \left| \frac{Y_{T+i} - Y_{T+i|T+i-h}}{Y_{T+i}} \right| \cdot 100\,\%.$$

The mean square error (MSE) *is defined as*

$$MSE := \frac{1}{N} \sum_{i=1}^{N} \left(Y_{T+i} - Y_{T+i|T+i-h} \right)^2$$

and the root mean square error (RMSE) *is defined as $RMSE := \sqrt{MSE}$.*

In the specification of the ESCov model, two overlaying seasonalities have to be taken into account: a within-week seasonality of length $d_1 = 24 \cdot 7 = 168$ hours and a within-day seasonality of length $d_2 = 24$ hours. We have to acknowledge that a yearly seasonality is also present, but we had to give up on modelling the intra-year seasonality altogether due to the lack of an appropriately long series of data. To specify the appropriate type of trend and seasonality, we compare the MSE of the models described in Tables 6.1 and 6.2 for a forecast step of $h = 24$, i. e. one day. We discard the exponential trend model from Table 6.3 because the observation series seems to be rather constant in level than of exponential growth. A model with additive seasonality (AS) turns out to be more appropriate than a model with multiplicative seasonality (MS). The trend type ADT is the broadest class of the trend models listed in Table 6.1 since it allows both a damped (for $\phi < 1$) and a non-damped trend (for $\phi = 1$). Therefore we deem model ADT-AS the most appropriate one for the data.

We apply model ADT-AS with double additive seasonality and possibly damped linear trend to the hourly electricity demand with and without using the hourly temperature as covariate. We estimate the smoothing and covariate parameters minimising the MSE of the one-hour-ahead through 24-hours-ahead forecasts, i. e. $h = 1, \dots, 24$, the 72-hours-ahead forecasts, i. e. $h = 72$, which is a 3-days-ahead forecast, and the 168-hours-ahead forecasts, i. e. one-week-ahead forecast, and the 240-hours-ahead forecasts, i. e. ten days ahead. The optimisation is done in R using the function `optim` under the method `"L-BFGS-B"`. This is a limited-memory modification of the BFGS quasi-Newton algorithm, which allows box constraints, see the `stats`-package in R (R Core Team 2014). The starting values $\boldsymbol{u}_0$ for the level, trend increment and seasonal factors are for both the ESCov method as well as the ES method chosen as follows: The initial trend increment Δ_0 is chosen as the mean of the observations in the second season (observations 169–336) minus the mean of the observations in the first season (observations 1–168) divided by $7 \cdot 24 \cdot \phi$. The initial level μ_0 is chosen as the mean of the first 168 observations minus $\Delta_0 \cdot 84$. The initial within-day season factors $e_{1,1,0}, \dots, e_{1,24,0}$ are estimated as the seasonal factors from a seasonal adjustment procedure of type Census I applied to the first 48 observations. The initial within-week season factors $e_{2,1,0}, \dots, e_{2,168,0}$ are estimated as the seasonal factors from a seasonal adjustment procedure of type Census I applied to the first 336 observations.

The estimated parameters, the MSE and the MAPE appear in Tables 6.6 and 6.7. For comparison we also report a naïve forecast, which we obtained by using the corresponding hour of the week, i. e. 168 hours, before (to compare with the results for $h = 1, 24, 72, 168$) or two weeks, i. e. 336 hours, before (to compare with the results for $h = 240$).

An important finding from the results is that varying lead times come with different sets of optimal parameters. In the case of the very short-term one-hour-ahead forecast, the smoothing parameter for the level is the rather high $\alpha_1 = 0.609$, which corresponds to a quick adaption of the forecast to level changes. This property gets less important for increasing lead times, where the level parameter is at its minimum. The trend parameter, which is almost always at its minimum, suggests that a trend component is not important for forecasts up to 10 days ahead. Consequently, the trend and therefore also the dampening of the trend seem to be of little added value. These findings apply to both ESCov and ES. In the presence of covariates, we find that the temperature coefficient is comparably low for lead time one-hour-ahead and increases with increasing lead time. Consequently, the improvement in the fit obtained by including the temperature as a covariate is negligible for one-hour-ahead forecasts. This is plausible since the weather does not change considerably from one hour to the next. Yet, the temperature coefficient increases with increasing lead time and leads to improvements in the fit of ESCov in comparison to ES. The increasing importance of the temperature covariate for bigger lead times in the ESCov model is compensated by a larger parameter for the intra-day seasonality in the case of ES. This demonstrates the ability of ES to adapt quickly to changes in the behaviour of a time series.

Figure 6.4 shows autocorrelation plots of the one-step-ahead forecast errors $\xi_1, \ldots, \xi_T$ up to lag 96 obtained for model ADT-AS for lead times $h = 1$ and $h = 24$. For $h = 1$, a pattern of autocorrelations alternating in sign is visible. In particular, the first-order autocorrelation of about 0.21 suggests to account for the autocorrelation of the errors in the model. For lead time $h = 24$, the autocorrelations are considerable and constantly decrease up to lag 72 while temporarily changing sign from lag 48 onwards. In general, while optimising the h-step forecasts with $h > 1$, autocorrelation does not decay as quickly as desired under a model assuming independent residuals. To address this problem, we fit the model ADT-AS under an inclusion of an AR(1) process on the errors, see Section 6.3.1. The estimated smoothing parameters, the covariate coefficient and the autoregressive parameter $\lambda \in [0; 1]$ for lead time $h = 1$ can be taken from Table 6.8. For lead times $h = 24, 72, 168, 240$, the autocorrelation parameter λ was estimated to 0 and hence brought no improvement, which is why we dispense with a report of the estimation results. The SSOE model under AR(1) errors has not been investigated empirically for other lead times.

The inclusion of an autoregressive process of order 1 on the errors for lead time $h = 1$ brings an improvement in the result of the fit and leads to smoothing parameters differing from those in Tables 6.6 and 6.7. Without the AR(1) process on the errors, higher parameters for the level as well as the two seasonalities seem to compensate for the autocorrelation in the errors. A similar result was observed by Taylor (2003b), who

Table 6.6: Estimated parameters of model with additive, potentially damped trend and additive season (ADT-AS) for exponential smoothing (ES) and exponential smoothing with covariates (ESCov) for lead times $1, \ldots, 14$ estimated by the hourly data for seven provinces in Emilia-Romagna aggregated. Covariate: average hourly temperature of the seven provinces of Emilia-Romagna. α_1 = parameter for level, α_2 = parameter for trend increment, ϕ = trend damping parameter, $\alpha_{1,3}$ = within-week seasonality, $\alpha_{2,3}$ = within-day seasonality, β = temperature coefficient.

h	model	α_1	α_2	ϕ	$\alpha_{1,3}$	$\alpha_{2,3}$	β	In-sample MSE (kWh^2)	In-sample MAPE (%)
1	ESCov	0.609	0.001	0.663	0.257	0.300	48.4	$6.38 \cdot 10^5$	1.42
1	ES	0.609	0.001	0.634	0.258	0.304	–	$6.40 \cdot 10^5$	1.43
2	ESCov	0.391	0.001	0.708	0.194	0.173	45.0	$9.88 \cdot 10^5$	1.80
2	ES	0.394	0.001	0.691	0.197	0.181	–	$9.91 \cdot 10^5$	1.81
3	ESCov	0.276	0.001	0.729	0.168	0.154	59.0	$1.21 \cdot 10^6$	2.03
3	ES	0.284	0.001	0.708	0.172	0.160	–	$1.22 \cdot 10^6$	2.03
4	ESCov	0.166	0.001	0.762	0.150	0.121	78.1	$1.39 \cdot 10^6$	2.17
4	ES	0.175	0.001	0.727	0.152	0.130	–	$1.40 \cdot 10^6$	2.18
5	ESCov	0.098	0.001	0.812	0.157	0.098	94.6	$1.50 \cdot 10^5$	2.23
5	ES	0.104	0.001	0.759	0.154	0.107	–	$1.52 \cdot 10^6$	2.25
6	ESCov	0.074	0.001	0.835	0.167	0.090	103.7	$1.57 \cdot 10^5$	2.28
6	ES	0.077	0.001	0.769	0.163	0.098	–	$1.60 \cdot 10^6$	2.30
7	ESCov	0.063	0.001	0.849	0.168	0.087	109.0	$1.64 \cdot 10^5$	2.33
7	ES	0.065	0.001	0.771	0.165	0.095	–	$1.66 \cdot 10^6$	2.35
8	ESCov	0.056	0.001	0.861	0.165	0.085	112.6	$1.69 \cdot 10^5$	2.37
8	ES	0.057	0.001	0.780	0.163	0.093	–	$1.72 \cdot 10^6$	2.39
9	ESCov	0.051	0.001	0.866	0.161	0.084	116.1	$1.73 \cdot 10^5$	2.40
9	ES	0.052	0.001	0.784	0.159	0.092	–	$1.76 \cdot 10^6$	2.42
10	ESCov	0.048	0.001	0.868	0.158	0.083	120.4	$1.76 \cdot 10^6$	2.43
10	ES	0.049	0.001	0.768	0.156	0.091	–	$1.79 \cdot 10^6$	2.45
11	ESCov	0.045	0.001	0.871	0.154	0.082	123.9	$1.79 \cdot 10^6$	2.45
11	ES	0.046	0.001	0.765	0.154	0.090	–	$1.83 \cdot 10^6$	2.47
12	ESCov	0.042	0.001	0.873	0.151	0.081	127.4	$1.82 \cdot 10^6$	2.47
12	ES	0.044	0.001	0.761	0.151	0.090	–	$1.85 \cdot 10^6$	2.49
13	ESCov	0.040	0.001	0.875	0.148	0.081	130.4	$1.84 \cdot 10^6$	2.49
13	ES	0.042	0.001	0.768	0.149	0.089	–	$1.88 \cdot 10^6$	2.51
14	ESCov	0.037	0.001	0.877	0.145	0.082	134.2	$1.87 \cdot 10^6$	2.50
14	ES	0.039	0.001	0.762	0.146	0.090	–	$1.90 \cdot 10^6$	2.53

Table 6.7: Estimated parameters of model with additive, potentially damped trend and additive season (ADT-AS) for exponential smoothing (ES) and exponential smoothing with covariates (ESCov) for lead times $15, \ldots, 24, 72, 168, 240$ estimated by the hourly data for seven provinces in Emilia-Romagna aggregated. Covariate: average hourly temperature of the seven provinces of Emilia-Romagna. α_1 = parameter for level, α_2 = parameter for trend increment, ϕ = trend damping parameter, $\alpha_{1,3}$ = within-week seasonality, $\alpha_{2,3}$ = within-day seasonality, β = temperature coefficient.

h	model	α_1	α_2	ϕ	$\alpha_{1,3}$	$\alpha_{2,3}$	β	In-sample MSE (kWh^2)	In-sample MAPE (%)
15	ESCov	0.034	0.001	0.879	0.140	0.083	139.1	$1.89 \cdot 10^6$	2.52
15	ES	0.037	0.001	0.740	0.143	0.091	–	$1.93 \cdot 10^6$	2.55
16	ESCov	0.030	0.001	0.884	0.136	0.087	143.2	$1.91 \cdot 10^6$	2.53
16	ES	0.035	0.001	0.742	0.140	0.093	–	$1.95 \cdot 10^6$	2.56
17	ESCov	0.025	0.001	0.896	0.131	0.094	165.4	$1.92 \cdot 10^6$	2.55
17	ES	0.032	0.001	0.710	0.136	0.098	–	$1.97 \cdot 10^6$	2.57
18	ESCov	0.021	0.001	0.904	0.128	0.104	154.8	$1.94 \cdot 10^6$	2.55
18	ES	0.028	0.001	0.708	0.132	0.106	–	$1.99 \cdot 10^6$	2.58
19	ESCov	0.017	0.001	0.919	0.125	0.120	158.0	$1.95 \cdot 10^6$	2.55
19	ES	0.024	0.001	0.750	0.127	0.121	–	$2.00 \cdot 10^6$	2.59
20	ESCov	0.010	0.124	0.779	0.122	0.142	158.2	$1.97 \cdot 10^6$	2.55
20	ES	0.019	0.001	0.819	0.123	0.143	–	$2.02 \cdot 10^6$	2.59
21	ESCov	0.007	0.092	0.830	0.121	0.178	165.4	$1.98 \cdot 10^6$	2.55
21	ES	0.015	0.001	0.828	0.120	0.175	–	$2.03 \cdot 10^6$	2.59
22	ESCov	0.004	0.145	0.804	0.119	0.219	168.5	$1.98 \cdot 10^6$	2.55
22	ES	0.010	0.001	0.858	0.117	0.223	–	$2.04 \cdot 10^6$	2.60
23	ESCov	0.001	0.001	0.997	0.119	0.269	175.3	$1.97 \cdot 10^6$	2.54
23	ES	0.004	0.042	0.825	0.114	0.277	–	$2.05 \cdot 10^6$	2.60
24	ESCov	0.001	0.001	0.997	0.120	0.271	186.2	$1.97 \cdot 10^6$	2.54
24	ES	0.001	0.001	0.998	0.113	0.315	–	$2.04 \cdot 10^6$	2.60
72	ESCov	0.001	0.001	0.997	0.118	0.284	168.2	$1.97 \cdot 10^6$	2.55
72	ES	0.001	0.001	0.886	0.111	0.348	–	$2.06 \cdot 10^6$	2.61
168	ESCov	0.002	0.164	0.733	0.117	0.332	219.5	$2.00 \cdot 10^6$	2.56
168	ES	0.001	0.001	0.998	0.108	0.365	–	$2.07 \cdot 10^6$	2.62
168	naïve forecast	–	–	–	–	–	–	$3.97 \cdot 10^6$	3.48
240	ESCov	0.001	0.001	0.998	0.126	0.291	231.3	$2.00 \cdot 10^6$	2.57
240	ES	0.001	0.001	0.998	0.114	0.349	–	$2.09 \cdot 10^6$	2.64
336	naïve forecast	–	–	–	–	–	–	$5.35 \cdot 10^6$	3.96

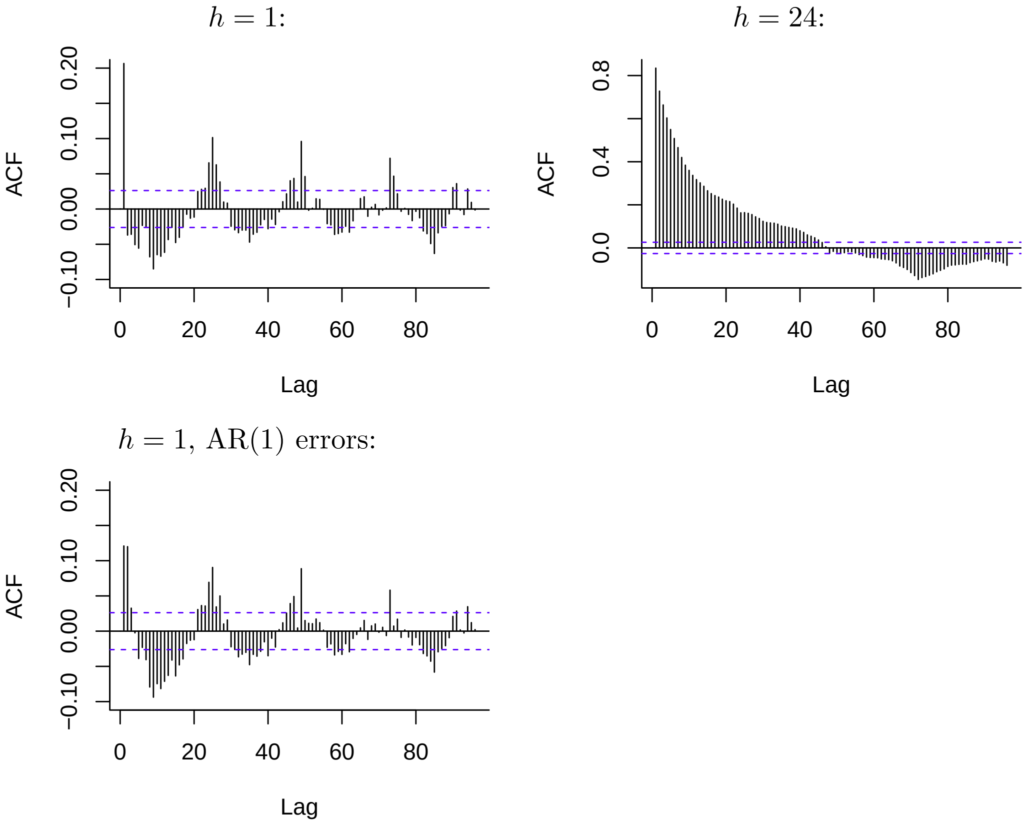


Figure 6.4: Autocorrelation plots of the one-step-ahead forecast errors under the ESCov model ADT-AS for lead times $h = 1$ without (top left) and with (bottom left) autocorrelation adjustment on the errors and $h = 24$ (top right) without autocorrelation adjustment. Dashed: confidence intervals of level 95 %.

applied ES with double multiplicative seasonality to half-hourly electricity data from England and Wales. The improvement on the first order autocorrelation of the errors gets visible in Fig. 6.4. The results show that it is worthwhile considering an autocorrelated residual model to improve upon the goodness of fit and forecast accuracy.

The autocorrelation in the residuals will not be eliminated by using the raw temperature data without a threshold. A preliminary simulation study showed that autocorrelation in the residuals is also visible in the covariate-free method, i. e. in the classical Holt-Winters method. Autocorrelation can be found in the residuals as soon as the model accounts for seasonality. Our supposition is that this is caused by an alternation of underestimation and overestimation in the model and is rather a characteristic of the seasonal ES model than of the data. As a consequence we can still see a seasonality in the ACF of the residuals.

Table 6.8: Parameters of model ADT-AS for ES and ESCov estimated by the hourly data for seven provinces in Emilia-Romagna aggregated including an AR(1) term on the errors. Covariate: average hourly temperature of the seven provinces of Emilia-Romagna. α_1 = parameter for level, α_2 = parameter for trend increment, ϕ = trend damping parameter, $\alpha_{1,3}$ = within-week seasonality, $\alpha_{2,3}$ = within-day seasonality, β = temperature coefficient. MSE in kWh^2, MAPE in %.

h	model	α_1	α_2	ϕ	$\alpha_{1,3}$	$\alpha_{2,3}$	β	λ	MSE	MAPE
1	ESCov	0.483	0.001	0.736	0.199	0.192	55.5	0.213	$6.06 \cdot 10^5$	1.37
1	ES	0.486	0.001	0.710	0.201	0.198	–	0.210	$6.08 \cdot 10^5$	1.37

6.9.2.2 Post-sample Analysis

We use the period 1 July 2006 to 25 July 2006 before the summer holiday period, in which many companies shut down their activities, for a post-sample evaluation. We report the MAPE, the MSE and the RMSE of the forecast as measures of forecast accuracy. See Definition 6.15 for the formulas of these measures. Here we have $T = 5592$, which corresponds to hour 23 on 30 June 2006, and $N = 25 \cdot 24 = 600$, i. e. the h-step-ahead forecasts for 600 hours are compared with the actual observations. The performance of the forecast is analysed for $h = 1$ hour, $h = 24$ hours (= 1 day), $h = 72$ hours (= 3 days), $h = 168$ hours (= 1 week) and $h = 240$ (= 10 days) ahead under the estimated parameters reported in Section 6.9.2.1. The hourly average temperature in the seven provinces in Emilia-Romagna, which we use as a covariate in the prediction, is based on the actually observed average hourly temperature values.

Table 6.9 lists the results of the post-sample analysis in the period 1 July 2006 to 25 July 2006. By the naïve forecast, an MAPE of 4.22 (lead times up to $h = 168$) or 3.61 (lead times up to $h = 336$) is achieved.

The post-sample forecast MAPEs for ESCov beat those of ES for all investigated lead times. While for one-hour-ahead forecasts the improvement in MAPE is only 1.5 %, it increases to almost 11 % for the 24-hours-ahead forecast and 15 % for one-week-ahead forecasts. This finding is in accordance with intuition: For forecasts from one hour to the next, the temperature does not play an important role in the forecast, but gains importance as a covariate for forecasts several days ahead. ESCov achieves a forecasting performance of more than 30 % better than the naïve forecast for several lead times in the post-sample period. The best result for the one-hour-ahead forecast is obtained by the ESCov model considering AR(1) residuals.

We analyse the performance of the prediction intervals obtained by the plug-in method described in Eq. (6.29) in the post-sample period 1 July 2006 to 25 July 2006. The

Table 6.9: Forecast accuracy of the one-hour-, 24-hours-, 72-hours-, 168-hours- and 240-hours-ahead forecasts in the post-sample period 1 July 2006 to 25 July 2006. MAPE in %, MSE in kWh^2, RMSE in kWh. The naïve forecast is the observed value of the same hour on the previous week (i. e. 168 hours before) or of two weeks before (i. e. 336 hours before), respectively.

h	model	MAPE	MSE	RMSE
1	ESCov	1.34	$6.4 \cdot 10^5$	800
1	ESCov with AR(1) residuals	1.32	$6.2 \cdot 10^5$	787
1	ES	1.36	$6.5 \cdot 10^5$	804
1	ES with AR(1) residuals	1.33	$6.2 \cdot 10^5$	789
24	ESCov	2.70	$2.7 \cdot 10^6$	1647
24	ES	3.02	$3.2 \cdot 10^6$	1781
72	ESCov	3.41	$3.9 \cdot 10^6$	1981
72	ES	4.04	$5.1 \cdot 10^6$	2254
168	ESCov	2.91	$3.3 \cdot 10^6$	1819
168	ES	3.41	$4.5 \cdot 10^6$	2120
168	naïve forecast	4.22	$6.8 \cdot 10^6$	2610
240	ESCov	2.63	$2.9 \cdot 10^6$	1702
240	ES	3.00	$3.4 \cdot 10^6$	1833
336	naïve forecast	3.61	$4.3 \cdot 10^6$	2082

parameter estimates for the SSOE model ADT-AS applied to the hourly electricity demand are the ones reported in Section 6.9.2.1, both for the case with and without the average hourly temperature as covariate. A plot of the 24-hours-ahead forecasts in the period 1 July 2006 to 25 July 2006 with level $\gamma = 95\,\%$ prediction intervals as well as the actual observations is provided in Fig. 6.5. During the analysed period, the actual observations mostly lie in the calculated prediction intervals for both models. The prediction intervals for ESCov fail in 10.2 %, those for ES in 11.3 % of the cases.

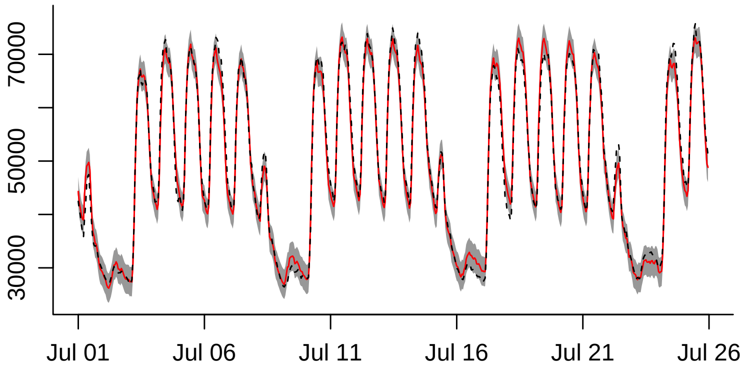


Figure 6.5: 24-hours-ahead load forecasts and actuals in the post-sample period 1 July 2006 to 25 July 2006 with level 95 % prediction intervals. Solid red line: 24-hours-ahead forecast, dashed black line: actual values, grey area: prediction intervals.

To visualise the coverage properties, we provide a plot of the differences between the actual loads and the load forecasts in Figure 6.6 both for ESCov and ES. The undercoverage is likely due to both the plug-in method and to the violation of the independence assumption of residuals.

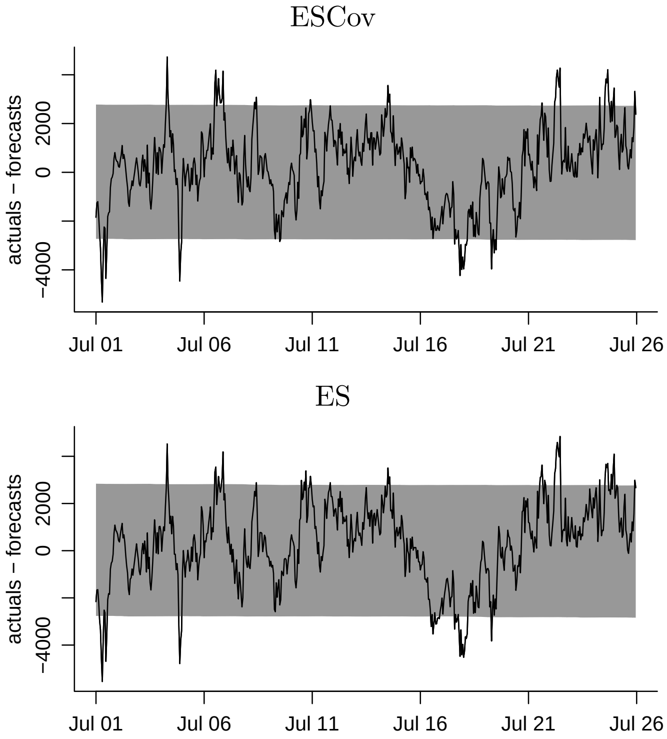


Figure 6.6: Difference between actuals and 24-hours-ahead load forecasts in the post-sample period 1 July 2006 to 25 July 2006 with level 95 % prediction intervals.

6.9.3 Renormalisation of Electricity Load Seasonalities

We investigate the effect of the renormalisation of the seasonal pattern as presented in Section 6.8 in the context of the electricity study. Since we used double additive seasonality for the model estimation of hourly electricity load data in Section 6.9.2, the renormalisation equations as presented in Definition 6.10 are applied and the results from Proposition 6.11 hold. The intra-day seasonality of length 24 and the intra-week seasonality of length 168 are both treated simultaneously.

We consider the model ADT-AS with the parameters optimised according to the one-hour-ahead forecast. Figure 6.7 displays the weekly deseasonalised (by applying a moving average of length 168) seasonality pattern for the classical scheme without renormalisation and the scheme after renormalisation. While the season pattern under the nonrenormalised scheme is below zero in average most of the time and rises to above zero in the end of the estimation period, the season pattern under the renormalised scheme stays clearly more stable around zero.

From Proposition 6.11 we know that for the case of additive seasonality the predictions are the same under the classical and the renormalised scheme. Furthermore, with the observation series ranging between about 20000 and 70000 and the seasonal component drift being small in absolute value (between about -10 and $+5$), the seasonal pattern of the renormalised model without applying a moving average of 168 cannot visibly be distinguished from the seasonal pattern under the classical scheme. Consequently, for the data at hand, even if one is concerned with the correct interpretation of the separate components, the renormalisation seems dispensable.

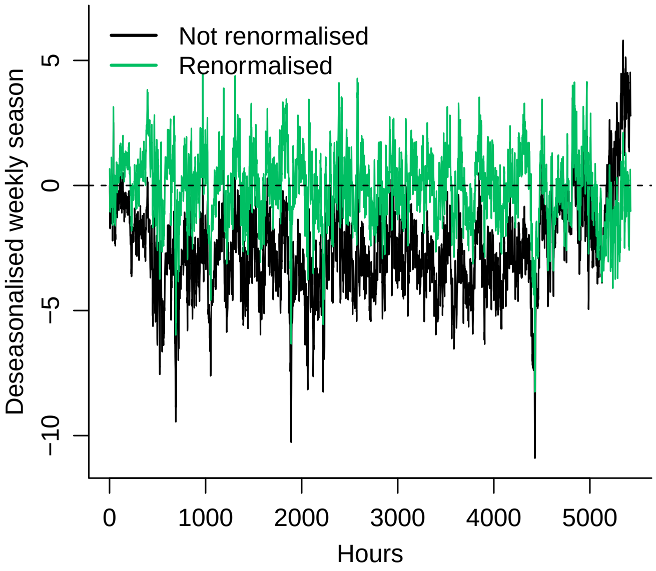


Figure 6.7: Deseasonalised weekly season pattern under the classical scheme and the renormalised scheme

6.9.4 Conclusion of the Load Forecasting Study

Our empirical study demonstrates the potential of exponential smoothing with covariates (ESCov) as a load forecasting methodology. The electricity forecasts turned out to be more precise under the use of meteorological variables, in particular temperature, than without. The gain in forecast accuracy was higher for prediction horizons in the range of one-day-ahead and more than for very short-term predictions of one-hour-ahead. The latter shows the strength of ordinary ES, which manages to adapt quickly to level changes if rather small time intervals are contemplated. Prediction intervals based on the SSOE model were applied by simply plugging in the estimated parameters into the equation for the prediction intervals constructed under the assumption of asymptotic normality. We found the intervals slightly too narrow in our analysis, which is consistent to Ord et al.'s (1997) observations. This shows the need for prediction intervals for ESCov which are robust against violations of the assumption. The temperature values were assumed to be known for the predictions. Considering the accuracy of temperature forecasts nowadays, this is justifiable.

The forecasting results for the analysed dataset were in general worse than the load forecasting results from other studies. This fact is mainly due to the small number of customers forming the portfolio of the energy vendor. Immigration and emigration of customers as well as an unusual behaviour of one bigger customer in a comparably small dataset can lead to considerable disturbances in the load series, which affect the forecast performance. Furthermore, the shortness of the series does not allow to properly estimate the influence of special days and special periods. Hence holiday periods were deleted in the sample, and the empirical analysis of Section 6.9.2 makes no use of public holidays and holiday periods as calendar covariates. This shows that considerable further efforts are required to integrate ESCov into an intelligent and automatic load forecasting system. Basically, the ESCov model allows the inclusion of such variables in form of covariates, and it is to be expected that this treatment will improve upon prediction accuracy for periods in which such days occur. However, to account for the effect of public holidays, the sample period has to be long enough so that the respective effect can be satisfactorily estimated.

6.10 Implementation and Numerical Issues

The ESCov method was implemented by the author of this thesis in the statistical computing environment R (R Core Team 2014). Several difficulties encountered during the implementation are reported in this section.

6.10.1 Parameter Optimisation

As described in Section 6.4, the smoothing and covariate parameters can be estimated by minimising a certain objective function, such as the mean square error (MSE) or mean absolute percentage error (MAPE) of the h-step-ahead forecasts, $h \in \{1, 2, \dots\}$. In our study, we optimised with respect to the MSE, which we found to be more stable than optimising with respect to the MAPE. This is in accordance to common practice in optimisation problems, e. g. least squares fitting in regression.

The parameter optimisation was done using the function `optim`, which is contained in the `stats`-package coming with the default R installation. It allows the optimisation method `"L-BFGS-B"`, in which lower and upper limits for the parameters can be specified. This is advantageous because the smoothing parameters $\alpha_1, \alpha_2, \phi, \alpha_{i,3},\ i = 1, \ldots, m$, take values between 0 and 1. Apart from the parameter bounds, starting values are requested for the parameters to initialise the search for the optimum.

The method `"L-BFGS-B"` follows a suggestion for solving large nonlinear optimisation problems with simple bounds by Byrd et al. (1995). Byrd et al. (1995) describe it as nearly as efficient as an unconstrained limited memory algorithm while being able to handle bounds on the parameters.

In our study, we have found the method `"L-BFGS-B"` quite useful and nearly always applicable. However, in certain situations it failed and did not find the optimal parameters. The failure could be identified by trying a different set of parameters, which turned out to show better results in terms of the MSE. The parameter sets resulting from the optimisation were consequently found to be sensitive with respect to the choice of the starting values for the optimisation. Hence, the optimisation result did not necessarily come with the minimum MSE. A possible explanation might be that the objective function is too complex and probably contains several local minima that the optimisation method is running into, which are not all global minima. This is clearly not desirable and hence a way is needed to overcome this problem.

We saw a possible solution in the implementation of a global optimisation method. The `optim` function in R offers the method `"SANN"`, a variant of simulated annealing to be found in Bélisle (1992). Simulated annealing is described by Bélisle (1992) as a Monte Carlo technique for solving optimisation problems. When the optimisation method `"L-BFGS-B"` failed or proved to be too sensitive with respect to the starting values, the method `"SANN"` usually came close to the optimum. However, it took considerably longer than the `"L-BFGS-B"` method. We might not have used the most appropriate optimisation control parameters, such as number of iterations or the convergence tolerance, but used only some relatively simple and few settings. So there might still be potential for improvement here.

More promising seems to be the optimisation method `"GenSA"`, which does not come with the standard installation of R, but as an additional package named `"GenSA"`. The optimisation function comes with the same name as the package itself, which was written by Xiang et al. (2013). The generalised simulated annealing method implemented in the "GenSA" package follows an algorithm introduced by Tsallis & Stariolo (1996). It is described by Xiang et al. (2013) as a method that searches for a global minimum when dealing with a very complex nonlinear objective function with a very large number of optima. Our experiences with the "GenSA" optimisation are very good. We found it to be considerably more efficient than the "SANN" method and mostly reliable in the results. Therefore we prefer "GenSA" over "SANN". The optimisation control parameters possibly leave also room for improvement, certainly depending on the data at hand.

In terms of computation time, both simulated annealing methods (naturally) took longer to compute the optimal parameters than the gradient method "L-BFGS-B". This encouraged us to look for a way to use the advantages of both methods to achieve good results: the precision from "GenSA" and the computation time from "L-BFGS-B". Hence, in our implementation of ESCov, we implemented a combination of the simulated annealing method with the gradient method. By applying the simulated annealing method in the first step, we hope to reach an area in the optimisation space where the objective function is more or less convex, such that the gradient method applied in the second step would find the optimum more easily and especially faster. With this approach, there is still the risk that the simulated annealing method applied in the beginning would not be applied sufficiently long to arrive at a region with the desired convexity properties or at a region which does not contain the global, but only a local optimum. However, the computation time that can be gained by this approach might make up for the loss in precision with respect to the objective function if the set of parameters found does not come with considerably worse results than the optimal one.

Nevertheless, we would like express a warning to not force the optimisation method to find the optimum by all means. When the optimisation fails, it should be taken as a warning that something might be wrong, such as a possible redundance of at least one smoothing or covariate parameter. As a combination of exponential smoothing with regression, ESCov adopts some of the characteristics that can be experienced in a regression context. Although finding the optimal smoothing and covariate parameters is not subject to an explicit formula including an inversion of the design matrix (something that would fail in the presence of collinearities in its columns), ESCov can encounter difficulties finding the optimal set of parameters if there are strong dependencies between the covariates. Likewise, the smoothing parameters associated with the states, in particular trend and season, can cause problems. If there is no trend in the data or the trend is explained by the covariates, the smoothing parameter α_2 is possibly superfluous. By nevertheless exposing α_2 to the optimisation routine might cause the parameter to not find its place in the parameter space. Similar holds for the seasonal component. We recommend to invest some effort into deliberate model selection, such that there is a smaller risk of a redundance of model parameters and implied optimisation problems.

6.10.2 Notation and Computation Time

The R code to compute ESCov has been developed with progress in theory and hence underwent a lot of changes in the course of time. One of the first approaches was to implement the procedure following the state transition formulas in Tables 6.1 to 6.3. A discrimination between the different models and hence a lot of `if` clauses were necessary to comply with the different models. The division of the models into the classes "linear SSOE", "partially linear SSOE" and "exponential trend models" and the introduction of the corresponding matrix notation using the components of Tables 6.4 and 6.5 allowed a more efficient and transparent way of programming. The computation time decreased considerably to at least half, possibly a third of the former computation time. Whether the gain in computation time is rightfully assigned to the matrix notation or just a gain in programming experience of the author of the code, is certainly difficult to distinguish. However, if not for the computation time, we advertise the matrix notation for reasons of transparency as well as an easier treatment of the starting values of the states. The latter goes well with the matrix notation and is less susceptible to programming errors.

One of the reasons for the popularity of ES is the comparably simple structure of the model. In former times, it had its attractiveness from the fact that for given smoothing parameters the current smoothed value could be calculated by a simple linear combination of the current observed value and the preceding smoothed value of level, trend increment or season. Hence, not more than the smoothed values of one season plus the smoothed values of level and trend increment have to be saved. However, this applies for given smoothing parameters only. In the early days of ES, rules of thumb were applied to choose an appropriate smoothing constant α_1 for the level. For example, Brown (1956) presents a table with values for α_1 in dependence of the total weight the n last periods carry when α_1 is chosen in a certain way. Nowadays, the demand for tables to choose appropriate values for the smoothing parameters from is low. Computers allow to apply numerical procedures to achieve smoothing parameters which are optimal in some sense. In our study, we chose the parameters such that the MSE of the h-step-ahead forecast was optimal. To achieve this, the optimisation procedure requires to compute the series of smoothed values from the first till the last observation for a number of sets of smoothing parameters. This repeatedly executed smoothing, in which the current smoothed value is dependent on the previously smoothed value, requires to programme `for` loops. They contribute most to the computation time, but cannot be avoided due to the definition of the method.

6.10.3 Initial States Values

In the empirical study of Section 6.9 we chose the initial level, trend increment and seasons fixed according to Census I based rules that were recommended for ES. In particular, we did not estimate the starting values for the states together with the smoothing and covariate parameters for fearing this would make it a too complex venture, see the remarks in Section 6.5. More recent numerical experiences have revealed that ESCov is far more sensitive with respect to the choice of initial state values than initially thought. The assumption that an inappropriate choice of starting values would smooth out after several rounds of smoothing mostly turned out to be true, but frequently resulted in a different set of smoothing and covariate parameters than under a different set of initial state values. We conjecture that the choice of the initial state values carries more importance for ESCov than for ES. In a first attempt, we estimated the initial level and initial trend increment simultaneously with the smoothing and covariate parameters while choosing the initial seasonal component with a Census I type procedure in advance. This approach worked considerably well. We are skeptical about also estimating the initial season simultaneously with the smoothing and covariate parameters. Especially in the case of the electricity data with hourly data and seasonalities of lengths 168 and 24, a large number of parameters would have to be estimated. Since we have already encountered problems in the optimisation of just the smoothing and covariate parameters, this is not unlikely to be doomed to failure and probably would come with a disproportionate computation time.

6.11 Conclusion and Outlook

We have considered the statistical time series method of exponential smoothing with covariates (ESCov) that has been found to be underpinned by a single source of error (SSOE) state-space model by Wang (2006) and extended the methodology to account for multiple seasonalities. The MMSE forecast and forecast variance under the linear and partially linear SSOE for ESCov have been presented. These schemes cover the most popular exponential smoothing (ES) variants. One of the attractive features of ES is the interpretability of its components. It can get partially lost with successive smoothing in the seasonal components, unless they are renormalised. We have adopted the renormalisation scheme of Roberts (1982), McKenzie (1986) and Archibald & Koehler (2003) and applied it to multiple seasonality models. Forecasts under ESCov rely on future covariate values. This naturally limits forecast horizons according to the availability of the covariate values.

The theoretical results of ESCov have been applied in an electricity load forecasting study. Covariates have been used in form of meteorological variables, in particular temperature. ESCov has demonstrated to perform well as a load forecasting methodology. We have carried out the optimisation of smoothing and covariate parameters in the ESCov model separately per forecast step. Considering the different parameter estimates, the approach appears to be viable. However, autocorrelation can be induced in the residuals, which can conflict with the assumptions on the residuals determining the forecast variance. This drawback of autocorrelated residuals as revealed by the study has been discovered by other authors, e. g. Chatfield (1978) and Taylor (2003b), for ES before. Also the inclusion of meteorological covariates could not make up for this in the present study. By accounting for autoregressive residuals of order 1, the prediction performance could be improved to a certain extent. This shows the potential of ESCov of including also higher order autocorrelation models for the residuals. Considering the involved formulas for the forecast and the forecast variance already in the case of $p = 1$, it is to be expected that the respective formulas become rather intractable unless the correlation structure is very simple. It therefore has to be postponed to future studies.

We have presented the MMSE forecast and variance for the linear, potentially damped, trend models and both additive and multiplicative as well as no seasonality. Formulas for the model-based forecast variance under an exponential trend ESCov model still have to be derived.

The state-space formulation of the ESCov model allows the derivation of the likelihood under certain assumptions. We have reviewed the approach by Ord et al. (1997) and Wang (2006) that holds for a forecast step of 1 in Section 6.5. The respective formula for the likelihood under larger forecast steps needs to be explicitely formulated. From there, likelihood-based measures as the AIC or BIC can be derived, which we find useful for model selection.

The part assigned to the covariate in the presented ESCov models consists of an additive term in the observation equation (6.15). Further methodological development of ESCov should also refine the multiplicative seasonal model by considering an alternative type of covariate influence. In the observation equation of the partially linear model considered in Section 6.3.2 and exemplified in Table 6.2, the seasonal coefficients have no effect on the covariate term $\boldsymbol{\beta}^\top \boldsymbol{x}$, which represents a mere additive shift in level. However, the seasonal coefficients might also act on the total level by $(\boldsymbol{\delta}_{\boldsymbol{\alpha},ne}^\top \boldsymbol{u}_{t-1,ne} + \boldsymbol{\beta}^\top \boldsymbol{x}) \prod_{i=1}^{m} e_{i,t-d_i}$. This observation equation is a case of a markedly different model where the covariates would have to be considered as states.

So far, we have applied simple rules to obtain initial values for the states (level, trend and seasons) in ESCov similar to ES without covariates. First tests have shown that methods proposed for ES show drawbacks when applied to ESCov, for there is a stronger interaction between the state components with the covariates than initially assumed. Refined methods to estimate the initial states in the presence of covariates have to be researched properly.

6.A Appendix

6.A.1 Proof of Proposition 6.3

To prove Proposition 6.3, we make use of the following proposition.

Proposition 6.16 (Recurrence Relation for Linear SSOEs)**.** *Consider the linear SSOE from Definition 6.2, and let $h > 0$. Then we have*

$$\begin{aligned} \boldsymbol{u}_{t+h} &= \sum_{l=0}^{h-1} \mathbf{G}_{\boldsymbol{\alpha}}^{l} \boldsymbol{w}_{\boldsymbol{\alpha}} \xi_{t+h-l} + \mathbf{G}_{\boldsymbol{\alpha}}^{h} \boldsymbol{u}_t, && (6.34) \\ Y_{t+h} &= \boldsymbol{\delta}_{\boldsymbol{\alpha}}^{\top} \left(\sum_{l=0}^{h-2} \mathbf{G}_{\boldsymbol{\alpha}}^{l} \boldsymbol{w}_{\boldsymbol{\alpha}} \xi_{t+h-l-1} + \mathbf{G}_{\boldsymbol{\alpha}}^{h-1} \boldsymbol{u}_t \right) + \boldsymbol{\beta}^{\top} \boldsymbol{x}_{t+h} + \xi_{t+h}. && (6.35) \end{aligned}$$

Proof. To prove Proposition 6.16, we use induction by h. For $h = 1$, the right-hand side of (6.34) amounts to $\boldsymbol{w}_{\boldsymbol{\alpha}} \xi_{t+1} + \mathbf{G}_{\boldsymbol{\alpha}} \boldsymbol{u}_t$, which equals $\boldsymbol{u}_{t+h} = \boldsymbol{u}_{t+1}$ by (6.9).

Let (6.34) be proven for $h > 0$. Then we have for $h + 1$ by Eq. (6.9)

$$\begin{aligned} \boldsymbol{u}_{t+h+1} &= \mathbf{G}_{\boldsymbol{\alpha}} \boldsymbol{u}_{t+h} + \boldsymbol{w}_{\boldsymbol{\alpha}} \xi_{t+h+1} \\ &= \mathbf{G}_{\boldsymbol{\alpha}} \left(\sum_{l=0}^{h-1} \mathbf{G}_{\boldsymbol{\alpha}}^{l} \boldsymbol{w}_{\boldsymbol{\alpha}} \xi_{t+h-l} + \mathbf{G}_{\boldsymbol{\alpha}}^{h} \boldsymbol{u}_t \right) + \boldsymbol{w}_{\boldsymbol{\alpha}} \xi_{t+h+1} \\ &= \sum_{l=0}^{h-1} \mathbf{G}_{\boldsymbol{\alpha}}^{l+1} \boldsymbol{w}_{\boldsymbol{\alpha}} \xi_{t+h-l} + \mathbf{G}_{\boldsymbol{\alpha}}^{h+1} \boldsymbol{u}_t + \mathbf{G}_{\boldsymbol{\alpha}}^{0} \boldsymbol{w}_{\boldsymbol{\alpha}} \xi_{t+h+1} \\ &= \sum_{m=1}^{h} \mathbf{G}_{\boldsymbol{\alpha}}^{m} \boldsymbol{w}_{\boldsymbol{\alpha}} \xi_{t+h+1-m} + \mathbf{G}_{\boldsymbol{\alpha}}^{0} \boldsymbol{w}_{\boldsymbol{\alpha}} \xi_{t+h+1} + \mathbf{G}_{\boldsymbol{\alpha}}^{h+1} \boldsymbol{u}_t \\ &= \sum_{m=0}^{h} \mathbf{G}_{\boldsymbol{\alpha}}^{m} \boldsymbol{w}_{\alpha} \xi_{t+h+1-m} + \mathbf{G}_{\boldsymbol{\alpha}}^{h+1} \boldsymbol{u}_t. \end{aligned}$$

Equation (6.35) is a direct consequence of (6.34) and (6.8). □

We proceed to the proof of Proposition 6.3.

From the properties of the conditional expectation we have $E[\boldsymbol{u}_t^m | \mathcal{H}_t] = \boldsymbol{u}_t^m$ for all $m \in \mathbb{N}$. By Proposition 6.16, $\boldsymbol{u}_{t+h}$ depends on the process history $\mathcal{H}_t$ up to time t only through $\boldsymbol{u}_t$, and on $\xi_{t+1}, \dots, \xi_{t+h}$. Because of the independence assumption on the errors, the assumption (6.10) and $E[\boldsymbol{u}_t | \mathcal{H}_t] = \boldsymbol{u}_t$, we obtain

$$\begin{aligned} E[\boldsymbol{u}_{t+h} | \mathcal{H}_t] &= E\Big[\sum_{l=0}^{h-1} \mathbf{G}_{\boldsymbol{\alpha}}^{l} \boldsymbol{w}_{\boldsymbol{\alpha}} \xi_{t+h-l} + \mathbf{G}_{\boldsymbol{\alpha}}^{h} \boldsymbol{u}_t \Big| \mathcal{H}_t \Big] \\ &= E\Big[\sum_{l=0}^{h-1} \mathbf{G}_{\boldsymbol{\alpha}}^{l} \boldsymbol{w}_{\boldsymbol{\alpha}} \xi_{t+h-l} \Big] + E[\mathbf{G}_{\boldsymbol{\alpha}}^{h} \boldsymbol{u}_t | \mathcal{H}_t] = \mathbf{G}_{\boldsymbol{\alpha}}^{h} \boldsymbol{u}_t. && (6.36) \end{aligned}$$

With (6.36) and (6.35) we obtain Eq. (6.11).

We obtain from the recurrence relation (6.35) and under the assumptions of Proposition 6.3

$$\begin{aligned} V[Y_{t+h}|\mathcal{H}_t] &= V\left[\boldsymbol{\delta}_{\boldsymbol{\alpha}}^\top \left(\sum_{l=0}^{h-2} \mathbf{G}_{\boldsymbol{\alpha}}^l \boldsymbol{w}_{\boldsymbol{\alpha}} \xi_{t+h-l-1} + \mathbf{G}_{\boldsymbol{\alpha}}^{h-1} \boldsymbol{u}_t\right) + \boldsymbol{\beta}^\top \boldsymbol{x}_{t+h} + \xi_{t+h} \middle| \mathcal{H}_t\right] \\ &= V\Big[\boldsymbol{\delta}_{\boldsymbol{\alpha}}^\top \sum_{l=0}^{h-2} \mathbf{G}_{\boldsymbol{\alpha}}^l \boldsymbol{w}_{\boldsymbol{\alpha}} \xi_{t+h-l-1}\Big] + V\Big[\boldsymbol{\delta}_{\boldsymbol{\alpha}}^\top \mathbf{G}_{\boldsymbol{\alpha}}^{h-1} \boldsymbol{u}_t |\mathcal{H}_t\Big] + V[\xi_{t+h}] \\ &= \sum_{l=0}^{h-2} \left(\boldsymbol{\delta}_{\boldsymbol{\alpha}}^\top \mathbf{G}_{\boldsymbol{\alpha}}^l \boldsymbol{w}_{\boldsymbol{\alpha}}\right)^2 \sigma_{\xi_{t+h-l-1}}^2 + \sigma_{\xi_{t+h}}^2. \end{aligned}$$

6.A.2 Proof of Proposition 6.4

To prove Proposition 6.4, we make use of the following proposition:

Proposition 6.17 (Recurrence Relation for Linear SSOEs with AR(1) Errors)**.** *Consider the linear SSOE from Definition 6.2 under AR(1) errors, i. e. Eq. (6.5) holds with independent errors (ε_t) that satisfy $E[\varepsilon_t] = 0$ and $V[\varepsilon_t] = \sigma_\varepsilon^2$. For each t and each $k > 0$ let $E[\xi_{t+k}|\mathcal{H}_t] = E[\xi_{t+k}|(\xi_s)_{s \le t}]$.*
For $k > 0$ we have

$$\xi_{t+k} = \lambda^k \xi_t + \sum_{i=0}^{k-1} \lambda^i \varepsilon_{t+k-i}, \tag{6.37}$$

and for $0 < k \le m$

$$\begin{aligned} \xi_{t+m}\xi_{t+k} &= \lambda^{k+m}\xi_t^2 + \xi_t \left(\lambda^m \sum_{i=0}^{k-1} \lambda^i \varepsilon_{t+k-i} + \lambda^k \sum_{j=0}^{m-1} \lambda^j \varepsilon_{t+m-j}\right) \\ &\qquad + \sum_{\substack{0 \le i \le k-l \\ 0 \le j \le m-1}} \lambda^{i+j} \varepsilon_{t+k-i}\varepsilon_{t+m-j}, \end{aligned} \tag{6.38}$$

$$E[\xi_{t+k}|\mathcal{H}_t] = \lambda^k \xi_t, \qquad E[\xi_{t+k}\xi_{t+m}|\mathcal{H}_t] = \lambda^{k+m}\xi_t^2 + \lambda^{m-k}\frac{1-\lambda^{2k}}{1-\lambda^2}\sigma_\varepsilon^2, \tag{6.39}$$

$$\mathrm{Cov}[\xi_{t+k}, \xi_{t+m}|\mathcal{H}_t] = \lambda^{m-k}\frac{1-\lambda^{2k}}{1-\lambda^2}\sigma_\varepsilon^2, \qquad V[\xi_{t+k}|\mathcal{H}_t] = \frac{1-\lambda^{2k}}{1-\lambda^2}\sigma_\varepsilon^2. \tag{6.40}$$

For $h > 0$ we have

$$\begin{aligned} Y_{t+h} &= \boldsymbol{\delta}_{\boldsymbol{\alpha}}^{\top} \left(\sum_{\ell=0}^{h-2} \mathbf{G}_{\boldsymbol{\alpha}}^{\ell} \boldsymbol{w}_{\boldsymbol{\alpha}} \left(\lambda^{h-\ell-1} \xi_t + \sum_{i=0}^{h-\ell-2} \lambda^i \varepsilon_{t+h-\ell-1-i} \right) + \mathbf{G}_{\boldsymbol{\alpha}}^{h-1} \boldsymbol{u}_t \right) \\ &\qquad + \boldsymbol{\beta}^{\top} \boldsymbol{x}_{t+h} + \lambda^h \xi_t + \sum_{j=0}^{h-1} \lambda^j \varepsilon_{t+h-j}, \qquad (6.41) \\ E[\boldsymbol{u}_{t+h} | \mathcal{H}_t] &= \sum_{\ell=0}^{h-1} \mathbf{G}_{\boldsymbol{\alpha}}^{\ell} \boldsymbol{w}_{\boldsymbol{\alpha}} \lambda^{h-\ell} \xi_t + \mathbf{G}_{\boldsymbol{\alpha}}^{h} \boldsymbol{u}_t. \qquad (6.42) \end{aligned}$$

Proof. For $k = 1$, (6.37) expresses the recurrence $\xi_{t+1} = \lambda \xi_t + \varepsilon_{t+1}$. Let (6.37) hold for a $k > 0$. Then for $k+1$ by the recurrence $\xi_{t+k+1} = \lambda \xi_{t+k} + \varepsilon_{t+k+1}$

$$\begin{aligned} \xi_{t+k+1} &= \lambda \xi_{t+k} + \varepsilon_{t+k+1} = \lambda \left(\lambda^k \xi_t + \sum_{i=0}^{k-1} \lambda^i \varepsilon_{t+k-i} \right) + \varepsilon_{t+k+1} \\ &= \lambda^{k+1} \xi_t + \sum_{i=0}^{k-1} \lambda^{i+1} \varepsilon_{t+k+1-(i+1)} + \varepsilon_{t+k+1} = \lambda^{k+1} \xi_t + \sum_{j=1}^{k} \lambda^j \varepsilon_{t+k+1-j} + \varepsilon_{t+k+1} \\ &= \lambda^{k+1} \xi_t + \sum_{j=0}^{k} \lambda^j \varepsilon_{t+k+1-j}. \end{aligned}$$

Equation (6.38) and the first part of Eq. (6.39) follow directly from (6.37). Consider the second part of Eq. (6.39). We have $k - i = m - j \iff j - i = m - k \iff j = m - k + i$, hence $E[\varepsilon_{t+k-i} \varepsilon_{t+m-j}] = \sigma_\varepsilon^2$ if $j - i = m - k$ and otherwise $E[\varepsilon_{t+k-i} \varepsilon_{t+m-j}] = 0$. Thus from (6.38)

$$\begin{aligned} &E[\xi_{t+k} \xi_{t+m} | \mathcal{H}_t] \\ &\quad = E\Big[\lambda^{k+m} \xi_t^2 + \xi_t \left(\lambda^m \sum_{i=0}^{k-1} \lambda^i \varepsilon_{t+k-i} + \lambda^k \sum_{j=0}^{m-1} \lambda^j \varepsilon_{t+m-j} \right) \\ &\qquad\qquad + \sum_{\substack{0 \le i \le k-1 \\ 0 \le j \le m-1}} \lambda^{i+j} \varepsilon_{t+k-i} \varepsilon_{t+m-j} \Big] \\ &\quad = \lambda^{k+m} \xi_t^2 + \sigma_\varepsilon^2 \sum_{i=0}^{k-1} \lambda^{i+m-k+i} = \lambda^{k+m} \xi_t^2 + \lambda^{m-k} \sigma_\varepsilon^2 \sum_{i=0}^{k-1} \lambda^{2i} \\ &\quad = \lambda^{k+m} \xi_t^2 + \lambda^{m-k} \sigma_\varepsilon^2 \frac{1 - \lambda^{2k}}{1 - \lambda^2}. \end{aligned}$$

Equation (6.40) follows from (6.39).

Equation (6.41) follows by inserting (6.37) into (6.35).

By (6.34) and (6.39) we obtain

$$\begin{aligned} E[\boldsymbol{u}_{t+h}|\mathcal{H}_t] &\overset{(6.34)}{=} E\Big[\sum_{\ell=0}^{h-1}\mathbf{G}_{\boldsymbol{\alpha}}^{\ell}\boldsymbol{w}_{\boldsymbol{\alpha}}\xi_{t+h-\ell}+\mathbf{G}_{\boldsymbol{\alpha}}^{h}\boldsymbol{u}_t|\mathcal{H}_t\Big] \\ &= E\Big[\sum_{\ell=0}^{h-1}\mathbf{G}_{\boldsymbol{\alpha}}^{\ell}\boldsymbol{w}_{\boldsymbol{\alpha}}\xi_{t+h-\ell}|\mathcal{H}_t\Big]+E\Big[\mathbf{G}_{\boldsymbol{\alpha}}^{h}\boldsymbol{u}_t|\mathcal{H}_t\Big] \\ &\overset{(6.39)}{=} \sum_{\ell=0}^{h-1}\mathbf{G}_{\boldsymbol{\alpha}}^{\ell}\boldsymbol{w}_{\boldsymbol{\alpha}}\lambda^{h-\ell}\xi_t+\mathbf{G}_{\boldsymbol{\alpha}}^{h}\boldsymbol{u}_t, \end{aligned}$$

i. e. Eq. (6.42). □

We can now prove Proposition 6.4.

It follows with (6.42)

$$E[Y_{t+h}|\mathcal{H}_t] = \boldsymbol{\delta}_{\boldsymbol{\alpha}}^{\top}\left(\sum_{\ell=0}^{h-2}\mathbf{G}_{\boldsymbol{\alpha}}^{\ell}\boldsymbol{w}_{\boldsymbol{\alpha}}\lambda^{h-\ell-1}\xi_t+\mathbf{G}_{\boldsymbol{\alpha}}^{h-1}\boldsymbol{u}_t\right)+\boldsymbol{\beta}^{\top}\boldsymbol{x}_{t+h}+\lambda^{h}\xi_t,$$

i. e. Eq. (6.13).

Now, we prove (6.14). We have

$$\begin{aligned} &\operatorname{Cov}\Big[\boldsymbol{\delta}_{\boldsymbol{\alpha}}^{\top}\sum_{\ell=0}^{h-2}\mathbf{G}_{\boldsymbol{\alpha}}^{\ell}\boldsymbol{w}_{\boldsymbol{\alpha}}\sum_{i=0}^{h-\ell-2}\lambda^{i}\varepsilon_{t+h-\ell-1-i},\sum_{j=0}^{h-1}\lambda^{j}\varepsilon_{t+h-j}\Big] \\ &\qquad = \sum_{\ell=0}^{h-2}\sum_{i=0}^{h-\ell-1}\sum_{j=0}^{h-1}\boldsymbol{\delta}_{\boldsymbol{\alpha}}^{\top}\mathbf{G}_{\boldsymbol{\alpha}}^{\ell}\boldsymbol{w}_{\boldsymbol{\alpha}}\lambda^{i+j}\operatorname{Cov}[\varepsilon_{t+h-\ell-1-i},\varepsilon_{t+h-j}]. \end{aligned}$$

We have $h-\ell-1-i=h-j \iff i=j-\ell-1$. Therefore

$$\operatorname{Cov}[\varepsilon_{t+h-\ell-1-i},\varepsilon_{t+h-j}] = V[\varepsilon_{t+h-j}] = \sigma_{\varepsilon}^{2}$$

if $i=j-\ell-1$ and $\operatorname{Cov}[\varepsilon_{t+h-\ell-1-i},\varepsilon_{t+h-j}]=0$ otherwise. Hence

$$\begin{aligned} &\operatorname{Cov}\Big[\boldsymbol{\delta}_{\boldsymbol{\alpha}}^{\top}\sum_{\ell=0}^{h-2}\mathbf{G}_{\boldsymbol{\alpha}}^{\ell}\boldsymbol{w}_{\boldsymbol{\alpha}}\sum_{i=0}^{h-\ell-2}\lambda^{i}\varepsilon_{t+h-\ell-1-i},\sum_{j=0}^{h-1}\lambda^{j}\varepsilon_{t+h-j}\Big] \\ &= \sum_{\ell=0}^{h-2}\sum_{j=\ell+1}^{h-1}\boldsymbol{\delta}_{\boldsymbol{\alpha}}^{\top}\mathbf{G}_{\boldsymbol{\alpha}}^{\ell}\boldsymbol{w}_{\boldsymbol{\alpha}}\lambda^{2j-\ell-1}\sigma_{\varepsilon}^{2} = \sum_{\ell=0}^{h-2}\boldsymbol{\delta}_{\boldsymbol{\alpha}}^{\top}\mathbf{G}_{\boldsymbol{\alpha}}^{\ell}\boldsymbol{w}_{\boldsymbol{\alpha}}\lambda^{\ell+1}\sigma_{\varepsilon}^{2}\sum_{j=\ell+1}^{h-1}(\lambda^{2})^{j-\ell-1} \\ &= \sigma_{\varepsilon}^{2}\sum_{\ell=0}^{h-2}\boldsymbol{\delta}_{\boldsymbol{\alpha}}^{\top}\mathbf{G}_{\boldsymbol{\alpha}}^{\ell}\boldsymbol{w}_{\boldsymbol{\alpha}}\lambda^{\ell+1}\sum_{k=0}^{h-\ell-2}(\lambda^{2})^{k} = \sigma_{\varepsilon}^{2}\sum_{\ell=0}^{h-2}\boldsymbol{\delta}_{\boldsymbol{\alpha}}^{\top}\mathbf{G}_{\boldsymbol{\alpha}}^{\ell}\boldsymbol{w}_{\boldsymbol{\alpha}}\lambda^{\ell+1}\frac{1-\lambda^{2(h-\ell-1)}}{1-\lambda^{2}}. \end{aligned}$$

Similarly, we have

$$\begin{aligned} V\Big[\boldsymbol{\delta}_{\boldsymbol{\alpha}}^{\top}\sum_{\ell=0}^{h-2}\mathbf{G}_{\boldsymbol{\alpha}}^{\ell}\boldsymbol{w}_{\boldsymbol{\alpha}}\sum_{i=0}^{h-\ell-2}\lambda^{i}\varepsilon_{t+h-\ell-1-i}\Big] &= \sum_{\ell=0}^{h-2}\left(\boldsymbol{\delta}_{\boldsymbol{\alpha}}^{\top}\mathbf{G}_{\boldsymbol{\alpha}}^{\ell}\boldsymbol{w}_{\boldsymbol{\alpha}}\right)^{2}\sigma_{\varepsilon}^{2}\sum_{i=0}^{h-\ell-2}\left(\lambda^{2}\right)^{i} \\ &= \sigma_{\varepsilon}^{2}\sum_{\ell=0}^{h-2}\left(\boldsymbol{\delta}_{\boldsymbol{\alpha}}^{\top}\mathbf{G}_{\boldsymbol{\alpha}}^{\ell}\boldsymbol{w}_{\boldsymbol{\alpha}}\right)^{2}\frac{1-\lambda^{2(h-\ell-1)}}{1-\lambda^{2}} \end{aligned}$$

and $\quad V\Big[\sum_{j=0}^{h-1}\lambda^{j}\varepsilon_{t+h-j}\Big] = \sigma_{\varepsilon}^{2}\sum_{j=0}^{h-1}\left(\lambda^{2}\right)^{j} = \sigma_{\varepsilon}^{2}\frac{1-\lambda^{2h}}{1-\lambda^{2}}.$

Collecting the latter three results and observing that in the conditional distribution under $\mathcal{H}_t$ the quantities $\boldsymbol{u}_t$, $\boldsymbol{\beta}^{\top}\boldsymbol{x}_{t+h}$ and ξ_t are constants, we obtain from (6.41) that

$$\begin{aligned} V[Y_{t+h}|\mathcal{H}_t] &= V\Big[\boldsymbol{\delta}_{\boldsymbol{\alpha}}^{\top}\sum_{\ell=0}^{h-2}\mathbf{G}_{\boldsymbol{\alpha}}^{\ell}\boldsymbol{w}_{\boldsymbol{\alpha}}\sum_{i=0}^{h-\ell-2}\lambda^{i}\varepsilon_{t+h-\ell-1-i}+\sum_{j=0}^{h-1}\lambda^{j}\varepsilon_{t+h-j}\Big] \\ &= \sigma_{\varepsilon}^{2}\left(2\sum_{\ell=0}^{h-2}\boldsymbol{\delta}_{\boldsymbol{\alpha}}^{\top}\mathbf{G}_{\boldsymbol{\alpha}}^{\ell}\boldsymbol{w}_{\boldsymbol{\alpha}}\lambda^{\ell+1}\frac{1-\lambda^{2(h-\ell-1)}}{1-\lambda^{2}}+\sum_{\ell=0}^{h-2}\left(\boldsymbol{\delta}_{\boldsymbol{\alpha}}^{\top}\mathbf{G}_{\boldsymbol{\alpha}}^{\ell}\boldsymbol{w}_{\boldsymbol{\alpha}}\right)^{2}\frac{1-\lambda^{2(h-\ell-1)}}{1-\lambda^{2}}\right. \\ &\qquad \left. +\frac{1-\lambda^{2h}}{1-\lambda^{2}}\right), \end{aligned}$$

i. e. formula (6.14).

6.A.3 Proof of Proposition 6.6

To prove Proposition 6.6, we make use of the following proposition.

Proposition 6.18 (Recurrence Relation for Partially Linear SSOEs)**.** *Consider the partially linear SSOE from Definition 6.5, and let $h > 0$. Then we have*

$$\boldsymbol{u}_{t+h,ne} = \sum_{\ell=0}^{h-1}\mathbf{G}_{\boldsymbol{\alpha},ne}^{\ell}\boldsymbol{w}_{\boldsymbol{\alpha},ne}\frac{\xi_{t+h-\ell}}{\prod_{i=1}^{m}e_{i,t-d_i+h-\ell}}+\mathbf{G}_{\boldsymbol{\alpha},ne}^{h}\boldsymbol{u}_{t,ne}, \tag{6.43}$$

$$\begin{aligned} Y_{t+h} &= \boldsymbol{\delta}_{\boldsymbol{\alpha},ne}^{\top}\left(\sum_{\ell=0}^{h-2}\mathbf{G}_{\boldsymbol{\alpha},ne}^{\ell}\boldsymbol{w}_{\boldsymbol{\alpha},ne}\frac{\xi_{t+h-\ell-1}}{\prod_{i=1}^{m}e_{i,t-d_i+h-\ell-1}}+\mathbf{G}_{\boldsymbol{\alpha},ne}^{h-1}\boldsymbol{u}_{t,ne}\right)\prod_{i=1}^{m}e_{i,t+h-d_i} \\ &\qquad +\boldsymbol{\beta}^{\top}\boldsymbol{x}_{t+h}+\xi_{t+h}. \end{aligned} \tag{6.44}$$

Proof. To prove (6.43), we use induction by h. For $h = 1$, the right-hand side of (6.43) amounts to $\boldsymbol{w}_{\boldsymbol{\alpha},ne}\xi_{t+1}/\prod_{i=1}^{m}e_{i,t-d_i+1}+\mathbf{G}_{\boldsymbol{\alpha},ne}\boldsymbol{u}_{t,ne}$, which equals $\boldsymbol{u}_{t+h,ne} = \boldsymbol{u}_{t+1,ne}$ by Eq. (6.16).

Let (6.43) be proven for $h > 0$. Then we have for $h + 1$ by Eq. (6.16)

$$\begin{aligned}
\boldsymbol{u}_{t+h+1,ne} &= \mathbf{G}_{\boldsymbol{\alpha},ne}\boldsymbol{u}_{t+h,ne} + \boldsymbol{w}_{\boldsymbol{\alpha},ne}\frac{\xi_{t+h+1}}{\prod_{i=1}^{m} e_{i,t-d_i+h+1}} \\
&= \mathbf{G}_{\boldsymbol{\alpha},ne}\left(\sum_{\ell=0}^{h-1}\mathbf{G}_{\boldsymbol{\alpha},ne}^{\ell}\boldsymbol{w}_{\boldsymbol{\alpha},ne}\frac{\xi_{t+h-\ell}}{\prod_{i=1}^{m} e_{i,t-d_i+h-\ell}} + \mathbf{G}_{\boldsymbol{\alpha},ne}^{h}\boldsymbol{u}_{t,ne}\right) \\
&\qquad + \boldsymbol{w}_{\boldsymbol{\alpha},ne}\frac{\xi_{t+h+1}}{\prod_{i=1}^{m} e_{i,t-d_i+h+1}} \\
&= \sum_{\ell=0}^{h-1}\mathbf{G}_{\boldsymbol{\alpha},ne}^{\ell+1}\boldsymbol{w}_{\boldsymbol{\alpha},ne}\frac{\xi_{t+h-\ell}}{\prod_{i=1}^{m} e_{i,t-d_i+h-\ell}} + \mathbf{G}_{\boldsymbol{\alpha},ne}^{h+1}\boldsymbol{u}_{t,ne} \\
&\qquad + \mathbf{G}_{\boldsymbol{\alpha},ne}^{0}\boldsymbol{w}_{\boldsymbol{\alpha},ne}\frac{\xi_{t+h+1}}{\prod_{i=1}^{m} e_{i,t-d_i+h+1}} \\
&= \sum_{k=1}^{h}\mathbf{G}_{\boldsymbol{\alpha},ne}^{k}\boldsymbol{w}_{\boldsymbol{\alpha},ne}\frac{\xi_{t+h+1-k}}{\prod_{i=1}^{m} e_{i,t-d_i+h+1-k}} + \mathbf{G}_{\boldsymbol{\alpha},ne}^{0}\boldsymbol{w}_{\boldsymbol{\alpha},ne}\frac{\xi_{t+h+1}}{\prod_{i=1}^{m} e_{i,t-d_i+h+1}} \\
&\qquad + \mathbf{G}_{\boldsymbol{\alpha},ne}^{h+1}\boldsymbol{u}_{t,ne} \\
&= \sum_{k=0}^{(h+1)-1}\mathbf{G}_{\boldsymbol{\alpha},ne}^{k}\boldsymbol{w}_{\boldsymbol{\alpha},ne}\frac{\xi_{t+h+1-k}}{\prod_{i=1}^{m} e_{i,t-d_i+h+1-k}} + \mathbf{G}_{\boldsymbol{\alpha},ne}^{h+1}\boldsymbol{u}_{t,ne}.
\end{aligned}$$

Equation (6.44) is a direct consequence of the observation equation (6.15) and Eq. (6.43).

□

We now proceed to the proof of Proposition 6.6.

From the properties of conditional expectation we have $E[\boldsymbol{u}_t^m|\mathcal{H}_t] = \boldsymbol{u}_t^m$ for all $m \in \mathbb{N}$. By Proposition 6.18, $\boldsymbol{u}_{t+h}$ depends on the process history $\mathcal{H}_t$ up to time t only through $\boldsymbol{u}_t$, and on $\xi_{t+1}, \ldots, \xi_{t+h}$. Because of the independence assumption on the errors, the assumption $\boldsymbol{u}_t = \boldsymbol{u}(t, \xi_t, \xi_{t-1}, ...)$, see Section 6.3, and $E[\boldsymbol{u}_t|\mathcal{H}_t] = \boldsymbol{u}_t$, we obtain

$$\begin{aligned}
E[\boldsymbol{u}_{t+h,ne}|\mathcal{H}_t] &= E\left[\sum_{\ell=0}^{h-1}\mathbf{G}_{\boldsymbol{\alpha},ne}^{\ell}\boldsymbol{w}_{\boldsymbol{\alpha},ne}\frac{\xi_{t+h-\ell}}{\prod_{i=1}^{m} e_{i,t-d_i+h-\ell}} + \mathbf{G}_{\boldsymbol{\alpha},ne}^{h}\boldsymbol{u}_{t,ne}\middle|\mathcal{H}_t\right] \\
&= E\left[\sum_{\ell=0}^{h-1}\mathbf{G}_{\boldsymbol{\alpha},ne}^{\ell}\boldsymbol{w}_{\boldsymbol{\alpha},ne}\frac{\xi_{t+h-\ell}}{\prod_{i=1}^{m} e_{i,t-d_i+h-\ell}}\middle|\mathcal{H}_t\right] + E\left[\mathbf{G}_{\boldsymbol{\alpha},ne}^{h}\boldsymbol{u}_{t,ne}|\mathcal{H}_t\right] \\
&= \mathbf{G}_{\boldsymbol{\alpha},ne}^{h}\boldsymbol{u}_{t,ne}.
\end{aligned}$$

With Eqs. (6.36) and (6.44) we obtain Eq. (6.18).

Hence we obtain from the recurrence relation (6.44) and under the assumptions from the proposition

$$\begin{aligned}
V[Y_{t+h}|\mathcal{H}_t] &= V\Big[\boldsymbol{\delta}^\top_{\boldsymbol{\alpha},ne}\Big(\sum_{\ell=0}^{h-2}\mathbf{G}^{\ell}_{\boldsymbol{\alpha},ne}\boldsymbol{w}_{\boldsymbol{\alpha},ne}\frac{\xi_{t+h-\ell-1}}{\prod_{i=1}^{m} e_{i,t-d_i+h-\ell-1}} + \mathbf{G}^{h-1}_{\boldsymbol{\alpha},ne}\boldsymbol{u}_{t,ne}\Big)\prod_{i=1}^{m} e_{i,t+h-d_i} \\
&\qquad + \boldsymbol{\beta}^\top\boldsymbol{x}_{t+h} + \xi_{t+h}\Big] \\
&= V\Big[\boldsymbol{\delta}^\top_{\boldsymbol{\alpha},ne}\sum_{\ell=0}^{h-2}\mathbf{G}^{\ell}_{\boldsymbol{\alpha},ne}\boldsymbol{w}_{\boldsymbol{\alpha},ne}\frac{\xi_{t+h-\ell-1}}{\prod_{i=1}^{m} e_{i,t-d_i+h-\ell-1}}\prod_{i=1}^{m} e_{i,t+h-d_i}\Big] \\
&\qquad + V\Big[\boldsymbol{\delta}^\top_{\boldsymbol{\alpha},ne}\mathbf{G}^{h-1}_{\boldsymbol{\alpha},ne}\boldsymbol{u}_{t,ne}\prod_{i=1}^{m} e_{i,t+h-d_i}|\mathcal{H}_t\Big] + V[\xi_{t+h}] \\
&= \sum_{\ell=0}^{h-2}\left(\boldsymbol{\delta}^\top_{\boldsymbol{\alpha},ne}\mathbf{G}^{\ell}_{\boldsymbol{\alpha},ne}\boldsymbol{w}_{\boldsymbol{\alpha},ne}\right)^2\left(\frac{\prod_{i=1}^{m} e_{i,t+h-d_i}}{\prod_{i=1}^{m} e_{i,t-d_i+h-\ell-1}}\right)^2\sigma^2_{\xi_{t+h-\ell-1}} + \sigma^2_{\xi_{t+h}},
\end{aligned}$$

i. e. Eq. (6.19).

□

6.A.4 Proof of Proposition 6.11

(a) We prove the assertion by induction in t. By assumption we have $\sum_{k=1}^{d_i}\widetilde{e}^{(k)}_{i,0} = 0$ for $i = 1, \dots, m$, and hence the assertion is valid for $s = 0$. The assumption is valid for $s = 1$ because

$$\begin{aligned}
\sum_{k=1}^{d_i}\widetilde{e}^{(k)}_{i,1} &= \sum_{k=1}^{d_i-1}\widetilde{e}^{(k)}_{i,1} + \widetilde{e}^{(d_i)}_{i,1} = \sum_{k=1}^{d_i-1}\left(\widetilde{e}^{(k+1)}_{i,0} - r_{i,1}\right) + \widetilde{e}^{(1)}_{i,0} + (1-\alpha_1)\alpha_{i,3}\widetilde{\xi}_1 - r_{i,1} \\
&= \sum_{k=1}^{d_i}\widetilde{e}^{(k)}_{i,0} - (d_i-1)r_{i,1} - r_{i,1} + \underbrace{(1-\alpha_1)\alpha_{i,3}\widetilde{\xi}_1}_{d_i r_{i,1}} \\
&= \underbrace{\sum_{k=1}^{d_i}\widetilde{e}^{(k)}_{i,0}}_{=0 \text{ by assumption}} \quad \underbrace{-d_i r_{i,1} + d_i r_{i,1}}_{=0} = 0.
\end{aligned}$$

We assume that the assertion is true for $s = 2, \dots, t$, i. e. $\sum_{k=1}^{d_i}\widetilde{e}^{(k)}_{i,s} = 0$. Then for time point $s = t+1$ and season $i = 1, \dots, m$:

$$\begin{aligned}
\sum_{k=1}^{d_i}\widetilde{e}^{(k)}_{i,t+1} &= \widetilde{e}^{(d_i)}_{i,t+1} + \sum_{k=1}^{d_i-1}\widetilde{e}^{(k)}_{i,t+1} \\
&= \widetilde{e}^{(1)}_{i,t} + (1-\alpha_1)\alpha_{i,3}\widetilde{\xi}_{t+1} - r_{i,t+1} + \sum_{k=1}^{d_i-1}\left(\widetilde{e}^{(k+1)}_{i,t} - r_{i,t+1}\right) \\
&= \widetilde{e}^{(1)}_{i,t} + (1-\alpha_1)\alpha_{i,3}\widetilde{\xi}_{t+1} - r_{i,t+1} - (d_i-1)r_{i,t+1} + \sum_{k=1}^{d_i-1}\widetilde{e}^{(k+1)}_{i,t}
\end{aligned}$$

$$\begin{aligned} &= \widetilde{e}^{(1)}_{i,t} + \sum_{k=1}^{d_i-1} \widetilde{e}^{(k+1)}_{i,t} + (1-\alpha_1)\alpha_{i,3}\widetilde{\xi}_{t+1} - d_i r_{i,t+1} \\ &= \underbrace{\sum_{k=1}^{d_i} \widetilde{e}^{(k)}_{i,t}}_{=0} + \underbrace{(1-\alpha_1)\alpha_{i,3}\widetilde{\xi}_{t+1} - d_i(1-\alpha_1)\alpha_{i,3}\widetilde{\xi}_{t+1}\cdot\frac{1}{d_i}}_{=0} \quad = \quad 0. \end{aligned}$$

Consequently, the assertion is also true for $s = t+1$ and the proof of (a) is complete.

For the proof of (b) we use induction in t. By definition we have $R_{i,0} = \frac{1}{d_i}\sum_{k=1}^{d_i} e^{(k)}_{i,0} = 0$ for $i = 1, \dots, m$. Hence, the assumption is valid for $s = 0$. For $R_{i,1}$ we have

$$\begin{aligned} R_{i,1} &= \frac{1}{d_i}\sum_{k=1}^{d_i} e^{(k)}_{i,1} = \frac{1}{d_i}\left(\sum_{k=1}^{d_i-1} e^{(k+1)}_{i,0} + e^{(1)}_{i,0} + (1-\alpha_1)\alpha_{i,3}\xi_1\right) \\ &= \underbrace{\frac{1}{d_i}\sum_{k=1}^{d_i} e^{(k)}_{i,0}}_{=R_{i,0}} + \underbrace{\frac{1}{d_i}(1-\alpha_1)\alpha_{i,3}\xi_1}_{=r_{i,1}}, \end{aligned}$$

and the assumption is valid for time point $s = 1$. In particular, since $\frac{1}{d_i}\sum_{k=1}^{d_i} e^{(k)}_{i,0} = 0$ by definition, we have $R_{i,0} = 0$ and $R_{i,1} = r_{i,1}$.

By $\widetilde{\mu}_0 = \mu_0, \widetilde{\Delta}_0 = \Delta_0$, and $\widetilde{e}^{(k)}_{i,0} = e^{(k)}_{i,0}$ for $k = 1, \dots, d_i$, we obtain $\widetilde{Y}_{1|0} = Y_{1|0}$, and consequently $\widetilde{\xi}_1 = Y_1 - \widetilde{Y}_{1|0} = Y_1 - Y_{1|0} = \xi_1$. Using these relations and $R_{i,1} = r_{i,1}$, assumption (c) is valid for time point $s = 1$ because

$$\begin{aligned} \widetilde{\mu}_1 &= \widetilde{\mu}_0 + \phi\widetilde{\Delta}_0 + \alpha_1\widetilde{\xi}_1 + \sum_{i=1}^{m} r_{i,1} \;=\; \mu_0 + \phi\Delta_0 + \alpha_1\xi_1 + \sum_{i=1}^{m} R_{i,1} \;=\; \mu_1 + \sum_{i=1}^{m} R_{i,1}, \\ \widetilde{\Delta}_1 &= \phi\widetilde{\Delta}_0 + \alpha_1\alpha_2\widetilde{\xi}_1 \;=\; \phi\Delta_0 + \alpha_1\alpha_2\xi_1 \;=\; \Delta_1, \\ \widetilde{e}^{(d_i)}_{i,1} &= \widetilde{e}^{(1)}_{i,0} + (1-\alpha_1)\alpha_{i,3}\widetilde{\xi}_1 - r_{i,1} = e^{(1)}_{i,0} + (1-\alpha_1)\alpha_{i,3}\xi_1 - r_{i,1} \\ &= e^{(d_i)}_{i,1} - r_{i,1} = e^{(d_i)}_{i,1} - R_{i,1}, \\ \widetilde{e}^{(k)}_{i,1} &= \widetilde{e}^{(k+1)}_{i,0} - r_{i,1} \;=\; e^{(k+1)}_{i,0} - R_{i,1}. \end{aligned}$$

Furthermore,

$$\begin{aligned} \widetilde{Y}_{h|0} &= \widetilde{\mu}_0 + \sum_{j=1}^{h}\phi^j\widetilde{\Delta}_0 + \sum_{i=1}^{m}\widetilde{e}^{(h)}_{i,0} + \boldsymbol{\beta}^\top\boldsymbol{x}_h \\ &= \mu_0 + \sum_{j=1}^{h}\phi^j\Delta_0 + \sum_{i=1}^{m} e^{(h)}_{i,0} + \boldsymbol{\beta}^\top\boldsymbol{x}_h = Y_{h|0}, \end{aligned}$$

and assumption (d) is valid for time point 0.

Assume that (b)–(d) are valid for time points $s = 1, \dots, t$. Then we have

$$\begin{aligned}
\widetilde{Y}_{t+1|t} &= \widetilde{\mu}_t + \phi\widetilde{\Delta}_t + \sum_{i=1}^{m} \widetilde{e}_{i,t}^{(1)} + \boldsymbol{\beta}^\top \boldsymbol{x}_{t+1} \\
&= \mu_t + \sum_{i=1}^{m} R_{i,t} + \phi\Delta_t + \sum_{i=1}^{m} e_{i,t}^{(1)} - \sum_{i=1}^{m} R_{i,t} + \boldsymbol{\beta}^\top \boldsymbol{x}_{t+1} \\
&= \mu_t + \phi\Delta_t + \sum_{i=1}^{m} e_{i,t}^{(1)} + \boldsymbol{\beta}^\top \boldsymbol{x}_{t+1} = Y_{t+1|t}
\end{aligned}$$

and hence $\widetilde{\xi}_{t+1} = \widetilde{Y}_{t+1|t} - Y_{t+1} = Y_{t+1|t} - Y_{t+1} = \xi_{t+1}$. Then for time point $s = t+1$ and $i = 1, \dots, m$ we have

$$\begin{aligned}
R_{i,t+1} &= \frac{1}{d_i} \sum_{k=1}^{d_i} e_{i,t+1}^{(k)} \\
&= \frac{1}{d_i} \left(\sum_{k=1}^{d_i-1} e_{i,t}^{(k+1)} + e_{i,t}^{(1)} + (1-\alpha_1)\alpha_{i,3}\xi_{t+1} \right) \\
&\overset{\widetilde{\xi}_{t+1}=\xi_{t+1}}{=} \frac{1}{d_i} \left(\sum_{k=2}^{d_i} e_{i,t}^{(k)} + e_{i,t}^{(1)} + (1-\alpha_1)\alpha_{i,3}\widetilde{\xi}_{t+1} \right) \\
&= \underbrace{\frac{1}{d_i} \sum_{k=1}^{d_i} e_{i,t}^{(k)}}_{=R_{i,t}} + \underbrace{\frac{1}{d_i}(1-\alpha_1)\alpha_{i,3}\widetilde{\xi}_{t+1}}_{=r_{i,t+1}}
\end{aligned}$$

and (b) is valid for time point $s = t+1$. This proves (b).

Under the assumption that (c) is valid for time points $s = 1, \dots, t$, we have for $i = 1, \dots, m$

$$\begin{aligned}
\widetilde{\mu}_{t+1} &= \widetilde{\mu}_t + \phi\widetilde{\Delta}_t + \alpha_1\widetilde{\xi}_{t+1} + \sum_{i=1}^{m} r_{i,t+1} \\
&\overset{\widetilde{\xi}_{t+1}=\xi_{t+1}}{=} \mu_t + \sum_{i=1}^{m} R_{i,t} + \phi\Delta_t + \alpha_1\xi_{t+1} + \sum_{i=1}^{m} r_{i,t+1} \\
&\overset{(b)}{=} \underbrace{\mu_t + \phi\Delta_t + \alpha_1\xi_{t+1}}_{\mu_{t+1}} + \sum_{i=1}^{m} R_{i,t+1}. \\
\widetilde{\Delta}_{t+1} &= \phi\widetilde{\Delta}_t + \alpha_1\alpha_2\widetilde{\xi}_{t+1} \overset{\widetilde{\xi}_{t+1}=\xi_{t+1}}{=} \phi\Delta_t + \alpha_1\alpha_2\xi_{t+1} = \Delta_{t+1}.
\end{aligned}$$

$$\begin{aligned}
\widetilde{e}^{(d_i)}_{i,t+1} &= \widetilde{e}^{(1)}_{i,t} + (1-\alpha_1)\alpha_{i,3}\widetilde{\xi}_{t+1} - r_{i,t+1} \\
&= e^{(1)}_{i,t} - R_{i,t} + (1-\alpha_1)\alpha_{i,3}\widetilde{\xi}_{t+1} - r_{i,t+1} \\
&\overset{\widetilde{\xi}_{t+1}=\xi_{t+1}}{=} e^{(1)}_{i,t} + (1-\alpha_1)\alpha_{i,3}\xi_{t+1} - \underbrace{(R_{i,t} + r_{i,t+1})}_{=R_{i,t+1}} \;=\; e^{(d_i)}_{i,t+1} - R_{i,t+1}, \\
\widetilde{e}^{(k)}_{i,t+1} &= \widetilde{e}^{(k+1)}_{i,t} - r_{i,t+1} = e^{(k)}_{i,t} - R_{i,t} - r_{i,t+1} \;=\; e^{(k)}_{i,t} - R_{i,t+1}.
\end{aligned}$$

This completes the proof of (c).

The proof of assertion (d) uses (c):

$$\begin{aligned}
\widetilde{Y}_{t+h|t} &= \widetilde{\mu}_t + \sum_{j=1}^{h}\phi^j\widetilde{\Delta}_t + \sum_{i=1}^{m}\widetilde{e}^{(h)}_{i,t} + \boldsymbol{\beta}^\top\boldsymbol{x}_{t+h} \\
&= \mu_t + \sum_{i=1}^{m} R_{i,t} + \sum_{j=1}^{h}\phi^j\Delta_t + \sum_{i=1}^{m}\left(e^{(h)}_{i,t} - R_{i,t}\right) + \boldsymbol{\beta}^\top\boldsymbol{x}_{t+h} \\
&= \mu_t + \sum_{i=1}^{m} R_{i,t} + \sum_{j=1}^{h}\phi^j\Delta_t + \sum_{i=1}^{m} e^{(h)}_{i,t} - \sum_{i=1}^{m} R_{i,t} + \boldsymbol{\beta}^\top\boldsymbol{x}_{t+h} \quad = \quad Y_{t+h|t}.
\end{aligned}$$

The proof of Proposition 6.11 is complete.

6.A.5 Proof of Proposition 6.14

(a) We prove the assertion by induction in t. By assumption we have $\sum_{k=1}^{d_i}\widetilde{e}^{(k)}_{i,0} = d_i$ for $i = 1,\dots,m$, and hence the assertion is valid for $s = 0$. The assumption is valid for $s = 1$ because

$$\begin{aligned}
\sum_{k=1}^{d_i}\widetilde{e}^{(k)}_{i,1} &= \sum_{k=1}^{d_i-1}\widetilde{e}^{(k)}_{i,1} + \widetilde{e}^{(d_i)}_{i,1} \\
&= \sum_{k=1}^{d_i-1}\widetilde{e}^{(k+1)}_{i,0}\Big/ r_{i,1} + \left(\widetilde{e}^{(1)}_{i,0} + \frac{(1-\alpha_1)\alpha_{i,3}\widetilde{\xi}_1}{\left(\widetilde{\mu}_0 + \phi\widetilde{\Delta}_0\right)}\right)\Big/ r_{i,t} \\
&= \sum_{k=2}^{d_i}\widetilde{e}^{(k)}_{i,0}\Big/ r_{i,1} + \frac{\widetilde{e}^{(1)}_{i,0}}{r_{i,t}} + \frac{r_{i,t}(1-\alpha_1)\alpha_{i,3}\widetilde{\xi}_1}{\left(\widetilde{\mu}_0 + \phi\widetilde{\Delta}_0\right)} \\
&= \left(\underbrace{\sum_{k=1}^{d_i}\widetilde{e}^{(k)}_{i,0}}_{=d_i} + \frac{(1-\alpha_1)\alpha_{i,3}\widetilde{\xi}_1}{\left(\widetilde{\mu}_0 + \phi\widetilde{\Delta}_0\right)}\right)\Big/ r_{i,t} \;=\; \frac{\left(d_i + \frac{(1-\alpha_1)\alpha_{i,3}\widetilde{\xi}_1}{\widetilde{\mu}_0+\phi\widetilde{\Delta}_0}\right)}{1 + \frac{(1-\alpha_1)\alpha_{i,3}\widetilde{\xi}_1}{d_i\left(\widetilde{\mu}_0+\phi\widetilde{\Delta}_0\right)}} \;=\; d_i.
\end{aligned}$$

We assume that the assertion is true for $s = 2,\dots,t$, i. e. $\sum_{k=1}^{d_i}\widetilde{e}^{(k)}_{i,s} = d_i$. Then for time point $s = t+1$ and season $i = 1, \ldots, m$:

$$\begin{aligned}
\sum_{k=1}^{d_i} \widetilde{e}_{i,t+1}^{(k)} &= \widetilde{e}_{i,t+1}^{(d_i)} + \sum_{k=1}^{d_i-1} \widetilde{e}_{i,t+1}^{(k)} \\
&= \left(\widetilde{e}_{i,t}^{(1)} + \frac{(1-\alpha_1)\alpha_{i,3}\widetilde{\xi}_{t+1}}{\widetilde{\mu}_{t-1} + \phi\widetilde{\Delta}_{t-1}} \right) \Big/ r_{i,t+1} + \sum_{k=1}^{d_i-1} \left(\widetilde{e}_{i,t}^{(k+1)} \Big/ r_{i,t+1} \right) \\
&= \left(\widetilde{e}_{i,t}^{(1)} + \frac{(1-\alpha_1)\alpha_{i,3}\widetilde{\xi}_{t+1}}{\widetilde{\mu}_{t-1} + \phi\widetilde{\Delta}_{t-1}} + \sum_{k=2}^{d_i} \widetilde{e}_{i,t}^{(k)} \right) \Big/ r_{i,t+1} \\
&= \left(\sum_{k=1}^{d_i} \widetilde{e}_{i,t}^{(k)} + \frac{(1-\alpha_1)\alpha_{i,3}\widetilde{\xi}_{t+1}}{\widetilde{\mu}_{t-1} + \phi\widetilde{\Delta}_{t-1}} \right) \Big/ r_{i,t+1} \\
&= \frac{d_i + \frac{(1-\alpha_1)\alpha_{i,3}\widetilde{\xi}_{t+1}}{\widetilde{\mu}_{t-1}+\phi\widetilde{\Delta}_{t-1}}}{1 + \frac{(1-\alpha_1)\alpha_{i,3}\widetilde{\xi}_{t+1}}{d_i\left(\widetilde{\mu}_{t-1}+\phi\widetilde{\Delta}_{t-1}\right)}} \quad = \quad d_i.
\end{aligned}$$

Consequently, the assertion is also true for $s = t+1$ and the proof of (a) is complete.

For the proof of (b) we use induction in t. By definition we have $R_{1,0} = \frac{1}{d_1}\sum_{k=1}^{d_1} e_{1,0}^{(k)} = 1$. For $R_{1,1}$ we have

$$\begin{aligned}
R_{1,1} &= \frac{1}{d_1}\sum_{k=1}^{d_1} e_{1,1}^{(k)} = \frac{1}{d_1}\left(\sum_{k=1}^{d_1-1} e_{1,0}^{(k+1)} + e_{1,0}^{(1)} + \frac{(1-\alpha_1)\alpha_{1,3}\xi_1}{\mu_0 + \phi\Delta_0} \right) \\
&= \underbrace{\frac{1}{d_1}\sum_{k=1}^{d_1} e_{1,0}^{(k)}}_{=1} + \frac{1}{d_1}\frac{(1-\alpha_1)\alpha_{1,3}\xi_1}{\mu_0 + \phi\Delta_0} \;=\; r_{1,0} \;=\; R_{1,0} r_{1,1},
\end{aligned}$$

and the assumption is valid for time point $s = 1$. In particular, since $\frac{1}{d_1}\sum_{k=1}^{d_1} e_{1,0}^{(k)} = 1$ by definition, we have $R_{1,0} = 1$ and $R_{1,1} = r_{1,1}$.

By $\widetilde{\mu}_0 = \mu_0, \widetilde{\Delta}_0 = \Delta_0$, and $\widetilde{e}_{1,0}^{(k)} = e_{1,0}^{(k)}$ for $k = 1, \ldots, d_1$, we obtain $\widetilde{Y}_{1|0} = Y_{1|0}$, and consequently $\widetilde{\xi}_1 = Y_1 - \widetilde{Y}_{1|0} = Y_1 - Y_{1|0} = \xi_1$. Using these relations and $R_{1,1} = r_{1,1}$, assumption (c) is valid for time point $s = 1$ because

$$\begin{aligned}
\widetilde{\mu}_1 &= \left(\widetilde{\mu}_0 + \phi\widetilde{\Delta}_0 + \frac{\alpha_1\widetilde{\xi}_1}{\widetilde{e}_0^{(1)}} \right) r_{1,1} \;=\; \left(\mu_0 + \phi\Delta_0 + \frac{\alpha_1\xi_1}{e_{1,0}^{(1)}} \right) R_{1,1} \;=\; \mu_1 R_{1,1}, \\
\widetilde{\Delta}_1 &= \left(\phi\widetilde{\Delta}_0 + \frac{\alpha_1\alpha_2\widetilde{\xi}_1}{\widetilde{\alpha}_0^{(1)}} \right) r_{1,1} \;=\; \left(\phi\Delta_0 + \frac{\alpha_1\alpha_2\xi_1}{e_{1,0}^{(1)}} \right) R_{1,1} \;=\; \Delta_1 R_{1,1}, \\
\widetilde{e}_{1,1}^{(d_1)} &= \left(\widetilde{e}_{1,0}^{(1)} + \frac{(1-\alpha_1)\alpha_{1,3}\widetilde{\xi}_1}{\widetilde{\mu}_0 + \phi\widetilde{\Delta}_0} \right) \Big/ r_{1,1} = \left(e_{1,0}^{(1)} + \frac{(1-\alpha_1)\alpha_{1,3}\xi_1}{\mu_0 + \phi\Delta_0} \right) \Big/ R_{1,1} \;=\; e_{1,1}^{(d_1)}/R_{1,1}, \\
\widetilde{e}_{1,1}^{(k)} &= \widetilde{e}_{1,0}^{(k+1)}/r_{1,1} \;=\; e_{1,0}^{(k+1)}/R_{1,1}.
\end{aligned}$$

Furthermore,

$$
\begin{aligned}
\widetilde{Y}_{1+h|1} &= \left(\widetilde{\mu}_1 + \sum_{j=1}^{h} \phi^j \widetilde{\Delta}_1\right) \widetilde{e}_{1,1}^{(h)} + \boldsymbol{\beta}^\top \boldsymbol{x}_{1+h} \\
&= \left(\mu_1 R_{1,1} + \sum_{j=1}^{h} \phi^j \Delta_1 R_{1,1}\right) e_{1,1}^{(h)}/R_{1,1} + \boldsymbol{\beta}^\top \boldsymbol{x}_{1+h} \\
&= \left(\mu_1 + \sum_{j=1}^{h} \phi^j \Delta_1\right) e_{1,1}^{(h)} + \boldsymbol{\beta}^\top \boldsymbol{x}_{1+h} \quad = \quad Y_{1+h|1},
\end{aligned}
$$

and assumption (d) is valid for time point $s = 0$.

Assume that (b)–(d) are valid for time points $s = 1, \ldots, t$. Then we have

$$
\begin{aligned}
\widetilde{Y}_{t+1|t} &= \left(\widetilde{\mu}_t + \phi \widetilde{\Delta}_t\right) \widetilde{e}_{1,t}^{(1)} + \boldsymbol{\beta}^\top \boldsymbol{x}_{t+1} \quad = \quad \left(\mu_t R_{1,t} + \phi \Delta_t \prod_{i=1}^{m} R_{i,t}\right) e_{1,t}^{(1)}/R_{1,t} + \boldsymbol{\beta}^\top \boldsymbol{x}_{t+1} \\
&= \left(\mu_t + \phi \Delta_t\right) e_{1,t}^{(1)} + \boldsymbol{\beta}^\top \boldsymbol{x}_{t+1} = Y_{t+1|t}
\end{aligned}
$$

and hence $\widetilde{\xi}_{t+1} = \widetilde{Y}_{t+1|t} - Y_{t+1} = Y_{t+1|t} - Y_{t+1} = \xi_{t+1}$.

Then for time point $s = t+1$ and $i = 1, \ldots, m$ we have

$$
\begin{aligned}
R_{1,t+1} &= \frac{1}{d_1} \sum_{k=1}^{d_1} e_{1,t+1}^{(k)} = \frac{1}{d_1} \left(\sum_{k=1}^{d_1 - 1} e_{1,t}^{(k+1)} + e_{1,t}^{(1)} + \frac{(1-\alpha_1)\alpha_{1,3}\xi_{t+1}}{\mu_t + \phi \Delta_t} \right) \\
&\overset{\widetilde{\xi}_{t+1} = \xi_{t+1}, (b)}{=} \frac{1}{d_1} \left(\sum_{k=2}^{d_1} e_{1,t}^{(k)} + e_{1,t}^{(1)} + \frac{(1-\alpha_1)\alpha_{1,3}\widetilde{\xi}_{t+1}}{\widetilde{\mu}_t / R_{1,t} + \phi \widetilde{\Delta}_t / R_{1,t}} \right) \\
&= \underbrace{\frac{1}{d_1} \sum_{k=1}^{d_1} e_{1,t}^{(k)}}_{= R_{1,t}} + R_{1,t} \frac{(1-\alpha_1)\alpha_{1,3}\widetilde{\xi}_{t+1}}{d_1(\mu_t + \phi \Delta_t)} \quad = \quad R_{1,t} r_{1,t+1}.
\end{aligned}
$$

Therefore, (b) is valid for time point $s = t+1$. This proves (b).

Under the assumption that (c) is valid for time points $s = 1, \dots, t$, we have for $m = 1$

$$
\begin{aligned}
\widetilde{\mu}_{t+1} &= \left(\widetilde{\mu}_t + \phi \widetilde{\Delta}_t + \frac{\alpha_1 \widetilde{\xi}_{t+1}}{\widetilde{e}^{(1)}_{1,t}} \right) r_{1,t+1} \\
&\overset{\widetilde{\xi}_{t+1} = \xi_{t+1}}{=} \left(\mu_t R_{1,t} + \phi \Delta_t R_{1,t} + \frac{\alpha_1 \xi_{t+1}}{e^{(1)}_{1,t}/R_{1,t}} \right) r_{1,t+1} \\
&\overset{(b)}{=} \left(\mu_t + \phi \Delta_t + \frac{\alpha_1 \xi_{t+1}}{e^{(1)}_{1,t}} \right) R_{1,t} r_{1,t+1} \;=\; \mu_{t+1} R_{1,t+1}. \\
\widetilde{\Delta}_{t+1} &= \left(\phi \widetilde{\Delta}_t + \frac{\alpha_1 \alpha_2 \widetilde{\xi}_{t+1}}{\widetilde{e}^{(1)}_{1,t}} \right) r_{1,t+1} \overset{\widetilde{\xi}_{t+1} = \xi_{t+1}}{=} \left(\phi \Delta_t R_{1,t} + \frac{\alpha_1 \alpha_2 \xi_{t+1}}{e^{(1)}_{1,t}/R_{1,t}} \right) r_{1,t+1} \\
&= \Delta_{t+1} R_{1,t} r_{1,t+1} \;=\; \Delta_{t+1} R_{1,t+1}. \\
\widetilde{e}^{(d_1)}_{1,t+1} &= \left(\widetilde{e}^{(1)}_{1,t} + \frac{(1-\alpha_1)\alpha_{1,3} \widetilde{\xi}_{t+1}}{\widetilde{\mu}_t + \phi \widetilde{\Delta}_t} \right) \Big/ r_{1,t+1} \\
&= \left(e^{(1)}_{1,t}/R_{1,t} + \frac{(1-\alpha_1)\alpha_{1,3} \widetilde{\xi}_{t+1}}{\mu_t R_{1,t} + \phi \mu_t R_{1,t}} \right) \Big/ r_{1,t+1} \\
&\overset{\widetilde{\xi}_{t+1} = \xi_{t+1}}{=} \left(e^{(1)}_{1,t} + \frac{(1-\alpha_1)\alpha_{1,3} \xi_{t+1}}{\mu_t + \phi \Delta_t} \right) \Big/ (R_{1,t} r_{1,t+1}) \;=\; e^{(d_1)}_{1,t+1}/R_{1,t+1}, \\
\widetilde{e}^{(k)}_{1,t+1} &= \widetilde{e}^{(k+1)}_{1,t}/r_{1,t+1} = e^{(k)}_{1,t}/(R_{1,t} r_{1,t+1}) \;=\; e^{(k)}_{1,t}/R_{1,t+1}.
\end{aligned}
$$

This completes the proof of (c).

The proof of assertion (d) uses (c):

$$
\begin{aligned}
\widetilde{Y}_{t+h|t} &= \left(\widetilde{\mu}_t + \sum_{j=1}^{h} \phi^j \widetilde{\Delta}_t \right) \widetilde{e}^{(h)}_{1,t} + \boldsymbol{\beta}^\top \boldsymbol{x}_{t+h} \\
&= \left(\mu_t R_{1,t} + \sum_{j=1}^{h} \phi^j \Delta_t R_{1,t} \right) e^{(h)}_{1,t}/R_{1,t} + \boldsymbol{\beta}^\top \boldsymbol{x}_{t+h} \\
&= \left(\mu_t + \sum_{j=1}^{h} \phi^j \Delta_t \right) e^{(h)}_{1,t} + \boldsymbol{\beta}^\top \boldsymbol{x}_{t+h} \quad = \quad Y_{t+h|t}.
\end{aligned}
$$

This completes the proof of Proposition 6.14.

7 Prediction Intervals for Exponential Smoothing with Covariates

7.1 Introduction

The core interest of statistical time series analysis is forecasting a future value Y_{T+h} of a time-indexed phenomenon, based on the observed process history $\mathcal{H}_T = (Y_T, Y_{T-1}, \ldots)$ up to time T. The forecast may be a *point forecast* $Y_{T+h|T}$, or an *interval forecast* $B_{T+h|T}$ under a prescribed confidence level γ such that $\mathrm{P}(Y_{T+h} \in B_{T+h|T}) \geq \gamma$.

The theory, software implementation and industrial practice of time series has strongly been concentrating on point forecasts. A large variety of time series models and related point forecasting algorithms have been suggested in the literature. Three approaches have had particular influence in industry: i) The autoregressive moving average (ARMA) and autoregressive integrated moving average (ARIMA) forecasting methodology introduced by Box & Jenkins (1970). The method is based on a precisely stated model which allows to derive optimum point forecasts. It has become a standard both in applied sciences and in industry, and has been implemented in numerous software solutions for forecasting purposes. ii) The FORSYS scheme by Lewandowski (1979, 1982) is less known in applied sciences, but very popular in various industrial branches, mainly due to its embedding in the forecasting systems for demand planning and logistics offered by the *Marketing Systems* company founded by R. Lewandowski. FORSYS has a clearly stated forecasting algorithm, but lacks an underlying time series model. iii) Exponential smoothing (ES) dates back to the work of Brown (1959) and Holt (1957) in the 1950s for purposes of inventory management, see Section 6.1 for the evolution of ES.

The issue of prediction intervals has been receiving considerably less attention than point forecasts. For long, prediction intervals had few significance in statistical software packages. Some, like Statistica, still do not support prediction intervals at all. In others, like SPSS, prediction intervals are supported by add-on modules. Poorness in prediction intervals continues to be a weak point of forecasting software, see the review by Küsters et al. (2006). Among popular textbooks, some neglect prediction intervals completely,

e. g. Brockwell & Davis (1996). Others, like Box et al. (1994), invest few effort into the subject, in comparison with the efforts made to build models and derive optimum point forecasts. However, the low priority given to prediction intervals in statistical methodology seems to be contrary to the interest of forecasting practitioners, see, for instance, the empirical survey conducted by Collopy & Armstrong (1992).

In the theory of statistical time series analysis, the by far most popular approach to prediction intervals is what Ord et al. (1997) call the *plug-in method* (*PIM*). Consider a parametric time series model indexed by a parameter vector $\boldsymbol{\theta}$. The model provides a parametric unbiased h-step-ahead forecast function $Y_{T+h|T} = P_{T,h,\mathcal{H}_T}(\boldsymbol{\theta})$, where $Y_{T+h|T}$ is the forecast for time $T+h$ made at time T, and $\mathcal{H}_T$ is the observable time series up to time T. The distribution $F_{\boldsymbol{\theta},T+h|T}$ of the forecast error $D_{T+h|T} = Y_{T+h} - Y_{T+h|T}$ is studied in dependence of $\boldsymbol{\theta}$. This is done, in particular, for the forecasting variance $\sigma^2_{T+h|T}(\boldsymbol{\theta})$. An estimator $\widehat{\boldsymbol{\theta}}$ for the parameter $\boldsymbol{\theta}$ is obtained by fitting the model to available observations $Y_1, \ldots, Y_T$ in the sense of minimising a fit criterion like the mean square error (MSE) or the mean absolute percentage error (MAPE). Then a prediction interval can be obtained by evaluating the estimated distribution $F_{\widehat{\boldsymbol{\theta}},T+h|T}$ of the forecast error $D_{T+h|T}$. The most popular approach is to start with a suitable normal distribution assumption in the base model, which leads to a normally distributed forecast error $D_{T+h|T}$. Then the PIM prediction interval $B_{T+h|T}$ of nominal confidence level γ is

$$B_{T+h|T} \quad = \quad \Big(Y_{T+h|T} - z\sigma_{T+h|T}(\widehat{\boldsymbol{\theta}});\; Y_{T+h|T} + z\sigma_{T+h|T}(\widehat{\boldsymbol{\theta}})\Big),$$

where $z = z_{N(0,1)}\left((1+\gamma)/2\right)$ is the $(1+\gamma)/2 \cdot 100\,\%$ quantile of the normal distribution $N(0,1)$. Section 7.4 considers the prediction interval of that type for exponential smoothing with covariates (ESCov) models.

It is crucial that the actual confidence level does not fall below the nominal confidence level γ, i. e. that $\mathrm{P}(Y_{T+h} \in B_{T+h|T}) \geq \gamma$ holds. In a simulation study based on tourist arrival time series, Kim et al. (2011) obtained relatively good actual confidence levels for PIM prediction intervals at a nominal 95 % level, but good coverages here are not the rule. Customary prediction intervals of the PIM type frequently offend against the coverage requirement, i. e. the actual confidence level is smaller than the nominal confidence level γ. The intervals use to be too narrow, suggesting a precision of inference which is actually not feasible at the stipulated level of confidence. This phenomenon was demonstrated by Makridakis et al. (1987) in a broad empirical study on the time series used in the 1982 M-competition (Makridakis et al. 1982). Chatfield (2001) lists 4 potential reasons:

1) wrongly identified model,
2) true model changing over time,
3) nonnormal error distribution,
4) insufficient account for the error in estimates of model parameters.

The reasons 1), 2) and 3) are of general nature and affect any method for calculating prediction intervals. The reason 4) is the specific problem of the PIM approach.

Several remedies against excessive narrowness of prediction intervals have been suggested, see Chatfield's (2001) review. To overcome problems 1)–3), methods to obtain prediction intervals can be used that either do not rely on a model or can do without normality of the errors. Empirical prediction intervals do not exploit model properties and should be rather robust against 1) and 2). Nonnormal error distributions can be dealt with by nonparametric quantile estimates, for instance. Popular alternative prediction intervals are based on simulation or bootstrapping, particularly when calculation time is not an issue or the calculation is fast.

With respect to problem 4), basically two approaches may be taken to account for the effect of the error in parameter estimates on the estimation of the forecast error distribution: i) Reduce model dependence by infiltrating parameter-free elements into the analysis of the forecast error distribution. Gardner's (1988) empirical prediction interval based on the Chebychev inequality, which we revise in Section 7.2.3, is an instance. ii) Evaluate the effect of parameter estimation error on the estimation of the forecast error distribution. Unfortunately, the nonlinear nature of time series models prevents against obtaining useful results in a straightforward manner. One way out is to resort to simulation methods. Another option is to transform the nonlinear problem into a linear scheme that can be dealt with by the techniques of regression analysis. This approach has been considered by Ord et al. (1997) in the context of ES.

This chapter collects a variety of prediction intervals applicable to ESCov of the types empirical, bootstrap, model-based and based on linear expansion. The latter applies a modified version of the approach of Ord et al. (1997) to ESCov. For time series analysis under covariates, the problem with PIM is particularly serious, since the coefficients of the external variables increase the number of parameters to be estimated, see the remarks in Section 7.5.3. Attention is restricted to the linear ESCov single source of error (SSOE) state-space model with one additive seasonality from Section 6.3.1 with independent and homoscedastic residuals ξ_s with constant variance $V[\xi_s] = \sigma_\xi^2$. With some more technical efforts, the results can be adapted to autocorrelated residuals and also to the case of multiplicative residuals analogously to the approach by Hyndman et al. (2002) under ES without covariates.

The study is organised as follows: Empirical prediction intervals are considered in Section 7.2. A bootstrap prediction interval is presented in Section 7.3. Section 7.4 deals with model-based prediction intervals exploiting the SSOE model for ESCov. Section 7.5 is concerned with prediction intervals derived by means of linear model theory and their application: Section 7.5.1 develops the linear expansion of the ESCov model. The statistical analysis of the linearised model follows in Section 7.5.2 and approximate prediction intervals based on the linearised model are developed in Section 7.5.3. Section 7.6 investigates the actual coverage probability of the intervals suggested by Section 7.5.3 in a simulation study based on time series of electricity load data from the Italian market. To illustrate the consequences for empirical forecasting, the approximate prediction intervals are applied to the prediction of electricity consumption in Section 7.7.

7.2 Empirical Prediction Intervals

Empirical prediction intervals do not make assumptions about the true underlying model and are basically always applicable. Chatfield (1993) suggested to consider empirical methods when model assumptions are in doubt or theoretical formulas not available. For a long time, they played an important role for ES methods when a model to derive theoretical prediction intervals was not yet available. For example, Chatfield & Yar (1991) constructed prediction intervals for the Holt-Winters smoothing method (linear trend, additive and multiplicative season) by stipulating that the smoothing algorithm leads to independent residuals. Instead of assuming the validity of a model (as in Section 7.4), empirical prediction intervals use the properties of the observed error distribution, see Chatfield (1993). All the prediction intervals presented in this section have in common that they rely on the assumption of at least i.i.d. forecast errors. The validity of this assumption cannot be expected to hold in general. E. g. in the case of the multiplicative Holt-Winters method, Chatfield & Yar (1991) found the errors to be dependent on the states, which would violate the assumption.

7.2.1 Prediction Intervals based on Empirical Quantiles

An easy way to construct prediction intervals is based on empirical quantiles of the forecast errors. They do not make assumptions about the underlying model and do not assume a certain error distribution.

Definition 7.1. *Let* $Y_{h|0}, Y_{h+1|1}, \ldots, Y_{T|T-h}$ *be predictions* h*-steps-ahead and* $Y_h, Y_{h+1}, \ldots, Y_T$ *the observed values. Let* $Z_h = Y_h - Y_{h|0}, Z_{h+1} = Y_{h+1} - Y_{h+1|1}, \ldots, Z_T = Y_T - Y_{T|T-h}$ *be the empirical* h*-step-ahead forecast errors and* $Y_{T+h|T}$ *be the* h*-step-ahead prediction. Then an empirical prediction interval of level* $\gamma \cdot 100\,\%$, $0 < \gamma < 1$, *for* Y_{T+h} *is given by*

$$B_{T+h|T} := \left(Y_{T+h|T} + z\left(\frac{1-\gamma}{2}\right); Y_{T+h|T} + z\left(\frac{1+\gamma}{2}\right) \right),$$

where $z(\alpha)$ *is the empirical* $\alpha \cdot 100\,\%$*-quantile of the empirical* h*-step-ahead forecast errors* $Z_h, Z_{h+1}, \ldots, Z_T$.

In the application of the above prediction interval, one needs to use an appropriate quantile estimation procedure to obtain the empirical quantiles. Common estimators of $\gamma \cdot 100\,\%$-quantiles, as e. g. the $\lfloor T\gamma \rfloor$th observation of the ordered sample $Z_{(1,T)} \leq Z_{(2,T)} \leq \ldots \leq Z_{(T,T)}$, where $\lfloor T\gamma \rfloor$ is the greatest integer smaller than $T\gamma$, frequently underestimate the true quantiles. Small sample sizes are particularly problematic, see, for example, the simulation results for a variety of distributions in Göb & Lurz (2015). Prediction intervals obtained in this way are therefore in danger of being too narrow. Unless the empirical error distribution is (nearly) symmetric, the prediction intervals are in general not symmetric.

7.2.2 Normal Approximation based Prediction Intervals

A simple symmetric prediction interval can be obtained by subtracting and adding a quantile of the normal distribution times the standard deviation of the empirical h-step-ahead forecast errors around the point forecast. The interval is an empirical interval because it is based on the empirical forecast errors and it is of parametric type because it assumes at least approximately normally distributed errors.

Definition 7.2. *Let* $Y_{h|0}, Y_{h+1|1}, \ldots, Y_{T|T-h}$ *be predictions* h*-steps-ahead and* $Y_h, Y_{h+1}, \ldots, Y_T$ *the observed values. Let* $Z_h = Y_h - Y_{h|0}, Z_{h+1} = Y_{h+1} - Y_{h+1|1}, \ldots, Z_T = Y_T - Y_{T|T-h}$ *be the empirical* h*-step-ahead forecast errors. Let the* h*-step-ahead point prediction be* $Y_{T+h|T}$ *and let* $\widehat{\sigma}_Z^2$ *be the empirical variance of the errors* $Z_h, Z_{h+1}, \ldots, Z_T$*, i. e.* $\widehat{\sigma}_Z^2 = \frac{1}{T-h} \sum_{i=h}^{T} (Z_i - \bar{Z})^2$. *Then an approximative prediction interval of level* $\gamma \cdot 100\,\%$, $0 < \gamma < 1$, *for* Y_{T+h} *based on the normal distribution, is given by*

$$B_{T+h|T} := \left(Y_{T+h|T} + z_{N(0,1)}\left(\frac{1-\gamma}{2}\right)\widehat{\sigma}_Z; Y_{T+h|T} + z_{N(0,1)}\left(\frac{1+\gamma}{2}\right)\widehat{\sigma}_Z \right),$$

where $z_{N(0,1)}(\alpha)$ *is the* $\alpha \cdot 100\,\%$*-quantile of the standard normal distribution* $N(0,1)$*.*

In case the distribution of the h-step-ahead forecast errors is actually a normal distribution, the prediction intervals obtained in this way are certainly a good choice. If, however, the error distribution is nonnormal, the prediction intervals might not have the desired performance. In a study by Gardner (1988), who applied the above interval to 111 time series from the M-competition by Makridakis et al. (1982), the interval underperformed in terms of coverage probability. Approximative normality of the errors might often be justifiable, but many times it will be violated (especially for small sample sizes). In this case, the applicability of the prediction interval has to be doubted.

The normal distribution is the most commonly used distribution to deliver the coefficient in the prediction interval. However, other distributions can optionally serve as models for the error distribution. Williams & Goodman (1971), for example, have found the gamma distribution to fit the forecast errors of their data well. If an appropriate error distribution is identified, prediction intervals can be constructed by replacing $z_{N(0,1)}((1-\gamma)/2)\,, z_{N(0,1)}((1+\gamma)/2)$ in Definition 7.2 by the corresponding distribution quantiles.

7.2.3 Prediction Intervals based on Chebychev's Inequality

The prediction intervals presented in this section follow Gardner (1988), who considers it dangerous to assume an arbitrary error distribution and therefore suggests a nonparametric prediction interval that is rather robust. It is of the same structure as the intervals from the previous section with the difference that it does not use a coefficient based on the normal distribution, but a coefficient based on Chebyshev's (1867) inequality.

The advantage of the Chebychev inequality are its weak assumptions about the underlying distribution. According to Chebychev's inequality, a random variable Y with finite mean μ and finite variance σ^2, regardless of the distribution of Y, fulfils

$$\mathrm{P}\left(\left|\frac{Y-\mu}{\sigma}\right| \geq \epsilon\right) \leq 1/\epsilon^2.$$

Based on this inequality, prediction intervals can be obtained.

Definition 7.3. *Let* $Y_{h|0}, Y_{h+1|1}, \ldots, Y_{T|T-h}$ *be predictions* h*-steps-ahead and* $Y_h, Y_{h+1}, \ldots,$ Y_T *be the observed values. Let* $Z_h = Y_h - Y_{h|0}, Z_{h+1} = Y_{h+1} - Y_{h+1|1}, \ldots, Z_T = Y_T - Y_{T|T-h}$ *be the empirical* h*-step-ahead forecast errors. Let the* h*-step-ahead point prediction be* $Y_{T+h|T}$ *and have variance* σ_Z^2*. A prediction interval for* Y_{T+h} *of level* $\gamma \cdot 100\,\%$

based on the Chebychev inequality is given by

$$B_{T+h|T} := \left(Y_{T+h|T} - \sqrt{\frac{1}{1-\gamma}}\sigma_Z;\, Y_{T+h|T} + \sqrt{\frac{1}{1-\gamma}}\sigma_Z\right).$$

The variance σ_Z^2 can be estimated by the empirical variance $\widehat{\sigma}_Z^2$ of the errors $Z_h, Z_{h+1}, \ldots, Z_T$, i. e. $\widehat{\sigma}_Z^2 = \frac{1}{T-h}\sum_{i=h}^{T}(Z_i - \bar{Z})^2$.

While in his study of a sample of 111 series taken from Makridakis et al. (1982), Gardner (1988) found the coverage probability of the normal distribution based intervals clearly too low, the Chebychev-based prediction intervals came close to their nominal confidence level. The rather broad Chebychev intervals better accounted for larger post-sample forecast errors and instability in the trend. However, Gardner (1988) indicated that the characteristics might depend on the frequency of the data (monthly, quarterly, yearly, ...) as well as the sample size.

Gardner (1988) described a way how to find the h-step-ahead forecast errors and forecasting variance by fitting a model optimised according to the one-step-ahead forecast and hence not finding separate parameter sets for each lead time. Two simple extensions of Gardner's (1988) approach can be made: 1) Fit the model separately for each lead time and use the empirical variance of the h-step-ahead predictions for the calculation of the prediction intervals. 2) Use estimates for the prediction variances $\sigma^2_{Y_{t+h}|\mathcal{H}_t}$ based on the respective SSOE model for ESCov from Propositions 6.3, 6.4 and 6.6 as estimates for σ_Z^2. The result is not model-free, but does at least not assume a particular parametric error distribution.

The difference in prediction interval width between a coefficient based on normal distribution and Chebychev's inequality is demonstrated when looking at an example given by Yar & Chatfield (1990): For a prescribed confidence level $\gamma = 0.95$, the coefficient $\sqrt{1/(1-\gamma)}$ in the Chebychev-based prediction interval takes the value 4.47, while the $(1+\gamma)/2$-quantile $z_{N(0,1)}$ of the normal distribution (see Section 7.2.2) takes the value 1.96. Consequently, the Chebychev-based prediction interval is more than double the size of the normal distribution based interval under a given prediction variance. Therefore, if actually certain distributions, like the normal distribution, can be found to adequately model the error distribution, the Chebychev-based prediction interval is very likely too conservative and therefore hardly of use in industrial or scientific practice.

As discussed by Gardner (1988), the Chebychev coefficients have the advantage that they better capture the usually larger post-sample errors in comparison to the smaller in-sample errors, on which the prediction intervals are based. Furthermore, the Chebychev-based prediction intervals have a better chance to deal with change points in the series.

However, for serious change points Chatfield (1993) argued that they might still not be wide enough and therefore preferred normal-distribution-based coefficients under the premise that the series remains rather stable and shows a behaviour in the future that is similar to the one in the past.

7.2.4 Prediction Intervals based on the Camp-Meidell Inequality

The inequality by Camp (1922) and Meidell (1922) is a generalisation of Chebychev's inequality relying on only few additional distribution assumptions. Göb & Lurz (2015) generalise and invert the inequality for the distribution function such as to obtain a coefficient that serves as a bound for a quantile. The essential characteristic of the bound is that it uses the roots of central moments of even order. A version of the Camp-Meidell inequality applied to quantiles is given in the following proposition.

Proposition 7.4. *Let X be a random variable symmetric around its mean μ_X, with distribution function F_X and finite central moments $m_{X,s} := E[(X-\mu_X)^s]$, $s \in \mathbb{N}$. Let $z_X(\rho)$ be the $\rho \cdot 100\,\%$-quantile fulfiling $F_X(z_X(\rho)) = \rho$ and let $r \in \mathbb{N}$. Then the following inequality holds:*

$$z_X(\rho) \quad \leq \quad z_{r,X}(\rho) \quad := \quad \begin{cases} \mu_X + (2\rho - 1) m_{X,2r}^{1/(2r)} (2r+1)^{1/(2r)} & \text{if } \rho \leq \frac{4r+1}{4r+2}, \\ \mu_X + \left(\frac{m_{X,2r}}{2(1-\rho)}\right)^{1/(2r)} \frac{2r}{2r+1} & \text{if } \rho > \frac{4r+1}{4r+2}. \end{cases}$$

Proof. See Göb & Lurz (2015). □

The above version of the Camp-Meidell inequality is valid for a symmetric random variable. A version for the more general class of mean-modal random variables is provided by Göb & Lurz (2015).

The Camp-Meidell inequality has been used so far mostly for $r = 1$, in which case it is often too inexact. Göb & Lurz (2015) find the approximation for most quantile orders to be superior under $r = 2$. Choosing $r = 2$ involves estimating roots of central moments up to order 4. The quantile approximation under $r \geq 3$ is better for some quantile orders under several common symmetric distributions, but does not necessarily bring sufficient improvement such as to justify the higher effort required to estimate roots of central moments of order 6 and higher. Hence, to use $r = 2$ is a good compromise. Quite frequently, central moment estimators and the roots thereof as used in practice are biased estimators. Göb & Lurz (2015) formulate how to obtain roots of central moments that are unbiased at least under a baseline distribution: The root $m_{X,2r}^{1/(2r)}$ is hereby replaced by the corrected estimator $R_{n,X,2r} := c_{n,2r} \left(\frac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X} \right)^{2r} \right)^{1/(2r)}$, where $c_{n,2r}$ is a correction factor causing $R_{n,X,2r}$ to be an unbiased estimator under the normal distribution. Inserting $R_{n,X,2r}$ into the equation in Proposition 7.4 results in the empirical quantile bound

$$\widehat{z}_{r,X}(\rho) \quad = \quad \begin{cases} \bar{X} + (2\rho - 1) R_{n,X,2r} (2r+1)^{1/(2r)} & \text{if } \rho \leq \frac{4r+1}{4r+2}, \\ \bar{X} + \frac{R_{n,X,2r}}{(2(1-\rho))^{1/(2r)}} \frac{2r}{2r+1} & \text{if } \rho > \frac{4r+1}{4r+2}, \end{cases} \tag{7.1}$$

where $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$.

The empirical Camp-Meidell bound (7.1) for a quantile turns out to be in fact a bound for the true quantile of common symmetric distributions most of the times, even for small sample sizes. It shows good coverage properties while not being overconservative. We can use it to construct nonparametric prediction intervals in the following way:

Definition 7.5. *Let $Y_{h|0}, Y_{h+1|1}, \ldots, Y_{T|T-h}$ be predictions h-steps-ahead and Y_h, Y_{h+1}, $\ldots$, Y_T be the observed values. Let $Z_h = Y_h - Y_{h|0}, Z_{h+1} = Y_{h+1} - Y_{h+1|1}, \ldots, Z_T = Y_T - Y_{T|T-h}$ be the empirical h-step-ahead forecast errors. Let the distribution of the errors be symmetric around their mean $E[Z]$ estimated by $\bar{Z} = \frac{1}{T-h+1} \sum_{i=h}^{T} Z_i$, and have finite central moments. A nonparametric prediction interval of level $\gamma \cdot 100\,\%$, $0 < \gamma < 1$, based on the Camp-Meidell inequality for Y_{T+h} is given by*

$$B_{T+h|T} \; := \; \left(2\bar{Z} - \widehat{z}_{r,Z}\left(\frac{1+\gamma}{2}\right) ; \widehat{z}_{r,Z}\left(\frac{1+\gamma}{2}\right) \right).$$

The prediction interval based on the Camp-Meidell inequality does not assume a certain distribution of the errors, but only that the error distribution is symmetric. As the prediction interval based on Chebychev's inequality it is nonparametric, but likely to be of more practical use than the Chebychev-based interval because the Camp-Meidell inequality is expected to deliver narrower bounds than the Chebychev inequality.

7.3 Bootstrap Prediction Intervals

Bootstrapping is a popular distribution-free approach of constructing prediction intervals. The approach is still dependent on the model, but does not assume a certain distribution of the forecast errors, as e. g. the normal distribution. Instead, the error distribution is derived numerically by resampling or simulating forecast errors. Apart from requiring fewer assumptions, bootstrapping methods have the advantage of being always applicable, see Chatfield (1993).

Let $Y_1, \ldots, Y_T$ be successive observations of a time series. Hyndman et al.'s (2002) approach to obtain bootstrap prediction intervals of level $\gamma \in (0;1)$ is to simulate 5000 future sample paths for $\{Y_{T+1}, \ldots, Y_{T+M}\}$ and take the empirical $(1-\gamma)/2 \cdot 100\,\%$- and $(1+\gamma)/2 \cdot 100\,\%$-quantiles of the simulated values at each forecasting step. They consider a parametric and an ordinary bootstrap approach:

a) For the parametric bootstrapping, future errors are sampled by assuming normally distributed errors.

b) For the ordinary bootstrapping, the errors are resampled from the empirical distribution of the fitted errors.

Obviously, approach a) does not prepare against violations of the normality assumption, which is often seen as the important advantage of bootstrapping in contrast to parametric methods, while approach b) does.

The steps of approach b) are described e. g. by Fan & Hyndman (2012):

1. Fit the model based on the historical data $Y_1, \ldots, Y_T$. From this model, T in-sample forecast errors $Z_h, Z_{h+1}, \ldots, Z_T$ are obtained and a point forecast $Y_{T+h|T}$ is derived.
2. Resample N times T errors from the set $Z_h, Z_{h+1}, \ldots, Z_T$ and from that create N artificial samples by inserting the errors into the fitted model from step 1.
3. Re-estimate N models with the N artificial samples from step 2.
4. Insert the original data $Y_1, \ldots, Y_T$ into the N models from step 3 and from them obtain N simulated forecasts $1, 2, \ldots, M$-steps-ahead.
5. Resample another set of errors and insert them into the original model from step 1 to obtain simulated actuals.
6. For each forecasting horizon $h \in \{1, \ldots, M\}$, calculate N differences between the simulated actuals from step 5 and the simulated forecasts from step 4. These are the N simulated forecast errors that for each forecasting horizon h serve to build the empirical forecast distribution.

Prediction intervals of level $\gamma \cdot 100\,\%$ for the h-step-ahead forecast, $h = 1, \ldots, M$, can be obtained by taking the empirical $(1-\gamma)/2 \cdot 100\,\%$- and $(1+\gamma)/2 \cdot 100\,\%$-quantiles of the empirical h-step-ahead forecast distribution of size N.

A bootstrap approach of type a) can be obtained by sampling h-step-ahead forecast errors from the normal distribution $N(0, \sigma^2)$ instead of from the empirical errors $Z_h, \ldots, Z_T$ in steps 2 and 5. Hereby, σ^2 needs to be replaced by an estimate of the h-step-ahead forecast variance.

Due to the potential heavy computational load associated with the bootstrapping approach, Fan & Hyndman (2012) also formulate an alternative bootstrapping approach specifically modified for their forecasting application, which we do not revise here.

7.4 Model-based Prediction Intervals of Plug-in Type

Before the foundation of ES methods on SSOE state-space models by Ord et al. (1997), theoretical prediction intervals for ES relied on the fact that some ES methods were optimal for certain ARIMA models, for which the theory was sufficiently evolved. The ES methods for which this applies are those with linear trend and additive seasonality. With the theory for these ES methods being available through the equivalent optimal ARIMA model, prediction intervals could be developed, as was done by Yar & Chatfield (1990). For the case of multiplicative seasonality, there does not exist an ARIMA model for which the ES method is optimal. It is due to the nonlinearity of the multiplicative seasonality method that forecasts cannot be represented as a linear combination of past values, see Chatfield & Yar (1991). Prediction intervals for the multiplicative seasonality case were therefore either of empirical type or based on approximations.

The situation changed when Ord et al. (1997) underpinned ES methods by the SSOE state-space model. The availability of a model allowed the derivation of theoretical prediction intervals for ES. In Hyndman et al. (2005), a broad survey of analytical expressions for the variance based on state-space models can be found. With Wang (2006) formulating the SSOE model for ESCov, the foundations for theoretical prediction intervals were laid for the covariate-processing method as well. The prediction intervals are of PIM type, i. e. the parameter estimates are inserted into the formulas for the theoretical prediction distribution and therefore treated as fixed, while uncertainty in the estimation of the parameters is not taken into account, see the remarks in the introduction of this chapter.

We consider prediction intervals which are based on the MMSE forecast and forecasting variance for the h-step-ahead observation Y_{T+h} under the SSOE model for ESCov. Since the interval relies on the distribution of the forecast, it is an instance of a parametric prediction interval. The computation of the prediction interval is performed conditional on the fitted SSOE model for ESCov, the availability of which is assumed in the subsequent definition.

Definition 7.6. *Let $Y_1, Y_2, \ldots, Y_T$ be the observed time series values. Consider the MMSE h-step-ahead forecast $Y_{T+h|T}$ and the forecasting variance $\sigma^2_{Y_{T+h}|\mathcal{H}_T}$, given the process history $\mathcal{H}_T$ up to time T. Let $0 < \gamma < 1$ be the confidence level. Then an (approximative) two-sided prediction interval for Y_{T+h} of level $\gamma \cdot 100\,\%$ is given by*

$$B_{T+h|T} := \left(Y_{T+h|T} - z_{N(0,1)}\left(\frac{1+\gamma}{2}\right)\sigma_{Y_{T+h}|\mathcal{H}_T};\; Y_{T+h|T} + z_{N(0,1)}\left(\frac{1+\gamma}{2}\right)\sigma_{Y_{T+h}|\mathcal{H}_T}\right),$$

where $z_{N(0,1)}(\alpha)$ is the $\alpha \cdot 100\,\%$-quantile of the standard normal distribution $N(0,1)$.

The MMSE h-step-ahead forecast and its variance can be read from Proposition 6.3 for the linear SSOE for ESCov under independent residuals, Proposition 6.4 for the linear SSOE for ESCov under AR(1) residuals and Proposition 6.6 for the partially linear SSOE for ESCov. The prediction interval defined in Definition 7.6 relies on the assumption that the h-step-ahead prediction $Y_{T+h|T}$ has either a normal distribution with mean Y_{T+h} and variance $\sigma^2_{Y_{T+h}|\mathcal{H}_T}$, or that $Y_{T+h|T}$ is approximately normally distributed with mean Y_{T+h} and variance $\sigma^2_{Y_{T+h}|\mathcal{H}_T}$.

Quantiles of the normal distribution are commonly used. Since the normal distribution is frequently too optimistic in real data applications, there is the risk that the above method delivers prediction intervals that are too small. Other parametric distribution quantiles z instead of $z_{N(0,1)}((1+\gamma)/2)$ are conceivable depending on the distribution of $Y_{T+h|T}$, as suggested e.g. by Chatfield (2001). The $(1+\gamma)/2 \cdot 100\,\%$-quantile of the Laplace distribution, for example, would yield a prediction interval which is also symmetric around the mean and would yield broader prediction intervals for γ close to 1 due to the fatter tails of the Laplace distribution.

The MMSE h-step-ahead forecast and its variance are functions of the smoothing parameters $\boldsymbol{\alpha}$ through the components $\mathbf{G}_{\boldsymbol{\alpha}}$, $\boldsymbol{\delta}_{\boldsymbol{\alpha}}$, $\boldsymbol{w}_{\boldsymbol{\alpha}}$ or $\mathbf{G}_{\boldsymbol{\alpha},ne}$, $\boldsymbol{\delta}_{\boldsymbol{\alpha},ne}$, $\boldsymbol{w}_{\boldsymbol{\alpha},ne}$, respectively, as well as the covariate parameter $\boldsymbol{\beta}$ and autoregressive parameter λ. Consequently, the prediction intervals rely on the knowledge of these parameters. The parameters are usually unknown in practice, which is why one would insert their estimated values $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\beta}}$ and $\hat{\lambda}$ instead. Using this approach, the uncertainty in the estimation of the parameters is not taken into account, also the uncertainty in the estimation of the initial state values is not. The prediction intervals might tend to be too narrow due to this reason. Chatfield (1993), however, deems the effect of the parameter uncertainty as having less effect on the overall uncertainty than that due to model specification and the effect of outliers and errors. For this reason, inserting the parameter estimates into the forecast equation seems justifiable in most cases, unless the number of estimated parameters is high in comparison to the length of the series.

For predictions under ESCov, perfect knowledge about the future covariate value $\boldsymbol{x}_{T+h}$ is assumed. Hence, $\boldsymbol{x}_{T+h}$ is treated as fixed. In reality, the future value often is a prediction itself that is associated with uncertainty or is only approximately known for some reason. Since $\boldsymbol{x}_{T+h}$ is not considered random in the ESCov model, its uncertainty is also not taken into account in the calculation of the prediction intervals.

7.5 Prediction Intervals based on Linear Expansion of the ESCov Model

We exploit linear model theory similarly as Ord et al. (1997) to derive prediction intervals for ESCov based on linearisation that account for uncertainty in the parameter estimates. We apply the approach to the ESCov model ADT-AS of linear damped trend and one additive seasonality, which is an instance of the linear SSOE model for ESCov from Section 6.3.1 with $m = 1$. A variety of specific model instances are obtained as reductions or special cases of the model ADT-AS, namely:

i) NT-NS: no trend, no season;

ii) AT-NS: additive trend, no season;

iii) ADT-NS: additive damped trend, no season;

iv) NT-AS: no trend, additive season;

v) AT-AS: additive trend, additive season;

vi) ADT-AS: additive damped trend, additive season.

The undamped trend AT is obtained by setting $\phi = 1$ in the formulas for the damped trend model. Models NT without trend and NS without season are obtained by omitting the respective components of the state and parameter vectors. We focus on the case of homoscedastic residuals (ξ_s) with constant variance $V[\xi_s] = \sigma_\xi^2$. See Sections 6.4 and 6.6 for an empirical fitting and forecasting under SSOE models for ESCov.

In the course of this and the subsequent sections, we use a notation for the smoothing parameters in ESCov that differs from the one in Chapter 6 and the previous sections of Chapter 7. Instead of α_1 (smoothing parameter for the level), α_2 (smoothing parameter for the trend increment) and $\alpha_{1,3} = \alpha_3$ (smoothing parameter for a single additive season), we use $\widetilde{\alpha}_1, \widetilde{\alpha}_2$ and $\widetilde{\alpha}_3$, where the matching equations are provided in Table 7.1. The parameter vector for the smoothing parameters is subsequently denoted as $\widetilde{\boldsymbol{\alpha}} = (\widetilde{\alpha}_1, \widetilde{\alpha}_2, \widetilde{\alpha}_3, \phi)^\top$.

Table 7.1: Smoothing parameters in linear ESCov model

Parameter	Notation up to Section 7.4	Notation in Sections 7.5–7.7
Smoothing parameter for level	α_1	$\widetilde{\alpha}_1 = \alpha_1$
Smoothing parameter for trend increment	α_2	$\widetilde{\alpha}_2 = \alpha_1\alpha_2$
Smoothing parameter for additive season	$\alpha_{1,3} = \alpha_3$	$\widetilde{\alpha}_3 = (1-\alpha_1)\alpha_3$

7.5.1 Linear Expansion of the ESCov Model

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_T, \ldots, Y_{T+h})^\top$ be a segment of a time series (Y_t) subject to a linear ESCov SSOE model as described in Section 6.3.1. In the respective empirical context, let $Y_1, \ldots, Y_T$ be the observed past observations, and let $Y_{T+1}, \ldots, Y_{T+h}$ be instances at future times $T+1, \ldots, T+h$. By the observation equation and state transition equations for ADT-AS in Table 6.1 for a single seasonality, each Y_s can be expressed as a function of the parameter vector $\boldsymbol{\theta} = (\widetilde{\boldsymbol{\alpha}}^\top, \boldsymbol{\beta}^\top)^\top$, the residuals $\xi_t = Y_t - Y_{t|t-1}$, $t = 1, \ldots, s$, and the initial state vector $\boldsymbol{u}_0$. Thus, conditioned on $\boldsymbol{u}_0$, we can write

$$\boldsymbol{Y} \quad = \quad \boldsymbol{H}(\boldsymbol{\theta}, \boldsymbol{\xi}) \quad = \quad \Big(H_s(\boldsymbol{\theta}, \boldsymbol{\xi})\Big)_{1 \le s \le T+h}, \tag{7.2}$$

where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_{T+h})^\top$ is the vector of residuals. For the model ADT-AS with one additive season of length d, we obtain the nonlinear functions

$$\begin{aligned} H_s(\boldsymbol{\theta}, \boldsymbol{\xi}) \quad &= \quad \sum_{m=1}^{s-1} \left(\widetilde{\alpha}_1 + \widetilde{\alpha}_2 \sum_{i=1}^{s-m} \phi^i + \widetilde{\alpha}_3 \mathbb{1}_{\{d(1+\lfloor \frac{s-m-1}{d} \rfloor)\}}(s-m) \right) \xi_m \\ &\quad + \mu_0 + \Delta_0 \sum_{i=1}^{s} \phi^i + \sum_{i=0}^{d-1} e_{-i} \mathbb{1}_{\{d(1+\lfloor \frac{s-1}{d} \rfloor)-s\}}(i) + \boldsymbol{\beta}^\top \boldsymbol{x}_s + \xi_s. \end{aligned} \tag{7.3}$$

The derivation of (7.3) is provided in Appendix 7.A, Section 7.A.1.

We intend to use methods from linear model theory to account for the effect of parameter and residual estimation on forecasts. Hence we follow Ord et al. (1997) and develop a linear expansion of $\boldsymbol{H}(\boldsymbol{\theta}, \boldsymbol{\xi})$ around $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\xi}})$, where the entries $\widehat{\xi}_1, \ldots, \widehat{\xi}_T$ are the estimates obtained from the observations $Y_1, \ldots, Y_T$, and where the future values are forecasted as $\widehat{\xi}_{T+1} = \ldots = \widehat{\xi}_{T+h} = 0$. Let k' be the dimension of $\boldsymbol{\theta}$, and let $N := T + h$. Let the

$N \times k'$ matrix $\mathbf{M}$ be defined by

$$\begin{aligned}\mathbf{M} \quad := \quad & \frac{\partial}{\partial\,\boldsymbol{\theta}} \boldsymbol{H}(\boldsymbol{\theta},\boldsymbol{\xi})|_{\substack{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}} \\ \boldsymbol{\xi}=\widehat{\boldsymbol{\xi}}}} \quad = \\ & \Bigg(\sum_{m=1}^{s-1} \widehat{\xi}_m,\ \sum_{m=1}^{s-1} \sum_{i=1}^{s-m} \widehat{\phi}^i \widehat{\xi}_m,\ \sum_{m=1}^{s-1} \mathbb{1}_{\{d(1+\lfloor \frac{s-m-1}{d} \rfloor)\}} (s-m)\widehat{\xi}_m, \qquad (7.4) \\ & \qquad \widehat{\widetilde{\alpha}}_2 \sum_{m=1}^{s-1} \widehat{\xi}_m \sum_{i=1}^{s-m} i\widehat{\phi}^{i-1} + \Delta_0 \sum_{m=1}^{s} m\widehat{\phi}^{m-1},\ \boldsymbol{x}_s \Bigg)_{1\le s\le T+h} .\end{aligned}$$

The derivation of $\frac{\partial}{\partial\boldsymbol{\theta}} \boldsymbol{H}(\boldsymbol{\theta},\boldsymbol{\xi})|_{\substack{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}} \\ \boldsymbol{\xi}=\widehat{\boldsymbol{\xi}}}}$ for the ESCov model ADT-AS can be found in Appendix 7.A, Section 7.A.1.

Let the $N \times N$ matrix $\mathbf{L}$ be defined by

$$\mathbf{L} \ := \ \frac{\partial}{\partial\,\boldsymbol{\xi}} \boldsymbol{H}(\boldsymbol{\theta},\boldsymbol{\xi})|_{\substack{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}} \\ \boldsymbol{\xi}=\widehat{\boldsymbol{\xi}}}}, \qquad (7.5)$$

where for $1 \le m \le s-1$

$$\frac{\partial}{\partial\,\xi_m} H_s(\boldsymbol{\theta},\boldsymbol{\xi}) \quad = \quad \widetilde{\alpha}_1 + \widetilde{\alpha}_2 \sum_{i=1}^{s-m} \phi^i + \widetilde{\alpha}_3 \mathbb{1}_{\{d(1+\lfloor \frac{s-m-1}{d} \rfloor)\}} (s-m), \qquad (7.6)$$

$\frac{\partial}{\partial\xi_s} H_s(\boldsymbol{\theta},\boldsymbol{\xi}) = 1$, $\frac{\partial}{\partial\xi_m} H_s(\boldsymbol{\theta},\boldsymbol{\xi}) = 0$ for $m \ge s+1$. By a first order Taylor expansion we obtain the approximation

$$\boldsymbol{Y} \ \approx \ \boldsymbol{H}(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\xi}}) + \mathbf{M}(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}) + \mathbf{L}(\boldsymbol{\xi}-\widehat{\boldsymbol{\xi}}) \ = \ \boldsymbol{H}(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\xi}}) - \mathbf{M}\widehat{\boldsymbol{\theta}} - \mathbf{L}\widehat{\boldsymbol{\xi}} + \underbrace{\mathbf{M}\boldsymbol{\theta} + \mathbf{L}\boldsymbol{\xi}}_{=:\,\boldsymbol{Z}}. \qquad (7.7)$$

Let $\boldsymbol{Y}_{\mathrm{p}} := (Y_1,\ldots,Y_T)^\top = \boldsymbol{H}_{\mathrm{p}}(\boldsymbol{\theta},\boldsymbol{\xi})$, $\boldsymbol{\xi}_{\mathrm{p}} := (\xi_1,\ldots,\xi_T)^\top$, $\boldsymbol{Z}_{\mathrm{p}} := (Z_1,\ldots,Z_T)^\top$ be the vectors corresponding to the observed past, and let $\boldsymbol{Y}_{\mathrm{f}} := (Y_{T+1},\ldots,Y_{T+h})^\top = \boldsymbol{H}_{\mathrm{f}}(\boldsymbol{\theta},\boldsymbol{\xi})$, $\boldsymbol{\xi}_{\mathrm{f}} := (\xi_{T+1},\ldots,\xi_{T+h})^\top$, $\boldsymbol{Z}_{\mathrm{f}} := (Z_{T+1},\ldots,Z_{T+h})^\top$ be the vectors corresponding to the unobserved future. We decompose the matrices $\mathbf{L}$, $\mathbf{M}$ by

$$\mathbf{L} \quad = \quad \begin{pmatrix} \underbrace{\mathbf{L}_{\mathrm{pp}}}_{T\times T} & \underbrace{\mathbf{O}}_{T\times h} \\ \underbrace{\mathbf{L}_{\mathrm{fp}}}_{h\times T} & \underbrace{\mathbf{L}_{\mathrm{ff}}}_{h\times h} \end{pmatrix}, \quad \mathbf{M} \quad = \quad \begin{pmatrix} \underbrace{\mathbf{M}_{\mathrm{p}}}_{T\times k'} \\ \underbrace{\mathbf{M}_{\mathrm{f}}}_{h\times k'} \end{pmatrix} \qquad (7.8)$$

where $\mathbf{L}$, $\mathbf{L}_{\mathrm{pp}}$, $\mathbf{L}_{\mathrm{ff}}$ are invertible as lower triangular matrices with 1 on the main diagonal, see Eqs. (7.5)–(7.6). Then

$$\begin{aligned} \boldsymbol{Z}_{\mathrm{p}} \quad &= \quad \mathbf{M}_{\mathrm{p}}\boldsymbol{\theta} + \mathbf{L}_{\mathrm{pp}}\boldsymbol{\xi}_{\mathrm{p}}, && (7.9) \\ \boldsymbol{Z}_{\mathrm{f}} \quad &= \quad \mathbf{M}_{\mathrm{f}}\boldsymbol{\theta} + \mathbf{L}_{\mathrm{fp}}\boldsymbol{\xi}_{\mathrm{p}} + \mathbf{L}_{\mathrm{ff}}\boldsymbol{\xi}_{\mathrm{f}}, && (7.10) \\ \boldsymbol{Y}_{\mathrm{f}} \quad &\approx \quad \boldsymbol{H}_{\mathrm{f}}(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\xi}}) - \mathbf{M}_{\mathrm{f}}\widehat{\boldsymbol{\theta}} - \mathbf{L}_{\mathrm{fp}}\widehat{\boldsymbol{\xi}}_{\mathrm{p}} - \mathbf{L}_{\mathrm{ff}}\widehat{\boldsymbol{\xi}}_{\mathrm{f}} + \boldsymbol{Z}_{\mathrm{f}}. && (7.11) \end{aligned}$$

7.5.2 Statistical Analysis of the Linear Expansion of the ESCov Model

On grounds of the approximation (7.11), the prediction of the future time series segment $\boldsymbol{Y}_{\mathrm{f}}$ amounts to predicting $\boldsymbol{Z}_{\mathrm{f}}$. The Eqs. (7.9) and (7.10) are linear regression equations with autocorrelated residual vectors $\mathbf{L}_{\mathrm{pp}}\boldsymbol{\xi}_{\mathrm{p}}$ and $\mathbf{L}_{\mathrm{fp}}\boldsymbol{\xi}_{\mathrm{p}}+\mathbf{L}_{\mathrm{ff}}\boldsymbol{\xi}_{\mathrm{f}}$, respectively, for an unknown parameter vector $\boldsymbol{\theta}$. Hence linear model theory can be used for inference on $\boldsymbol{Z}_{\mathrm{f}}$. Ord et al. (1997) undertake a Bayesian analysis. We use a frequentist approach, which provides additional insight into the problem. See the Appendix 7.A, Section 7.A.2, for a derivation of the subsequently presented results.

Considering Eq. (7.9), we obtain from linear model theory (Christensen 1996) the quantity

$$\widehat{\sigma}^2_{\xi,\mathrm{LM}} \quad = \quad \frac{1}{T-r}(\mathbf{L}_{\mathrm{pp}}^{-1}\boldsymbol{Z}_{\mathrm{p}})^\top[\mathbf{I}-\boldsymbol{\Sigma}]\mathbf{L}_{\mathrm{pp}}^{-1}\boldsymbol{Z}_{\mathrm{p}} \tag{7.12}$$

as an unbiased estimator of the residual variance σ^2_ξ (see Section 6.4), where r is the rank of $\mathbf{M}_{\mathrm{p}}$, and where

$$\boldsymbol{\Sigma} \quad = \quad \mathbf{L}_{\mathrm{pp}}^{-1}\mathbf{M}_{\mathrm{p}}\Big((\mathbf{L}_{\mathrm{pp}}^{-1}\mathbf{M}_{\mathrm{p}})^\top(\mathbf{L}_{\mathrm{pp}}^{-1}\mathbf{M}_{\mathrm{p}})\Big)^{-1}(\mathbf{L}_{\mathrm{pp}}^{-1}\mathbf{M}_{\mathrm{p}})^\top\ . \tag{7.13}$$

The vector $P^\star(\boldsymbol{Z}_{\mathrm{p}})$ of best linear unbiased predictors for the entries of the future vector $\boldsymbol{Z}_{\mathrm{f}}$ is

$$P^\star(\boldsymbol{Z}_{\mathrm{p}}) \;=\; \Big((\mathbf{M}_{\mathrm{f}}-\mathbf{L}_{\mathrm{fp}}\mathbf{L}_{\mathrm{pp}}^{-1}\mathbf{M}_{\mathrm{p}})\Big((\mathbf{L}_{\mathrm{pp}}^{-1}\mathbf{M}_{\mathrm{p}})^\top(\mathbf{L}_{\mathrm{pp}}^{-1}\mathbf{M}_{\mathrm{p}})\Big)^{-1}\mathbf{M}_{\mathrm{p}}^\top(\mathbf{L}_{\mathrm{pp}}^\top)^{-1}+\mathbf{L}_{\mathrm{fp}}\Big)\,\mathbf{L}_{\mathrm{pp}}^{-1}\boldsymbol{Z}_{\mathrm{p}}. \tag{7.14}$$

If the residuals $\xi_1,\ldots,\xi_{T+h}$ have the normal distribution $N(0,\sigma^2_\xi)$, then the limits of a level γ prediction interval for Z_{T+k}, $k=1,\ldots,h$, are

$$P^\star(\boldsymbol{Z}_{\mathrm{p}})_k \quad \mp \quad z\cdot\widehat{\sigma}_{\xi,\mathrm{LM}}\cdot\sqrt{u_{kk}}, \tag{7.15}$$

where $z=z_{t(T-r)}((1+\gamma)/2)$ is the $((1+\gamma)/2)\cdot 100\,\%$ quantile of the central t-distribution $t(T-r)$, and where u_{kk} is the k-th diagonal entry of the $h\times h$ matrix

$$\mathbf{U} \;:=\; \mathbf{L}_{\mathrm{ff}}\mathbf{L}_{\mathrm{ff}}^\top+(\mathbf{M}_{\mathrm{f}}-\mathbf{L}_{\mathrm{fp}}\mathbf{L}_{\mathrm{pp}}^{-1}\mathbf{M}_{\mathrm{p}})\Big((\mathbf{L}_{\mathrm{pp}}^{-1}\mathbf{M}_{\mathrm{p}})^\top(\mathbf{L}_{\mathrm{pp}}^{-1}\mathbf{M}_{\mathrm{p}})\Big)^{-1}(\mathbf{M}_{\mathrm{f}}-\mathbf{L}_{\mathrm{fp}}\mathbf{L}_{\mathrm{pp}}^{-1}\mathbf{M}_{\mathrm{p}})^\top. \tag{7.16}$$

By Bayesian heuristics, Ord et al. (1997) achieve an interval similar to (7.15) of the type point forecast $\mp\, z\cdot\widehat{\sigma}_\xi\cdot\sqrt{u_{kk}}$, with the essential difference that $z=z_{N(0,1)}((1+\gamma)/2)$ is the $(1+\gamma)/2\cdot 100\,\%$ quantile of the normal distribution $N(0,1)$. At the same variance estimation $\widehat{\sigma}_\xi$, the interval (7.15) is always broader than the interval suggested by Ord et al. (1997), where the difference is particularly noticeable for a small number T of observations.

7.5.3 Approximate Prediction Intervals

If the ESCov model defined by the observation and state transition equations for ADT-AS from Table 6.1 is warranted, the k-step-ahead point forecast should be the MMSE predictor $Y_{T+k|T}$ as defined in Table 6.1 with $m = 1$ for one seasonality. The results obtained from the linearisation in Section 7.5.1 are used to account for the dispersion of the forecast around Y_{T+k}. By Section 7.5.1, the dispersion of a forecast for Y_{T+k} is approximately the dispersion of a forecast for Z_{T+k} based on the linear model (7.9) and (7.10). Hence, to obtain an approximate level γ prediction interval for Y_{T+k}, we replace in the prediction interval (7.15) the point forecast $P^{\star}(\boldsymbol{Z}_{\mathrm{p}})_k$ by the point forecast $Y_{T+k|T}$. The general scheme for the prediction interval limits is $Y_{T+k|T} \mp c\widehat{\sigma}$, where $\widehat{\sigma}^2$ is an estimate of the one-step-ahead forecasting variance. We have two choices for the estimator $\widehat{\sigma}^2$: 1) the estimator $\widehat{\sigma}^2_{\xi,\mathrm{ES}}$, see formula (6.21) in Section 6.4; 2) the estimator $\widehat{\sigma}^2_{\xi,\mathrm{LM}}$, see formula (7.12). In the PIM intervals, the coefficient c is chosen as the $(1+\gamma)/2 \cdot 100\,\%$ quantile $c = z_{N(0,1)}((1+\gamma)/2)$ of the normal distribution $N(0,1)$, see Definition 7.6. The analysis of Section 7.5.2 leading to formula (7.15) establishes the coefficient $c = z \cdot u_{kk}$ where u_{kk} is the k-th diagonal entry of the $h \times h$ matrix $\mathbf{U}$ defined by (7.16), and where $z = z_{t(T-r)}((1+\gamma)/2)$ is the $(1+\gamma)/2 \cdot 100\,\%$ quantile of the central t-distribution $t(T-r)$. Ord et al. (1997) also use the coefficient $c = z \cdot u_{kk}$ but with $z = z_{N(0,1)}((1+\gamma)/2)$. To obtain sufficient evidence about the best choice, we will subsequently consider the intervals I1, I2, I3, I4 resulting from the four combinations of the two coefficients with the two available variance estimators:

I1) *t*-distribution interval with standard deviation estimated by the linearisation estimator $\boldsymbol{\widehat{\sigma}_{\xi,\mathrm{LM}}}$: Limits $Y_{T+k|T} \pm z\widehat{\sigma}_{\xi,\mathrm{LM}}\sqrt{u_{kk}}$, z is the $(1+\gamma)/2 \cdot 100\,\%$-quantile of the t-distribution $t(T-r)$, see formula (7.15) with $P^{\star}(\boldsymbol{Z}_{\mathrm{p}})_k$ replaced by $Y_{T+k|T}$.

I2) Normal distribution interval with standard deviation estimated by the SSOE plug-in estimator $\boldsymbol{\widehat{\sigma}_{\xi,\mathrm{ES}}}$: Limits $Y_{T+k|T} \pm z\widehat{\sigma}_{\xi,\mathrm{ES}}$, z is the $(1+\gamma)/2 \cdot 100\,\%$-quantile of the normal distribution $N(0,1)$, see Definition 7.6.

I3) *t*-distribution interval with standard deviation estimated by the SSOE plug-in estimator $\boldsymbol{\widehat{\sigma}_{\xi,\mathrm{ES}}}$: Limits $Y_{T+k|T} \pm z\widehat{\sigma}_{\xi,\mathrm{ES}}\sqrt{u_{kk}}$, z is the $(1+\gamma)/2 \cdot 100\,\%$-quantile of the t-distribution $t(T-r)$, see Eq. (7.15) with $P^{\star}(\boldsymbol{Z}_{\mathrm{p}})_k$ replaced by $Y_{T+k|T}$.

I4) Normal distribution interval with standard deviation estimated by the linearisation estimator $\boldsymbol{\widehat{\sigma}_{\xi,\mathrm{LM}}}$: Limits $Y_{T+k|T} \pm z\widehat{\sigma}_{\xi,\mathrm{LM}}$, z is the $(1+\gamma)/2 \cdot 100\,\%$-quantile of the normal distribution $N(0,1)$, see Definition 7.6.

To understand the inadequacy of the customary plug-in approach I2 for ESCov, consider the case that the vector $\widetilde{\boldsymbol{\alpha}}$ of smoothing parameters is known. It remains to estimate the covariate coefficient $\boldsymbol{\beta}$. Consider to approach this estimation problem by MSE minimisation as considered in Section 6.4. By Eqs. (7.2) and (7.3), the resulting estimator $\widehat{\boldsymbol{\beta}}$ is a generalised least squares estimator under a linear model with correlated residuals. In this case, the ES variance estimator $\widehat{\sigma}^2_{\xi,\mathrm{ES}}$ essentially coincides with the linear model variance estimator $\widehat{\sigma}^2_{\xi,\mathrm{LM}}$, see Eq. (7.12), and the exact prediction interval under normally distributed residuals is (7.15). In (7.15), however, different from the naïve scheme from Definition 7.6, the estimator $\widehat{\sigma_\xi}$ in form of $\widehat{\sigma}^2_{\xi,\mathrm{LM}}$ has an additional coefficient $\sqrt{u_{kk}} > 1$. Hence the interval designed by Definition 7.6 is definitely too narrow.

7.6 Simulation Study based on Electricity Load Data

We investigate the actual coverage probability of intervals of nominal confidence level γ of the types I1, I2, I3, I4 suggested by Section 7.5.3 in a simulation study based on electricity load data. The subsequent section describes the underlying data and the design of the simulation studies. The results are discussed in Section 7.6.2.

7.6.1 Italian Electricity Load Data

The simulation study is conducted on the basis of two datasets of daily electricity loads in Italy over nine weeks, i. e. 63 days. The first dataset, subsequently addressed as electricity data type E1, covers the daily loads of a consumer from 9 May 2005 to 10 July 2005. The second dataset, electricity data type E2, covers the daily loads of a consumer from 1 September 2005 to 2 November 2005. Each considered period contains one public holiday on a weekday, the 2nd of June and the 1st of September, respectively, on which the electricity consumption is unusually low compared to usual weekdays. To avoid disturbance of the parameter estimation by these unusual days, we replace these days' electricity loads by the averages of the values of the two adjacent days.

In both cases, the regional average temperature in °C on day t is chosen as the covariate x_t. To simplify the evaluation of the prediction intervals for the considered forecasting period over the days 57 to 63, the temperatures $x_{57}, \ldots, x_{63}$ are assumed to be known precisely at the forecasting time $t = 56$. For load forecasting in field practice, the uncertainty in temperature forecasts has to be taken into account. However, temperature forecasts have become remarkably exact so that the remaining uncertainty will not affect the accuracy of prediction intervals substantially. Teisberg et al. (2005) demonstrate that the accuracy of load forecasts based on meteorological forecasts is only marginally inferior to the accuracy of load forecasts based on the exact knowledge of weather characteristics.

Both datasets are analysed by ESCov. The first eight weeks are used to estimate the smoothing and covariate parameters as well as the initial values for the states and the variance of the one-step-ahead forecast errors. On the basis of these parameters, the simulation is initialised. This results in a sample size of $n = 56$. The covariates in form of the average daily temperatures on the remaining seven days are needed for the prediction of the electricity load of one to seven days ahead.

We apply the ESCov model ADT-AS, i. e. the model with damped linear trend and additive seasonality. The values of the parameters $\widetilde{\boldsymbol{\alpha}}$, β minimising the MSE of the k-days-ahead forecast, the starting values μ_0, Δ_0, $e_{-d+1}, \ldots, e_0$ for the level, trend increment and season and the variance σ_ξ^2 of the one-step-ahead forecast errors that are used to initialise the simulation are provided in Tables 7.3 and 7.4. For each of the two datasets, we consider the prediction lead times $k = 1, 2, \ldots, 7$ days. Given a set $(\widetilde{\boldsymbol{\alpha}}, \beta, \mu_0, \Delta_0, e_{-d+1}, \ldots, e_0, \sigma_\xi^2)$ of simulation parameters, we obtain a time series $Y_1, \ldots, Y_T, Y_{T+1}, \ldots, Y_{T+h}$, $T = 56$, by first simulating the vector $(\xi_1, \xi_2, \ldots, \xi_T, \xi_{T+1}, \ldots, \xi_{T+h})^\top$ of independent residuals, and then inserting them into Eq. (7.3). We consider two distributions for the residuals ξ_i, both with mean 0 and with variance σ_ξ^2 found from the ESCov estimation, see Tables 7.3 and 7.4: i) normal distribution, and ii) Laplace distribution as an alternative with much more weight on the tails than the normal distribution. For the simulated series $Y_1, \ldots, Y_T$, we reestimate the parameters of the ESCov model ADT-AS. On the grounds of the estimated model we calculate the prediction intervals I1, I2, I3 and I4 for a nominal confidence level of $\gamma = 0.95$ and evaluate the empirical coverage by observing whether the true value Y_{T+k} is contained in the prediction interval. For each setting, we carry out 5000 simulation runs. In total, we have $2 \times 7 \times 2$ settings with 5000 simulation runs each. A summary of the simulation design is provided in Table 7.2.

Table 7.2: Simulation design

Factor	Levels
datasets	electricity data type E1, electricity data type E2
prediction lead times k	$1, 2, \ldots, 7$
distribution of the errors ξ_i	normal distribution, Laplace distribution
confidence level γ	0.95
simulation runs	5000

Table 7.3: Simulation parameters for electricity data E1 under ESCov model ADT-AS.

k	$\widetilde{\alpha}_1$	$\widetilde{\alpha}_2$	$\widetilde{\alpha}_3$	ϕ	β	μ_0	Δ_0	$e_{-d+1}, \dots, e_0$	σ_ξ^2
1	0.874	0.173	0.126	0.025	830.6	17469.5	3488.1	1868.4, 1370.9, 1148.5, 1305.8, 650.8, -2695.5, -3648.9	3149110
2	0.495	0.073	0.344	0.015	1290.0	9999.1	5622.5	1868.4, 1370.9, 1148.5, 1305.8, 650.8, -2695.5, -3648.9	3539957
3	0.484	0.098	0.348	0.014	1410.4	7964.0	6203.9	1868.4, 1370.9, 1148.5, 1305.8, 650.8, -2695.5, -3648.9	3702301
4	0.503	0.056	0.381	0.013	1462.0	5993.6	6766.9	1868.4, 1370.9, 1148.5, 1305.8, 650.8, -2695.5, -3648.9	3933766
5	0.550	0.052	0.352	0.012	1456.6	5234.5	6983.8	1868.4, 1370.9, 1148.5, 1305.8, 650.8, -2695.5, -3648.9	4008116
6	0.550	0.100	0.325	0.013	1450.6	5569.4	6888.1	1868.4, 1370.9, 1148.5, 1305.8, 650.8, -2695.5, -3648.9	3930932
7	0.001	$1.921 \cdot 10^{-4}$	0.316	0.011	1661.5	1520.1	8045.0	1868.4, 1370.9, 1148.5, 1305.8, 650.8, -2695.5, -3648.9	7624992

Table 7.4: Simulation parameters for electricity data E2 under ESCov model ADT-AS.

k	$\widetilde{\alpha}_1$	$\widetilde{\alpha}_2$	$\widetilde{\alpha}_3$	ϕ	β	μ_0	Δ_0	$e_{-d+1}, \dots, e_0$	σ_ξ^2
1	0.604	$6.04 \cdot 10^{-4}$	0.396	0.982	44.0	22627.3	-120.9	1350.3, 1565.8, -242.0, -1645.8, -708.2, -776.2, 456.0	520420.5
2	0.528	$5.27 \cdot 10^{-4}$	0.472	0.984	97.3	22626.7	-120.7	1350.3, 1565.8, -242.0, -1645.8, -708.2, -776.2, 456.0	583832.1
3	0.486	$4.86 \cdot 10^{-4}$	0.452	0.984	127.4	22626.6	-120.7	1350.3, 1565.8, -242.0, -1645.8, -708.2, -776.2, 456.0	668798.9
4	0.021	0.021	0.244	0.968	81.8	22633.7	-122.7	1350.3, 1565.8, -242.0, -1645.8, -708.2, -776.2, 456.0	1042202
5	0.023	0.023	0.211	0.965	89.3	22634.9	-123.1	1350.3, 1565.8, -242.0, -1645.8, -708.2, -776.2, 456.0	1069170
6	0.468	$8.64 \cdot 10^{-3}$	0.171	0.974	144.6	22630.9	-121.9	1350.3, 1565.8, -242.0, -1645.8, -708.2, -776.2, 456.0	814094
7	0.827	$5.32 \cdot 10^{-3}$	0.049	0.979	159.4	22628.9	-121.3	1350.3, 1565.8, -242.0, -1645.8, -708.2, -776.2, 456.0	903297.2

7.6.2 Results of the Simulation Study

The results of the simulation study described in Section 7.6.1 are provided for electricity data type E1 in Table 7.5 for normally distributed residuals and in Table 7.6 for Laplace distributed residuals. The results for the electricity data E2 can be found in Tables 7.7 and 7.8, respectively. The tables provide the estimator $\widehat{p}$ of the coverage probability from 5000 independent observations and the two-sided Clopper & Pearson confidence limits at the level 0.99 for the true coverage probability p. The comparison of confidence limits shows whether differences between the four methods are significant at the level $0.01 = 1 - 0.99$. In general, the estimated coverage probabilities decrease with increasing prediction lead time.

We review the results for the four methods in detail: **I1)**: The estimated coverage probabilities meet the nominal confidence level or tend to be only slightly too narrow for small prediction lead times k, but underperform for increasing k. **I2)**: The plug-in method, which ignores the uncertainty in the parameter estimation, is considerably worse than I1 and underestimates the nominal confidence level for all prediction lead times. **I3)**: This interval outperforms all others in terms of the coverage probability. Apart from a few exceptions for $k = 2$, the improvement is significant at the level 0.01. **I4)**: These intervals show the worst performance for the conducted simulation study and produce prediction intervals which are clearly too narrow.

Overall, the best results are obtained by method I3. The nominal confidence level of $\gamma = 0.95$ is met well for small prediction lead times k and only slightly underestimated for bigger lead times. Interval I4 is not competitive and shows a very bad performance. For the plug-in interval I2, the hypothesis is confirmed that these intervals tend to be too narrow. Apparently, method I3 is astonishingly robust against deviations from the normality assumption used in Section 7.5.2 for the derivation of the linear model results.

Table 7.5: Estimates $\widehat{p}$ of coverage probabilities and two-sided Clopper & Pearson confidence limits at the level 0.99 from 5000 simulation runs, electricity data type E1, normally distributed residuals.

	I1			I2			I3			I4		
k	p_L	$\widehat{p}$	p_U	p_L	$\widehat{p}$	p_U	p_L	$\widehat{p}$	p_U	p_L	$\widehat{p}$	p_U
1	0.9419	0.9504	0.9580	0.9174	0.9274	0.9366	0.9357	0.9446	0.9527	0.9242	0.9338	0.9426
2	0.9306	0.9398	0.9481	0.9037	0.9144	0.9243	0.9445	0.9528	0.9602	0.8793	0.8912	0.9023
3	0.9210	0.9308	0.9398	0.9020	0.9128	0.9228	0.9475	0.9556	0.9628	0.8508	0.8638	0.8761
4	0.8940	0.9052	0.9156	0.8835	0.8952	0.9061	0.9301	0.9394	0.9478	0.8064	0.8208	0.8346
5	0.8633	0.8758	0.8876	0.8631	0.8756	0.8874	0.9189	0.9288	0.9379	0.7483	0.7642	0.7796
6	0.8253	0.8392	0.8524	0.8307	0.8444	0.8574	0.8896	0.9010	0.9116	0.6948	0.7116	0.7280
7	0.9372	0.9460	0.9540	0.9227	0.9324	0.9413	0.9436	0.9520	0.9595	0.9138	0.9240	0.9334

Table 7.6: Estimates $\widehat{p}$ of coverage probabilities and two-sided Clopper & Pearson confidence limits at the level 0.99 from 5000 simulation runs, electricity data type E1, Laplace distributed residuals.

	I1			I2			I3			I4		
k	p_L	$\widehat{p}$	p_U	p_L	$\widehat{p}$	p_U	p_L	$\widehat{p}$	p_U	p_L	$\widehat{p}$	p_U
1	0.9229	0.9326	0.9415	0.9034	0.9142	0.9242	0.9191	0.9290	0.9381	0.9094	0.9198	0.9294
2	0.9165	0.9266	0.9358	0.8990	0.9100	0.9202	0.9306	0.9398	0.9482	0.8754	0.8874	0.8987
3	0.9094	0.9198	0.9294	0.8929	0.9042	0.9147	0.9368	0.9456	0.9536	0.8556	0.8684	0.8805
4	0.8835	0.8952	0.9061	0.8758	0.8878	0.8991	0.9254	0.9350	0.9437	0.8101	0.8244	0.8381
5	0.8687	0.8810	0.8926	0.8668	0.8792	0.8909	0.9172	0.9272	0.9364	0.7684	0.7838	0.7987
6	0.8266	0.8404	0.8536	0.8338	0.8474	0.8603	0.8911	0.9024	0.9130	0.7008	0.7176	0.7339
7	0.9189	0.9288	0.9379	0.9098	0.9202	0.9298	0.9237	0.9334	0.9422	0.8984	0.9094	0.9196

7.7 Application of Prediction Intervals in Forecasting

We apply the prediction intervals of types I1, I2, I3, I4 introduced in Section 7.5.3 to the two electricity load time series E1 and E2, which provided the basis for the simulation study considered in Section 7.6. It is interesting to compare forecasts retrospectively with the realised values. To this end, we take the 7 last days of the available data as the forecasting period, i. e. 4 July to 10 July 2005 for E1 and 27 October to 2 November 2005 for E2. The model ADT-AS was fitted to the remaining data, i. e. 9 May to 3 July 2005 for E1 and 1 September to 26 October 2005 for E2.

Figures 7.1 and 7.2 show the prediction intervals at the nominal level 95 % together with the observed daily electricity demand values (white line). The widths of the prediction intervals for the seven forecasts are displayed by Figs. 7.3 and 7.4. As to be expected, the

Table 7.7: Estimates $\widehat{p}$ of coverage probabilities and two-sided Clopper & Pearson confidence limits at the level 0.99 from 5000 simulation runs, electricity data type E2, normally distributed residuals.

	I1			I2			I3			I4		
k	p_L	$\widehat{p}$	p_U	p_L	$\widehat{p}$	p_U	p_L	$\widehat{p}$	p_U	p_L	$\widehat{p}$	p_U
1	0.9233	0.9330	0.9418	0.8938	0.9050	0.9154	0.9178	0.9278	0.9370	0.8990	0.9100	0.9202
2	0.9032	0.9140	0.9240	0.8695	0.8818	0.8933	0.9254	0.9350	0.9437	0.8253	0.8392	0.8524
3	0.8887	0.9002	0.9109	0.8579	0.8706	0.8826	0.9214	0.9312	0.9402	0.7864	0.8014	0.8158
4	0.9276	0.9370	0.9456	0.9039	0.9146	0.9245	0.9357	0.9446	0.9527	0.8938	0.9050	0.9154
5	0.9216	0.9314	0.9403	0.8967	0.9078	0.9181	0.9329	0.9420	0.9502	0.8837	0.8954	0.9063
6	0.8326	0.8462	0.8592	0.8132	0.8274	0.8410	0.8791	0.8910	0.9021	0.7131	0.7296	0.7457
7	0.7304	0.7466	0.7624	0.7520	0.7678	0.7831	0.8291	0.8428	0.8559	0.5799	0.5980	0.6159

Table 7.8: Estimates $\widehat{p}$ of coverage probabilities and two-sided Clopper & Pearson confidence limits at the level 0.99 from 5000 simulation runs, electricity data type E2, Laplace distributed residuals.

	I1			I2			I3			I4		
k	p_L	$\widehat{p}$	p_U	p_L	$\widehat{p}$	p_U	p_L	$\widehat{p}$	p_U	p_L	$\widehat{p}$	p_U
1	0.9091	0.9196	0.9293	0.8856	0.8972	0.9080	0.9066	0.9172	0.9270	0.8894	0.9008	0.9115
2	0.9045	0.9152	0.9251	0.8804	0.8922	0.9033	0.9278	0.9372	0.9458	0.8496	0.8626	0.8749
3	0.8913	0.9026	0.9132	0.8575	0.8702	0.8822	0.9293	0.9386	0.9471	0.7926	0.8074	0.8216
4	0.9178	0.9278	0.9370	0.8978	0.9088	0.9190	0.9244	0.9340	0.9428	0.8898	0.9012	0.9118
5	0.9223	0.9320	0.9409	0.9045	0.9152	0.9251	0.9318	0.9410	0.9493	0.8948	0.9060	0.9164
6	0.8417	0.8550	0.8676	0.8235	0.8374	0.8507	0.8881	0.8996	0.9103	0.7400	0.7560	0.7715
7	0.7245	0.7408	0.7567	0.7477	0.7636	0.7790	0.8256	0.8394	0.8526	0.5676	0.5858	0.6038

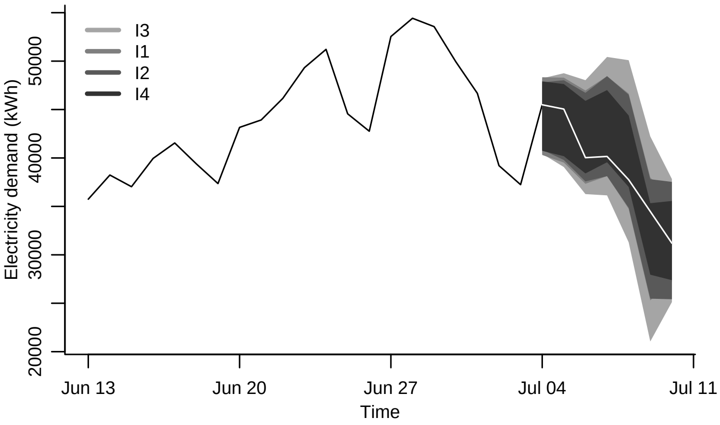


Figure 7.1: 7-day-forecast for electricity data E1

ascending order of widths essentially replicates the ascending order of actual coverage probabilities found by the simulation study, in the succession I4, I2, I1, I3. The widths of I1, I2, I3 are clearly associated. At moderate forecasting lead times $1 \leq k \leq 4$, the width of the best interval I3 does not differ too much from the widths of the worse intervals. At larger forecasting lead times, the width of the best interval I3 can be considerably higher than the widths of worse intervals. Not surprisingly, the improved actual coverage probability can sometimes entail a considerable loss in forecasting precision. In Fig. 7.2, for example, the standard deviation $\widehat{\sigma}_{\xi,\mathrm{ES}}$ at lead time 7 is almost three times higher than $\widehat{\sigma}_{\xi,\mathrm{LM}}$. Interval I1 compensates this by the factor $\sqrt{u_{kk}}$ in Eq. (7.15) being close to 3. Combining $\widehat{\sigma}_{\xi,\mathrm{ES}}$ and $\sqrt{u_{kk}}$ in the interval I3 leads to an inflation of its length. However, in particular for large forecasting lead times, the simulation study shows that the apparent precision of classical plug-in intervals like I2 is a fallacious illusion, bought with a drastic loss of intended forecasting reliability.

7.8 Conclusion and Outlook

We have presented in this chapter a number of prediction intervals applicable to exponential smoothing with covariates (ESCov): Prediction intervals exploiting the empirical error distribution have been considered in the form of prediction intervals based on empirical quantiles, normal quantiles and on two statistical inequalities – the Chebychev and Camp-Meidell inequalities. Apart from the one using normal distribution quantiles, they are all of non-parametric type. A bootstrap approach as well as a prediction interval of plug-in type (PIM) based on the ESCov SSOE model has been presented. The bootstrap prediction interval takes uncertainty in the estimation of parameters into

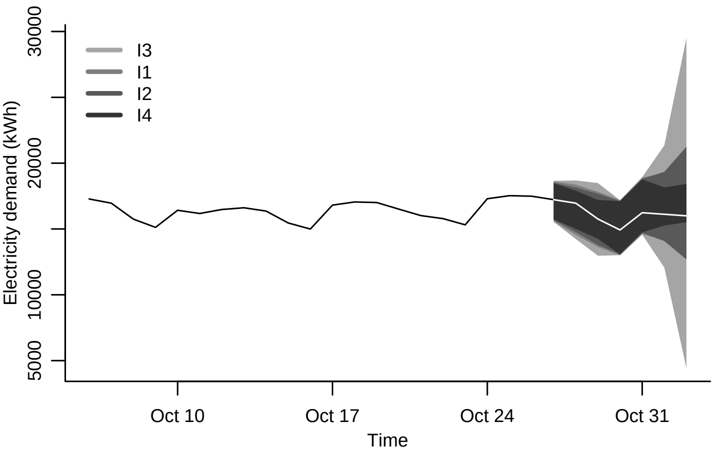


Figure 7.2: 7-day-forecast for electricity data E2

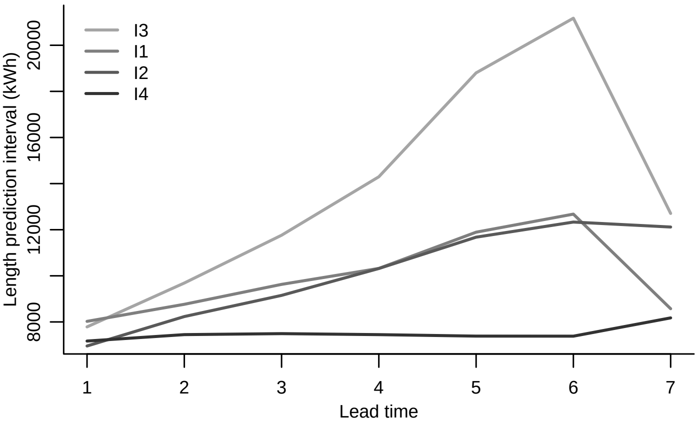


Figure 7.3: Widths of prediction intervals for the 7-day-forecast for electricity data E1

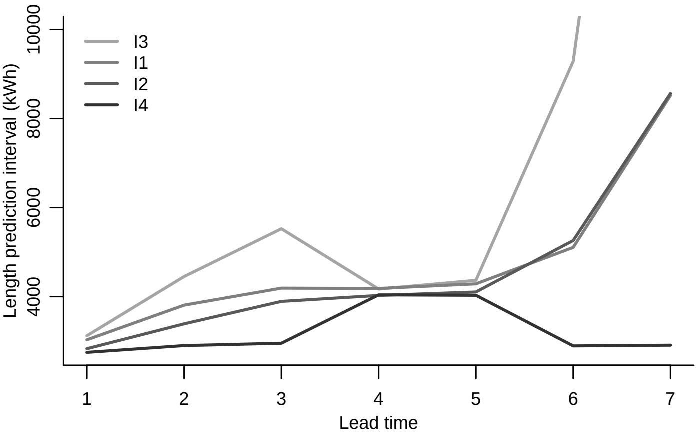


Figure 7.4: Widths of prediction intervals for the 7-day-forecast for electricity data E2

account. It is not of analytical nature and therefore involves extensive calculations. The PIM type prediction interval on the other hand is easy to calculate, but does not provide for parameter estimation uncertainty. A prediction interval that seems to overcome both remedies is the prediction interval based on linear expansion of the ESCov model. The idea has been adopted from Ord et al. (1997), who exploited linear model theory to derive prediction intervals for exponential smoothing without covariates, and modified to be applicable for ESCov. An empirical comparison of model-based prediction intervals based on the PIM method and on the linearisation method and combinations thereof has been performed in form of a simulation study and applied to daily electricity load data from an Italian energy vendor.

The results of the simulation study confirm the reservations about plug-in prediction intervals. Different from the findings of Kim et al. (2011) for exponential smoothing without covariates, the plug-in method performs rather bad under exponential smoothing with covariates. The plug-in method's lack of accounting for the uncertainty in estimating the covariate coefficients is the most likely reason for this behaviour. In contrast, the intervals derived from a linearisation of the underlying ESCov model perform very well. The simulation study revealed the best results in terms of the coverage probability for the t-distribution interval I3 with the standard deviation estimated by the SSOE plug-in estimator $\widehat{\sigma}_{\xi,\mathrm{ES}}$. For smaller lead times, the actual coverage probability is close to the nominal confidence level. For larger lead times, the loss in coverage is significantly less than for the other three prediction intervals considered.

From the point of view of empirical practice, the linearisation intervals have clear advantages over competing methods like simulation or resampling. The linearisation method is easily implemented and operates very economically in computation time. Taking the findings of Section 7.7 into consideration, we recommend to calculate both the intervals I3 as well as I1 for the practice and in general to trust interval I3. In case of an unfavourable constellation of the involved factors and a very broad prediction interval I3, which considerably lacks forecasting information, we recommend to use interval I1.

The results are promising, but still limited in empirical and in theoretical respects. In an empirical respect, more simulation studies are necessary to corroborate the superiority of the linearisation method and to investigate its behaviour. In a theoretical respect, we have considered a model with additive trend, additive season and additive residual. Future studies should consider the respective multiplicative model versions. It will also be necessary to get more analytical insight into the linearisation intervals. This problem may tentatively be approached by first considering simpler model instances, e. g. a mere local level model without trend and seasonality.

Several other prediction intervals presented in this chapter have not been investigated empirically. Further simulation studies should include them, too. In particular, the empirical prediction interval that exploits the Camp-Meidell inequality and relies on the estimation of central moments is a rather new idea. Experience in the application of the interval on real data is therefore still missing. Under heteroscedastic errors, the method is probably inadequate, but we expect it to be a competitive nonparametric alternative in applications, particularly when model-based methods are not applicable or normality assumptions do not hold. Empirical studies need to underpin the conjecture.

7.A Appendix

7.A.1 Derivation of the Components of the Linear Model Expansion of ESCov Model ADT-AS from Section 7.5.1

We derive the components $Y_s = H_s(\boldsymbol{\theta}, \boldsymbol{\xi})$, $\mathbf{M}$ and $\mathbf{L}$ for the ESCov model ADT-AS. The parameter vector is $\boldsymbol{\theta}^\top = (\widetilde{\alpha}_1, \widetilde{\alpha}_2, \widetilde{\alpha}_3, \phi, \boldsymbol{\beta}^\top)$.

Letting $t = 0$ and $h = s$ in the observation recursion (6.35) provides the recursion

$$\begin{aligned}
Y_s &= \boldsymbol{\delta}_{\widetilde{\boldsymbol{\alpha}}}^\top \left(\sum_{\ell=0}^{s-2} \mathbf{G}_{\widetilde{\boldsymbol{\alpha}}}^{\ell} \boldsymbol{w}_{\widetilde{\boldsymbol{\alpha}}} \xi_{s-\ell-1} + \mathbf{G}_{\widetilde{\boldsymbol{\alpha}}}^{s-1} \boldsymbol{u}_0 \right) + \boldsymbol{\beta}^\top \boldsymbol{x}_s + \xi_s \\
&= \boldsymbol{\delta}_{\widetilde{\boldsymbol{\alpha}}}^\top \left(\sum_{\ell=1}^{s-1} \mathbf{G}_{\widetilde{\boldsymbol{\alpha}}}^{\ell-1} \boldsymbol{w}_{\widetilde{\boldsymbol{\alpha}}} \xi_{s-\ell} + \mathbf{G}_{\widetilde{\boldsymbol{\alpha}}}^{s-1} \boldsymbol{u}_0 \right) + \boldsymbol{\beta}^\top \boldsymbol{x}_s + \xi_s \\
&= \boldsymbol{\delta}_{\widetilde{\boldsymbol{\alpha}}}^\top \left(\sum_{\ell=2}^{s-1} \mathbf{G}_{\widetilde{\boldsymbol{\alpha}}}^{\ell-1} \boldsymbol{w}_{\widetilde{\boldsymbol{\alpha}}} \xi_{s-\ell} + \boldsymbol{w}_{\widetilde{\boldsymbol{\alpha}}} \xi_{s-1} + \mathbf{G}_{\widetilde{\boldsymbol{\alpha}}}^{s-1} \boldsymbol{u}_0 \right) + \boldsymbol{\beta}^\top \boldsymbol{x}_s + \xi_s \\
&= \boldsymbol{\delta}_{\widetilde{\boldsymbol{\alpha}}}^\top \left(\sum_{\ell=1}^{s-2} \mathbf{G}_{\widetilde{\boldsymbol{\alpha}}}^{\ell} \boldsymbol{w}_{\widetilde{\boldsymbol{\alpha}}} \xi_{s-\ell-1} + \boldsymbol{w}_{\widetilde{\boldsymbol{\alpha}}} \xi_{s-1} + \mathbf{G}_{\widetilde{\boldsymbol{\alpha}}}^{s-1} \boldsymbol{u}_0 \right) + \boldsymbol{\beta}^\top \boldsymbol{x}_s + \xi_s \\
&= \boldsymbol{\delta}_{\widetilde{\boldsymbol{\alpha}}}^\top \left(\sum_{m=1}^{s-2} \mathbf{G}_{\widetilde{\boldsymbol{\alpha}}}^{s-m-1} \boldsymbol{w}_{\widetilde{\boldsymbol{\alpha}}} \xi_{m} + \boldsymbol{w}_{\widetilde{\boldsymbol{\alpha}}} \xi_{s-1} + \mathbf{G}_{\widetilde{\boldsymbol{\alpha}}}^{s-1} \boldsymbol{u}_0 \right) + \boldsymbol{\beta}^\top \boldsymbol{x}_s + \xi_s, \qquad (7.17)
\end{aligned}$$

where $\mathbf{G}_{\widetilde{\boldsymbol{\alpha}}}^0 = \mathbf{I}$. From (7.17) we obtain for $s \geq 1$

$$\frac{\partial}{\partial \xi_m} Y_s = \begin{cases} \boldsymbol{\delta}_{\widetilde{\boldsymbol{\alpha}}}^\top \mathbf{G}_{\widetilde{\boldsymbol{\alpha}}}^{s-m-1} \boldsymbol{w}_{\widetilde{\boldsymbol{\alpha}}} & \text{for } 1 \leq m \leq s-1, \\ 1 & \text{for } m = s, \\ 0 & \text{for } m \geq s+1. \end{cases} \qquad (7.18)$$

From Table 6.4 we obtain the subsequent results on $\boldsymbol{\delta}_{\widetilde{\boldsymbol{\alpha}}}^\top \mathbf{G}_{\widetilde{\boldsymbol{\alpha}}}^{\ell}$, $\mathbf{G}_{\widetilde{\boldsymbol{\alpha}}}^{\ell} \boldsymbol{w}_{\widetilde{\boldsymbol{\alpha}}}$, $\boldsymbol{\delta}_{\widetilde{\boldsymbol{\alpha}}}^\top \mathbf{G}_{\widetilde{\boldsymbol{\alpha}}}^{\ell} \boldsymbol{w}_{\widetilde{\boldsymbol{\alpha}}}$ and $\boldsymbol{\delta}_{\widetilde{\boldsymbol{\alpha}}}^\top \boldsymbol{w}_{\widetilde{\boldsymbol{\alpha}}}$.

$$\boldsymbol{\delta}_{\widetilde{\boldsymbol{\alpha}}}^\top \mathbf{G}_{\widetilde{\boldsymbol{\alpha}}}^{\ell} = (1, \phi, \mathbf{0}_{d-1}, 1) \mathbf{G}_{\widetilde{\boldsymbol{\alpha}}}^{\ell} = (1, \sum_{i=1}^{\ell+1} \phi^i, \boldsymbol{v}_\ell^\top),$$

where

$$\boldsymbol{v}_\ell = (v_{\ell,j})_{i \leq j \leq d} \text{ with } v_{\ell,j} = \begin{cases} 1 & \text{if } j = d(1 + \lfloor \frac{\ell}{d} \rfloor) - \ell, \\ 0, & \text{otherwise,} \end{cases}$$

$$\mathbf{G}_{\widetilde{\boldsymbol{\alpha}}}^{\ell} \boldsymbol{w}_{\widetilde{\boldsymbol{\alpha}}} = \begin{pmatrix} \widetilde{\alpha}_1 + \widetilde{\alpha}_2 \sum_{i=1}^{\ell} \phi^i \\ \widetilde{\alpha}_2 \phi \\ \widetilde{\alpha}_3 \boldsymbol{q}_\ell \end{pmatrix},$$

where

$$\boldsymbol{q}_\ell \;=\; (q_{\ell,j})_{1\le j\le d} \text{ with } q_{\ell,j} \;=\; \begin{cases} 1 & \text{if } j=\ell-\lfloor \frac{\ell}{d}\rfloor d+1, \\ 0 & \text{otherwise,} \end{cases}$$

$$\mathbf{G}_{\widetilde{\boldsymbol{\alpha}}}^{s-1}\boldsymbol{u}_0 \;=\; \mathbf{G}_{\widetilde{\boldsymbol{\alpha}}}^{s-1}\begin{pmatrix} \mu_0 \\ \Delta_0 \\ e_0 \\ \vdots \\ e_{-d+1} \end{pmatrix} \;=\; \mu_0+\Delta_0\sum_{i=1}^{s}\phi^i+\sum_{i=0}^{d-1}e_{-i}\mathbb{1}_{\{d(1+\lfloor\frac{s-1}{d}\rfloor)-s\}}(i),$$

where the indicator function provides $\mathbb{1}_{\{d(1+\lfloor\frac{s-1}{d}\rfloor)-s\}}(i)=1$ if $i=d(1+\lfloor\frac{s-1}{d}\rfloor)-s$, and otherwise $\mathbb{1}_{\{d(1+\lfloor\frac{s-1}{d}\rfloor)-s\}}(i)=0$. For $\ell=1,2,\ldots$ we have

$$\boldsymbol{\delta}_{\widetilde{\boldsymbol{\alpha}}}^\top\mathbf{G}_{\widetilde{\boldsymbol{\alpha}}}^{\ell}\boldsymbol{w}_{\widetilde{\boldsymbol{\alpha}}} \;=\; (1,\phi,\mathbf{0}_{d-1},1)\mathbf{G}_{\widetilde{\boldsymbol{\alpha}}}^{\ell}\begin{pmatrix} \widetilde{\alpha}_1 \\ \widetilde{\alpha}_2 \\ \widetilde{\alpha}_3 \\ \mathbf{0}_{d-1} \end{pmatrix} \;=\; \widetilde{\alpha}_1+\widetilde{\alpha}_2\sum_{i=1}^{\ell+1}\phi^i+\widetilde{\alpha}_3\mathbb{1}_{\{d(1+\lfloor\frac{\ell}{d}\rfloor)-s\}}(\ell), \tag{7.19}$$

where the indicator function provides $\mathbb{1}_{\{d(1+\lfloor\frac{\ell}{d}\rfloor)-1\}}(\ell)=1$ if $\ell=d(1+\lfloor\frac{\ell}{d}\rfloor)-1$, and otherwise $\mathbb{1}_{\{d(1+\lfloor\frac{\ell}{d}\rfloor)-1\}}(\ell)=0$. Finally

$$\boldsymbol{\delta}_{\widetilde{\boldsymbol{\alpha}}}^\top\boldsymbol{w}_{\widetilde{\boldsymbol{\alpha}}} \;=\; (1,\phi,\mathbf{0}_{d-1},1)\begin{pmatrix} \widetilde{\alpha}_1 \\ \widetilde{\alpha}_2 \\ \widetilde{\alpha}_3 \\ \mathbf{0}_{d-1} \end{pmatrix} \;=\; \widetilde{\alpha}_1+\widetilde{\alpha}_2\phi.$$

The latter result can also be obtained by letting $\ell=0$ in (7.19). Hence from (7.17)

$$\begin{aligned} Y_s \;=\; H_s(\boldsymbol{\theta},\boldsymbol{\xi}) \;&=\; \sum_{m=1}^{s-1}\left(\widetilde{\alpha}_1+\widetilde{\alpha}_2\sum_{i=1}^{s-m}\phi^i+\widetilde{\alpha}_3\mathbb{1}_{\{d\left(1+\lfloor\frac{s-m-1}{d}\rfloor\right)\}}(s-m)\right)\xi_m \\ &\quad+\mu_0+\Delta_0\sum_{i=1}^{s}\phi^i+\sum_{i=0}^{d-1}e_{-i}\mathbb{1}_{\{d\left(1+\lfloor\frac{s-1}{d}\rfloor\right)-s\}}(i)+\boldsymbol{\beta}^\top\boldsymbol{x}_s+\xi_s, \end{aligned}$$

i. e. Eq. (7.3). Furthermore,

$$\frac{\partial}{\partial \xi_m} Y_s = \begin{cases} \widetilde{\alpha}_1 + \widetilde{\alpha}_2 \sum_{i=1}^{s-m} \phi^i + \widetilde{\alpha}_3 \mathbb{1}_{\{d(1+\lfloor \frac{s-m-1}{d} \rfloor)\}}(s-m) & \text{for } 1 \leq m \leq s-1, \\ 1 & \text{for } m = s, \\ 0 & \text{for } m \geq s+1. \end{cases}$$

We find by taking derivatives in the succession $\frac{\partial}{\partial \widetilde{\alpha}_1} Y_s$, $\frac{\partial}{\partial \widetilde{\alpha}_2} Y_s$, $\frac{\partial}{\partial \widetilde{\alpha}_3} Y_s$, $\frac{\partial}{\partial \phi} Y_s$, $\frac{\partial}{\partial \beta_i} Y_s$

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{H}(\boldsymbol{\theta}, \boldsymbol{\xi})|_{\substack{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}} \\ \boldsymbol{\xi}=\widehat{\boldsymbol{\xi}}}} = & \Bigg(\sum_{m=1}^{s-1} \widehat{\xi}_m, \sum_{m=1}^{s-1} \sum_{i=1}^{s-m} \widehat{\phi}^i \widehat{\xi}_m, \sum_{m=1}^{s-1} \mathbb{1}_{\{d(1+\lfloor \frac{s-m-1}{d} \rfloor)\}}(s-m) \widehat{\xi}_m, \\ & \widehat{\widetilde{\alpha}}_2 \sum_{m=1}^{s-1} \widehat{\xi}_m \sum_{i=1}^{s-m} i \widehat{\phi}^{i-1} + \Delta_0 \sum_{m=1}^{s} m \widehat{\phi}^{m-1}, \ \boldsymbol{x}_s \Bigg)_{1 \leq s \leq T+h}, \end{aligned}$$

i. e. Eq. (7.4), where the sums $\sum_{m=1}^{1-1}$ are defined as 0.

We find $\mathbf{L} = (\boldsymbol{l}_s^\top)_{1 \leq s \leq T+h}$, where the vectors $\boldsymbol{l}_s \in \mathbb{R}^{T+h}$ are defined by $\boldsymbol{l}_1^\top := (1, 0, \ldots, 0)$ and

$$\boldsymbol{l}_s^\top := (\widetilde{\widehat{\alpha}}_1 + (s-1)\widetilde{\widehat{\alpha}}_2, \widetilde{\widehat{\alpha}}_1 + (s-2)\widetilde{\widehat{\alpha}}_2, \ldots, \widetilde{\widehat{\alpha}}_1 + \widetilde{\widehat{\alpha}}_2, 1, 0, \ldots, 0)$$

for $2 \leq s \leq T + h$.

7.A.2 Derivation of the Linear Model Results of Section 7.5.2

Let $\mathbf{W}_{\text{pp}} := \text{Cov}[\boldsymbol{Z}_{\text{p}}]$. Then

$$\mathbf{W}_{\text{pp}} = \text{Cov}[\boldsymbol{Z}_{\text{p}}] = \text{Cov}[\boldsymbol{\varepsilon}_{\text{p}}] = \text{Cov}[\mathbf{L}_{\text{pp}} \boldsymbol{\xi}_{\text{p}}] = \sigma_\xi^2 \mathbf{L}_{\text{pp}} \mathbf{L}_{\text{pp}}^\top. \tag{7.20}$$

From (7.9) and (7.10) we obtain

$$\text{Cov}[\boldsymbol{Z}_{\text{p}}, \boldsymbol{Z}_{\text{f}}] = \text{Cov}[\mathbf{L}_{\text{pp}} \boldsymbol{\xi}_{\text{p}}, \mathbf{L}_{\text{fp}} \boldsymbol{\xi}_{\text{p}}] = \sigma_\xi^2 \mathbf{L}_{\text{pp}} \mathbf{L}_{\text{fp}}^\top. \tag{7.21}$$

Let

$$\mathbf{D} := \mathbf{W}_{\text{pp}}^{-1} \text{Cov}[\boldsymbol{Z}_{\text{p}}, \boldsymbol{Z}_{\text{f}}] = (\mathbf{L}_{\text{pp}}^\top)^{-1} \mathbf{L}_{\text{pp}}^{-1} \mathbf{L}_{\text{pp}} \mathbf{L}_{\text{fp}}^\top = (\mathbf{L}_{\text{pp}}^\top)^{-1} \mathbf{L}_{\text{fp}}^\top. \tag{7.22}$$

We use linear model theory results from Christensen (1996) to derive the results of Section 7.5.2. The formula (7.12) for the unbiased variance estimate is obtained from

Christensen (1996, p. 31). From Christensen (1996, Theorem 12.2.3) it follows that the vector $P^\star(\boldsymbol{Z}_\mathrm{p})$ of best linear unbiased predictors for the components of $\boldsymbol{Z}_\mathrm{f}$ has the form

$$\begin{aligned} P^\star(\boldsymbol{Z}_\mathrm{p}) &= \Big[(\mathbf{M}_\mathrm{f}-\mathbf{D}^\top\mathbf{M}_\mathrm{p})\Big(\mathbf{M}_\mathrm{p}^\top\mathbf{W}_\mathrm{pp}^{-1}\mathbf{M}_\mathrm{p}\Big)^{-1}\mathbf{M}_\mathrm{p}^\top\mathbf{W}_\mathrm{pp}^{-1}+\mathbf{D}^\top\Big]\boldsymbol{Z}_\mathrm{p} \\ &= \Big[(\mathbf{M}_\mathrm{f}-\mathbf{L}_\mathrm{fp}\mathbf{L}_\mathrm{pp}^{-1}\mathbf{M}_\mathrm{p})\Big((\mathbf{L}_\mathrm{pp}^{-1}\mathbf{M}_\mathrm{p})^\top(\mathbf{L}_\mathrm{pp}^{-1}\mathbf{M}_\mathrm{p})\Big)^{-1}\mathbf{M}_\mathrm{p}^\top(\mathbf{L}_\mathrm{pp}^\top)^{-1}\mathbf{L}_\mathrm{pp}^{-1}+\mathbf{L}_\mathrm{fp}\mathbf{L}_\mathrm{pp}^{-1}\Big]\boldsymbol{Z}_\mathrm{p}. \end{aligned}$$

This proves Eq. (7.14).

Let the residuals $\xi_1,...,\xi_{T+h}$ have the normal distribution $N(0,\sigma_\xi^2)$. To justify the prediction interval (7.15), we consider the distribution of the ratios

$$\frac{Z_{T+k}-P^\star(\boldsymbol{Z}_\mathrm{p})_k}{\widehat{\sigma}_{\xi,\mathrm{LM}}\sqrt{u_{kk}}} \quad \text{for } k=1,...,h. \tag{7.23}$$

From Eqs. (7.9), (7.10), (7.13) and (7.14) we obtain

$$\boldsymbol{Z}_\mathrm{f}-P^\star(\boldsymbol{Z}_\mathrm{p}) = \mathbf{M}_\mathrm{f}\boldsymbol{\theta}+\mathbf{L}_\mathrm{ff}\boldsymbol{\xi}_\mathrm{f}+\mathbf{L}_\mathrm{fp}\boldsymbol{\Sigma}\boldsymbol{\xi}_\mathrm{p}-\mathbf{M}_\mathrm{f}\Big((\mathbf{L}_\mathrm{pp}^{-1}\mathbf{M}_\mathrm{p})^\top(\mathbf{L}_\mathrm{pp}^{-1}\mathbf{M}_\mathrm{p})\Big)^{-1}\mathbf{M}_\mathrm{p}^\top(\mathbf{L}_\mathrm{pp}^\top)^{-1}\boldsymbol{\xi}_\mathrm{p}. \tag{7.24}$$

From (7.13) it is easy to see that $\boldsymbol{\Sigma}^2=\boldsymbol{\Sigma}$ and hence

$$\Big[\mathbf{L}_\mathrm{fp}\boldsymbol{\Sigma}-\mathbf{M}_\mathrm{f}\Big((\mathbf{L}_\mathrm{pp}^{-1}\mathbf{M}_\mathrm{p})^\top(\mathbf{L}_\mathrm{pp}^{-1}\mathbf{M}_\mathrm{p})\Big)^{-1}\mathbf{M}_\mathrm{p}^\top(\mathbf{L}_\mathrm{pp}^\top)^{-1}\Big][\mathbf{I}-\boldsymbol{\Sigma}] = \mathbf{O}.$$

From the latter result, Eq. (7.24), the independence of $\boldsymbol{\xi}_\mathrm{p}$, $\boldsymbol{\xi}_\mathrm{f}$ and $\boldsymbol{\Sigma}^\top=\boldsymbol{\Sigma}$ we get

$$\begin{aligned} &\mathrm{Cov}[\boldsymbol{Z}_\mathrm{f}-P^\star(\boldsymbol{Z}_\mathrm{p}),[\mathbf{I}-\boldsymbol{\Sigma}]\mathbf{L}_\mathrm{pp}^{-1}\boldsymbol{Z}_\mathrm{p}] \\ &\quad= \Big[\mathbf{L}_\mathrm{fp}\boldsymbol{\Sigma}-\mathbf{M}_\mathrm{f}\Big((\mathbf{L}_\mathrm{pp}^{-1}\mathbf{M}_\mathrm{p})^\top(\mathbf{L}_\mathrm{pp}^{-1}\mathbf{M}_\mathrm{p})\Big)^{-1}\mathbf{M}_\mathrm{p}^\top(\mathbf{L}_\mathrm{pp}^\top)^{-1}\Big]\mathrm{Cov}[\boldsymbol{\xi}_\mathrm{p}][\mathbf{I}-\boldsymbol{\Sigma}] \\ &\quad= \mathbf{O}. \end{aligned}$$

Because of the normality assumption on the residuals $\xi_1,...,\xi_{T+h}$, the latter result demonstrates the independence of $\boldsymbol{Z}_\mathrm{f}-P^\star(\boldsymbol{Z}_\mathrm{p})$ and $[\mathbf{I}-\boldsymbol{\Sigma}]\mathbf{L}_\mathrm{pp}^{-1}\boldsymbol{Z}_\mathrm{p}$. With Eq. (7.12) we obtain the independence of $\boldsymbol{Z}_\mathrm{f}-P^\star(\boldsymbol{Z}_\mathrm{p})$ and $\widehat{\sigma}_{\xi,\mathrm{LM}}^2$. Letting

$$\mathbf{K} := \mathbf{L}_\mathrm{fp}\boldsymbol{\Sigma}-\mathbf{M}_\mathrm{f}\Big((\mathbf{L}_\mathrm{pp}^{-1}\mathbf{M}_\mathrm{p})^\top(\mathbf{L}_\mathrm{pp}^{-1}\mathbf{M}_\mathrm{p})\Big)^{-1}\mathbf{M}_\mathrm{p}^\top(\mathbf{L}_\mathrm{pp}^\top)^{-1},$$

matrix algebra shows that

$$\mathbf{K}\mathbf{K}^\top = (\mathbf{M}_\mathrm{f}-\mathbf{L}_\mathrm{fp}\mathbf{L}_\mathrm{pp}^{-1}\mathbf{M}_\mathrm{p})\Big((\mathbf{L}_\mathrm{pp}^{-1}\mathbf{M}_\mathrm{p})^\top(\mathbf{L}_\mathrm{pp}^{-1}\mathbf{M}_\mathrm{p})\Big)^{-1}(\mathbf{M}_\mathrm{f}-\mathbf{L}_\mathrm{fp}\mathbf{L}_\mathrm{pp}^{-1}\mathbf{M}_\mathrm{p})^\top.$$

Hence with definition (7.16) and (7.24) we see that $\boldsymbol{Z}_\mathrm{f}-P^\star(\boldsymbol{Z}_\mathrm{p})$ has a normal distribution with zero mean and with

$$\mathrm{Cov}[\boldsymbol{Z}_\mathrm{f}-P^\star(\boldsymbol{Z}_\mathrm{p})] = \sigma_\xi\mathbf{U}.$$

Christensen (1996, p. 31) states that $(T-r)\widehat{\sigma}^2_{\xi,\mathrm{LM}}/\sigma^2_\xi$ has the central χ^2-distribution $\chi^2(T-r)$. Because of the independence of numerator and denominator, we can conclude that the ratios

$$\frac{Z_{T+k} - P^{\star}(\boldsymbol{Z}_{\mathrm{p}})_k}{\widehat{\sigma}_{\xi,\mathrm{LM}}\sqrt{u_{kk}}}$$

have the central t-distribution $t(T-r)$.

8 Summary and Outlook

This thesis has dealt with confidence intervals and prediction under prior information and covariates. In the first part, the general theory of minimum volume confidence intervals for distribution parameters has been presented. The crucial idea behind the approach is to exploit the duality between confidence regions and prediction regions. By determining prediction regions containing points of maximum prediction likelihood ratio, confidence regions can be found such that the whole measurement and prediction space is of minimum weighted volume. Prior knowledge on the parameter Y of interest can be expressed by stipulating an appropriate distribution on Y. The theory has been applied to the probability parameter p of a binomial distribution and the expectation λ of a Poisson distribution. Prior knowledge has been expressed by means of a beta distribution $\mathrm{Beta}(p_0, p_1, a, b)$ with shape parameters $a, b > 0$ on the support $[p_0; p_1]$ in the case of the binomial probability p and a gamma distribution $\mathrm{Gamma}(\kappa, \vartheta)$ with shape and scale parameters $\kappa, \vartheta > 0$ for the Poisson expectation λ. The resulting two-sided confidence intervals are of frequentist type and in that sense always exact, i. e. the pointwise coverage is at least a prespecified confidence level γ. In contrast to existing exact confidence intervals like those by Clopper & Pearson (1934), they assign different weights to the confidence intervals in dependence on the outcomes $x = 0, 1, \ldots$, and the intervals are seeked to be shorter in weighted average.

With the binomial distribution and the Poisson distribution, two discrete distributions have been investigated for which underdispersion and equidispersion, respectively, hold. The theory of minimum volume confidence intervals under prior information should be applied to other distributions as well, including the negative binomial distribution as an instance of a distribution showing overdispersion.

Bayesian credibility intervals with special focus on binomial intervals have been examined and compared with the two-sided frequentist confidence intervals of minimum weighted volume. Although both approaches make use of prior information in the form of the conjugate distribution for the binomial – the beta distribution –, the approaches are markedly different. While the frequentist intervals are concerned with maximising the prediction likelihood ratio and the confidence intervals are derived from the prediction regions, the Bayesian highest posterior density (HPD) intervals maximise the density of the posterior distribution. The HPD credibility intervals do not fulfil the typical frequentist criterion of exactness with respect to a pointwise coverage of at least the stipulated credibility level β. Their coverage probability functions look very different from the frequentist ones. It could be shown, however, that under the restriction that both shape parameters of the beta prior distribution on $[0;1]$ are at most 1, a β can be found such that the pointwise coverage of the corresponding level β HPD interval is at least a prescribed level γ. The equivalent statement for the Poisson credibility intervals is still pending and will not be able to be achieved by means of the proof idea in the binomial case because it does not come with the comfort of having a prediction space of finite cardinality. Further work in context with shortest Poisson confidence limits involves their investigation from the point of view of limiting properties of the binomial interval.

The minimum volume confidence intervals for a probability have been used in a two-sided confidence interval of Stringer type, which is a procedure to estimate the mean in zero-inflated populations. Originally, it was proposed by Stringer (1963) in its one-sided version to be applied in audit sampling while making use of exact one-sided binomial confidence bounds. In the two-sided version presented in this thesis, the use of the shortest confidence intervals causes the Stringer interval to lead to a considerable reduction in length under certain prior information distributions. Statistical audit procedures based on this interval would require a minimum sample size of sometimes more than 20 % lower in contrast to intervals without prior information. Under more extreme prior information, the two-sided Stringer interval holds indifference and acceptance properties close to the one-sided version; yet the two-sided version has the clear advantage of being able to lead to the rejection of the population.

The presented two-sided Stringer interval is applicable under the assumption that the random tainting $Y = (U - W)/U$ ranges between 0 and 1 with probability 1 and the de facto value W is smaller or equal to the book value U in an auditing population. This assumption of overstatement is valid for certain types of audit populations. Further work has to be done to develop a confidence interval for the mean in zero-inflated populations that is applicable in the presence of both over- and understatement errors. Poisson confidence intervals instead of binomial confidence intervals should be used in the two-sided Stringer interval to possibly find out about limiting characteristics of the binomial version. Further simulation and empirical studies should support the interval's conservativeness and applicability.

Prediction under covariates has been considered in the context of time series analysis in the second part of this thesis. Exponential smoothing with covariates (ESCov) as a forecasting method that combines the history-based method of exponential smoothing with an additive covariate term, is supposed to improve the univariate prediction by additional exogenous information. The single source of error (SSOE) state-space model underlying ESCov has been formulated for multiple seasonalities. The minimum mean square error prediction and its variance have been presented in a linear and a partially linear SSOE version, covering the most popular methods of linear, potentially damped trend and additive or multiplicative seasonality models. The scheme by Roberts (1982), McKenzie (1986) and Archibald & Koehler (2003) to renormalise seasonal components in exponential smoothing such that they remain well interpretable has been transferred to the multiple seasonality case for ESCov. In a study of forecasting the hourly electricity load in Northern Italy, the double seasonality ESCov model has been applied with the temperature as covariate. ESCov has shown good performance as a load forecasting methodology. The influence of the covariate has been found to be more pronounced for larger forecasting horizons than for very short-term forecasting horizons of only a few hours ahead.

Various types of prediction intervals for ESCov have been described in the last chapter of this thesis. The frequently encountered behaviour of plug-in prediction intervals being too narrow could be supported in a simulation study. A prediction interval for the linear ESCov model has been formulated that exploits the theory of linear models to account also for the uncertainty in the estimation of the parameters.

The interaction between the states and the covariates in ESCov has not yet been fully investigated. In particular, the choice of the initial state values seems to have influence on the estimation of the smoothing and covariate parameters, which needs to be examined. Further theoretical work needs to be done by modifying the SSOE model for ESCov to account for higher-order autoregressive residuals. Only the case of AR(1) errors has so far been considered. Alternative models need to be investigated where the covariates are treated as states in the SSOE model. Further research should also include a more detailed comparison of ESCov with ARIMAX methods. With respect to prediction intervals for ESCov, further work should include linear model based prediction intervals for the multiplicative SSOE versions. Broader empirical and simulation studies should be performed to investigate the performance of the presented prediction intervals for ESCov.

Bibliography

Abramowitz, M., & Stegun, I. A. (Eds.) (1972). *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*. New York: Dover.

Agresti, A., & Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, *52*(2), 119–126.

Agresti, A., & Min, Y. (2001). On small-sample confidence intervals for parameters in discrete distributions. *Biometrics*, *57*(3), 963–971.

Agresti, A., & Min, Y. (2005). Frequentist performance of Bayesian confidence intervals for comparing proportions in 2x2 contingency tables. *Biometrics*, *61*, 515–523.

AICPA (1981). *Statement on Auditing Standards 39 - Audit Sampling*. American Institute of Certified Public Accountants. Auditing Standards Board.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Anderson, R., & Teitlebaum, A. D. (1973). Dollar-unit sampling. *Canadian Chartered Accountant*, *102*(4), 30–39.

Archibald, B. C., & Koehler, A. B. (2003). Normalization of seasonal factors in Winters' methods. *International Journal of Forecasting*, *19*, 143–148.

Athanasopoulos, G., & Hyndman, R. J. (2008). Modelling and forecasting Australian domestic tourism. *Tourism Management*, *29*, 19–31.

Ba, A., Sinn, M., Goude, Y., & Pompey, P. (2012). Adaptive learning of smoothing functions: Application to electricity load forecasting. In P. Bartlett, F. Pereira, C. Burges, L. Bottou, & K. Weinberger (Eds.) *Advances in Neural Information Processing Systems*, vol. 25, (pp. 2519–2527). Cambridge, Massachusetts: MIT Press.

Bayarri, M. J., & Berger, J. O. (2004). The interplay of Bayesian and frequentist analysis. *Statistical Science*, *19*(1), 58–80.

Bayes, T., & Price, R. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society*, *53*, 370–418.

Bélisle, C. J. P. (1992). Convergence theorems for a class of simulated annealing algorithms on Rd. *Journal of Applied Probability*, *29*(4), 885–895.

Berg, N. (2006). A simple Bayesian procedure for sample size determination in an audit of property value appraisals. *Real Estate Economics*, *34*(1), 133–155.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. New York, Berlin, Heidelberg: Springer Series in Statistics, Springer-Verlag.

Berger, J. O., Boukai, B., & Wang, Y. (1997). Unified frequentist and Bayesian testing of a precise hypothesis. *Statistical Science*, *12*(3), 133–160.

Bickel, P. J. (1992). Inference and auditing: The Stringer bound. *International Statistical Review*, *60*(2), 197–209.

Blaker, H. (2000). Confidence curves and improved exact confidence intervals for discrete distributions. *Canadian Journal of Statistics*, *28*(4), 783–798.

Blocher, E., & Robertson, J. C. (1976). Bayesian sampling procedures for auditors: Computer-assisted instruction. *The Accounting Review*, *51*(2), 359–363.

Blyth, C. R., & Hutchinson, D. W. (1960). Table of Neyman-shortest unbiased confidence intervals for the binomial parameter. *Biometrika*, *47*(3/4), 381–391.

Blyth, C. R., & Still, H. A. (1983). Binomial confidence intervals. *Journal of the American Statistical Association*, *78*(381), 108–116.

Bowman, K. O., & Shenton, L. R. (1988). *Properties of estimators for the gamma distribution*. New York: Marcel Dekker.

Box, G. E. P., & Jenkins, G. M. (1970). *Time Series Analysis, Forecasting and Control*. San Francisco: Holden Day.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1994). *Time Series Analysis: Forecasting and Control*. New Jersey: Prentice Hall, 3rd ed.

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley Publishing Company, Inc.

Box, G. E. P., & Tiao, G. C. (1975). Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical Association*, *70*(349), 70–79.

Brockwell, P. J., & Davis, R. A. (1996). *Introduction to Time Series and Forecasting*. New York: Springer, 2nd ed.

Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, *16*(2), 101–133.

Brown, R. G. (1956). Exponential smoothing for predicting demand. Tech. rep., Arthur D. Little, Inc., Cambridge 42, Massachusetts.

Brown, R. G. (1959). *Statistical forecasting for inventory control*. New York: McGraw-Hill.

Brown, R. G. (1963). *Smoothing, forecasting and prediction of discrete time series*. Englewood Cliffs, NJ: Prentice-Hall.

Bunn, D. W. (1982). Short-term forecasting: A review of procedures in the electricity supply industry. *The Journal of the Operational Research Society*, *33*(6), 533–545.

Burdick, R. K., & Reneau, J. H. (1978). The impact of different error distributions on the performance of selected sampling estimators in accounting populations. *Proceedings of the Business and Economical Statistics Section, American Statistical Association*, (pp. 779–781).

Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. Y. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, *16*(6), 1190–1208.

Camp, B. H. (1922). A new generalization of Tschebycheff's statistical inequality. *Bulletin of the American Mathematical Society*, *28*, 427–432.

Casella, G., & Berger, R. L. (2001). *Statistical Inference*. Duxbury Advanced Series, 2nd ed.

Casella, G., & Robert, C. (1988). Refining Poisson confidence intervals. *Canadian Journal of Statistics*, *17*, 45–57.

Chakhchoukh, Y., Panciatici, P., & Bondon, P. (2009). Robust estimation of SARIMA models: Application to short-term load forecasting. In *IEEE/SP 15th Workshop on Statistical Signal Processing, 2009. SSP '09*, (pp. 77–80). Cardiff, UK.

Chakhchoukh, Y., Panciatici, P., & Mili, L. (2011). Electric load forecasting based on statistical robust methods. *IEEE Transactions on Power Systems*, *26*(3), 982–991.

Chaloner, K. M., & Duncan, G. T. (1983). Assessment of a beta prior distribution: PM elicitation. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *32*(1/2), 174–180.

Chatfield, C. (1978). The Holt-Winters forecasting procedures. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *27*(3), 264–279.

Chatfield, C. (1993). Calculating interval forecasts. *Journal of Business & Economic Statistics*, *11*(2), 121–135.

Chatfield, C. (2001). Prediction intervals for time series. In J. S. Armstrong (Ed.) *Principles of Forecasting: A Handbook for Practitioners and Researchers*, vol. 30, (pp. 475–494). Springer.

Chatfield, C., & Yar, M. (1991). Prediction intervals for multiplicative Holt-Winters. *International Journal of Forecasting*, *7*(1), 31–37.

Chebyshev, P. (1867). Des valeurs moyennes. *Journal de mathématique pure et appliquée*, *12*, 177–184.

Cho, H., Goude, Y., Brossat, X., & Yao, Q. (2013). Modelling and forecasting daily electricity load curves: A hybrid approach. *Journal of the American Statistical Association*, *108*(501), 7–21.

Christensen, R. (1996). *Plane Answers to Complex Questions: The Theory of Linear Models*. New York, Berlin, Heidelberg: Springer-Verlag, 2nd ed.

Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, *26*(4), 404–413.

Cochran, W. G. (1963). *Sampling Techniques*. New York: John Wiley & Sons, Inc., 2nd ed.

Collopy, F., & Armstrong, J. S. (1992). Expert opinions about extrapolation and the mystery of the overlooked discontinuities. *International Journal of Forecasting*, *8*(4), 575–582.

Congdon, P. (2006). *Bayesian Statistical Modelling*. John Wiley & Sons Inc.

Corless, J. C. (1972). Assessing prior distributions for applying Bayesian statistics in auditing. *The Accounting Review*, *47*(3), 556–566.

CPSMR (1988). *Statistical Models and Analysis in Auditing: A Study of Statistical Models and Methods for Analyzing Nonstandard Mixtures of Distributions in Auditing*. Washington, D. C.: National Academy Press.

Crow, E. L. (1956). Confidence intervals for a proportion. *Biometrika*, *43*, 423–435.

Crow, E. L., & Gardner, R. S. (1959). Confidence intervals for the expectation of a Poisson variable. *Biometrika*, *46*(3/4), 441–453.

Datta, G. S., & Ghosh, J. K. (1995). On priors providing frequentist validity for Bayesian inference. *Biometrika*, *82*(1), 37–45.

Datta, G. S., Mukerjee, R., Ghosh, M., & Sweeting, T. J. (2000). Bayesian prediction with approximate frequentist validity. *The Annals of Statistics*, *28*(5), 1414–1426.

Datta, G. S., & Sweeting, T. J. (2005). Probability matching priors. *Handbook of statistics*, *25*, 91–114.

de Jager, N. G., Pap, G., & van Zuijlen, M. C. A. (1997). Facts, phantasies, and a new proposal concerning the Stringer bound. *Computers Math. Applic.*, *33*(5), 37–54.

Deming, W. E. (1960). *Sample Design in Business Research*. John Wiley & Sons.

Diebold, F. X. (2012). Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold-Mariano tests. Tech. rep., National Bureau of Economic Research.

Diego J. Pedregal, J. R. T. (2010). Mid-term hourly electricity forecasting based on a multi-rate approach. *Energy Conversion and Management*, *51*(1), 105–111.

Dordonnat, V., Koopman, S. J., & Ooms, M. (2012). Dynamic factors in periodic time-varying regressions with an application to hourly electricity load modelling. *Computational Statistics & Data Analysis*, *56*(11), 3134–3152.

Dordonnat, V., Koopman, S. J., Ooms, M., Dessertaine, A., & Collet, J. (2008). An hourly periodic state space model for modelling French national electricity load. *International Journal of Forecasting*, *24*(4), 566–587.

Fan, S., & Hyndman, R. J. (2012). Short-term load forecasting based on a semi-parametric additive model. *IEEE Transactions on Power Systems*, *27*(1), 134–141.

Fienberg, S. E., Neter, J., & Leitch, R. A. (1977). Estimating the total overstatement error in accounting populations. *Journal of the American Statistical Association*, *72*(358), 295–302.

Fisher, R. A. (1930). Inverse probability. *Proceedings of the Cambridge Philosophical Society*, *26*, 528–535.

Fraser, D. A. S. (2011). Is Bayes posterior just quick and dirty confidence? *Statistical Science*, *26*(3), 299–316.

Fraser, D. A. S., Reid, N., Marras, E., & Yi, G. Y. (2010). Default priors for Bayesian and frequentist inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(5), 631–654.

Gardner, E. S. (1985). Exponential smoothing: The state of the art. *Journal of Forecasting*, *4*(1), 1–28.

Gardner, E. S. (1988). A simple method of computing prediction intervals for time series forecasts. *Management Science*, *34*(4), 541–546.

Gardner, E. S. (2006). Exponential smoothing: The state of the art – Part II. *International Journal of Forecasting*, *22*(4), 637–666.

Gardner, E. S., & McKenzie, E. (1985). Forecasting trends in time series. *Management Science*, *31*(10), 1237–1246.

Gardner, E. S., & McKenzie, E. (1988). Model identification in exponential smoothing. *The Journal of the Operational Research Society*, *39*(9), 863–867.

Garthwaite, P. H., & O'Hagan, A. (2000). Quantifying expert opinion in the UK water industry: An experimental study. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *49*(4), 455–477.

Garwood, F. (1936). Fiducial limits for the Poisson distribution. *Biometrika*, *28*(3/4), 437–442.

Geisser, S. (1984). On prior distributions for binary trials. *The American Statistician*, *38*(4), 244–247.

George, E. I., Makov, U. E., & Smith, A. F. M. (1993). Conjugate likelihood distributions. *Scandinavian Journal of Statistics*, *20*(2), 147–156.

Gillet, P. R. (2000). Monetary unit sampling: A belief-function implementation for audit and accounting applications. *International Journal of Approximate Reasoning*, *25*, 43–70.

Göb, R., & Lurz, K. (2014). Design and analysis of shortest two-sided confidence intervals for a probability under prior information. *Metrika*, *77*(3), 389–413.

Göb, R., & Lurz, K. (2015). The use of inequalities of Camp-Meidell type in nonparametric statistical process monitoring. In S. Knoth, & W. Schmid (Eds.) *Frontiers in Statistical Quality Control 11*, (pp. 163–182). Berlin, Heidelberg: Springer International Publishing.

Göb, R., Lurz, K., & Pievatolo, A. (2013a). Electrical load forecasting by exponential smoothing with covariates. *Applied Stochastic Models in Business and Industry*, *29*, 629–645.

Göb, R., Lurz, K., & Pievatolo, A. (2013b). Rejoinder to the discussions of the paper on "Electrical load forecasting by exponential smoothing with covariates". *Applied Stochastic Models in Business and Industry*, *29*, 652–658.

Göb, R., Lurz, K., & Pievatolo, A. (2014). More accurate prediction intervals for exponential smoothing with covariates with applications in electrical load forecasting and sales forecasting. *Quality and Reliability Engineering International*.

Godfrey, J., & Neter, J. (1984). Bayesian bounds for monetary unit sampling in accounting and auditing. *Journal of Accounting Research*, (pp. 497–525).

Godfrey, J. T., & Andrews, R. W. (1982). A finite population Bayesian model for compliance testing. *Journal of Accounting Research*, *20*(2), 304–315.

Goodfellow, J. L., Loebbecke, J. K., & Neter, J. (1974). Some perspectives on CAV sampling plans. *Part I, CA Magazine (October 1974)*, (pp. 23–30).

Guy, D. M., Carmichael, D. R., & Whittington, R. (2002). *Audit Sampling: An Introduction*. John Wiley & Sons, Inc., 5th ed.

Hald, A. (1981). *Statistical theory of sampling inspection by attributes*. London: Academic Press.

Ham, J., Losell, D., & Smieliauskas, W. (1985). An empirical study of error characteristics in accounting populations. *Accounting Review*, *60*(3), 387–406.

Hansen, M. H., & Hurwitz, W. N. (1943). On the theory of sampling from finite populations. *The Annals of Mathematical Statistics*, *14*(4), 333–362.

Hewitt, E., & Stromberg, K. (1969). *Real and Abstract Analysis*. Berlin, Heidelberg, New York: Springer.

Hinman, J., & Hickey, E. (2009). Modeling and forecasting short-term electricity load using regression analysis. Tech. rep., Illinois State University.

Hogarth, R. M. (1975). Cognitive processes and the assessment of subjective probability distributions. *Journal of the American Statistical Association*, *70*(350), 271–289.

Holt, C. C. (1957). Forecasting seasonals and trends by exponentially weighted moving averages. *ONR Memorandum*, *52*. Pittsburgh, PA7 Carnegie Institute of Technology; Available from the Engineering Library, University of Texas at Austin.

Hyndman, R. J., Koehler, A. B., Ord, J. K., & Snyder, R. D. (2005). Prediction intervals for exponential smoothing using two new classes of state space models. *Journal of Forecasting*, *24*, 17–37.

Hyndman, R. J., Koehler, A. B., Snyder, R. D., & Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting*, *18*, 439–454.

Hyndman, R. J., Ord, A. B. B. K. K., & Snyder, R. D. (2008). Normalizing seasonal components. In *Forecasting with Exponential Smoothing*, Springer Series in Statistics, (pp. 123–136). Springer Berlin Heidelberg.

IAASB (2013). *Handbook of International Quality Control, Auditing, Review, Other Assurance, and Related Services Pronouncements*, vol. I. International Auditing and Assurance Standards Board.

Icerman, R. C., & Hillison, W. A. (1990). Distributions of audit-detected errors partitioned by internal control. *Journal of Accounting, Auditing & Finance*, *5*(4), 527–543.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, *186*(1007), 453–461.

Jenkinson, D. (2005). The elicitation of probabilities – a review of the statistical literature. Tech. rep., University of Sheffield, Sheffield, UK.

Johnson, J. R., Leitch, R. A., & Neter, J. (1981). Characteristics of errors in accounts receivable and inventory audits. *The Accounting Review*, *56*(2), 270–293.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1994). *Continuous Univariate Distributions*, vol. 1. Wiley, 2nd ed.

Johnson, N. L., Kotz, S., & Kemp, A. W. (1992). *Univariate Discrete Distributions*. John Wiley & Sons, Inc.

Kabaila, P., & Byrne, J. (2001). Exact short Poisson confidence intervals. *The Canadian Journal of Statistics*, *29*(1), 99–106.

Kadane, J. B., Dickey, J. M., Winkler, R. L., Smith, W. S., & Peters, S. C. (1980). Interactive elicitation of opinion for a normal linear model. *Journal of the American Statistical Association*, *75*(372), 845–854.

Kadane, J. B., & Wolfson, L. J. (1998). Experiences in elicitation. *The Statistician*, *47*(1), 3–19.

Kaplan, R. S. (1973). Statistical sampling in auditing with auxiliary information estimators. *Journal of Accounting Research*, *11*(2), 238–258.

Kass, R. E., & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, *91*(435), 1343–1370.

Kendrick, T. (2009). *Identifying and Managing Project Risk: Essential Tools for Failure-Proofing Your Project*. New York: AMACOM, 2nd ed.

Kerman, J. (2011). Neutral noninformative and informative conjugate beta and gamma prior distributions. *Electronic Journal of Statistics*, *5*, 1450–1470.

Khotanzad, T. J. T. . R. F. W. . A. (2007). The economic value of temperature forecasts in electricity generation. *Bulletin of the American Meteorological Society*, *86*, 1765–1771.

Kim, J. H., Wong, K., Athanasopoulos, G., & Liu, S. (2011). Beyond point forecasting: Evaluation of alternative prediction intervals for tourist arrivals. *International Journal of Forecasting*, *27*(3), 887–901.

Krishnamoorthy, K., & Peng, J. (2007). Some properties of the exact and score methods for binomial proportion and sample size calculation. *Communications in Statistics – Simulation and Computation*, *36*, 1171–1186.

Küsters, U., McCullough, B., & Bell, M. (2006). Forecasting software: Past, present and future. *International Journal of Forecasting*, *22*(3), 599–615. Twenty five years of forecasting.

Lawton, R. (1998). How should additive Holt-Winters estimates be corrected? *International Journal of Forecasting*, *14*, 393–403.

Leitch, R. A., Neter, J., Plante, R., & Sinha, P. (1982). Modified multinomial bounds for larger numbers of errors in audits. *Accounting Review*, *57*, 384–400.

Leslie, D. A., Teitlebaum, A. D., & Anderson, R. J. (1979). *Dollar-Unit Sampling: A Practical Guide for Auditors*. Copp Clark Pitman Toronto.

Lewandowski, R. (1979). *La Prévision à Court Terme*. Paris: Dunod.

Lewandowski, R. (1982). Practitioners' forum. Sales forecasting by forsys. *Journal of Forecasting*, *1*(2), 205–214.

Lewis, N. D. C. (2004). *Operational Risk with Excel and VBA: Applied Statistical Methods for Risk Management*. John Wiley and Sons.

Lindley, D. V. (1958). Fiducial distributions and Bayes' theorem. *Journal of the Royal Statistical Society. Series B (Methodological)*, (pp. 102–107).

Lindley, D. V. (1972). *Bayesian Statistics, A Review*. SIAM.

Loebbecke, J. K., & Neter, J. (1975). Considerations in choosing statistical sampling procedures in auditing. *Journal of Accounting Research*, *13*, 38–52.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., & Winkler, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting*, *1*(2), 111–153.

Makridakis, S., Hibon, M., Lusk, E., & Belhadjali, M. (1987). Confidence intervals: An empirical investigation of the series in the M-competition. *International Journal of Forecasting*, *3*(3), 489–508.

Marchand, É., & Strawderman, W. E. (2013). On Bayesian credible sets, restricted parameter spaces and frequentist coverage. *Electronic Journal of Statistics*, *7*, 1419–1431.

Marchand, É., Strawderman, W. E., Bosa, K., & Lmoudden, A. (2008). On the frequentist coverage of Bayesian credible intervals for lower bounded means. *Electronic Journal of Statistics*, *2*, 1028–1042.

Matsumura, E. M., Plante, R., Tsui, K.-W., & Kannan, P. (1991). Comparative performance of two multinomial-based methods for obtaining lower bounds on the total overstatement error in accounting populations. *Journal of Business & Economic Statistics*, *9*(4), 423–429.

McKenzie, E. (1986). Technical note – Renormalization of seasonals in Winters' forecasting systems: Is it necessary? *Operations Research*, *34*(1), 174–176.

Meidell, M. B. (1922). Sur un problème du calcul des probabilités et les statistiques mathématiques. *Comptes Rendus*, *175*, 806–808.

Meikle, G. R. (1972). *Statistical Sampling in an Audit Context*. Toronto: Canadian Institute of Chartered Accountants.

Molenaar, W. (1970). Approximations to the Poisson, binomial and hypergeometric distribution functions. *MC Tracts*, *31*, 1–160.

Neter, J., Johnson, J. R., & Leitch, R. A. (1985). Characteristics of dollar-unit taints and error rates in accounts receivable and inventory. *The Accounting Review*, *60*(3), 488–499.

Neter, J., Leitch, R. A., & Fienberg, S. E. (1978). Dollar unit sampling: Multinomial bounds for total overstatement and understatement errors. *The Accounting Review*, *53*(1), 77–93.

Neter, J., & Loebbecke, J. K. (1975). *Behavior of Major Statistical Estimators in Sampling Accounting Populations: An Empirical Study*. New York: American Institute of Certified Public Accountants.

Neter, J., & Loebbecke, J. K. (1977). On the behavior of statistical estimators when sampling accounting populations. *Journal of the American Statistical Association*, *72*(359), 501–507.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, *97*(4), 558–625.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society, Series A, Mathematical and Physical Sciences*, *236*(767), 333–380.

Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthese*, *36*(1), 97–131.

O'Hagan, A. (1998). Eliciting expert beliefs in substantial practical applications. *The Statistician*, *47*(1), 21–35.

Ord, J. K., Koehler, A. B., & Snyder, R. D. (1997). Estimation and prediction for a class of dynamic nonlinear statistical models. *Journal of the American Statistical Association*, *92*(440), 1621–1629.

Pap, G., & van Zuijlen, M. C. A. (1995). The Stringer bound in case of uniform taintings. *Computers Math. Applic.*, *29*(10), 51–59.

Pap, G., & van Zuijlen, M. C. A. (1996). On the asymptotic behaviour of the Stringer bound. *Statistica Neerlandica*, *50*(3), 367–389.

Papalexopoulos, A., & Hesterberg, T. (1990). A regression-based approach to short-term system load forecasting. *IEEE Transactions on Power Systems*, *5*(4), 1535–1547.

Park, J. H., Park, Y. M., & Lee, K. Y. (1991). Composite modeling for adaptive short-term load forecasting. *IEEE Transactions on Power Systems*, *6*(2).

Pegels, C. C. (1969). Exponential forecasting: Some new variations. *Management Science*, *15*(5), 311–315.

Plante, R., Neter, J., & Leitch, R. A. (1984). A lower multinomial bound for the total overstatement error in accounting populations. *Management Science*, *30*(1), 37–50.

Plante, R., Neter, J., & Leitch, R. A. (1985). Comparative performance of multinomial, cell, and Stringer bounds. *Auditing: A Journal of Practice & Theory*, *5*, 40–56.

Przyborowski, J., & Wileński, H. (1935). Statistical principles of routine work in testing clover seed for dodder. *Biometrika*, *27*(3/4), 273–292.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
URL http://www.R-project.org

Ramage, J. G., Krieger, A. M., & Spero, L. L. (1979). An empirical study of error characteristics in audit populations. *Journal of Accounting Research*, *17*, 72–102.

Ramanathan, R., Engle, R., Granger, C. W., Vahid-Araghi, F., & Brace, C. (1997). Short-run forecasts of electricity loads and peaks. *International Journal of Forecasting*, *13*(2), 161–174.

Reneau, J. H. (1978). CAV bounds in dollar unit sampling: Some simulation results. *Accounting Review*, (pp. 669–680).

Ricker, W. E. (1937). The concept of confidence or fiducial limits applied to the Poisson frequency distribution. *Journal of the American Statistical Association*, *32*(198), 349–356.

Roberts, S. A. (1982). A general class of Holt-Winters type forecasting models. *Management Science*, *28*(7), 808–820.

Ross, T. D. (2003). Accurate confidence intervals for binomial proportion and Poisson rate estimation. *Computers in Biology and Medicine*, *33*, 509–531.

Rousseau, J. (2000). Coverage properties of one-sided intervals in the discrete case and application to matching priors. *Annals of the Institute of Statistical Mathematics*, *52*(1), 28–42.

Ruckdeschel, P., Kohl, M., Stabla, T., & Camphausen, F. (2006). S4 classes for distributions. *R News*, *6*(2), 2–6.

Sahai, H., & Khurshid, A. (1993). Confidence intervals for the mean of a Poisson distribution – A review. *Biometrical Journal*, *35*(7), 857–867.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Scricciolo, C. (1999). Probability matching priors: A review. *Statistical Methods and Applications*, *8*(1), 83–100.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310.

Soares, L. J., & Medeiros, M. C. (2005). Modeling and forecasting short-term electric load demand: A two-step methodology. Tech. rep., Pontifícia Universidade Católica do Rio de Janeiro, Textos para Discussão.

Soares, L. J., & Medeiros, M. C. (2008). Modeling and forecasting short-term electricity load: A comparison of methods with an application to Brazilian data. *International Journal of Forecasting*, *24*(4), 630–644.

Sterne, T. E. (1954). Some remarks on confidence or fiducial limits. *Biometrika*, *41*, 275–278.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 44–47.

Stringer, K. W. (1963). Practical aspects of statistical sampling in auditing. *Proceedings of Business and Economic Statistics Section, American Statistical Association*.

Sugiura, N. (1978). Further analysts of the data by Akaike's information criterion and the finite corrections: Further analysts of the data by Akaike's. *Communications in Statistics – Theory and Methods*, *7*(1), 13–26.

Taylor, J. W. (2003a). Exponential smoothing with a damped multiplicative trend. *International Journal of Forecasting*, *19*, 715–725.

Taylor, J. W. (2003b). Short-term electricity demand forecasting using double seasonal exponential smoothing. *The Journal of the Operational Research Society*, *54*(8), 799–805.

Taylor, J. W. (2010). Triple seasonal methods for short-term electricity demand forecasting. *European Journal of Operational Research*, *204*, 139–152.

Taylor, J. W., de Menezes, L. M., & McSharry, P. E. (2006). A comparison of univariate methods for forecasting electricity demand up to a day ahead. *International Journal of Forecasting*, *22*, 1–16.

Taylor, J. W., & McSharry, P. E. (2007). Short-term load forecasting methods: An evaluation based on European data. *IEEE Transactions on Power Systems*, *22*(4), 2213–2219.

Teisberg, T. J., Weiher, R. F., & Khotanzad, A. (2005). The economic value of temperature forecasts in electricity generation. *Bulletin of the American Meteorological Society*, *86*(12), 1765–1771.

Teitlebaum, A. D. (1973). Dollar-unit sampling in auditing. In *National Meeting of the American Statistical Association, Montreal*.

Teitlebaum, A. D., & Robinson, C. F. (1975). The real risks in audit sampling. *Journal of Accounting Research*, *13*, 70–91.

Thatcher, A. R. (1964). Relationships between Bayesian confidence limits for predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, *26*(2), 176–210.

Tibshirani, R. (1989). Noninformative priors for one parameter of many. *Biometrika*, *76*(3), 604–608.

Tsallis, C., & Stariolo, D. A. (1996). Generalized simulated annealing. *Physica A: Statistical Mechanics and its Applications*, *233*(1–2), 395–406.

Tsui, K.-W., Matsumura, E. M., & Tsui, K.-L. (1985). Multinomial-Dirichlet bounds for dollar-unit sampling in auditing. *Accounting Review*, *60*(1), 76–96.

Uhlmann, W. (1982). *Statistische Qualitätskontrolle: eine Einführung*, vol. 7. Teubner.

von Collani, E., & Dräger, K. (2001). *Binomial Distribution Handbook for Scientists and Engineers*. Birkhäuser Boston.

von Collani, E., & Dumitrescu, M. (2001). Complete Neyman measurement procedure. *Metrika*, *54*, 111–130.

von Collani, E., Dumitrescu, M., & Lepenis, R. (2001). Neyman measurement and prediction procedures. *Economic Quality Control*, *16*(1), 109–132.

Walls, L., & Quigley, J. (2001). Building prior distributions to support Bayesian reliability growth modelling using expert judgement. *Reliability Engineering and System Safety*, *74*(2), 117–128.

Wang, H. (2009). Exact average coverage probabilities and confidence coefficients of confidence intervals for discrete distributions. *Statistics and Computing*, *19*(2), 139–148.

Wang, S. (2006). *Exponential Smoothing for Forecasting and Bayesian Validation of Computer Models*. Ph.D. thesis, Georgia Institute of Technology.

Wasserman, L. (2011). Frasian inference. *Statistical Science*, *26*(3), 322–325.

Welch, B. L., & Peers, H. W. (1963). On formulae for confidence points based on integrals of weighted likelihoods. *Journal of the Royal Statistical Society. Series B (Methodological)*, *25*(2), 318–329.

Weron, R. (2006). *Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach (The Wiley Finance Series)*. Wiley.

Williams, W. H., & Goodman, M. L. (1971). A simple method for the construction of empirical confidence limits for economic forecasts. *Journal of the American Statistical Association*, *66*(336), 752–754.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, *22*(158), 209–212.

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management Science*, *6*(3), 324–342.

Woodroofe, M. (1986). Very weak expansions for sequential confidence levels. *The Annals of Statistics*, *14*(3), 1049–1067.

Xiang, Y., Gubian, S., Suomela, B., & Hoeng, J. (2013). Generalized simulated annealing for global optimization: The GenSA package for R. *The R Journal*, *5*(1), 13–29.

Yar, M., & Chatfield, C. (1990). Prediction intervals for the Holt-Winters forecasting procedure. *International Journal of Forecasting*, *6*, 127–137.