Prediction and interpolation of time series by state space models

Jouni Helske
Department of Mathematics and Statistics
University of Jyväskylä

Abstract

A large amount of data collected today is in the form of a time series. In order to make realistic inferences based on time series forecasts, in addition to point predictions, prediction intervals or other measures of uncertainty should be presented. Multiple sources of uncertainty are often ignored due to the complexities involved in accounting them correctly. In this dissertation, some of these problems are reviewed and some new solutions are presented. A state space approach is also advocated for an efficient and flexible framework for time series forecasting, which can be used for combining multiple types of traditional time series and other models.

Acknowledgements

I wish to thank my supervisor, Professor Jukka Nyblom, for guiding me in my statistical journey from master's studies to the PhD thesis. His early steering of me towards computational statistics and software development, together with shared expertise from a broad class of statistics, has given me valuable experiences beyond the scope of my thesis.

I would also like to thank the Emil Aaltonen foundation for providing me the funding for the main part of my postdoctoral studies. I thank also the Finnish Doctoral Programme in Stochastics and Statistics (FDPSS) and the University of Jyväskylä for additional funding at the beginning of my studies.

I thank Professors Juha Alho and Pentti Saikkonen for reviewing my thesis. The department of Mathematics and Statistics provided the facilities, and I thank the administration staff who helped in many aspects of my studies. Intriguing statistical and non-statistical discussions with the fellow PhD students and the staff of Statistics in the coffee room, at lunch tables, and at the corridors of the department have taught and inspired me in countless ways. Special thanks to Salme Kärkkäinen for introducing me to my co-authors Petri Ekholm and Kristian Meissner from the Finnish Environment Institute, who I wish to thank especially for giving interesting perspectives to other fields of science. I also wish to thank Sara Taskinen for introducing me to Perttu Luukko, who I had a pleasure to work with in a side project resulting in a paper and an R package. And finally my deepest gratitude goes to the best co-author both in work and personal life, the love of my life, Satu.

I am grateful for the support given by relatives and friends, who might have believed in me in my unusual path of studies. This has been an interesting nonlinear time series full of significant events, the most significant one being my daughter, Aini. Thank you for setting things in perspective and for all those wonderful bursts of happiness you generate in people around you.

Jyväskylä, September 2015

Jouni Helske

List of original publications

This thesis consists of an introductory part and publications listed below.

I Helske, J. and Nyblom, J. (2015). Improved frequentist prediction intervals for autoregressive models by simulation. In Koopman, S. J. and Shephard, N., editors, *Unobserved Components and Time Series Econometrics*. Oxford University Press. In press.

II Helske, J. and Nyblom, J. (2014). Improved frequentist prediction intervals for ARMA models by simulation. In Knif, J. and Pape, B., editors, *Contributions to Mathematics, Statistics, Econometrics, and Finance: Essays in Honour of Professor Seppo Pynnönen*, number 296 in Acta Wasaensia, pages 71–86. University of Vaasa.

III Helske, J., Nyblom, J., Ekholm, P., and Meissner, K. (2013). Estimating aggregated nutrient fluxes in four Finnish rivers via Gaussian state space models. *Environmetrics*, 24(4):237– 247.

IV Helske, J. (2015). KFAS: Exponential family state space models in R. Submitted.

The author of this dissertation has been the primary author in all of the original publications. The original problem and the solution of paper [I] was suggested by the second author and further developed by the first author in [I] and [II]. The problem of paper [III] was suggested by the third and fourth authors, and the solutions were formulated by the first author with helpful discussions with the second author.

Contents

Chapter 1

Introduction

Assuming that we are not dealing with completely deterministic systems, no matter what statistical method or model we use, our predictions contain some uncertainty. The uncertainties arise from several sources and their effects on the final results should be carefully considered. In addition to point predictions, accompanying measures of uncertainty should be presented. In the traditional time series prediction, the underlying uncertainties relating to intermediate steps needed for the final forecasts are commonly ignored, which leads to results which seem more accurate than they really are. Ignoring the uncertainty regarding the chosen model and its parameters is a typical example. Dismissing these kind of issues can lead to prediction intervals which have coverage probabilities considerably smaller than the nominal level. This is a known problem, and, for example, Chatfield (1995, 1996) strongly criticizes this approach, stating that it is not enough to just perform diagnostic checks on the best fitting model, but also the process of model selection needs to be assessed. Unfortunately no general solutions exist.

Several categories of uncertainty in predictions have been made. Clements and Hendry (1999) present a detailed taxonomy of forecast errors, but here I use coarser classification similar to the ones in Chatfield (2000) and Alho and Spencer (2005). In addition to uncertainty caused by random variation of the process, common sources of uncertainty are:

1. True data-generating mechanisms are not known, and the chosen statistical model is only an approximation of the truth.

2. During our observation period or in the future which we are trying to forecast, the underlying data-generating processes can change in a way that is unaccounted by our model.

3. Parameters of the chosen model need to be estimated, often from the same data which was used in the model selection.

4. Data contains outliers or it is otherwise of poor quality, affecting the model identification, parameter estimation and forecasting.

5. Wrong distributional assumptions and data transformations.

This dissertation considers mainly the issues (3)–(5). After introducing the basic concepts of time series prediction in Chapter 2, autoregressive integrated moving average (ARIMA)

models and a general state space modeling framework is introduced in Chapter 3. In Chapter 4, effects of uncertainties relating to parameter estimation are discussed in detail. Finally, some additional problems are discussed in Chapter 5.

Chapter 2

Time series prediction

Define a time series $y_1, y_2, \ldots$ and a model $p(y_1, \ldots, y_n)$ for all $n = 1, 2, \ldots$ with p being a generic notation for a density or a probability mass function. Assume that we have observed the time series $y_1, \ldots, y_n$ for some n, and we wish to make predictions of future values given the past, i.e., we are interested in conditional densities

$$p(y_{n+h}|y_1, \ldots, y_n), \quad h = 1, 2, \ldots. \tag{2.1}$$

These conditional densities are often called predictive densities.

A typical choice for a point forecast of y_{n+h} is the conditional mean $\hat{y}_{n+h} = \mathrm{E}(y_{n+h}|y_1, \ldots, y_n)$, which solves the minimization problem

$$\min_f \int [y_{n+h} - f(y_1, \ldots, y_n)]^2 p(y_{n+h}|y_1, \ldots, y_n) \mathrm{d}y_{n+h}, \tag{2.2}$$

for all (measurable) functions f. Equivalently, the conditional mean $\hat{y}_{n+h}$ minimizes the mean square prediction error $\mathrm{E}[(y_{n+h} - f(y_1, \ldots, y_n))^2]$. For the proof, see, for example, (Pollock, 1999). Alternative choices include the conditional median $\bar{y}_{n+h}$, which minimizes the mean absolute deviation, and the conditional mode which gives the highest predictive density value. If the predictive density $p(y_{n+h}|y_1, \ldots, y_n)$ is symmetric around its mean with finite variance, then all three are equal. An important special case is the Gaussian density.

Sometimes the interest is not in the future values y_{n+h}, but in the missing intermediate values $y_{m+1}, \ldots, y_{m+h}$, $1 < m \leq m + h < n$. Thus our conditional densities of interest are

$$p(y_{m+j}|y_1, \ldots, y_m, y_{m+h+1}, \ldots, y_n), \quad j = 1, \ldots, h. \tag{2.3}$$

Denote the prediction error as $e_{n+h} = y_{n+h} - \hat{y}_{n+h}$. Under a Gaussian predictive density, it is straightforward to construct a $100(1 - 2\alpha)\%$ prediction interval for y_{n+h} as

$$y_{n+h} = \hat{y}_{n+h} \pm z_\alpha \sqrt{\mathrm{E}[e_{n+h}^2]}, \tag{2.4}$$

where z_α is the percentage point of the standard normal distribution with a proportion α above it. The future value y_{n+h} is expected to lie between the upper and lower limits of the interval with probability $1 - 2\alpha$. Note that given the unbiased forecast,

$$\mathrm{E}[e_{n+h}^2] = \mathrm{Var}[e_{n+h}] = \mathrm{Var}[y_{n+h}|y_1, \ldots, y_n]. \tag{2.5}$$

In practice, $p(y_1, \ldots, y_n)$ depends on unknown parameters and is actually $p(y_1, \ldots, y_n|\psi)$, where ψ is the parameter vector. Effects of unknown parameters will be discussed in Chapter 4.

Chapter 3

Time series models for continuous and count data

3.1 Autoregressive integrated moving average models

The unified modeling approach of ARIMA processes via a so called Box–Jenkins approach (Box and Jenkins, 1970) made them enormously popular in the 1970s. In the Box–Jenkins approach, after choosing the appropriate member of the class of ARIMA models based on the autocorrelation and partial autocorrelation functions (perhaps together with some information criteria such as AIC), the model parameters are estimated, and then diagnostics checks based on the model residuals are performed. If the model seems inadequate, alternative members of ARIMA models are tested until the forecaster is satisfied with the final model. The forecasts based on the final model are then computed using conditional expectations and variances given by the weights relating to the infinite-order moving average process presentation of the model, with the implicit assumption that the chosen model and its parameters are correct (Box and Jenkins, 1970). ARIMA models can be extended to handle seasonal patterns and exogenous variables, as well as multivariate series, making the ARIMA models applicable to the broad range of forecasting problems.

The univariate ARIMA(p,d,q) model without seasonal or exogenous variables can be written as

$$y_t^* = \phi_1 y_{t-1}^* + \ldots + \phi_p y_{t-p}^* + \xi_t + \theta_1 \xi_{t-1} + \ldots + \theta_q \xi_{t-q}, \tag{3.1}$$

where $y_t^* = \Delta^d y_t$ with Δ being a difference operator, and $\xi_t \sim \mathrm{N}(0, \sigma^2)$. Here p is the order of autoregressive part, d is the number of differencing and q is the order of the moving average part.

Although commonly used, even as a black-box approach, the Box–Jenkins modeling approach is not without problems. Real life time series are often non-stationary, and differencing possibly with transformations is needed, which affects the subsequent analysis. Choosing the correct order for autoregressive and moving average processes can also be a difficult task, and often there are multiple equally good candidate models. Harvey (1989, Section 2.6.4), Durbin (2000) and Chatfield (2004, Sections 5.2.5 and 5.4) discuss the practical problems regarding the ARIMA modeling, the main message perhaps being that considerable statistical experience is needed when using ARIMA models.

The point predictions and variances of the prediction errors can be obtained using the

model definition (3.1) and the so called ψ-weights (Box and Jenkins, 1970, Chapter 5) of the infinite moving average process representation of (3.1). However, the forecasts, as well as the log-likelihood of the model, can also be efficiently obtained using the state space approach with the Kalman filter recursions (Box and Jenkins, 1970, Chapter 5). This approach also allows missing values in time series. The state space models and the Kalman filter will be discussed in the next Section.

3.2 Linear Gaussian state space models

The linear Gaussian state space model can be written as

$$\begin{aligned} y_t &= Z_t\alpha_t + \epsilon_t, \qquad \text{(observation equation)} \\ \alpha_{t+1} &= T_t\alpha_t + R_t\eta_t, \qquad \text{(state equation)} \end{aligned} \tag{3.2}$$

where $\epsilon_t \sim N(0, H_t), \eta_t \sim N(0, Q_t)$ and $\alpha_1 \sim N(a_1, P_1)$ independently of each other. Here the vector y_t contains the observations at time t, whereas α_t is the vector of the latent state process at time t. The system matrices Z_t, T_t, and R_t, together with the covariance matrices H_t and Q_t depend on the particular model definition, and often some of these matrices contain unknown (hyper)parameters ψ which need to be estimated. If a particular matrix such as Z_t does not depend on t, it is said to be time-invariant, i.e., $Z_t = Z$ for all t. The prior state distribution $N(a_1, P_1)$ can be informative or (partially) non-informative, in which case P_1 can be decomposed to $P_* + \kappa P_\infty$ with $\kappa \to \infty$. Here P_* corresponds to the (possibly weakly) informative part of the initial state distribution, and P_∞ is a diagonal matrix with ones corresponding to the diffuse elements of the state vector and zeros elsewhere. See Koopman (1997) and Koopman and Durbin (2003) for details of this exact diffuse initialization.

Hyndman et al. (2008) advocates the use of the so called innovations representation of (3.2) where the disturbances of observation and state equations are perfectly correlated and not perfectly independent like in (3.2). This has some computational and other benefits especially in the case of exponential smoothing (Hyndman et al., 2008; Chatfield et al., 2001). But the general form of (3.2) also contains the special case of perfectly correlated ϵ and η, as the disturbances ϵ can be straightforwardly augmented to states α. Obviously other correlation structures than the perfect correlation are now also possible.

The ARIMA(p,d,q) model with stationary initial distribution for differenced series with $r = \max(p, q+1)$ can be written as a state space model by defining

$$Z' = \begin{pmatrix} 1_{d+1} \\ 0 \\ \vdots \\ 0 \end{pmatrix}, H = 0, T = \begin{pmatrix} U_d & 1_d' & 0 & \cdots & 0 \\ 0 & \phi_1 & 1 & & 0 \\ \vdots & & & \ddots & \\ \vdots & \phi_{r-1} & 0 & & 1 \\ 0 & \phi_r & 0 & \cdots & 0 \end{pmatrix}, R = \begin{pmatrix} 0_d \\ 1 \\ \theta_1 \\ \vdots \\ \theta_{r-1} \end{pmatrix},$$

$$\alpha_t = \begin{pmatrix} y_{t-1} \\ \vdots \\ \Delta^{d-1} y_{t-1} \\ y_t^* \\ \phi_2 y_{t-1}^* + \ldots + \phi_r y_{t-r+1}^* + \theta_1 \eta_t + \ldots + \theta_{r-1} \eta_{t-r+2} \\ \vdots \\ \phi_r y_{t-1}^* + \theta_{r-1} \eta_t \end{pmatrix}, Q = \sigma^2,$$

$$a_1 = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, P_{*,1} = \begin{pmatrix} 0 & 0 \\ 0 & S_r \end{pmatrix}, P_{\infty,1} = \begin{pmatrix} I_d & 0 \\ 0 & 0 \end{pmatrix}, \eta_t = \xi_{t+1},$$

where $\phi_{p+1} = \ldots = \phi_r = \theta_{q+1} = \ldots = \theta_{r-1} = 0$, 1_{d+1} is a $1 \times (d+1)$ vector of ones, U_d is $d \times d$ upper triangular matrix of ones and S_r is the covariance matrix of stationary elements of α_1. The elements of the initial state vector α_1, which correspond to the differenced values $y_0, \ldots, \Delta^{d-1} y_0$, are treated as diffuse. The covariance matrix S_r can be computed by solving the linear equation $(I - T \otimes T)\mathrm{vec}(S_r) = \mathrm{vec}(RR')$.

The state space representation of stationary ARMA process ($d = 0$) was first given in Harvey (1981), which has been subsequently used by many others, including Chib and Greenberg (1994) in a Bayesian framework. Durbin and Koopman (2012, p.137–139) generalized the formulation for non-stationary case.

Many other models, such as structural time series (Harvey, 1989), linear mixed models (Sallas and Harville, 1981; Tsimikas and Ledolter, 1997), and exponential smoothing methods (Hyndman et al., 2008), can be formulated as a state space model by carefully defining the system matrices of (3.2), and different kinds of models can also be combined straightforwardly. For example the ARIMA model and the linear regression model can be combined by adding regression coefficients into the state vector and explanatory variables into Z_t. Using diffuse initialization for the regression coefficients, the uncertainty corresponding their estimation is automatically accounted for.

3.2.1 Kalman filter and smoother

The main algorithms for the inference of Gaussian state space models are the Kalman filtering and smoothing recursions. From the Kalman filtering algorithm we obtain the one-step-ahead predictions and the prediction errors

$$\begin{aligned} a_{t+1} &= \mathrm{E}(\alpha_{t+1} | y_t, \ldots, y_1), \\ v_t &= y_t - \mathrm{E}(y_t | y_{t-1}, \ldots, y_1), \end{aligned} \tag{3.3}$$

and their covariance matrices

$$\begin{aligned} P_{t+1} &= \mathrm{Var}(\alpha_{t+1} | y_t, \ldots, y_1), \\ F_t &= \mathrm{Var}(v_t). \end{aligned} \tag{3.4}$$

Complete recursive formulas for the Kalman filtering can be written as follows.

$$\begin{aligned} v_t &= y_t - Z_t a_t \\ F_t &= Z_t P_t Z_t' + H_t \\ K_t &= P_t Z_t' \\ a_{t+1} &= T_t(a_t + K_t F_t^{-1} v_t) \\ P_{t+1} &= T_t(P_t - K_t F_t^{-1} K_t') T_t' + R_t Q_t R_t, \end{aligned} \tag{3.5}$$

where $K_t = \mathrm{Cov}(a_t, y_t | \ldots)$. The derivation of these formulas can be found, for example, in Durbin and Koopman (2012, Section 4.3.1), who show that these results are also valid without the normality assumption, as the a_{t+1} is the minimum variance linear unbiased estimate of α_{t+1} given $y_t, \ldots, y_1$. The same a_{t+1} is also the minimum variance linear posterior mean estimate. Therefore, given the hyperparameters ψ, the resulting predictive distributions are Bayesian posterior distributions given the prior distribution $N(a_1, P_1)$.

Using the results of the Kalman filtering, we establish the state smoothing equations (Durbin and Koopman, 2012, Section 4.4)

$$\begin{aligned} r_{t-1} &= Z_t' F_t^{-1} v_t + L_t' r_t \\ N_{t-1} &= Z_t' F_t^{-1} Z_t + L_t' N_t L_t \\ L_t &= T_t - T_t K_t F_t^{-1} Z_t' \\ \hat{\alpha}_t &= a_t + P_t r_t \\ V_t &= P_t - P_t N_t P_t. \end{aligned} \tag{3.6}$$

The results from the Kalman filtering are used for extrapolation, whereas the smoothing results can be used for interpolation. In both cases the unknown future and past values can be set as missing values. The missing values are automatically handled by properly implementing the Kalman filtering and smoothing algorithms. Assume that we wish to predict the future values $y_{n+1}, \ldots y_{n+h}$ given the observations up to time n. This is achieved by setting $Z_t = 0$ in the Kalman filter recursions for $t = n+1, \ldots, n+h$, which together with the model definition (3.2) give

$$\begin{aligned} a_{n+j} &= T_{n+j-1} a_{n+j-1} \\ P_{n+j} &= T_{n+j-1} P_{n+j-1} T_{n+j-1}' + R_{n+j-1} Q_{n+j-1} R_{n+j-1} \\ \mathrm{E}[y_{n+j}] &= Z_{n+j} a_{n+j} \\ \mathrm{Var}[y_{n+j}] &= Z_{n+j} P_{n+j} Z_{n+j}' + H_{n+j}, \end{aligned} \tag{3.7}$$

for $j = 1, \ldots, h$ (Durbin and Koopman, 2012, Sections 4.10 and 4.11). For the interpolation problem, the same adjustment is made also for the smoothing algorithm at times $t = m + 1, \ldots, m + h$, giving

$$\begin{aligned} r_{m+j} &= T_{m+j+1}' r_{m+j+1} \\ N_{m+j} &= T_{m+j+1}' N_{m+j+1} T_{m+j+1} \\ \hat{\alpha}_{m+j} &= a_{m+j+1} + P_{m+j+1} r_{m+j+1} \\ V_{m+j} &= P_{m+j+1} - P_{m+j+1} N_{m+j+1} P_{m+j+1} \\ \mathrm{E}[y_{m+j}] &= Z_{m+j+1} \hat{\alpha}_{m+j+1} \\ \mathrm{Var}[y_{m+j}] &= Z_{m+j+1} V_{m+j+1} Z_{m+j+1}' + H_{m+j+1}, \end{aligned} \tag{3.8}$$

for $j = 1, \ldots, h$ (Durbin and Koopman, 2012, Section 4.10).

The log-likelihood of an arbitrary model of form (3.2) can be obtained directly from the one-step-ahead prediction errors and their variances by

$$\log L = \text{constant} - \frac{1}{2}\sum_{t=1}^{n}(\log |F_t| + v_t' F_t^{-1} v_t).$$

In the case of a (partly) diffuse initial state distribution, adjustments for the filtering and smoothing recursions, as well as and log-likelihood computation are needed, see, e.g., Durbin and Koopman (2012, Chapter 5).

3.3 Exponential family state space models

State space models can also be extended to non-Gaussian cases. An important special case are exponential family state space models, where the state equation retains its linear Gaussian form, but the observation equation has the general form

$$p(y_t|\theta_t) = p(y_t|Z_t\alpha_t),$$

where $\theta_t = Z_t\alpha_t$ is the signal and $p(y_t|\theta_t)$ is the observational density. The signal θ_t is the linear predictor which is connected to the expected value $\mathrm{E}[y_t] = \mu_t$ via a link function $l(\mu_t) = \theta_t$. In the R package KFAS presented in Article IV, possible choices for observational distributions are Gaussian, Poisson, binomial, negative binomial and gamma distributions. Note that it is possible to define a multivariate model where each series has different distribution.

Denote $\alpha = (\alpha_1^\top, \ldots, \alpha_n^\top)^\top$, $y = (y_1^\top, \ldots, y_n^\top)^\top$ and $\theta = (\theta_1^\top, \ldots, \theta_n^\top)^\top$. In order to make inferences of the exponential family models, we first find a Gaussian model which has the same conditional posterior mode as $p(\theta|y)$ (Durbin and Koopman, 2000). This is done using an iterative process with Laplace approximation of $p(\theta|y)$, where the updated estimates for θ_t are computed via the Kalman filtering and smoothing from the approximating Gaussian model. In the approximating Gaussian model the observation equation is replaced by

$$\tilde{y}_t = Z_t\alpha_t + \epsilon_t, \quad \epsilon_t \sim N(0, H_t),$$

where the pseudo-observations $\tilde{y}_t$ variances H_t are based on the first and second derivatives of $\log p(y_t|\theta_t)$ with respect to θ_t (Durbin and Koopman, 2000).

Final estimates $\hat{\theta}_t$ correspond to the posterior mode of $p(\theta|y)$. In the Gaussian case the mode is also the mean. In the other cases supported by KFAS the difference between the mode and the mean is often negligible, and even the conditional variance estimate obtained from the Kalman smoother using the approximating model provides a relatively good approximation of the true conditional variance. This method is closely related to the iterative reweighted least squares (IRLS) method used in a generalized linear model framework (McCullagh and Nelder, 1989). Consequently, we can write a generalized linear model in a state space form, and the Kalman filter algorithm for the corresponding approximating Gaussian model gives results which are identical to the IRLS based analysis.

Instead of the linear predictor θ, we are usually more interested in μ_t. As the link function is non-linear, the direct transformation $\hat{\mu}_t = l^{-1}(\hat{\theta}_t)$ introduces some bias. To solve this problem, a simulation approach based on importance sampling can be used, which allows

us to correct these approximation errors (Durbin and Koopman, 2012, Chapter 11). With the importance sampling technique we can also compute the log-likelihood and the smoothed estimates for $f(\alpha)$, where f is an arbitrary function of states.

In the importance sampling scheme, we first find the approximating Gaussian model, simulate the states α^i from this Gaussian model and then compute the corresponding weights

$$w_i = \frac{p(y|\alpha^i)}{g(y|\alpha^i)},$$

where $p(y|\alpha^i)$ represents the conditional non-Gaussian density of the original observations and $g(y|\alpha^i)$ is the conditional Gaussian density of the pseudo-observations $\tilde{y}$. These weights are then used for computing

$$\mathrm{E}(f(\alpha)|y) = \frac{\sum_{i=1}^N f(\alpha^i) w_i}{\sum_{i=1}^N w_i}.$$

The log-likelihood function for the non-Gaussian model can be written as (Durbin and Koopman, 2012, p. 272)

$$\begin{aligned}\log L(y) &= \log \int p(\alpha, y) \mathrm{d}\alpha \\ &= \log L_g(y) + \log E_g\left[\frac{p(y|\theta)}{g(y|\theta)}\right],\end{aligned}$$

where $L_g(y)$ is the log-likelihood of the Gaussian approximating model and the expectation is taken with respect to the Gaussian density $g(\alpha|y)$. The expectation can be approximated by importance sampling as

$$\log E_g\left[\frac{p(y|\theta)}{g(y|\theta)}\right] \approx \log \frac{1}{N}\sum_{i=1}^N w_i, \tag{3.9}$$

or without any simulations as

$$\log E_g\left[\frac{p(y|\theta)}{g(y|\theta)}\right] \approx \log\left[\frac{p(y|\hat{\theta})}{g(y|\hat{\theta})}\right], \tag{3.10}$$

which is often sufficient at least for preliminary analysis.

For non-Gaussian exponential family models in the context of generalized linear models, a typical way of obtaining *confidence* intervals of the forecast is to compute confidence intervals in the scale of a linear predictor and then transform to the scale of observations, and the issue of prediction intervals is often dismissed. For obtaining proper prediction intervals in the case of non-Gaussian state space models, the following algorithm is used in KFAS.

(1) Draw N replicates of linear predictor θ from the approximating Gaussian density $g(\theta|y)$ with importance weights $p(y|\theta)/g(y|\theta)$. Denote this sample $\tilde{\theta}^1, \ldots, \tilde{\theta}^N$ as $\tilde{\theta}$

(2) Using the importance weights as sampling probabilities, draw a sample of size N with replacement from $\tilde{\theta}$. We know have N independent draws from $p(\theta|y)$.

(3) For each $\tilde{\theta}^i$ sampled in step (2), take a random sample of y^i from the observational distribution $p(y|\theta^i)$.

(4) Compute the prediction intervals as empirical quantiles from $y^1, \ldots, y^N$.

Assuming all the model parameters are known, these intervals coincide with the one obtained from Bayesian analysis using the same priors for states.

Chapter 4

Unknown parameters of the model

The unknown model parameters ψ are commonly estimated by the maximum likelihood method. In a traditional plug-in solution it is then assumed that the maximum likelihood estimate $\hat{\psi} = \psi$, i.e., it is assumed that the model parameters are known exactly. Using the plug-in approach, we disregard the uncertainties relating to our estimate $\hat{\psi}$. Nevertheless, this is often viewed as an acceptable approach on the basis that given the unbiased and consistent estimation method, the parameter uncertainty diminishes as the sample size increases. Still, in many applications the length of the time series can be short, say 50, and in those cases the parameter uncertainty can have significant effects, for example, when computing prediction intervals. This is especially true when combined with other sources of uncertainty, such as structural breaks (Clements and Hendry, 1999). It is also not uncommon to have biased $\hat{\psi}$, especially for short series or if the model selection is based on the same data as the parameter estimation (Phillips, 1979; Chatfield, 1995, 2000). Using the same series for the parameter estimation and forecasting can also produce non-Gaussian predictive distributions, even for Gaussian models (Phillips, 1979; Chatfield, 2000).

Although often disregarded, it seems that parameter uncertainty might be the most often tackled problem relating to too narrow prediction intervals. A large number of solutions are presented in the literature, see, for example, references in Article I. Many solutions focus on finding more accurate estimates for the prediction mean square error, which can then be plugged into Equation 2.4 in place of $\mathrm{E}[e_{n+h}^2]$, still assuming normality of errors and a symmetrical prediction interval around the point forecast. Another option is to obtain the prediction intervals as quantiles of simulated future observations.

In some cases the forecasting method automatically takes into account the uncertainties of parameter estimation. For example, analytical formulas for prediction intervals in linear regression take account of the uncertainty of estimating the model parameters, given the model is correct. In the time series context the inference is often done assuming both the model and its parameters are correct, as deriving similar analytical expressions for the prediction intervals is not as straightforward. These derivations are usually based on large sample properties of parameters, and some order of approximation is used. For example, Vidoni (2009) derives expressions for he prediction intervals for autoregressive models which take account of the parameter estimation uncertainty, but even for a simple first order autoregressive model the resulting formulas are rather complicated.

4.1 Bootstrap approaches

Instead of analytical formulas, bootstrap methods can be used. Although rather straightforward with serially independent data, bootstrapping of time series is somewhat more complicated, especially in a non-stationary case. In a general state space modeling framework, Rodriguez and Ruiz (2009, 2012) and Pfeffermann and Tiller (2005) present several bootstrap schemes, named as parametric and nonparametric bootstrap. In the parametric bootstrap, new observations are obtained by simulating the disturbance terms of the model, and thus rely heavily on the distributional and structural assumptions of the selected model. The classical nonparametric bootstrap is based on resampling the observations, but as we are dealing with dependent observations, the nonparametric versions in a time series context are based on resampling the standardized residuals obtained from the Kalman filter, which are then used to construct new observations. Therefore even these non-parametric versions are heavily based on the assumption that our model is correct, even though they do not make assumptions about the distribution of the residuals. In a linear mixed models framework, Morris (2002) shows how this type of semiparametric bootstrap is inconsistent, resulting in the underestimation of the variation in the data, especially in small samples. Due to the similarity of mixed models and state space models (Sallas and Harville, 1981; Tsimikas and Ledolter, 1997), these results are likely to hold also in the case of time series models where the resampling of residuals is used.

Using a simple example, Andrews (2000) shows that bootstrapping is not consistent when some of the parameters lie on the boundaries of the parameter space. This is a common issue in time series models. For example, some variances in structural time series can be estimated as zero, or the estimated autoregressive parameters are very close to the boundaries stationary region. The complex likelihood functions can also pose other problems in applying bootstrap methods to time series. The likelihood function often contains multiple maxima, and thus the estimation routine can be sensitive to the initial values. As the parameter estimation procedure is repeated for all bootstrap samples, one must be sure that the proper maximum is found for each replication. So, in theory, one should try several initial values for each bootstrap series, which increases the computational burden.

One of the early methods for dealing with parameter uncertainty in the context of prediction intervals for autoregressive models is given by Thombs and Schucany (1990). Their method is based on a nonparametric bootstrap. Thombs and Schucany (1990) claim that in terms of coverage probabilities, there was no distinction between models which had parameters near the boundaries of stationary regions and those which were well within the stationary region. This comparison is not shown, and it is unclear if they are referring only to their method, or to the standard plug-in method as well. The effect of stationary constraints for the standard plug-in method is illustrated in Articles I and II, where it is shown that, for example, in the case of the first order autoregressive process, the coverage probabilities depend on sample size, forecast horizon, as well as the value of the autoregressive coefficient.

Overall, the general applicability of bootstrap based solutions in the time series context seem to be somewhat questionable due to possible consistency issues and a heavy computational burden due to the repeated estimation of model parameters.

4.2 Bayesian approaches

A Bayesian approach offers yet another way of dealing with parameter uncertainty. After specifying prior distributions for the unknown parameters, a predictive distribution of the future observation can be obtained, which incorporates all the prior and posterior information into our predictions. Analytical formulas for posterior predictive distribution in time series context are rarely available, so one must rely on simulation techniques. Two common approaches are importance sampling (Ripley, 1987) and Markov chain Monte Carlo (MCMC) (see Gelman et al. (2013) for extensive introduction to MCMC and Bayesian data analysis in general). Perhaps the first implementations of MCMC sampling approach for Bayesian time series analysis are presented in Chib and Greenberg (1994) and Frühwirth-Schnatter (1994). Using the state space formulation of ARMA models as in Harvey (1981), Chib and Greenberg (1994) present a Bayesian approach for regression models with stationary ARMA errors focusing on parameter estimation (instead of prediction), whereas Frühwirth-Schnatter (1994) considers a linear Gaussian state space models with unknown variance parameters (again focusing more on parameter estimation). Durbin and Koopman (2000) present an importance sampling scheme for non-Gaussian state space modeling. It should be noted that although straightforward in principle, both the MCMC and the importance sampling methods are computationally intensive and considerable care is needed in analysing the results. For example in MCMC methods, checks on converge and mixing of Markov chains need to be performed, and the non-degeneracy of importance weights needs to be checked in the importance sampling approach.

The choice of prior distributions for parameters in Bayesian approach is difficult in some situations. For example, in structural time series it is possible that some of the disturbances have zero variance (so that for example, slope term is estimated as a constant), so prior distributions should allow this. Gelman (2006) compares prior distributions for variances of hierarchical models, and shows that commonly used noninformative priors based on the inverse-Gamma distribution are actually not noninformative if near-zero variances are possible in the light of the data. In these cases the inverse-Gamma prior is very sensitive to the choice of hyperparameters of the distribution. Similar problems are also expected in the case of the inverse-Wishart prior for general covariance matrices (Gelman, 2006).

In the Bayesian paradigm the coverage probabilities obtained from the Bayesian analysis are exactly correct if one accepts the chosen prior distribution, but the frequentist coverage probabilities defined as average coverage probabilities over the future realizations do not necessary coincide with the chosen nominal level. Article I presents a transparent way of computing accurate (in the frequentist sense) prediction intervals for Gaussian autoregressive models. The method is based on a Bayesian framework but focuses on frequentist coverage probabilities of the obtained intervals. The one-step-ahead prediction interval based on uniform priors for autoregressive parameters and $\log \sigma$ can be derived analytically, whereas h-step-ahead intervals can be obtained by a simple importance sampling scheme. With importance sampling, other prior distributions can also be entertained. The method produces considerably better prediction intervals than the plug-in method, especially when the estimated autoregressive paramaters are well within the stationary region. Article II generalizes the method to ARIMA models with exogenous variables. The extension of the method presented in Articles I and II to general Gaussian state space models is also straightforward and is presented in the next Section.

4.2.1 Bayesian prediction intervals via importance sampling

The Bayesian predictive density of y_{n+h} for Gaussian state space models is defined as the conditional density of y_{n+h} given $Y_n = (y_1, \ldots, y_n)$. It is obtained by integration as

$$p(y_{n+h}|Y_n) = \int p(y_{n+h}|\psi, Y_n)\, p(\psi, |Y_n)\, \mathrm{d}\psi, \tag{4.1}$$

where $p(\psi, |Y_n)$ is the posterior density of hyperparameters ψ. As the disturbances ϵ and η are assumed to be Gaussian, the density $p(y_{n+h}|\psi, Y_n)$ is also Gaussian. Then the conditional probability that the future value y_{n+h} is smaller than some b is given by

$$P(y_{n+h} \leq b|Y_n) = E\left[\Phi\left(\frac{b - \mathrm{E}(y_{n+h}|Y_n, \psi)}{\sqrt{\mathrm{Var}(y_{n+h}|Y_n, \psi)}}\right) \,\middle|\, Y_n\right], \tag{4.2}$$

where $E(\cdot|Y_n)$ refers to expectation with respect to the posterior distribution of ψ, and Φ is the cumulative distribution function of the standard normal distribution. Thus the prediction interval with nominal coverage probability of $(1 - 2\alpha)$ could be found by solving $P(y_{n+h} \leq b_{\text{lower}}|Y_n) = \alpha$ and $P(y_{n+h} \leq b_{\text{upper}}|Y_n) = 1 - \alpha$ with respect to b_α and $b_{1-\alpha}$. In general, these cannot be solved analytically but can be efficiently approximated via importance sampling. The Gaussian large sample approximation density $g(\psi_i|Y_n)$ is used as an approximating posterior for parameters ψ, which works relatively well when parameter estimates are well within the boundaries of the parameter space. There are no general restrictions to the prior density $p(\psi)$, but it is often desirable to constraint the possible values of simulated ψ (for example to positive values), and this can be done by specifying the prior accordingly. The algorithm for computing the Bayesian prediction intervals is as follows.

(1) Draw ψ_i from $N(\hat{\psi}, \widehat{\Sigma})$, where $\widehat{\psi}$ is the maximum likelihood estimate and $\widehat{\Sigma}$ is its approximate large sample covariance matrix.

(2) Run the Kalman filter with ψ_i to obtain $p(Y_n|\psi_i)$, the likelihood of the model, as well as $\mathrm{E}(y_{n+h}|Y_n, \psi_i)$ and $\mathrm{Var}(y_{n+h}|Y_n, \psi_i)$.

(3) Compute the weight

$$w_i = \frac{p(\psi_i)p(Y_n|\psi_i)}{g(\psi_i|Y_n)}.$$

(4) Repeat (1)–(3) independently N times.

(5) Compute the weighted average

$$\bar{P}_N(b) = \frac{\sum_{i=1}^N w_i \Phi\left(\frac{b - \mathrm{E}(y_{n+h}|Y_n, \psi_i)}{\sqrt{\mathrm{Var}(y_{n+h}|Y_n, \psi_i)}}\right)}{\sum_{i=1}^N w_i}$$

(6) Find such values b_{lower} and b_{upper} that $\bar{P}_N(b_{\text{lower}}) = \alpha$ and $\bar{P}_N(b_{\text{upper}}) = 1 - \alpha$.

Here only the model parameters are simulated, whereas in the traditional MCMC sampling approach the future observations and hidden states are simulated as well. Thus we need only

the Kalman filtering, compared to MCMC approaches such as one given in Chib and Greenberg (1994), where the smoothing algorithm is also needed in order to simulate the states given the data. Also the prediction interval is obtained by solving numerically two simple equations instead of empirical quantiles of simulated future observations. Thus the proposed method should be computationally more efficient. Standard errors of prediction limits can also be obtained by a straightforward extension of results in Article I.

An alternative importance sampling method could be obtained by computing weighted averages of forecasts and forecast variances, and then these could be used in the traditional prediction interval formula (2.4), assuming the symmetry of the prediction interval. This assumption is not used in the proposed algorithm. Simulation experiments seem to suggest that this root-finding approach gives somewhat more accurate results than the method using the formula (2.4), at least in the case of structural time series and ARIMA models.

Chapter 5

Additional problems in prediction

5.1 Data transformations

Modeling is often done using transformed series, for example, in order to make model components additive or to make the series more normally distributed. If the transformation is non-linear like the logarithmic transformation, back-transforming the point forecasts of the transformed series to the original scale induces some bias. In the case of the logarithms, the bias depends on the forecast horizon and the variance of the forecasted series (Harvey, 1989). In general, back-transforming the forecasted mean produces a forecast of the median in the original scale (Chatfield, 1993). For prediction intervals, the interval on the transformed scale and the original scale have the same nominal coverage probabilities, but if the interval for the transformed variable was symmetric, the back-transformed interval is asymmetric (Chatfield, 1993).

The simulation smoothing algorithms (for example, Durbin and Koopman (2002)) can be used to simulate states (or disturbances) of the state space model from their conditional distributions $p(\alpha_t|y_1, \ldots, y_n)$. These simulated samples can then be efficiently used to compute arbitrary estimates of interest, such as point and interval forecasts of observations in a non-transformed scale. Simulated series also exhibit the correct (as defined by the model) dependence structure between different time points, and thus they can be also used for computing, for example, non-linear aggregated estimates, such as expected yearly totals and their standard errors from log-transformed time series. This approach was used in Article III for estimating yearly totals for nutrient fluxes in the presence of missing data when the modeling was done in a logarithmic scale.

As noted in Section 3.2.1, forecasting future observations with state space models is done by adding missing observations to the end of our series, and the Kalman filter is run using this extended model. If the forecast horizon is one, we can simulate the observation from its marginal density $p(y_{n+1}|y_1, \ldots, y_n)$ obtained from the Kalman filter. But for longer horizons, simulating from marginal distributions does not yield proper paths of future observations as the time dependency of the future values is lost. Thus even in this case one would typically run a simulation smoothing algorithm in order to get simulated series of future observations in the original scale. In addition to the simulation smoother of Durbin and Koopman (2002), an R package KFAS presented in Article IV contains a method for simulating states (and observations if one augments ϵ to the state vector) from consecutive predictive distributions $p(\alpha_t|y_1, \ldots, y_{t-1})$, $t = 1, \ldots, n$ in a way that the correct time dependency of states is preserved.

This is achieved by omitting the smoothing phase of the algorithm of Durbin and Koopman (2002) (i.e., only filtering is performed). Thus, this simulation filtering method can be used to obtain predictive distributions of the form $p(y_{n+1}, \ldots, y_{n+h}|y_1, \ldots, y_n)$ in a computationally somewhat more efficient way than the simulation smoothers for long forecasting horizons.

5.2 Model uncertainty

As is the case with model parameters, the model itself is rarely known a priori. Whatever method we use for model selection, we are always incorporating some degree of uncertainty to our predictions, as even the best fitting model is just an approximation of the truth.

One way of dealing with model uncertainty is to define multiple models, which are all used in subsequent analysis, instead of just choosing the best fitting model. Although the final results might not be as interpretable as the forecasts based on a single model, this might be a non-issue if the goal is to produce accurate forecasts in a black-box manner. In Bayesian model averaging (BMA) (Draper, 1995; Hoeting et al., 1999) multiple models with chosen prior probabilities are combined, and thus the final forecasts are weighted averages of forecasts from all models considered. BMA has been successfully used in many fields of statistics, and there is easy-to-use software available for certain types of models, such as BMA package (Raftery et al., 2014) for generalized linear models and survival analysis. Yet there seems to be no standard way of using BMA for time series modeling. The computation of Bayesian factors used in computing the predictive distributions is usually not analytically tractable, although several approximations are available, at least for certain types of models and priors (see, e.g., Raftery (1995) and Hoeting et al. (1999)). As a comment to Hoeting et al. (1999), Clyde (1999) states that specifying the prior distributions on both the parameters and the model space is perhaps the most difficult aspect of BMA, and "In many problems, the subjective elicitation of prior hyperparameters is extremely difficult". Also, there are multiple ways to choose non-informative priors for parameters, which can have unintended influences on posterior model probabilities (Clyde, 1999; Gelman, 2006). Clyde (1999), Draper (1999) and George (1999) also warn about the idea of equally weighted models a priori, which can lead to surprising results if the chosen set of models contains multiple models which give practically identical forecasts. This results in a higher prior probability for this particular model set at the expense of other models, even though one actually wants to use non-informative model priors. The cited writers deal with regression analysis, but the same situation arises also in a time series context, where multiple seemingly unrelated models can produce very similar forecasts.

There are also some non-Bayesian model combining methods. One can think of multiple ways of weighting the forecasts from different models, and thus the problem of model selection is essentially replaced with the choice of the weighting scheme between competing models (one could perhaps argue that the same is also true for the Bayesian model averaging, where the choice of the model is replaced with the choice of priors and a set of models). Burnham and Anderson (2002) suggest using Akaike weights (computed from AIC or AICc) as model averaging weights in linear regression settings. Chatfield (1996) suggests that even a simple average of forecasts can perform reasonably well.

If the goal is to find a single model for prediction, parsimonious and flexible models should be preferred over complex models (especially if the model selection is based on data and not exogenous information). As the model complexity increases, the danger of overfitting and

model misspecification increases. Complex models can also be less robust to structural breaks. Chatfield (2004) argues that models with explanatory variables can often produce a better in-sample fit than models without external information, but are more easily misspecified (for example, variables which have no out-of-sample predictive power are erroneously included), and are more sensitive to model assumptions. Armstrong (2001) states that complex models should only be used if they are well supported by theory, and we have a large amount of high quality data available from the past. As an example, Armstrong (2001) reviewed 32 selected studies and found that only in five cases the more complex forecasting method gave more accurate results than exponential smoothing method. Also in general, the accuracy of out-of-sample forecasts is likely to be worse than that of in-sample forecasts. Therefore, the performance of the selected model should be tested with the data which was not used in the model selection and parameter estimation.

A time series cross-validation (Hyndman and Athanasopoulos, 2014) can be used to assess the accuracy of forecasts. In the time series cross-validation, observations are added to the training set sequentially starting with some initial series. For example, only observations up to time t are used in forecasting the observation $t + h$ (where h is the forecast horizon), and the mean square error or other measure of accuracy is computed. Then the model is estimated again using the training set augmented with the observation y_{t+1}, and y_{t+1+h} is forecasted. A similar method can also be used for assessing the sensitivity of the model selection procedure and the parameter estimation.

In Article III the focus was on interpolation and aggregated estimates. As part of the model validation, a thinning experiment was used where parts of the data were excluded from the parameter estimation, and the aggregated estimate of the excluded data was computed. The exclusion was done randomly and not sequentially, as the interest was in interpolation as opposed to extrapolation.

Summary of original publications

Article I considers the effects of parameter estimation uncertainty in the prediction interval computations of autoregressive models. An importance sampling approach based on a Bayesian framework is introduced, which takes account of the uncertainty of parameter estimates. Thus the method produces prediction intervals which are closer to the nominal level (in the frequentist sense) than the standard plug-in approach, where this uncertainty is ignored. The effects of different weakly informative priors are compared. Because the method seems to work at least as well as the plug-in method, it could be used as a default method for the computation of prediction intervals for autoregressive models. Compared to MCMC based methods, the suggested method is more straightforward to implement and understand, and an estimate for the error due to simulation can also be easily obtained. The method is also computationally more efficient, as there is no need to sample future observations but only model parameters. Method for checking the average coverage probabilities is also presented, which can be used to assess the accuracy of the chosen prior.

Article II extends the theory of Article I to ARMA models with explanatory variables by defining the model in a Gaussian state space form, and several prior distributions are again compared. As in Article I, the method performs significantly better than the standard plug-in method. The results suggest that simple uniform prior with stationarity and invertibility constraints is a good default prior for these types of models.

Article III considers the estimation of yearly nutrient fluxes in four Finnish rivers in southern Finland. Due to the sparse recordings of nutrient concentrations, missing daily data need to be interpolated before aggregated estimates of yearly fluxes can be made. A logarithmic transformation is also used in order to linearize the relationship between concentrations and water flow which induces bias. A state space modeling approach is used for simulating missing observations in the original scale, which are then used to produce yearly point and interval estimates for phosphorus and nitrogen fluxes. Simulation experiments are used to examine the effects of the model uncertainty, violations of normality assumptions, and highly sparse data. Comparison to ordinary regression model is also considered, showing how using an unrealistic model produces seemingly more accurate estimates and leads to completely wrong and unrealistic inference regarding the effects of yearly sample sizes.

Article IV introduces the R package KFAS for exponential family state space modeling. KFAS supports state space models with the observations from Gaussian, Poisson, binomial, negative binomial and gamma distributions. After introducing the basic theory behind the state space modeling in Gaussian and non-Gaussian cases, examples of different types of models in a state

space form is provided. An application to alcohol related deaths in Finland is used to illustrate the functionality of the package, and comparison to alternative Bayesian modeling framework using INLA package is presented. Although commercial software with similar features exists, KFAS is a unique R package with its support for multivariate exponential family state space models, with optional use of exact diffuse initialization and importance sampling approaches.

Bibliography

Alho, J. and Spencer, B. (2005). *Statistical Demography and Forecasting*. Springer, New York.

Andrews, D. W. K. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, 68(2):399–405.

Armstrong, J. S. (2001). *Principles of forecasting: A handbook for researchers and practitioners*, volume 30. Springer Science & Business Media, New York.

Box, G. E. P. and Jenkins, G. M. (1970). *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco, First edition.

Burnham, K. P. and Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. Springer, New York.

Chatfield, C. (1993). Calculating interval forecasts. *Journal of Business & Economic Statistics*, 11:121–135.

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 158:419–466.

Chatfield, C. (1996). Model uncertainty and forecast accuracy. *Journal of Forecasting*, 15:495–508.

Chatfield, C. (2000). *Time-series forecasting*. Chapman & Hall/CRC, Boca Raton.

Chatfield, C. (2004). *The analysis of time series: an introduction*. Chapman & Hall/CRC, Boca Raton, Sixth edition.

Chatfield, C., Koehler, A. B., Ord, J. K., and Snyder, R. D. (2001). A new look at models for exponential smoothing. *The Statistician*, 20:147–159.

Chib, S. and Greenberg, E. (1994). Bayes inference in regression models with arma (p, q) errors. *Journal of Econometrics*, 64(1-2):183–206.

Clements, M. P. and Hendry, D. F. (1999). *Forecasting Non-stationary Economic Time Series*. The MIT Press, Cambridge.

Clyde, M. (1999). Comment on "Bayesian model averaging: A tutorial". *Statistical Science*, 14(4):pp. 401–404.

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):pp. 45–97.

Draper, D. (1999). Comment on "Bayesian model averaging: A tutorial". *Statistical Science*, 14(4):pp. 405–409.

Durbin, J. (2000). The foreman lecture: The state space approach to time series analysis and its potential for official statistics (with discussion). *Australian & New Zealand Journal of Statistics*, 42(1):1–23.

Durbin, J. and Koopman, S. J. (2000). Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives. *Journal of Royal Statistical Society B*, 62:3–56.

Durbin, J. and Koopman, S. J. (2002). A simple and efficient simulation smoother for state space time series analysis. *Biometrika*, 89:603–615.

Durbin, J. and Koopman, S. J. (2012). *Time series analysis by state space methods*. Oxford University Press, New York, Second edition.

Frühwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *Journal of Time Series Analysis*, 15(2):183–202.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1:1–19.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC Press, Third edition.

George, E. I. (1999). Comment on "Bayesian model averaging: A tutorial". *Statistical Science*, 14(4):pp. 409–412.

Harvey, A. C. (1981). *Time Series Models*. Philip Allan, London.

Harvey, A. C. (1989). *Forecasting, structural time series models and the Kalman Filter*. Cambridge University Press.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, 14(4):382–417.

Hyndman, R. J. and Athanasopoulos, G. (2014). *Forecasting: Principles and practice*. Otexts.com.

Hyndman, R. J., Koehler, A. B., Ord, J. K., and Snyder, R. D. (2008). *Forecasting with exponential smoothing. The State Space Approach*. Springer.

Koopman, S. J. (1997). Exact initial kalman filtering and smoothing for nonstationary time series models. *Journal of the American Statistical Association*, 92(440):1630–1638.

Koopman, S. J. and Durbin, J. (2003). Filtering and smoothing of state vector for diffuse state-space models. *Journal of Time Series Analysis*, 24:85–98.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*. London: Chapman & Hall, Second edition.

Morris, J. S. (2002). The BLUPs are not "best" when it comes to bootstrapping. *Statistics & Probability Letters*, 56(4):425 – 430.

Pfeffermann, D. and Tiller, R. (2005). Bootstrap approximation to prediction MSE for state-space models with estimated parameters. *Journal of Time Series Analysis*, 26(6):893–916.

Phillips, P. (1979). The sampling distribution of forecasts from a first-order autoregression. *Journal of Econometrics*, 9(3):241–261.

Pollock, D. (1999). *Handbook of Time Series Analysis, Signal Processing, and Dynamics*. Academic Press, London.

Raftery, A. (1995). Bayesian model selection in social research. *Sociological Methodology*, 25:111–196.

Raftery, A., Hoeting, J., Volinsky, C., Painter, I., and Yeung, K. Y. (2014). *BMA: Bayesian Model Averaging*. R package version 3.18.1.

Ripley, B. D. (1987). *Stochastic Simulation*. John Wiley & Sons, Inc., New York, NY, USA.

Rodriguez, A. and Ruiz, E. (2009). Bootstrap prediction intervals in state-space models. *Journal of Time Series Analysis*, 30(2):167–178.

Rodriguez, A. and Ruiz, E. (2012). Bootstrap prediction mean squared errors of unobserved states based on the Kalman filter with estimated parameters. *Computational Statistics & Data Analysis*, 56(1):62 – 74.

Sallas, W. M. and Harville, D. A. (1981). Best linear recursive estimation for mixed linear models. *Journal of the American Statistical Association*, 76(376):860–869.

Thombs, L. A. and Schucany, W. R. (1990). Bootstrap prediction intervals for autoregression. *Journal of the American Statistical Association*, 85(410):486–492.

Tsimikas, J. V. and Ledolter, J. (1997). Mixed model representation of state space models: New smoothing results and their application to REML estimation. *Statistica Sinica*, 7:973–991.

Vidoni, P. (2009). A simple procedure for computing improved prediction intervals for autoregressive models. *Journal of Time Series Analysis*, 30(6):577–590.

I

Helske, J. and Nyblom, J. (2015). "Improved frequentist prediction intervals for autoregressive models by simulation". In Koopman, S. J. and Shephard, N., editors, *Unobserved Components and Time Series Econometrics*. Oxford University Press. In press.

Improved Frequentist Prediction Intervals for Autoregressive Models by Simulation

Jouni Helske* and Jukka Nyblom
Department of Mathematics and Statistics
University of Jyväskylä

Abstract

It is well known that the so called plug-in prediction intervals for autoregressive processes, with Gaussian disturbances, are too narrow, i.e. the coverage probabilities fall below the nominal ones. However, simulation experiments show that the formulas borrowed from the ordinary linear regression theory yield one-step prediction intervals, which have coverage probabilities very close to what is claimed. From a Bayesian point of view the resulting intervals are posterior predictive intervals when uniform priors are assumed for both autoregressive coefficients and logarithm of the disturbance variance. This finding opens the path how to treat multi-step prediction intervals which are obtained easily by simulation either directly from the posterior distribution or using importance sampling. A notable improvement is gained in frequentist coverage probabilities. An application of the method to forecasting the annual gross domestic product growth in the United Kingdom and Spain is given for the period 2002–2011 using the estimation period 1962–2001.

Key words: Importance sampling; Jeffreys's prior distribution; Predictive distribution; Multi-step predictions; Uniform prior distribution.

*Corresponding author: Jouni Helske, jouni.helske@jyu.fi

1 Introduction

A traditional approach to time series forecasting usually involves a selection of a family of suitable models, e.g. the class of autoregressive integrated moving average (ARIMA) models. Then using different model selection criteria within the family based of autocorrelation and partial autocorrelation functions together with formal criteria such as Akaike or Bayesian information criterion, the analyst chooses a suitable "best fitting" representative from the model family, estimates the parameters, makes diagnostic checks, and if he is happy with his choice computes predictions and the prediction intervals. The prediction intervals are usually computed as if the chosen model were correct and the parameters completely known, with no reference to the process of model selection. Chatfield (1993, 1996) has strongly criticized the omission of model uncertainty in forecasting. Clements and Hendry (1999, sections 1.3.8 and 2.2) introduce a detailed taxonomy of forecast errors, and stress the effects of the structural breaks in time series forecasting. As a remedy they propose robustifying forecasts for example by differencing and intercept corrections.

It is the common view of references given in the previous paragraph that the parameter uncertainty is often a minor source of prediction errors in practical applications when the sample size is large enough. Clements and Hendry (1999, p. 128) remark that although the parameter uncertainty is unlikely to lead to serious forecast failure it may have larger effect in conjunction with model misspecification. Nevertheless, we believe that it is justified to handle also this part of the model uncertainty. In textbooks it is a common topic, see for example Harvey (1993, p. 58-59). Here we show how to make corrections in a fairly simple way under autoregressive (AR) models.

Several proposals have been made for improving prediction intervals when parameters are estimated. One group of solutions focus on finding a more accurate prediction mean squared error in the presence of estimation; see, for example Phillips (1979), Fuller and Hasza (1981), Ansley and Kohn (1986), Quenneville and Singh (2000), and Pfeffermann and Tiller (2005). Both analytic and bootstrap approaches are tried.

Barndorff-Nielsen and Cox (1996) give general results for prediction intervals in the presence of estimated parameters. These results are further developed for time series models by Vidoni (2004, 2009). Unfortunately fairly complicated expressions appear already in rather simple models. Bootstrap solutions are given by several authors; see for example Beran (1990),

Masarotto (1990), Grigoletto (1998), Kim (2004), Pascual, Romo, and Ruiz (2004), Clements and Kim (2007), Kabaila and Syuhada (2008), and Rodriguez and Ruiz (2009).

Our aim is to find solutions for prediction interval problems using a Bayesian viewpoint with a requirement of objectivity for our predictions. Berger (2006) strongly encourages the use of the term "objective Bayes" in such situations. Operationally "objectivity" in our application means adopting priors which produce prediction intervals which exhibit approximately same coverage probabilities in both Bayesian and frequentist sense.

There is a vast literature on matching posterior and frequentist inferences to some degree of approximation. Methods based on invariance arguments, on information criteria or divergence measures as well as on asymptotic expansions are tried. The introduction section of Fraser et al. (2010) gives a short review on these as well as a list of relevant references to their own works and others. The starting point of their approach is to replace p value computations in a sample space with with posterior integration over the parameter space. Matching the Bayesian and frequentist coverage probabilities of prediction intervals in regression models with independent cases are treated in Datta and Mukerjee (2003) and Datta and Mukerjee (2004). A common feature in all these approaches is that often the so called Jeffreys's prior or its modification shows up as a solution. As we will see it happens also in this article. In a recent article of Arellano and Bonhomme (2009), where the main issue is the bias reduction in panel data models, the authors use a prior related to Jeffreys's prior. A predecessor in the bias reduction with the help of Jeffreys's prior is given by Firth (1993).

Early examples using the Bayesian approach in autoregressive models are given by Zellner (1971, p. 188) and Chow (1974). Later Broemeling and Land (1984) showed, among other things, that using a normal-gamma prior for the parameters the one-step ahead prediction follows a t distribution (understood as a location-scale family). They further deduced that all predictions up to k step ahead have a joint predictive density which consists of the product of k univariate t densities. Thompson and Miller (1986) propose a simulation method for computing prediction intervals based on the Bayesian arguments called "sampling the future". This differs from our approach. We compute the prediction interval by simulating directly the posterior prediction probability which is more accurate and considerably less time consuming. Liu (1994) develops an approximation to their method, and Snyder, Ord, and Koehler (2001) make an approximate extension of it to ARIMA mod-

els. They also compare several types of prediction intervals in terms of the frequentist coverage probabilities.

After the emergence of Markov Chain Monte Carlo simulation the Bayesian paradigm has gained increasing popularity in time series and econometrics; see for example Geweke (2005) and Prado and West (2010) and references therein. A thorough exposition of the Bayesian forecasting is given by Geweke and Whiteman (2006). Nevertheless, in forecasting under AR model either direct simulation or importance sampling is fast and sufficient. Based on independent simulation replicates it also renders rather simple formulas for assessing the Monte Carlo simulation error.

2 Motivation

Consider an AR(p) process

$$y_t = \beta_0 + \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p} + \varepsilon_t, \quad t = 1, \ldots, n, \ldots, \tag{2.1}$$

where the errors ε_t are independently drawn from $N(0, \sigma^2)$, and coefficients $\beta_j, j = 0, \ldots, p$, are arbitrary fixed values. Assume that we have observed $y_{-p+1}, \ldots, y_n$. Write $\boldsymbol{x}_t = (1, y_{t-1}, \ldots, y_{t-p})'$, and let $\boldsymbol{X}$ be the matrix with rows $\boldsymbol{x}_t'$, $t = 1, \ldots, n$. Further let $\boldsymbol{Y} = (y_1, \ldots, y_n)'$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$. Then the model (2.1) can be written as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. If we condition on the starting values $\boldsymbol{y}_0' = (y_{-p+1}, \ldots, y_0)$ we are led to the conditional likelihood

$$L_c(\boldsymbol{\beta}, \sigma) = (2\pi)^{-\frac{n}{2}} \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \right). \tag{2.2}$$

The conditional maximum likelihood estimates for $\boldsymbol{\beta}$ coincide with the least squares estimates

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{X}'\boldsymbol{Y}, \tag{2.3}$$

and the maximum likelihood estimate for σ^2, corrected by the degrees of freedom, is

$$s^2 = \frac{(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})}{n - p - 1}. \tag{2.4}$$

The predictive value for y_{n+1} is $\boldsymbol{x}_{n+1}'\widehat{\boldsymbol{\beta}}$, and the standard prediction interval with approximate coverage probability $1 - 2\alpha$ is

$$\boldsymbol{x}_{n+1}'\widehat{\boldsymbol{\beta}} \pm s z_\alpha, \tag{2.5}$$

where z_α is the α quantile of the standard normal distribution. In practice, the true coverage may be considerably below the nominal value $1-2\alpha$.

Let us suppose for a moment that we have an ordinary regression model with some truly exogeneous variables. Then using the same notation, $\boldsymbol{X}$ and $\boldsymbol{x}_{n-1}$, as before the exact coverage probability is obtained by

$$\boldsymbol{x}'_{n+1}\widehat{\boldsymbol{\beta}} \pm st_{\alpha,n-p-1}\sqrt{1+\boldsymbol{x}'_{n+1}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1}}, \tag{2.6}$$

As is well known, the extra factor here compared to (2.5) involving the exogeneous variable takes into account the estimation error in regression coefficients. In addition, a minor correction is done by replacing the normal quantile with that from Student's t with $n-p-1$ degrees of freedom. Although the assumptions of AR models do not satisfy the assumptions leading to (2.6), our simulations show that the very same intervals (2.6) have practically correct coverage probabilities also under the AR models.

Combining the formulas (2.2)–(2.4) with the identity

$$\begin{aligned}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) &= (\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}) \\ &\quad + (\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})'(\boldsymbol{X}'\boldsymbol{X})(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\end{aligned}$$

the conditional likelihood function (2.2) can be written as

$$\begin{aligned}L_c(\boldsymbol{\beta},\sigma) &= (2\pi)^{-\frac{n}{2}}\sigma^{-n}\exp\left(-\frac{(n-p-1)s^2}{2\sigma^2}\right) \\ &\quad \times \exp\left(-\frac{1}{2\sigma^2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})'(\boldsymbol{X}'\boldsymbol{X})(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right).\end{aligned} \tag{2.7}$$

Until now we have had the frequentist approach. But it is also illuminating to see the prediction interval (2.6) from a Bayesian point of view, where $\boldsymbol{\beta}$ and σ are now treated as a random vector and variable. If we multiply the likelihood $L_c(\boldsymbol{\beta},\sigma)$ by the improper prior $p(\boldsymbol{\beta},\sigma)=1/\sigma$, we find, as is well known in the ordinary regression, that *a posteriori*

$$\begin{aligned}\left.\frac{(n-p-1)s^2}{\sigma^2}\,\right|\,\boldsymbol{Y} &\sim \chi^2(n-p-1), \\ \boldsymbol{\beta}\,|\,\boldsymbol{Y},\sigma &\sim N(\widehat{\boldsymbol{\beta}},\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}).\end{aligned} \tag{2.8}$$

These together lead to the result that the interval (2.6) is a Bayesian predictive interval with exact coverage probability $1-2\alpha$. The corresponding

frequentist coverage probability is, however, only approximate

$$P\left(y_{n+1} \in \boldsymbol{x}'_{n+1}\widehat{\boldsymbol{\beta}} \pm st_{\alpha,n-p-1}\sqrt{1+\boldsymbol{x}'_{n+1}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1}}\,\middle|\,\boldsymbol{\beta},\sigma\right) \approx 1-2\alpha. \tag{2.9}$$

However, in section 5 we will find that the approximation (2.9) is very good when the coefficients are well within the stationarity region. The approximation seems to be worst in the nearly unit root cases. The explanation is likely to be as follows. If the coefficients are not too close to the boundary of the stationarity region, the sampling distribution of $\widehat{\boldsymbol{\beta}}$ is approximately as in in (2.8) when the roles of $\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}}$ are interchanged. On the other hand this not true in the nearly unit root case. So the question arises what is "good" prior in AR models in general. It turns out that using a certain modification of Jeffreys's prior seems to be preferred to uniform prior in nearly unit root models. One solution to the nearly unit root problem is provided by Clements and Hendry (1999, p. 92) who recommend in some cases to impose a unit root albeit it were not warranted by the unit root test.

At this point it might be interesting to recall the vigorous debate that broke out in the early 1990's between Peter Phillips and Christopher Sims on the value of unit root econometrics as such and its relation to Bayesian approach in statistics and econometrics, see Sims and Uhlig (1991), Phillips (1991), Sims (1991) and their references. The unit root inference is not an issue here. The Bayesian controversy focused on the choice of an appropriate prior distribution. Nevertheless, we do not want to revive this controversy here simply because we are not so much interested in inferences on autoregressive coefficients themselves but rather in prediction intervals. The role of the prior in this article is to produce a good weighting scheme for the predictive distribution.

3 Predictive distributions

3.1 Uniform prior

In this section we develop prediction formulas under the Bayesian paradigm employing the noninformative uniform prior for $(\boldsymbol{\beta}, \log\sigma)$, i.e. the prior takes the form $p(\boldsymbol{\beta},\sigma) = 1/\sigma$, for $\beta \in \mathbb{R}^{p+1}$ and $\sigma > 0$. We have already seen that when $k = 1$ the predictive distribution is Student's t. But for $k > 1$

the predictive distribution is not known to be any common distribution. Therefore we have to rely on simulation.

We use a generic notation $p(\cdot)$ for a density. The Bayesian predictive density of y_{n+k} is defined as the conditional density of y_{n+k} given $\boldsymbol{Y}$. It is obtained by integration as

$$p(y_{n+k} \,|\, \boldsymbol{Y}) = \int p(y_{n+k} \,|\, \boldsymbol{\beta}, \sigma, \boldsymbol{Y})\, p(\boldsymbol{\beta}, \sigma \,|\, \boldsymbol{Y})\, \mathrm{d}\boldsymbol{\beta}\, \mathrm{d}\sigma, \tag{3.1}$$

where $p(\boldsymbol{\beta}, \sigma \,|\, \boldsymbol{Y})$ is the posterior density of $(\boldsymbol{\beta}, \sigma)$. In AR models with normally distributed errors the density $p(y_{n+k} \,|\, \boldsymbol{\beta}, \sigma, \boldsymbol{Y})$ can be written explicitly. Recall the prediction formulas given $\boldsymbol{\beta}, \sigma^2, \boldsymbol{Y}$ (see for example Box et al. (2008, chap. 5)) and designate $\hat{y}_n(k) = E(y_{n+k} \,|\, \boldsymbol{Y}, \boldsymbol{\beta})$, $k = 0, \pm 1, \pm 2, \ldots$. Note that when $k \leq 0$ we have $\hat{y}_n(k) = y_{n+k}$. The equation (2.1) immediately yields recursion for predicted values

$$\hat{y}_n(k) = \beta_0 + \beta_1\, \hat{y}_n(k-1) + \cdots + \beta_p\, \hat{y}_n(k-p), \quad k = 1, 2, \ldots.$$

It is also plain from (2.1) that for some constants $\psi_1, \psi_2, \ldots$ depending on $\boldsymbol{\beta}$ we have

$$y_{n+k} - \hat{y}_n(k) = \varepsilon_{n+k} + \psi_1\, \varepsilon_{n+k-1} + \cdots + \psi_{k-1}\, \varepsilon_{n+1}.$$

Box et al. (2008) gives recursions for the coefficients ψ_j as follows

$$\psi_j = \beta_1\, \psi_{j-1} + \cdots + \beta_p\, \psi_{j-p}, \quad j \geq 1,\ \psi_0 = 1,\ \text{and } \psi_j = 0 \text{ for } j < 0.$$

We find $\psi_1 = \beta_1, \psi_2 = \beta_1^2 + \beta_2$ and so on. The prediction error variance, given $\boldsymbol{\beta}, \sigma^2$, is then $\sigma^2(1 + \psi_1^2 + \cdots + \psi_{k-1}^2) = \sigma^2 v^2(\boldsymbol{\beta})$.

In the further development it is useful to introduce a more detailed notation $\hat{y}_n(k; \boldsymbol{\beta}) = E(y_{n+k} \,|\, \boldsymbol{Y}, \boldsymbol{\beta})$. Then

$$y_{n+k} \,|\, \boldsymbol{\beta}, \sigma^2, \boldsymbol{Y} \ \sim N(\hat{y}_n(k; \boldsymbol{\beta}), \sigma^2 v_k^2(\boldsymbol{\beta})).$$

Combining this with (3.1) and changing the order of integration we find

$$P(y_{n+k} \leq b \,|\, \boldsymbol{Y}) = E\left[\Phi\left(\frac{b - \hat{y}_n(k; \boldsymbol{\beta})}{\sigma v_k(\boldsymbol{\beta})} \right) \,\middle|\, \boldsymbol{Y} \right], \tag{3.2}$$

where Φ is the cumulative distribution function of the standard normal distribution. When $k > 1$ we cannot do the integration involved analytically. Nevertheless, a Monte Carlo solution is straightforward. We can proceed as follows:

1. Draw independent q_i, $i = 1, \ldots, N$ from $\chi^2(n-p-1)$, and let $\sigma_i^2 = (n-p-1)s^2/q_i$.

2. Draw $\boldsymbol{\beta}_i$ from $N(\widehat{\boldsymbol{\beta}}, \sigma_i^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$, independently for $i = 1, \ldots, N$.

3. Compute the average

$$\bar{P}_N(b) = \frac{1}{N}\sum_{i=1}^{N} \Phi\left(\frac{b - \hat{y}_n(k;\boldsymbol{\beta}_i)}{\sigma_i v_k(\boldsymbol{\beta}_i)}\right). \tag{3.3}$$

The prediction interval is then found by solving separately both $\bar{P}_N(b) = \alpha$ and $\bar{P}_N(b) = 1-\alpha$. Let the solutions be $\hat{b}_\alpha, \hat{b}_{1-\alpha}$ respectively. Then $(\hat{b}_\alpha, \hat{b}_{1-\alpha})$ is the prediction interval with posterior coverage probability $1-2\alpha$ when N is large. Broemeling and Land (1984) noted that multi-step predictions can be constructed through a sequence of one-step predictions each step involving t distribution. However, they do not suggest any computational method how to utilize this in practice. Thompson and Miller (1986) continued the work of Broemeling and Land by proposing the Bayesian simulation of the future values $y_{n+1}, \ldots, y_{n+k}$. The prediction limits are then derived from the quantiles of the simulated values. This is more time consuming as well as being less accurate than what we suggest here.

3.2 Prediction With General Prior

Combine the conditional likelihood (2.2) and a general prior $p(\boldsymbol{\beta}, \sigma)/\sigma$. Then by (2.7) and (2.8) the joint posterior is

$$\begin{aligned} p(\boldsymbol{\beta}, \sigma \,|\, \boldsymbol{Y}) \;\propto\; & \sigma^{-(n-p)} e^{-(n-p-1)s^2/(2\sigma^2)} \\ & \times \sigma^{-p-1} \exp\left(-\frac{1}{2\sigma^2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\boldsymbol{X}'\boldsymbol{X})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right) \\ & \times p(\boldsymbol{\beta}, \sigma). \end{aligned}$$

It can be evaluated using importance sampling. In the simulation algorithm only the item 3 need be changed to

3' Compute the weighted average

$$\bar{P}_{N,w}(b) = \frac{\sum_{i=1}^{N} w_i \Phi\left(\frac{b - \hat{y}_n(k;\boldsymbol{\beta}_i)}{\sigma_i v_k(\boldsymbol{\beta}_i)}\right)}{\sum_{i=1}^{N} w_i}, \tag{3.4}$$

$$w_i = p(\boldsymbol{\beta}_i, \sigma_i). \tag{3.5}$$

The prediction interval is then solved as before.

It is also possible to incorporate the starting values $(y_{-p+1}, \ldots, y_0) = \boldsymbol{y}_0'$ into the likelihood. Denote $\mathrm{cov}(\boldsymbol{y}_0 \,|\, \boldsymbol{\beta}, \sigma) = \sigma^2 \boldsymbol{V} = \sigma^2 \boldsymbol{V}(\boldsymbol{\beta})$, and $E(y_i \,|\, \boldsymbol{\beta}) = \mu = \mu(\boldsymbol{\beta})$, $i = -p+1, \ldots, 0$. The matrix $\boldsymbol{V}$ can be obtained from the Yule-Walker equations, see Box et al. (2008, p. 58), and $\mu = \beta_0/(1-\beta_1-\cdots-\beta_p)$. To obtain the full likelihood, the conditional likelihood (2.2) is just multiplied by

$$p(\boldsymbol{y}_0 \,|\, \beta, \sigma) = (2\pi)^{-\frac{p}{2}} \sigma^{-p} |\boldsymbol{V}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2} (\boldsymbol{y}_0 - \mu \boldsymbol{1})' \boldsymbol{V}^{-1} (\boldsymbol{y}_0 - \mu \boldsymbol{1}) \right).$$

This leads to changing the weights in step 3'. The new weights are

$$w_i \;\; = \;\; I(\boldsymbol{\beta}_i \in \mathbb{R} \times S_p)\, p(\boldsymbol{\beta}_i, \sigma_i) p(\boldsymbol{y}_0 \,|\, \boldsymbol{\beta}_i, \sigma_i) \tag{3.6}$$

where $I(\cdot)$ is an indicator and S_p is the stationarity region of AR(p) process. The first coordinate of $\boldsymbol{\beta}_i$ can, of course, take any real value.

3.3 Standard Errors for Monte Carlo Prediction Limits

Here we give simple formulas for approximate standard errors when computing the prediction limits by Monte Carlo simulation. Let b_α be such that $P(y_{n+k} \le b_\alpha \,|\, \boldsymbol{Y}) = \alpha$ in (3.2), and $\hat{b}_\alpha$ be such that $\bar{P}_N(\hat{b}_\alpha) = \alpha$ in (3.3). The first order Taylor expansion at $\hat{b}_\alpha$ leads to

$$\hat{b}_\alpha - b_\alpha \approx \frac{\alpha - \bar{P}_N(b_\alpha)}{\bar{P}_N'(\hat{b}_\alpha)},$$

where $\bar{P}_N'$ is the derivative of $\bar{P}_N$. The variance of the numerator on the right side can be estimated from the sample values by S^2/N where

$$S^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left[\Phi\left(\frac{\hat{b}_\alpha - \hat{y}_n(k; \boldsymbol{\beta}_i)}{\sigma_i v_k(\boldsymbol{\beta}_i)} \right) - \alpha \right]^2.$$

Define

$$\text{s.e.}(\hat{b}_\alpha) = \frac{S/\sqrt{N}}{\bar{P}_N'(\hat{b}_\alpha)} = \frac{S}{\sum_{i=1}^{N} \frac{1}{\sigma_i v_k(\boldsymbol{\beta}_i)} \varphi\left(\frac{\hat{b}_\alpha - \hat{y}_n(k;\boldsymbol{\beta}_i)}{\sigma_i v_k(\boldsymbol{\beta}_i)} \right) / \sqrt{N}},$$

where φ is the density of the standard normal distribution. Omitting the details we can show that $(\hat{b}_\alpha - b_\alpha)/\text{s.e.}(\hat{b}_\alpha)$ tends to the standard normal distribution as $N \to \infty$.

In case we use a weighted average as in (3.5) and (3.6), a similar technique leads to the standard error

$$\begin{aligned} \text{s.e.}(\hat{b}_\alpha) &= \frac{S}{\sum_{i=1}^N \frac{w_i}{\sigma_i v_k(\boldsymbol{\beta}_i)} \varphi\left(\frac{\hat{b}_\alpha - \hat{y}_n(k;\boldsymbol{\beta}_i)}{\sigma_i v_k(\boldsymbol{\beta}_i)}\right)/\sqrt{N}}, \\ S^2 &= \frac{1}{N-1}\sum_{i=1}^N \left[\alpha w_i - w_i \Phi\left(\frac{\hat{b}_\alpha - \hat{y}_n(k;\boldsymbol{\beta}_i)}{\sigma_i v_k(\boldsymbol{\beta}_i)}\right)\right]^2. \end{aligned}$$

4 Priors

Here we give some examples of the priors that are likely to give improvements for frequentist coverage probabilities. A popular principle to generate priors, often improper, is to follow Jeffreys's rule which leads to the square root of the determinant of the information matrix. Applying this to the conditional likelihood (2.2) we first find that

$$\frac{\partial^2 L_c}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = -\frac{1}{\sigma^2}\sum_{t=1}^n \boldsymbol{x}_t \boldsymbol{x}_t'.$$

The information matrix is obtained by taking the expectation given the parameters and changing the sign. Assuming stationarity yields

$$-E\left[\frac{\partial^2 L_c}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right] = \frac{n}{\sigma^2} E\left[\boldsymbol{x}_1 \boldsymbol{x}_1'\right] = \frac{n}{\sigma^2}\begin{pmatrix} 1 & \mu \mathbf{1}' \\ \mu \mathbf{1} & \boldsymbol{V} + \mu^2 \mathbf{1}\mathbf{1}' \end{pmatrix}.$$

The determinant of the matrix involved is easily seen to equal $|\boldsymbol{V}|$ (note that $\boldsymbol{V}$ depends only on β_j, $j = 1, \ldots, p$). Now we take the convention that the parameter groups $\{\sigma\}$, $\{\beta_0\}$ and $\{\beta_1, \ldots, \beta_p\}$ are independent *a priori*, and that $\log \sigma$, β_0 are uniform. This leads to the prior we here call Jeffreys's prior.

$$p_J(\boldsymbol{\beta}, \sigma)/\sigma = I(\sigma > 0) I(\boldsymbol{\beta} \in \mathbb{R} \times S_p) \sigma^{-1} \sqrt{|\boldsymbol{V}(\boldsymbol{\beta})|}. \tag{4.1}$$

The uniform prior over the stationarity region is

$$p_U(\boldsymbol{\beta}, \sigma)/\sigma = I(\sigma > 0) I(\boldsymbol{\beta} \in \mathbb{R} \times S_p) \sigma^{-1}.$$

In our simulations we combine these priors with the full likelihood, i.e. the weights are $p(\boldsymbol{y}_0 \,|\, \boldsymbol{\beta}_i, \sigma_i) p_J(\boldsymbol{\beta}_i, \sigma_i)$ and $p(\boldsymbol{y}_0 \,|\, \boldsymbol{\beta}_i, \sigma_i) p_U(\boldsymbol{\beta}_i, \sigma_i)$ for Jeffreys's and uniform on the stationarity region priors, respectively. Note that when Jeffreys's prior is used the determinant $|\boldsymbol{V}|$ cancels when forming the weights.

For the stationary AR(1) model Jeffreys's marginal prior density of β_1 is

$$\begin{aligned} p_J(\beta_1) &= \frac{1}{\pi\sqrt{1-\beta_1^2}}, & |\beta_1| < 1 \\ &= 0, & \text{otherwise,} \end{aligned}$$

and for AR(2) the marginal prior density of (β_1, β_2) is

$$\begin{aligned} p_J(\beta_1, \beta_2) &= \frac{1}{(1+\beta_2)\sqrt{(1-\beta_2)^2 - \beta_1^2}}, & \beta_1 + \beta_2 < 1,\ \beta_2 - \beta_1 < 1, \\ & & |\beta_2| < 1, \\ &= 0, & \text{otherwise.} \end{aligned}$$

Note that the first density is proper whereas the second one is not; the latter property is true for all AR(p) with $p > 1$.

Without a stationarity restriction Berger and Yang (1994) defined an alternative proper marginal prior density for the AR(1) model as

$$\begin{aligned} p_R(\beta_1) &= \frac{1}{2\pi\sqrt{1-\beta_1^2}}, & |\beta_1| < 1, \\ &= \frac{1}{2\pi|\beta_1|\sqrt{\beta_1^2 - 1}}, & |\beta_1| > 1, \end{aligned} \tag{4.2}$$

which they call a reference prior. It is easily seen that the reference prior is invariant under the transformation $\beta_1 \to 1/\beta_1$. Note that the reference prior should be combined with the conditional likelihood (2.2), and then it produces peaks also in the posterior at $\beta_1 = \pm 1$. There is no use to combine it with the full likelihood because then it reduces to Jeffreys's prior. There seems to be no generalization available of the reference prior to higher order models.

5 Simulation experiments

In this section we compare different priors as regards their ability to match the frequentist coverage probabilities to the Bayesian ones. Thus, we now

turn back to the frequentist interpretation of the probability. Suppose that we have a realization from an autoregressive process with parameters $\boldsymbol{\beta}, \sigma$ which are fixed but unknown. Despite of this we wish to compute the prediction interval $(b_\alpha, b_{1-\alpha})$ via a Bayesian route from the probability distribution (3.2). Although the posterior coverage probability is exactly $1 - 2\alpha$, the corresponding frequentist probability usually is not. It refers to infinite sequence of new realizations from the same model with parameters $\boldsymbol{\beta}, \sigma$, and it is defined by the probability

$$P(b_\alpha \leq y_{n+k} \leq b_{1-\alpha} \,|\, \boldsymbol{\beta}, \sigma),$$

where all $y_{n+k}, b_\alpha, b_{1-\alpha}$ are random. In short, the frequentist coverage probability is an average coverage probability over the realizations. The conditional frequentist coverage probability $P(b_\alpha \leq y_{n+k} \leq b_{1-\alpha} \,|\, \boldsymbol{Y}, \boldsymbol{\beta}, \sigma)$ is a random variable, and in an actual application we do not know this probability. In our simulation experiment the values $b_\alpha, b_{1-\alpha}$ are replaced by $\hat{b}_\alpha, \hat{b}_{1-\alpha}$ obtained from (3.3) (apart from the case of the uniform prior with $k = 1$).

The chosen priors are uniform, uniform on the stationarity region, Jeffreys's prior (4.1) and the reference prior (4.2) of Berger and Yang (1994) for AR(1). In all comparisons we have used 50,000 replicates, where each replicate is a realization from a stationary AR(p) process of length $n + p$. Within each replicate the Monte Carlo sample size is $N = 50$. In this type of simulation experiments N need not be large, because the main variation in coverage probabilities is due to the different replicates. Nevertheless, in an actual application N should be considerably larger as is seen in the next section. All computations are done in the R environment (R Development Core Team, 2012).

Start with AR(1). Figure 5.1 shows the coverage probabilities of one-step ahead prediction intervals (2.6) based on the t distribution for several values of β_1, when $n = 30$. Here, and in all other experiments, we are aiming for the coverage probability of 0.9. For negative β_1 the coverage probabilities tend to be slightly above the nominal value whereas when β_1 is close to 1, they drop below the nominal value. However, note that even for $\beta_1 = 0.9$ the coverage probability is 0.894. The standard errors of coverage probabilities are less than 3×10^{-4} in all cases. Although not shown in the figure the prediction intervals appear to have approximately equal tail probabilities. The coverage probabilities of the standard prediction intervals stay below 0.88.

Figure 5.2 shows the coverage probabilities of multi-step prediction for AR(1) processes with different prior choices and four different parameter

values for β_1. The standard error of coverage probability was less than 7×10^{-4} in all cases. The reference prior seems to be slightly preferred compared to others. Especially, when β_1 is near to 1, the reference prior leads to coverage probabilities that are closest to the nominal frequentist coverage probability. Elsewhere the priors give almost the same coverage probabilities.

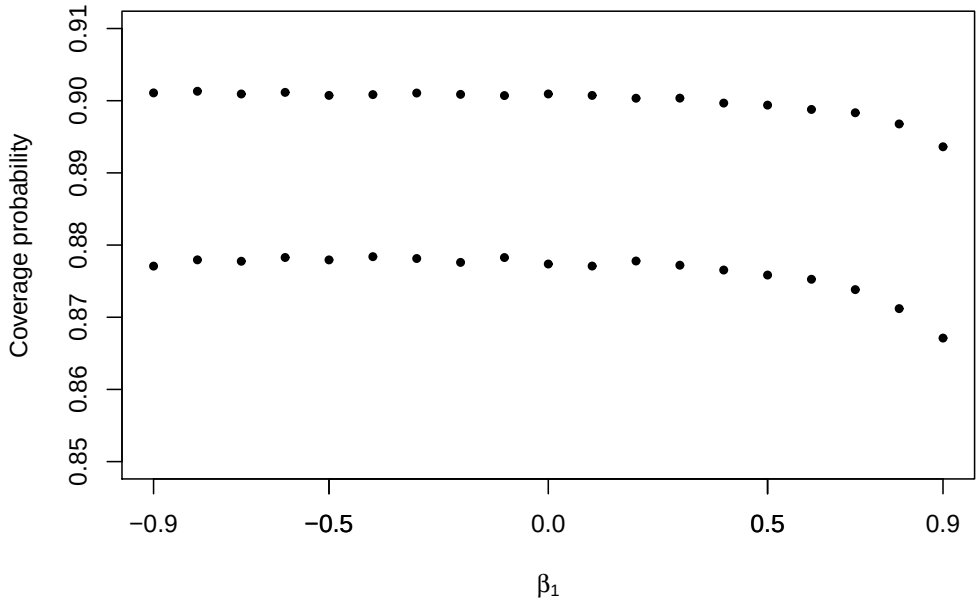

Figure 5.1: Coverage probabilities of one-step ahead predictions in AR(1) models when $n = 30$. The nominal coverage is 0.9. Dots are the estimated values, based on 50000 replicates. The upper dots relate to the Bayesian intervals with uniform prior and the lower ones to standard intervals.

In order to examine the differences between plug-in method and the Bayesian method with different priors in the case of AR(2) processes, we use nine different parameter combinations for β_1 and β_2. The parameters are defined through the roots of the characteristic equation

$$\beta(r) = 1 - \beta_1 r - \beta_2 r^2 = 0.$$

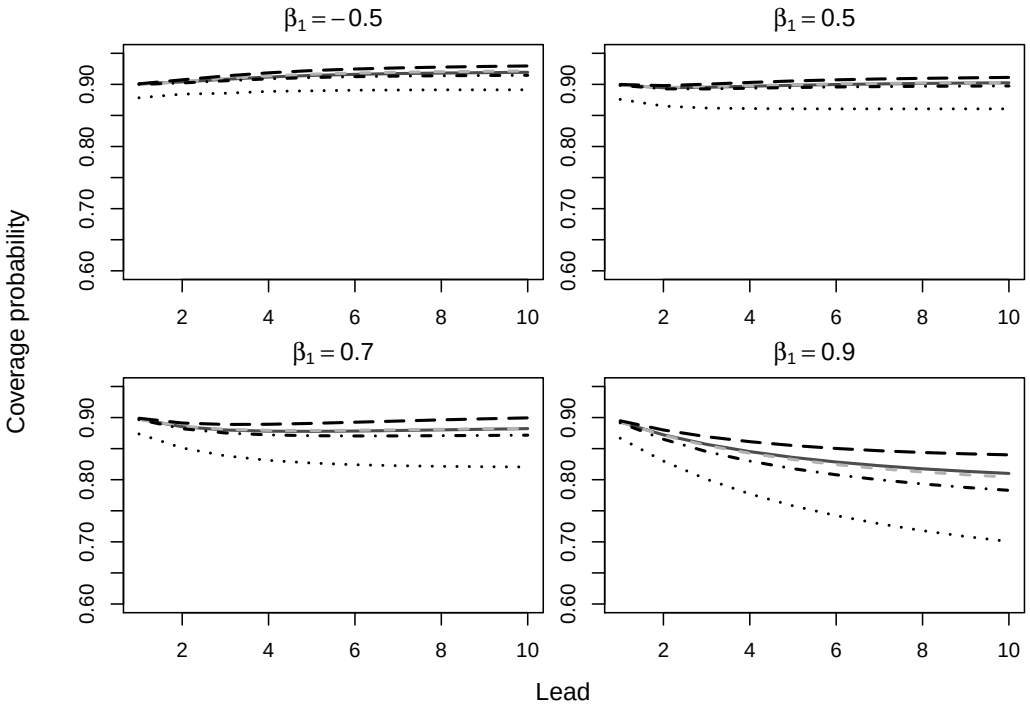

Figure 5.2: Coverage probabilities of multi-step ahead predictions in AR(1) models with different values of β_1 when $n = 30$. The nominal coverage is 0.9. The top dashed black line is based on the reference prior and the bottom dotted line represents the standard coverage probabilities. The lines based on the uniform (solid dark grey line), Jeffreys (light grey dashed line) and uniform stationary (black dot-and-dash line) prior are almost indiscernible apart from the case $\beta_1 = 0.9$.

Let the roots be r_1 and r_2. Then the parameters β_1 and β_2 can be written as

$$\beta_1 = \frac{1}{r_1} + \frac{1}{r_2},\ \beta_2 = -\frac{1}{r_1 r_2}.$$

The reciprocals of the roots and the corresponding parameters β_1 and β_2 are in Table 5.1.

Figure 5.3 shows the coverage probabilities for each process with the nominal coverage probability of 0.9 and $n = 30$. The standard error of the coverage probability was less than 5×10^{-4} in all cases. The Bayesian methods perform much better in all cases. Figures 5.4 and 5.5 show the spectral densities and the associated autocorrelation functions of the corresponding

processes. The spectral densities are scaled such that the total density integrates to 1.

Table 5.1: The AR(2) models used in the simulation experiments.

r_1^{-1}	r_2^{-1}	β_1	β_2
0.9	0.5	1.4	−0.45
0.9	−0.5	0.4	0.45
−0.9	0.5	−0.4	0.45
−0.9	−0.5	−1.4	−0.45
0.5	0.5	1.0	−0.25
0.5	−0.5	0	0.25
$0.9\exp(\frac{i}{5})$	$0.9\exp(-\frac{i}{5})$	1.76	−0.81
$0.9\exp(\frac{i\pi}{2})$	$0.9\exp(-\frac{i\pi}{2})$	0	−0.81
$0.9\exp(\frac{i3\pi}{4})$	$0.9\exp(-\frac{i3\pi}{4})$	−1.27	−0.81

As in the case of AR(1) process, the standard plug-in method gives much smaller coverage probabilities than the Bayesian method. We further see that when the mass of the spectral density function in Figure 5.4 is mostly on the smaller frequencies, corresponding to the slowly decaying autocorrelation function in Figure 5.5, the coverage probabilities stay under the nominal coverage probability. Furthermore, Jeffreys's prior performs slightly better than the other priors. The plug-in method is very poor, especially for longer forecast horizons.

When the mass of the spectral density function is mostly peaked on the medium or high frequencies and the corresponding autocorrelation functions alternate, the forecast horizon does not seem to affect the coverage probabilities neither for the Bayesian nor for the plug-in methods. The Bayesian methods are almost exactly equal to the nominal coverage or somewhat exceed it, while the plug-in method is staying clearly below the nominal coverage. In the case where the spectral density function is very flat, the Bayesian methods give coverage probabilities slightly over the nominal level, while the plug-in method stays below. Assuming stationarity with uniform prior seems to provide coverage probabilities little below those of the uniform and Jeffreys's priors.

6 Annual gross domestic product growth

As an empirical example we applied our method to forecasting the annual gross domestic product (GDP) changes (in percentages) in the United Kingdom (UK) and Spain (Figure 6.1). The data is from World Bank's databank (*http://databank.worldbank.org*). Observations from 1962–2001 are used for finding an appropriate model for each series. Then the series are forecast for 10 years ahead 2002-2011. In the following we first pretend, as far as possible, that we do not know what has happened in the forecast period. Afterwards when future has been revealed, we try to learn from it.

Based on the autocorrelation and partial autocorrelation functions, we assume an AR(1) model for both series. The least squares estimates of the autoregressive coefficients are 0.35 for the UK and 0.65 for Spain. The estimates for β_0 are 1.77 (UK) and 1.25 (Spain) and for σ 1.93 (UK) and 1.83 (Spain).

The residuals of the Spanish series are well behaved. There is neither apparent autocorrelation nor deviation from normality. As for the residuals

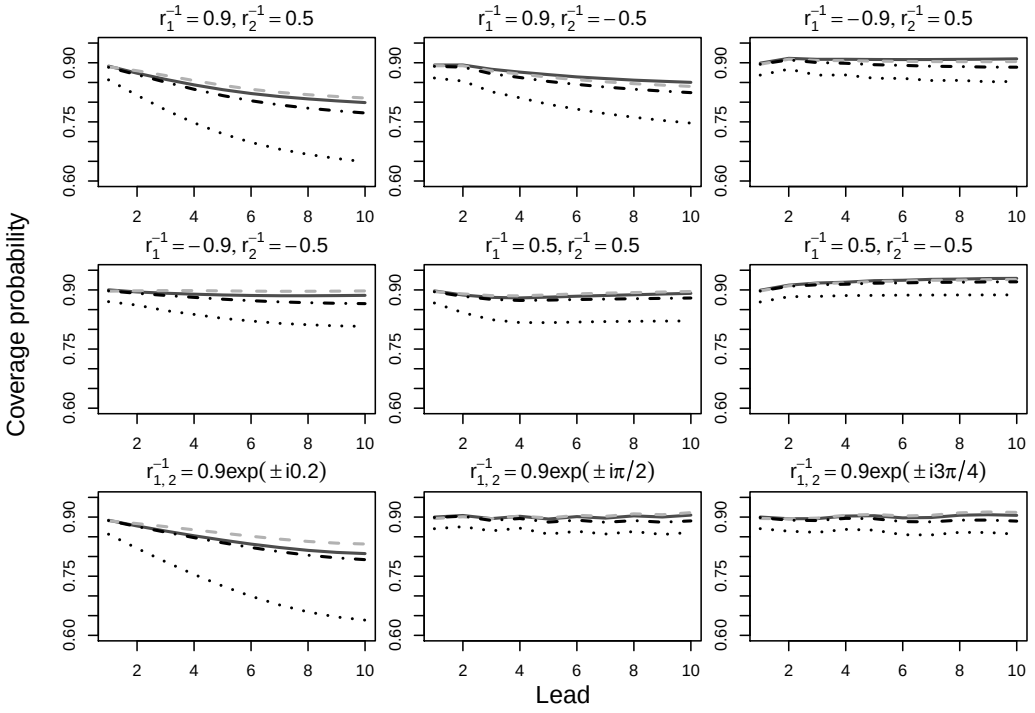

Figure 5.3: The coverage probabilities of the nine AR(2)-processes given in Table 5.1, with $n = 30$. The nominal coverage is 0.9. The black dotted line shows the coverage probabilities of the traditional plug-in method. The other lines are related to the Bayesian methods; the solid dark grey line corresponds to the uniform prior, the light grey dash to Jeffreys's prior, and the dot-and-dash line to the uniform stationary prior.

from the UK series, there appears to be two fairly large negative residuals which stem from the large drops in 1974 and 1980 (i.e. 2 in 40 years). Therefore we should not be surprised to see at least one aberrant value in the forecast period of ten years. No autocorrelation is left in the residuals of the UK series. Granted, we should take into account possible outlying values, but it would require models outside autoregressive family, which is beyond the scope of this article. Therefore we forecast with the estimated model also in this case. In summary, we expect that the Spanish GDP is likely to stay within 90% interval during the forecast period with possibly at most one minor violation. But we are more uncertain whether this holds for the GDP of the UK.

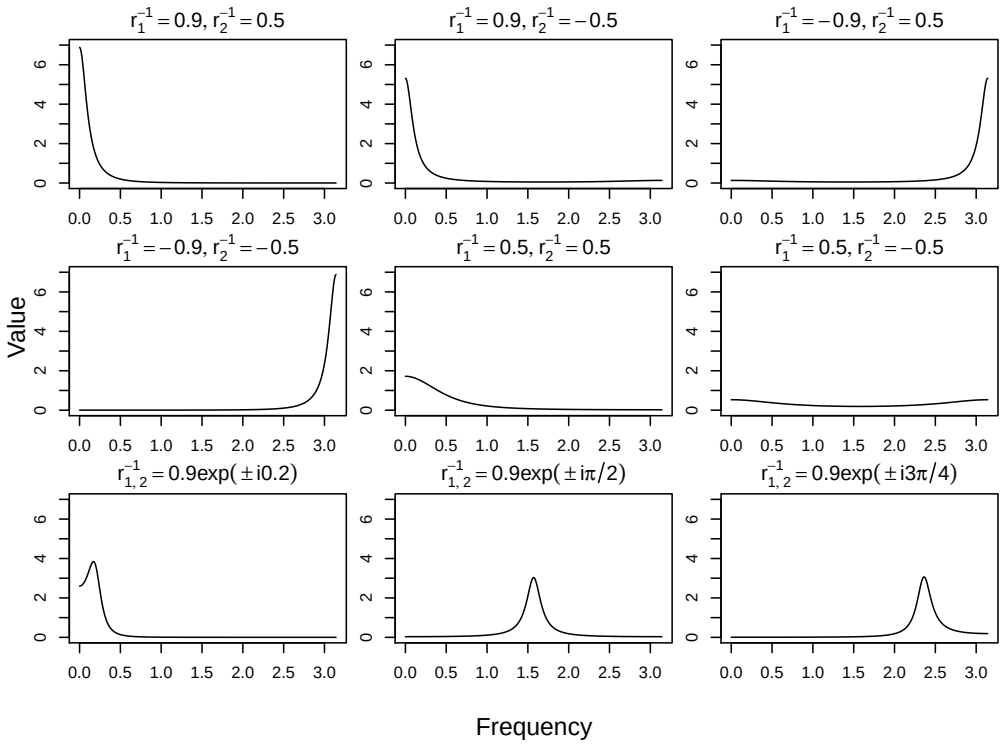

Figure 5.4: The scaled spectral density functions of the nine AR(2)-processes from Table 5.1.

Now, of course, we know that the preceding reasoning is too optimistic. In 2001 there were not any indication of the financial crisis starting in 2008. Figure 6.1 shows the actual values, the 90% prediction intervals and the point predictions for both series. The latter values in the Bayesian forecasting are posterior medians, computed by setting $\alpha = 0.5$. As expected, the Bayesian method produces wider prediction intervals than the standard plug-in method for both series and substantially wider for the Spanish series. We expected at most one minor violation in the Spanish series, whereas we see one large drop below the 90% prediction limit. In the UK series there are one value at the boundary (in 2008) and one significantly below the boundary (in 2009). The former is plausible bearing in mind the drops in 1974 and 1980, whereas latter is exceptionally low. But in both cases the main body of values in 2002–2011 especially those in 2010 and 2011 are well within boundaries. In summary, the values in 2009 are exceptionally low in both

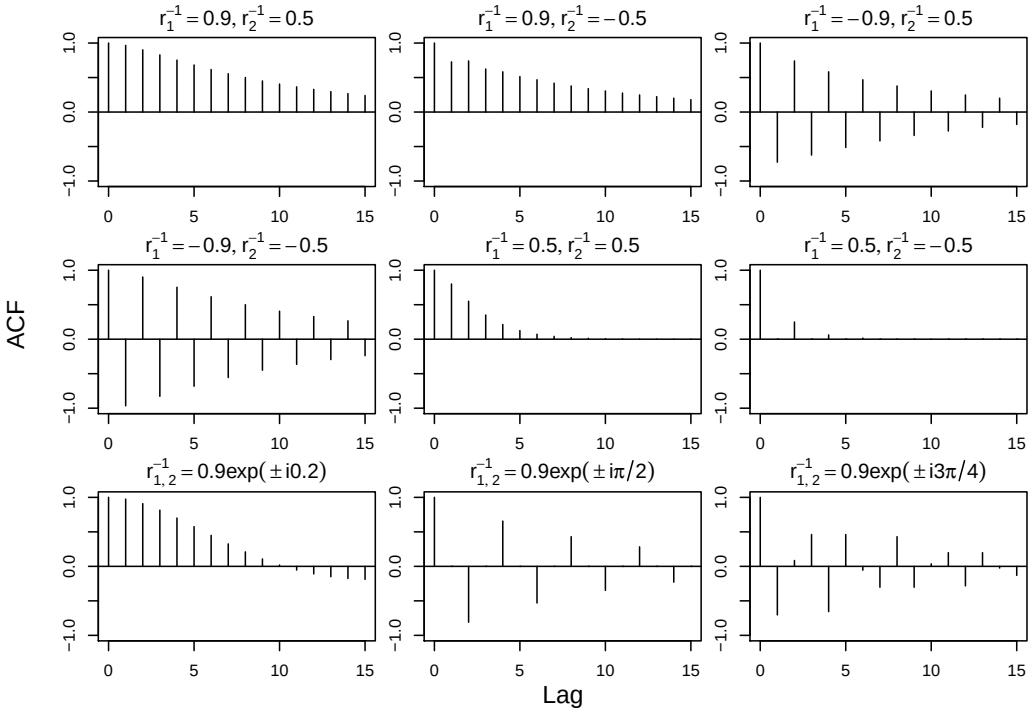

Figure 5.5: The autocorrelation functions of the nine AR(2)-processes from Table 5.1.

series compared to what our model predicts.

There is also one more thing we can learn. Although the differences between the Bayesian intervals are mostly negligible, we see that Jeffreys's prior leads to significantly higher upper limit for the Spanish series. The same is true also for the the point forecasts (i.e. medians). A closer examination of the importance weights using Jeffreys's prior shows that large values are associated with a large values of the autoregressive coefficient. This seems to be due to the large difference between the first observation and the stationary mean, leading to doubts on the stationarity of the series. Our conclusion is that the uniform prior combined with the conditional likelihood is more preferable to Jeffreys's prior in this case.

What should we think of the drops in 2009 occurring in both series? Looking at the series as such and ignoring other information they seem to be single unlikely events rather than the sign of a structural break. There

are at least two arguments for this. Firstly, the values after 2009 (in 2010 and 2011) fit rather well into the main body of the data. Secondly, we computed prediction bands such that the observed values of the year 2009 are lying exactly on the boundary. The corresponding coverage probabilities are $1 - \alpha$ with $\alpha = 0.0016$ for the UK and with $\alpha = 0.0078$ for the Spanish series. Moreover the values in 2009 are also extremely aberrant compared with those in the 40 year period 1962–2011. In summary, the values in 2009 are highly improbable in the light of the rest of the data.

Nevertheless, the accumulated information on world economy and the crisis concerning the euro countries till the end of 2012 makes us think that the assumption of a structural break should still be considered seriously. If we are to forecast from 2012 onwards we should carefully explore the methods suggested by Clements and Hendry (1999).

Table 6.1 gives numeric information on the predictions and prediction intervals related to our application under the assumption that our model is correct. Firstly, it contains the actual prediction limits with their Monte Carlo standard errors obtained from the formulas in section 3.3. Secondly, it shows the coverage probabilities for both models in case they were true. The forecasting horizon is $k = 1, 10$ and the nominal coverage is 0.90. The two Bayesian methods give one-step ahead prediction intervals which are practically correct. When $k = 10$, the coverage probabilities for the UK slightly differ from the nominal ones. The standard method gives intervals which have coverage probabilities below the nominal level in both cases. These comparisons concern the chosen models only, not the actual series: The reported coverage probabilities are averages corresponding to AR(1) models with coefficients 0.35 and 0.65. This explains why the actual intervals for the Spanish series with uniform and Jeffrey's priors are of considerably different widths, though their coverage probabilities are close to each other on the average. The deviance may also be related to the aberrant starting value.

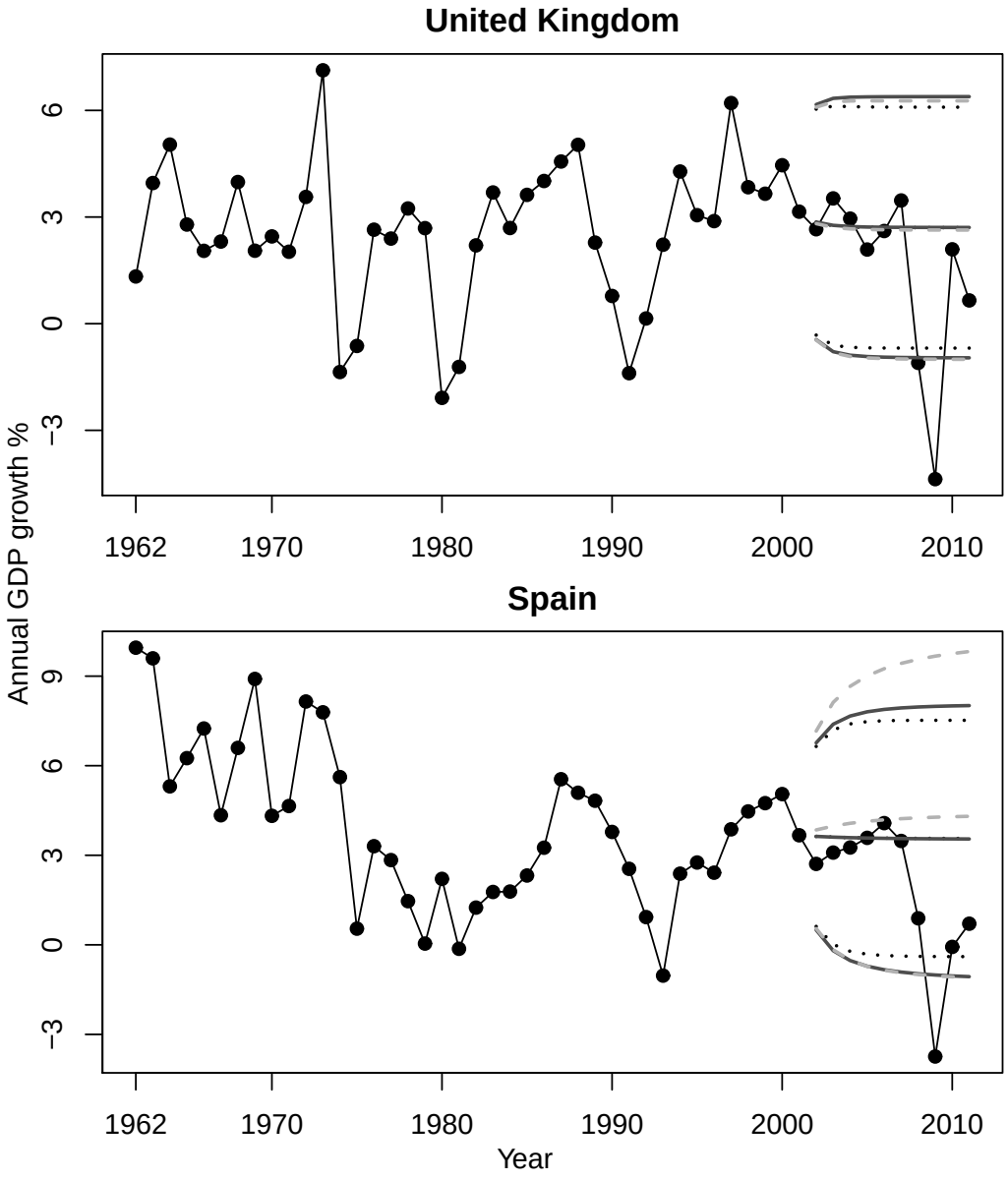

Figure 6.1: The annual GDP 1962–2001 in the UK and Spain together with the point predictions, actual values and 90% prediction intervals for the years 2002–2011. Solid grey lines represents intervals and point estimates computed by the Bayesian method with uniform prior, the grey dashed lines corresponds to the Bayesian method with Jeffreys's prior and black dotted line corresponds to the standard plug-in method.

Table 6.1: Coverage probabilities and prediction limits, with standard errors, related to the fitted models for the UK and SPA GDP series. The forecast horizon is 2002–2011, and the nominal coverage probability is 0.9.

	Uniform prior		Jeffreys's prior		Plug-in	
	$k = 1$	$k = 10$	$k = 1$	$k = 10$	$k = 1$	$k = 10$
			United Kingdom			
Coverage	0.900	0.906	0.900	0.907	0.883	0.881
$\hat{b}_{1-\alpha}$	6.164	6.388	6.089	6.269	6.037	6.091
s.e.$(\hat{b}_{1-\alpha})$	0.002	0.003	0.002	0.003	–	–
$\hat{b}_{\alpha}$	-0.448	-0.960	-0.462	-1.003	-0.321	-0.687
s.e.$(\hat{b}_{\alpha})$	0.002	0.003	0.002	0.003	–	–
			Spain			
Coverage	0.899	0.892	0.899	0.895	0.881	0.850
$\hat{b}_{1-\alpha}$	6.766	8.014	7.159	9.825	6.647	7.521
s.e.$(\hat{b}_{1-\alpha})$	0.002	0.005	0.003	0.021	–	–
$\hat{b}_{\alpha}$	0.498	-1.066	0.546	-1.097	0.616	-0.395
s.e.$(\hat{b}_{\alpha})$	0.002	0.006	0.003	0.011	–	–

We have used the sample size of $N = 100,000$ in our Monte Carlo simulation, and therefore the standard errors are fairly small. Standard errors should decrease at the rate $1/\sqrt{N}$ that is also confirmed by our experiments. With $N = 1000$ the standard errors are approximately tenfold compared to those in Table 6.1.

We have also compared the probabilities that the future value lies below the lower limit or above the upper limit. They seem to be approximately equal. Thus, our method seems to produce equal tail prediction intervals.

7 Discussion

We have shown the benefits of the Bayesian approach to prediction interval calculations under autoregressive schemes. Our message to the practitioners is that there are appropriate prior distributions leading to improved prediction intervals compared to those obtained by the common plug-in method. It has turned out that the uniform and Jeffreys's priors meet most practical goals for such intervals. Jeffreys's prior might have a slight advantage as regards coverage probabilities, although the dependence on the initial observations may have detrimental effects if the starting values are too far from their mean. Our simulation method is straightforward and easy to understand and to implement. An estimate for the Monte Carlo error can also be obtained.

It is plain that when the length of the time series increases, the parameter uncertainty decreases and thus also the coverage probabilities get closer to the nominal level. For example in AR(1) case, with $n = 100$, $h = 10$, and $\beta_1 = 0.5$, the coverage probability of the plug-in method roughly achieves the nominal probability. But any guidelines when to use the simulation method proposed here, instead of the plug-in method, seems to require massive Monte Carlo experimenting. The reason is that the outcome depends on many variables: the length of the series and forecasting horizon, the order of the model and the chosen coefficients. We believe that the proposed approach could be taken as a default method, because it is computationally so light and hardly ever gives results worse than the plug-in method.

The prediction intervals are computed under the Gaussian assumption. We have made some simulation experiments under AR(1) and AR(2) models, where the true errors are from Student's t distribution with 5 degrees of freedom and from Laplace distribution, but we still compute the prediction

intervals under Gaussian assumption. In both cases the coverage probabilities are smaller than in the case where the true errors are Gaussian, but usually the difference is about one percentage point or less. Although such a small experiment does not afford any general conclusions, it makes us to conjecture that the method is fairly robust against minor deviation from normality.

Although we have handled univariate AR processes only, the method could be extended to general ARIMA and vector autoregressive models, but more careful considerations of prior distributions are then needed. In addition, these complex models can be more sensititive to structural breaks and other issues discussed by Clements and Hendry (1999).

Finally, when dealing with a particular data set, some evidence of the adopted prior is obtained by simulating the estimated model as we have done in section 6.

Acknowledgements

The authors thank the editors of the volume and two anonymous referees for their comments and constructive criticism which has led to considerable improvements. We also thank the participants of the Conference in honour of Andrew Harvey's 65 year for the discussion and comments. Jouni Helske thanks for the financial support from Emil Aaltonen Foundation.

References

C. F. Ansley and R. Kohn. Prediction mean squared error for state space models with estimated parameters. *Biometrika*, 73:467–473, 1986.

M. Arellano and S. Bonhomme. Robust priors in nonlinear panel data models. *Econometrica*, 77:489–536, 2009.

O. E. Barndorff-Nielsen and D. R. Cox. Prediction and asymptotics. *Bernoulli*, 2:319–340, 1996.

R. Beran. Calibrating prediction regions. *Journal of the American Statistical Association*, 85:715–723, 1990.

J. Berger. The case of objective Bayesian analysis. *Bayesian Analysis*, 1: 385–402, 2006.

J. O. Berger and R. Yang. Noninformative priors and Bayesian testing for the AR(1) model. *Econometric Theory*, 10:461–482, 1994.

G. E. P. Box, G. M. Jenkins, and G. C. Reinsel. *Time Series Analysis: Forecasting and Control*. Wiley, Hoboken, Fourth edition, 2008.

L. Broemeling and M. Land. On forecasting with univariate autoregressive processes: A Bayesian approach. *Communications in Statistics – Theory and Methods*, 13:1305–1320, 1984.

C. Chatfield. Calculating interval forecasts. *Journal of Business & Economic Statistics*, 11:121–135, 1993.

C. Chatfield. Model uncertainty and forecast accuracy. *Journal of Forecasting*, 15:495–508, 1996.

G.C. Chow. Multiperiod predictions from stochastic difference equations by Bayesian methods. In S. E. Fienberg and A. Zellner, editors, *Studies in Bayesian Econometrics and Statistics*, chapter 8, pages 313–324. North-Holland, Amsterdam, 1974.

M. P. Clements and D. F. Hendry. *Forecasting Non-stationary Economic Time Series*. The MIT Press, Cambridge, 1999.

M. P. Clements and J. H. Kim. Bootstrap prediction intervals for autoregressive time series. *Computational Statistics & Data Analysis*, 51:3580–3594, 2007.

G. Datta and R. Mukerjee. Probability matching priors for predicting a dependent variable with application to regression models. *Annals of the Institute of Statistical Mathematics*, 55:1–6, 2003.

G. Datta and R. Mukerjee. *Probability Matching Priors: Higher Order Asymptotics*. Springer, New York, 2004.

D. Firth. Bias reduction of maximum likelihood estimates. *Biometrika*, 80: 27–38, 1993.

D. A. S. Fraser, N. Reid, E. Marras, and G. Y. Yi. Default priors for Bayesian and frequentist inference. *Journal of the Royal Statistical Society, Series B*, 72:631–654, 2010.

W. A Fuller and D. P. Hasza. Properties of predictors for autoregressive time series. *Journal of the American Statistical Association*, 76:155–161, 1981.

J. Geweke. *Contemporary Bayesian Econometrics and Statistics*. Wiley, Hoboken, 2005.

J. Geweke and C. Whiteman. Bayesian forecasting. In C. W. J. Elliott, G. Granger and A. Timmermann, editors, *Handbook of Economic Forecasting*, volume 1, chapter 1, pages 3–80. Elsevier B.V., Amsterdam, 2006.

M. Grigoletto. Bootstrap prediction intervals for autoregressions: Some alternatives. *International Journal of Forecasting*, 14:447–456, 1998.

A. C. Harvey. *Time Series Models*. Harvester Wheatsheaf, New York, Second edition, 1993.

P. Kabaila and K. Syuhada. Improved prediction limits for AR(p) and ARCH(p) processes. *Journal of Time Series Analysis*, 29:213–223, 2008.

J. H. Kim. Bootstrap prediction intervals for autoregression using asymptotically mean-unbiased estimators. *International Journal of Forecasting*, 20: 85–97, 2004.

S. I. Liu. Multiperiod Bayesian forecasts for AR models. *Annals of the Institute of Statistical Mathematics*, 46:429–452, 1994.

G. Masarotto. Bootstrap prediction intervals for autoregressions. *International Journal of Forecasting*, 6:229–239, 1990.

L. Pascual, J. Romo, and E. Ruiz. Bootstrap predictive inference for ARIMA processes. *Journal of Time Series Analysis*, 25:449–465, 2004.

D. Pfeffermann and R. Tiller. Bootstrap approximation to prediction MSE for state-space models with estimated parameters. *Journal of Time Series Analysis*, 26:893–916, 2005.

P. C. B. Phillips. The sampling distribution of forecasts from a first-order autoregression. *Journal of Econometrics*, 9:241–261, 1979.

P. C. B. Phillips. To criticize the critics: An objective Bayesian analysis of stochastic trends. *Journal of Applied Econometrics*, 6:333–364, 1991.

R. Prado and M. West. *Time Series—Modeling, Computation, and Inference*. Chapman & Hall/CRC, Boca Raton, 2010.

B. Quenneville and A. C. Singh. Bayesian prediction mean squared error for state space models with estimated parameters. *Journal of Time Series Analysis*, 21:219–236, 2000.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

A. Rodriguez and E. Ruiz. Bootstrap prediction intervals in state-space models. *Journal of Time Series Analysis*, 30:167–178, 2009.

C. A Sims. Comment by Christopher A. Sims on 'To criticize the critics', by Peter C. B. Phillips. *Journal of Applied Econometrics*, 6:423–434, 1991.

C. A Sims and H. Uhlig. Understanding unit rooters: A helicopter tour. *Econometrica*, 59:1591–1599, 1991.

R. D. Snyder, J. K. Ord, and A. B. Koehler. Prediction intervals for ARIMA models. *Journal of Business & Economic Statistics*, 19:217–225, 2001.

P. A. Thompson and R. B. Miller. Sampling the future: A Bayesian approach to forecasting to univariate time series models. *Journal of Business & Economic Statistics*, 4:427–436, 1986.

P. Vidoni. Improved prediction intervals for stochastic process models. *Journal of Time Series Analysis*, 25:137–154, 2004.

P. Vidoni. A simple procedure for computing improved prediction intervals for autoregressive models. *Journal of Time Series Analysis*, 30:577–590, 2009.

A. Zellner. *An Introduction to Bayesian Inference in Econometrics*. Wiley, New York, 1971.

II

Helske, J. and Nyblom, J. (2014). "Improved frequentist prediction intervals for ARMA models by simulation". In Knif, J. and Pape, B., editors, *Contributions to Mathematics, Statistics, Econometrics, and Finance: Essays in Honour of Professor Seppo Pynnönen*, number 296 in Acta Wasaensia, pages 71–86. University of Vaasa.

IMPROVED FREQUENTIST PREDICTION INTERVALS FOR ARMA MODELS BY SIMULATION

Jouni Helske and Jukka Nyblom
University of Jyväskylä

1 Introduction

In a traditional approach to time series forecasting, prediction intervals are usually computed as if the chosen model were correct and the parameters of the model completely known, with no reference to the uncertainty regarding the model selection and parameter estimation. The parameter uncertainty may not be a major source of prediction errors in practical applications, but its effects can be substantial if the series is not too long. The problems of interval prediction are discussed in depth in Chatfield (1993, 1996) and Clements & Hendry (1999).

Several proposals have been made for improving prediction intervals when parameters are estimated. One group of solutions focus on finding a more accurate prediction mean squared error in the presence of estimation; e.g. see Phillips (1979), Fuller & Hasza (1981), Ansley & Kohn (1986), Quenneville & Singh (2000), and Pfeffermann & Tiller (2005). Both analytic and bootstrap approaches are tried. Barndorff-Nielsen & Cox (1996) give general results for prediction intervals in the presence of estimated parameters. These results are further developed for time series models by Vidoni (2004, 2009). Bootstrap solutions are given by several authors; see for example Beran (1990), Masarotto (1990), Grigoletto (1998), Kim (2004), Pascual, Romo & Ruiz (2004), Clements & Kim (2007), Kabaila & Syuhada (2008), and Rodriguez & Ruiz (2009).

Here we show how to take into account the parameter uncertainty in a fairly simple way under autoregressive moving average (ARMA) models. We construct prediction intervals having approximately correct frequentist coverage probability, i.e. an average coverage probability over the realizations is approximately correct under the true parameter values. Due to the uncertainty in parameter estimation, the traditional plug-in method usually provides prediction intervals with average coverage probabilities falling below the nominal level. Our proposed method is based on Bayesian approach. Therefore the coverage probability is exactly correct if one is ready to accept the chosen prior distribution. But our aim is to find such priors that yield approximately correct coverage probabilities also in the frequentist sense. As a computational device the fairly simple importance sampling is employed in poste-

rior calculations. The method is an extension of the approach proposed by Helske & Nyblom (2013) for pure autoregressive models. The paper is organized as follows. Sections 2 and 3 derive general results, and section 4 applies them to ARMA models. Section 5 discusses prior distributions. Section 6 compares the plug-in method to Bayesian solutions by means of simulation experiments. Section 7 presents an application to real data. Section 8 concludes.

2 The model

We start with a fairly general linear model and later apply the results to ARMA models. Assume that the observations $y_1, \ldots, y_n$ are stacked in a vector $\boldsymbol{y}$ satisfying the model

$$\boldsymbol{y} \,|\, \boldsymbol{\psi}, \sigma, \boldsymbol{\beta} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{V}_{\boldsymbol{\psi}}), \tag{1}$$

where $\boldsymbol{X}$ is the $n \times k$ matrix of fixed regressors with rows $\boldsymbol{x}_t' = (x_{t1}, \ldots, x_{tk})$, and $\sigma^2 \boldsymbol{V}_{\boldsymbol{\psi}}$ is the covariance matrix depending on the parameters $(\psi_1, \ldots, \psi_r)' = \boldsymbol{\psi}$. We assume that $\boldsymbol{X}$ is of full rank k. The error vector is defined as $\boldsymbol{\epsilon} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$. Plainly $\boldsymbol{\epsilon} \sim \mathrm{N}(0, \sigma^2 \boldsymbol{V}_{\boldsymbol{\psi}})$. Next recall the well known identity

$$\begin{aligned}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})' \boldsymbol{V}_{\boldsymbol{\psi}}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) &= (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\psi}})' \boldsymbol{V}_{\boldsymbol{\psi}}^{-1} (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\psi}}) \\ &\quad + (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{\psi}}) \boldsymbol{X}' \boldsymbol{V}_{\boldsymbol{\psi}}^{-1} \boldsymbol{X} (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{\psi}}),\end{aligned}$$

where

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{\psi}} = (\boldsymbol{X}' \boldsymbol{V}_{\boldsymbol{\psi}}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}' \boldsymbol{V}_{\boldsymbol{\psi}}^{-1} \boldsymbol{y}.$$

The estimate $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\psi}}$ is the generalized least squares estimate for $\boldsymbol{\beta}$ when $\boldsymbol{\psi}$ is known. Define also

$$S_{\boldsymbol{\psi}}^2 = (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\psi}})' \boldsymbol{V}_{\boldsymbol{\psi}}^{-1} (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\psi}}).$$

Then the likelihood can be written as

$$\begin{aligned}p(\boldsymbol{y} \,|\, \boldsymbol{\psi}, \boldsymbol{\beta}, \sigma) &= (2\pi)^{-\frac{n}{2}} \sigma^{-n} |\boldsymbol{V}_{\boldsymbol{\psi}}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})' \boldsymbol{V}_{\boldsymbol{\psi}}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \right) \\ &= (2\pi)^{-\frac{n}{2}} \sigma^{-n} |\boldsymbol{V}_{\boldsymbol{\psi}}|^{-\frac{1}{2}} \exp\left(-\frac{S_{\boldsymbol{\psi}}^2}{2\sigma^2} \right) \\ &\quad \times \exp\left(-\frac{1}{2\sigma^2} (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{\psi}}) \boldsymbol{X}' \boldsymbol{V}_{\boldsymbol{\psi}}^{-1} \boldsymbol{X} (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{\psi}}) \right).\end{aligned}$$

Although our main purpose is to derive frequentist prediction intervals, we use the Bayes approach in their construction. Therefore, assume now that the parameters $\boldsymbol{\beta}$, σ and $\boldsymbol{\psi}$ are random and have a joint prior distribution. Moreover, $\boldsymbol{\psi}$ is indepen-

dent from $\boldsymbol{\beta}$ and σ with $(\boldsymbol{\beta}, \log\sigma)$ having the improper uniform prior distribution. Let $p(\boldsymbol{\psi})$ be the prior of $\boldsymbol{\psi}$. Then the joint prior is of the form $p(\boldsymbol{\psi})/\sigma$. These assumptions lead to the joint posterior density

$$\begin{aligned} p(\boldsymbol{\beta}, \boldsymbol{\psi}, \sigma \,|\, \boldsymbol{y}) &\propto p(\boldsymbol{\psi})\sigma^{-n-1}|\boldsymbol{V}_{\boldsymbol{\psi}}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})'\boldsymbol{V}_{\boldsymbol{\psi}}^{-1}(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})\right) \\ &\quad \times \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\boldsymbol{X}'\boldsymbol{V}_{\boldsymbol{\psi}}^{-1}\boldsymbol{X}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\right) \\ &\propto p(\boldsymbol{\psi})|\boldsymbol{V}_{\boldsymbol{\psi}}|^{-\frac{1}{2}}\sigma^{-(n-k+1)} \exp\left(-\frac{S_{\boldsymbol{\psi}}^2}{2\sigma^2}\right) \qquad (2) \\ &\quad \times\sigma^{-k} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\boldsymbol{X}'\boldsymbol{V}_{\boldsymbol{\psi}}^{-1}\boldsymbol{X}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\right). \qquad (3) \end{aligned}$$

Let us factorize the posterior as

$$p(\boldsymbol{\psi}, \sigma, \boldsymbol{\beta} \,|\, \boldsymbol{y}) = p(\boldsymbol{\psi} \,|\, \boldsymbol{y})p(\sigma \,|\, \boldsymbol{\psi}, \boldsymbol{y})p(\boldsymbol{\beta} \,|\, \boldsymbol{\psi}, \sigma, \boldsymbol{y}).$$

The formula (2)–(3) yield the conditional posteriors

$$\begin{aligned} \boldsymbol{\beta} \,|\, \boldsymbol{\psi}, \sigma, \boldsymbol{y} &\sim N\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\psi}}, \sigma^2(\boldsymbol{X}'\boldsymbol{V}_{\boldsymbol{\psi}}^{-1}\boldsymbol{X})^{-1}\right), \\ \left.\frac{S_{\boldsymbol{\psi}}^2}{\sigma^2}\right| \boldsymbol{\psi}, \boldsymbol{y} &\sim \chi^2(n-k). \end{aligned}$$

For $\boldsymbol{\psi}$, the marginal posterior is

$$p(\boldsymbol{\psi} \,|\, \boldsymbol{y}) \propto p(\boldsymbol{\psi})|\boldsymbol{V}_{\boldsymbol{\psi}}|^{-\frac{1}{2}}|\boldsymbol{X}'\boldsymbol{V}_{\boldsymbol{\psi}}^{-1}\boldsymbol{X}|^{-\frac{1}{2}}S_{\boldsymbol{\psi}}^{-(n-k)}, \qquad (4)$$

whenever the right side is integrable. In section 4, $\boldsymbol{\psi}$ and the related covariance matrix $\boldsymbol{V}_{\boldsymbol{\psi}}$ are specified through an appropriate ARMA model.

3 Bayesian prediction intervals

Assume that the future observations $y_{n+1}, y_{n+2}, \ldots$ come from the same model (1) with known values $\boldsymbol{x}_{n+1}, \boldsymbol{x}_{n+2}, \ldots$. Let

$$\begin{aligned} E(y_{n+h} \,|\, \boldsymbol{y}, \boldsymbol{\beta}, \sigma, \boldsymbol{\psi}) &= \hat{y}_{n+h|n}(\boldsymbol{\beta}, \boldsymbol{\psi}) \qquad (5) \\ \mathrm{var}(y_{n+h} \,|\, \boldsymbol{y}, \boldsymbol{\beta}, \sigma, \boldsymbol{\psi}) &= \sigma^2 v_{n+h|n}^2(\boldsymbol{\psi}). \qquad (6) \end{aligned}$$

Then

$$y_{n+h} \,|\, \boldsymbol{y}, \boldsymbol{\beta}, \sigma, \boldsymbol{\psi} \sim N(\hat{y}_{n+h|n}(\boldsymbol{\beta}, \boldsymbol{\psi}), \sigma^2 v_{n+h|n}^2(\boldsymbol{\psi})), \quad h = 1, 2, \ldots,$$

where for simplicity of notation the dependence on $\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_{n+h}$ is not explicitly shown. Then the Bayesian prediction intervals boils down to computing posterior probabilities of the form

$$P(y_{n+h} \leq b|\boldsymbol{y}) = E\left[\Phi\left(\frac{b - \hat{y}_{n+h|n}(\boldsymbol{\beta}, \boldsymbol{\psi})}{\sigma v_{n+h|n}(\boldsymbol{\psi})}\right) \,\middle|\, \boldsymbol{y}\right],$$

where $E(\cdot \mid \boldsymbol{y})$ refers to expectation with respect to the posterior distribution of $(\boldsymbol{\beta}, \sigma, \boldsymbol{\psi})$.

In practice the computation is accomplished by simulation. Suppose that we have the maximum likelihood estimate $\widehat{\boldsymbol{\psi}}$ and its approximate large sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$. Then we employ the following importance sampling for computing prediction intervals:

(i) Draw $\boldsymbol{\psi}_j$ from $N(\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\Sigma}})$, and compute the weight

$$w_j = \frac{p(\boldsymbol{\psi}_j \mid \boldsymbol{y})}{g(\boldsymbol{\psi}_j)},$$

where $p(\boldsymbol{\psi}_j \mid \boldsymbol{y})$ is defined in (4) and

$$g(\boldsymbol{\psi}_j) \propto \exp\left(-\frac{1}{2}(\boldsymbol{\psi}_j - \widehat{\boldsymbol{\psi}})'\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\psi}_j - \widehat{\boldsymbol{\psi}})\right).$$

(ii) Draw $q_j \sim \chi^2(n-k)$ independently from $\boldsymbol{\psi}_j$, and let $\sigma_j^2 = S_{\boldsymbol{\psi}_j}^2/q_j$.

(iii) Draw $\boldsymbol{\beta}_j \sim N(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\psi}_j}, \sigma_j^2(\boldsymbol{X}'\boldsymbol{V}_{\boldsymbol{\psi}_j}^{-1}\boldsymbol{X})^{-1})$.

(iv) Repeat (i)–(iii) independently for $j = 1, \ldots, N$.

(v) Compute the weighted average

$$\bar{P}_N(b) = \frac{\sum_{j=1}^N w_j \Phi\left(\frac{b - \hat{y}_{n+h|n}(\boldsymbol{\beta}_j, \boldsymbol{\psi}_j)}{\sigma_j v_{n+h|n}(\boldsymbol{\psi}_j)}\right)}{\sum_{j=1}^N w_j}. \tag{7}$$

(vi) Find the values b_α and $b_{1-\alpha}$ such that $\bar{P}_N(b_\alpha) = \alpha$ and $\bar{P}_N(b_{1-\alpha}) = 1 - \alpha$. When N is large $(b_\alpha, b_{1-\alpha})$ yields a prediction interval with coverage probability $1 - 2\alpha$.

4 Regression with ARMA errors

The regression model with ARMA errors is defined by the equations

$$y_t = \beta_1 x_{t1} + \cdots + \beta_k x_{tk} + \epsilon_t, \tag{8}$$

$$\epsilon_t = \phi_1 \epsilon_{t-1} + \ldots + \phi_p \epsilon_{t-p} + \xi_t + \theta_1 \xi_{t-1} + \ldots + \theta_q \xi_{t-q}, \tag{9}$$

where ξ_t are independent for all t and drawn from $N(0, \sigma^2)$. Thus, the process $\{\epsilon_t\}$ is ARMA(p, q) that we assume stationary and invertible. This is a special case of the model in section 2 with $\boldsymbol{\psi}' = (\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)$. Let $r = \max(p, q+1)$. For notational convenience we add zeros to either autoregressive or moving average parameters such that we have $\phi_1, \ldots \phi_r$ and $\theta_1, \ldots, \theta_{r-1}$. Of course, if $r = 1$ there are no moving average parameters. Following Durbin & Koopman (2001, pp. 46–47) the model (8)–(9) can be put into a state space form as

$$y_t = \boldsymbol{z}_t' \boldsymbol{\alpha}_t, \tag{10}$$

$$\boldsymbol{\alpha}_{t+1} = \boldsymbol{T}\boldsymbol{\alpha}_t + \boldsymbol{R}\xi_{t+1}, \tag{11}$$

where $\boldsymbol{z}_t' = (\boldsymbol{x}_t', 1, 0, \ldots, 0)$,

$$\boldsymbol{\alpha}_t = \begin{pmatrix} \boldsymbol{\beta}_t \\ \epsilon_t \\ \phi_2 \epsilon_{t-1} + \ldots + \phi_r \epsilon_{t-r+1} + \theta_1 \xi_t + \ldots + \theta_{r-1} \xi_{t-r+2} \\ \vdots \\ \phi_r \epsilon_{t-1} + \theta_{r-1} \xi_t \end{pmatrix},$$

$$\boldsymbol{T} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0}' & \phi_1 & 1 & & 0 \\ \vdots & \vdots & & \ddots & \\ \boldsymbol{0}' & \phi_{r-1} & 0 & & 1 \\ \boldsymbol{0}' & \phi_r & 0 & \cdots & 0 \end{pmatrix}, \quad \boldsymbol{R} = \begin{pmatrix} \boldsymbol{0} \\ 1 \\ \theta_1 \\ \vdots \\ \theta_{r-1} \end{pmatrix}.$$

Note that this formulation implies that actually $\boldsymbol{\beta}_t$ is constant $\boldsymbol{\beta}$. The initial distribution for $\boldsymbol{\alpha}_1$ is $N(\boldsymbol{0}, \boldsymbol{P}_1)$ with

$$\boldsymbol{P}_1 = \begin{pmatrix} \kappa \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Gamma} \end{pmatrix}, \tag{12}$$

where $\kappa \boldsymbol{I}$ corresponds to $\boldsymbol{\beta}_1$, and $\boldsymbol{\Gamma}$ is the covariance matrix of the stationary ARMA component of $\boldsymbol{\alpha}_t$.

Let $\boldsymbol{T}_{\boldsymbol{\phi}}$ and $\boldsymbol{R}_{\boldsymbol{\theta}}$ be the blocks of $\boldsymbol{T}$ and $\boldsymbol{R}$, respectively, related to the ARMA

process. Then $\boldsymbol{\Gamma}$ satisfies $\boldsymbol{\Gamma} = \boldsymbol{T}_{\boldsymbol{\phi}}\boldsymbol{\Gamma}\boldsymbol{T}_{\boldsymbol{\phi}}' + \boldsymbol{R}_{\boldsymbol{\theta}}\boldsymbol{R}_{\boldsymbol{\theta}}'$ and is given by

$$\mathrm{vec}(\boldsymbol{\Gamma}) = (\boldsymbol{I} - \boldsymbol{T}_{\boldsymbol{\phi}} \otimes \boldsymbol{T}_{\boldsymbol{\phi}})^{-1}\mathrm{vec}(\boldsymbol{R}_{\boldsymbol{\theta}}\boldsymbol{R}_{\boldsymbol{\theta}}'),$$

see Durbin & Koopman (2001, p. 112). The $\mathrm{vec}(\cdot)$ notation stands for the column-wise transformation of a matrix to a vector.

The initial distribution for $\boldsymbol{\beta}_1$ is actually defined through the limit $\kappa \to \infty$ which corresponds to the improper constant prior for $\boldsymbol{\beta}$ assumed in section 2. Durbin & Koopman (2001, Ch. 5) gives the updating formulas under this assumption called diffuse initialization. Thus, the Kalman filter together with the diffuse initialization automatically yields the values

$$\begin{aligned} E(\boldsymbol{\beta}_{n+1} \mid \boldsymbol{y}, \sigma, \boldsymbol{\psi}) &= \widehat{\boldsymbol{\beta}}_{\boldsymbol{\psi}}, \\ \mathrm{cov}(\boldsymbol{\beta}_{n+1} \mid \boldsymbol{y}, \sigma, \boldsymbol{\psi}) &= \sigma^2(\boldsymbol{X}'\boldsymbol{V}_{\boldsymbol{\psi}}^{-1}\boldsymbol{X})^{-1}. \end{aligned}$$

Additionally the Kalman filter gives the prediction errors

$$e_{t|t-1} = y_t - E(y_t \mid y_1, \ldots, y_{t-1}, \sigma, \boldsymbol{\psi}), \quad t = 1, \ldots, n,$$

and their variances

$$\mathrm{var}(e_{t|t-1}) = \mathrm{var}(y_t \mid y_1, \ldots, y_{t-1}, \sigma, \boldsymbol{\psi}) = \sigma^2 v_{t|t-1}^2, \quad t = 1, \ldots, n.$$

Due to the improper uniform prior of $\boldsymbol{\beta}$, i.e. the diffuse initialization, some variances $v_{t|t-1}^2 \to \infty$, as $\kappa \to \infty$ (Durbin & Koopman, 2001, sect. 5.2.1). Let $F = \{t \mid v_{t|t-1}^2 \text{ is finite}, t = 1, \ldots, n\}$. Then given $\boldsymbol{\psi}$ we have, by the results of Durbin & Koopman (2001, sect. 7.2.1), that

$$\begin{aligned} \sum_{t \in F} \frac{e_{t|t-1}^2}{v_{t|t-1}^2} &= S_{\boldsymbol{\psi}}^2, \\ \prod_{t \in F} v_{t|t-1}^2 &= |\boldsymbol{V}_{\boldsymbol{\psi}}|^{-\frac{1}{2}} |\boldsymbol{X}'\boldsymbol{V}_{\boldsymbol{\psi}}^{-1}\boldsymbol{X}|^{-\frac{1}{2}}. \end{aligned}$$

Because $\boldsymbol{X}$ is of rank k, the number of finite variances is $n - k$. We have now all elements for the algorithm of section 3 except the prior $p(\boldsymbol{\psi})$ that is discussed in the next section.

5 Jeffreys's rule for priors

Good candidates for the prior meeting our purposes is found by Jeffreys's rule which leads to the square root of the determinant of the Fisher information matrix. Apart from an additive constant, the log-likelihood is here

$$\ell(\boldsymbol{\beta}, \sigma, \boldsymbol{\psi}) = -n \log \sigma - \frac{1}{2} \log |\boldsymbol{V}_{\boldsymbol{\psi}}| - \frac{1}{2\sigma^2} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})' \boldsymbol{V}_{\boldsymbol{\psi}}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

A straightforward calculation gives the information matrix

$$\begin{aligned}
\boldsymbol{I}(\boldsymbol{\beta}, \sigma, \boldsymbol{\psi}) &= \begin{bmatrix} \frac{1}{\sigma^2}(\boldsymbol{X}'\boldsymbol{V}_{\boldsymbol{\psi}}^{-1}\boldsymbol{X}) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{2n}{\sigma^2} & \frac{1}{\sigma}\boldsymbol{I}'_{21}(\boldsymbol{\psi}) \\ \mathbf{0} & \frac{1}{\sigma}\boldsymbol{I}_{21}(\boldsymbol{\psi}) & \boldsymbol{I}_{22}(\boldsymbol{\psi}), \end{bmatrix}, \\
[\boldsymbol{I}_{21}(\boldsymbol{\psi})]_i &= \operatorname{trace}\left(\boldsymbol{V}_{\boldsymbol{\psi}}^{-1} \frac{\partial \boldsymbol{V}_{\boldsymbol{\psi}}}{\partial \psi_i} \right), \quad i = 1, \ldots, r \\
[\boldsymbol{I}_{22}(\boldsymbol{\psi})]_{ij} &= \frac{1}{2} \operatorname{trace}\left(\boldsymbol{V}_{\boldsymbol{\psi}}^{-1} \frac{\partial \boldsymbol{V}_{\boldsymbol{\psi}}}{\partial \psi_i} \boldsymbol{V}_{\boldsymbol{\psi}}^{-1} \frac{\partial \boldsymbol{V}_{\boldsymbol{\psi}}}{\partial \psi_j} \right), \quad i, j = 1, \ldots, r.
\end{aligned}$$

Hence,

$$|\boldsymbol{I}(\boldsymbol{\beta}, \sigma, \boldsymbol{\psi})|^{\frac{1}{2}} = \frac{1}{\sigma^{k+1}} \left| \boldsymbol{X}'\boldsymbol{V}_{\boldsymbol{\psi}}^{-1}\boldsymbol{X} \right|^{\frac{1}{2}} \left| \boldsymbol{I}_{22}(\boldsymbol{\psi}) - (2n)^{-1} \boldsymbol{I}_{21}(\boldsymbol{\psi}) \boldsymbol{I}_{21}(\boldsymbol{\psi})' \right|^{\frac{1}{2}}. \quad (13)$$

Because we want the joint prior to be of the form $p(\boldsymbol{\psi})/\sigma$, we insert $k = 0$ in (13) and define

$$p(\boldsymbol{\psi}) \propto \left| \boldsymbol{X}'\boldsymbol{V}_{\boldsymbol{\psi}}^{-1}\boldsymbol{X} \right|^{\frac{1}{2}} \left| \boldsymbol{I}_{22}(\boldsymbol{\psi}) - (2n)^{-1} \boldsymbol{I}_{21}(\boldsymbol{\psi}) \boldsymbol{I}_{21}(\boldsymbol{\psi})' \right|^{\frac{1}{2}}. \quad (14)$$

With this specification $p(\boldsymbol{\psi})/\sigma$ is called here the exact joint Jeffreys prior. Note that this prior depends on the sample size n. The approximate joint prior of the same form is obtained with

$$p(\boldsymbol{\psi}) \propto \left| \boldsymbol{X}'\boldsymbol{V}_{\boldsymbol{\psi}}^{-1}\boldsymbol{X} \right|^{\frac{1}{2}} |\boldsymbol{J}_{\boldsymbol{\psi}}|^{\frac{1}{2}}, \quad (15)$$

where

$$\boldsymbol{J}_{\boldsymbol{\psi}} = \lim_{n \to \infty} n^{-1} \left(\boldsymbol{I}_{22}(\boldsymbol{\psi}) - (2n)^{-1} \boldsymbol{I}_{21}(\boldsymbol{\psi}) \boldsymbol{I}_{21}(\boldsymbol{\psi})' \right).$$

Substituting either (14) or (15) to (4) we find that the determinant $|\boldsymbol{X}'\boldsymbol{V}_{\boldsymbol{\psi}}^{-1}\boldsymbol{X}|$ cancels.

Box et al. (2008, Ch. 7) gives useful results for the ARMA(p, q) models. We find that $\boldsymbol{J}_{\boldsymbol{\psi}}^{-1}/n$ is the large sample covariance matrix of the maximum likelihood estimate $\widehat{\boldsymbol{\psi}}$. In the pure AR model we have $|\boldsymbol{V}_{\boldsymbol{\psi}}| = |\boldsymbol{J}_{\boldsymbol{\psi}}|$, although the matrices are different. For the pure MA models the same determinant equation is approximately true, but the same does not apply to the mixed models. The marginal Jeffreys priors

are obtained by dropping off the factor $\left|\boldsymbol{X}'\boldsymbol{V}_{\boldsymbol{\psi}}^{-1}\boldsymbol{X}\right|^{\frac{1}{2}}$ in (14) and (15).

The numerical evaluation of the posteriors involves the determinant $|\boldsymbol{V}_{\boldsymbol{\psi}}|$, the inverse $\boldsymbol{V}_{\boldsymbol{\psi}}^{-1}$ and the partial derivatives of $\boldsymbol{V}_{\boldsymbol{\psi}}$. For short series the determinant and the inverse can be calculated directly. For longer series we can use the formulas provided by Lin & Ho (2008). The partial derivatives can be found recursively as follows. Recall the state space representation (10)–(11) and the initial covariance matrix $\boldsymbol{\Gamma}$ in (12). Due to stationarity of the process $\{\boldsymbol{\alpha}_t\}$ we find that $\mathrm{cov}(\boldsymbol{\alpha}_{t+s}, \boldsymbol{\alpha}_t) = \boldsymbol{T}^s\boldsymbol{P}_1$, where the block $\boldsymbol{T}_{\boldsymbol{\phi}}^s\boldsymbol{\Gamma}$ corresponds the autocovariance matrix of the ARMA process. The position $(1,1)$ of this matrix shows $\mathrm{cov}(y_{t+s}, y_t)$. We find the partial derivatives recursively for the autoregressive parameters

$$\frac{\partial(\boldsymbol{T}_{\boldsymbol{\phi}}^s\boldsymbol{\Gamma})}{\partial\phi_j} = \frac{\partial\boldsymbol{T}_{\boldsymbol{\phi}}}{\partial\phi_j}\boldsymbol{T}_{\boldsymbol{\phi}}^{s-1}\boldsymbol{\Gamma} + \boldsymbol{T}_{\boldsymbol{\phi}}\frac{\partial(\boldsymbol{T}_{\boldsymbol{\phi}}^{s-1}\boldsymbol{\Gamma})}{\partial\phi_j}, \quad s = 1, 2, \ldots.$$

For moving average parameters we have

$$\frac{\partial(\boldsymbol{T}_{\boldsymbol{\phi}}^s\boldsymbol{\Gamma})}{\partial\theta_j} = \boldsymbol{T}_{\boldsymbol{\phi}}^s\frac{\partial\boldsymbol{\Gamma}}{\partial\theta_j}, \quad s = 1, 2, \ldots.$$

Because $\boldsymbol{\Gamma}$ satisfies $\boldsymbol{\Gamma} = \boldsymbol{T}_{\boldsymbol{\phi}}\boldsymbol{\Gamma}\boldsymbol{T}_{\boldsymbol{\phi}}' + \boldsymbol{R}_{\boldsymbol{\theta}}\boldsymbol{R}_{\boldsymbol{\theta}}'$, we find by differentiating on both sides that

$$\begin{aligned}
\frac{\partial\boldsymbol{\Gamma}}{\partial\phi_j} &= \boldsymbol{T}_{\boldsymbol{\phi}}\frac{\partial\boldsymbol{\Gamma}}{\partial\phi_j}\boldsymbol{T}_{\boldsymbol{\phi}}' + \frac{\partial\boldsymbol{T}_{\boldsymbol{\phi}}}{\partial\phi_j}\boldsymbol{\Gamma}\boldsymbol{T}_{\boldsymbol{\phi}}' + \boldsymbol{T}_{\boldsymbol{\phi}}\boldsymbol{\Gamma}\frac{\partial\boldsymbol{T}_{\boldsymbol{\phi}}'}{\partial\phi_j}, \\
\frac{\partial\boldsymbol{\Gamma}}{\partial\theta_j} &= \boldsymbol{T}_{\boldsymbol{\phi}}\frac{\partial\boldsymbol{\Gamma}}{\partial\theta_j}\boldsymbol{T}_{\boldsymbol{\phi}}' + \frac{\partial\boldsymbol{R}_{\boldsymbol{\theta}}}{\partial\theta_j}\boldsymbol{R}_{\boldsymbol{\theta}} + \boldsymbol{R}_{\boldsymbol{\theta}}\frac{\partial\boldsymbol{R}_{\boldsymbol{\theta}}'}{\partial\theta_j}.
\end{aligned}$$

which implies that

$$\begin{aligned}
\mathrm{vec}\left(\frac{\partial\boldsymbol{\Gamma}}{\partial\phi_j}\right) &= (\boldsymbol{I} - \boldsymbol{T}_{\boldsymbol{\phi}} \otimes \boldsymbol{T}_{\boldsymbol{\phi}})^{-1}\mathrm{vec}\left(\frac{\partial\boldsymbol{T}_{\boldsymbol{\phi}}}{\partial\phi_j}\boldsymbol{\Gamma}\boldsymbol{T}_{\boldsymbol{\phi}}' + \boldsymbol{T}_{\boldsymbol{\phi}}\boldsymbol{\Gamma}\frac{\partial\boldsymbol{T}_{\boldsymbol{\phi}}'}{\partial\phi_j}\right), \\
\mathrm{vec}\left(\frac{\partial\boldsymbol{\Gamma}}{\partial\theta_j}\right) &= (\boldsymbol{I} - \boldsymbol{T}_{\boldsymbol{\phi}} \otimes \boldsymbol{T}_{\boldsymbol{\phi}})^{-1}\mathrm{vec}\left(\frac{\partial\boldsymbol{R}_{\boldsymbol{\theta}}}{\partial\theta_j}\boldsymbol{R}_{\boldsymbol{\theta}} + \boldsymbol{R}_{\boldsymbol{\theta}}\frac{\partial\boldsymbol{R}_{\boldsymbol{\theta}}'}{\partial\theta_j}\right).
\end{aligned}$$

6 Simulation experiments for ARMA models

Recall that our primary goal is to improve frequentist coverage probabilities in interval prediction. For that purpose we have conducted simulation experiments to find out the benefits of the Bayesian approach especially in relation to the standard plug-in method. The latter method yields the well known intervals

$$\hat{y}_{n+h|n}(\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\beta}}) \pm z_\alpha\hat{\sigma}v_{n+h|n}(\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\beta}}), \quad \hat{\sigma} = S^2/(n-k), \tag{16}$$

see (5) and (6).

In all simulations the length of the time series is 50, and the regression part consists of the constant term $\beta_1 = \beta$ only, i.e. $\boldsymbol{X} = (1, \ldots, 1)'$. The affine linear transformation on the observations $y_i \mapsto a + cy_i$ yields the same transformation on the limits $b_\alpha \mapsto a + cb_\alpha$ in item (vi) of section 3. Therefore we can set in simulations, without loss of generality, $\sigma = 1$, and $\beta = 0$. We simulate 5000 replicates from a given ARMA process with fixed coefficients, and from each realization we estimate the parameters by maximum likelihood, and compute the prediction intervals using the plug-in method (16) as well as the Bayesian interval from the formula (7) with $N = 100$. Because the main variation in simulations is between series, the sample size in computing the prediction interval need not be large. Because in simulation we know all the parameters we can compute the frequentist conditional coverage probability

$$P(b_\alpha \leq y_{n+h} \leq b_{1-\alpha} \,|\, \boldsymbol{y}, \beta = 0, \sigma = 1, \boldsymbol{\psi}),$$

where $\boldsymbol{\psi}$ specifies the parameters used in a simulation, and the limits $b_\alpha, b_{1-\alpha}$ are fixed. Averaging these probabilities over the 5000 replications of $\boldsymbol{y}$ from the same model, gives us a good estimate of the frequentist coverage probability

$$P(b_\alpha \leq y_{n+h} \leq b_{1-\alpha} \,|\, \beta = 0, \sigma = 1, \boldsymbol{\psi}),$$

where all $y_{n+h}, b_\alpha, b_{1-\alpha}$ are random. This frequentist coverage probability is used when we compare the plug-in method and the five different Bayesian methods. The joint priors $p(\boldsymbol{\psi})/\sigma$ used in the experiment are defined through $p(\boldsymbol{\psi})$ as follows:

- Uniform prior $p(\boldsymbol{\psi}) \propto 1$.
- Approximate joint Jeffreys's prior $p(\boldsymbol{\psi}) \propto |\boldsymbol{X}'\boldsymbol{V}_{\boldsymbol{\psi}}^{-1}\boldsymbol{X}|^{\frac{1}{2}}|\boldsymbol{J}_{\boldsymbol{\psi}}|^{\frac{1}{2}}$.
- Approximate marginal Jeffreys's prior $p(\boldsymbol{\psi}) \propto |\boldsymbol{J}_{\boldsymbol{\psi}}|^{\frac{1}{2}}$.
- Exact joint Jeffreys's prior

$$p(\boldsymbol{\psi}) \propto \left|\boldsymbol{X}'\boldsymbol{V}_{\boldsymbol{\psi}}^{-1}\boldsymbol{X}\right|^{\frac{1}{2}} \left|\boldsymbol{I}_{22}(\boldsymbol{\psi}) - (2n)^{-1}\boldsymbol{I}_{21}(\boldsymbol{\psi})\boldsymbol{I}_{21}(\boldsymbol{\psi})'\right|^{\frac{1}{2}}.$$

- Exact marginal Jeffreys's prior

$$p(\boldsymbol{\psi}) \propto \left|\boldsymbol{I}_{22}(\boldsymbol{\psi}) - (2n)^{-1}\boldsymbol{I}_{21}(\boldsymbol{\psi})\boldsymbol{I}_{21}(\boldsymbol{\psi})'\right|^{\frac{1}{2}}.$$

All the five priors above are constrained onto the the stationarity and invertibility regions. Figure 1 shows the coverage probabilities of one step ahead prediction intervals for ARMA(1,1) processes with varying values of ϕ and θ. In all cases the

Bayesian methods are superior to the plug-in method, and the differences between priors are rather small. The drop in the curves occurs in the neighborhood of $\phi + \theta = 0$ which corresponds to the white noise process, i.e. the parameters are then unidentified. Also the nearly white noise processes yield unstable estimates for ϕ and θ.

The Figure 2 shows the results for the ten step ahead predictions, where again the plug-in method stays below the nominal level in all cases. On the other hand, the coverage probabilities of the Bayesian method is somewhat over the nominal level in most cases, except when the autoregressive parameter ϕ is near the bounds of the stationary region. Also the variation between different priors is somewhat larger here than in the one step ahead predictions. In most cases the uniform prior is the closest to the nominal level. The variation due to the moving average part is smaller here than in the one step ahead predictions.

In Figure 3 the coverage probabilities of ARMA(2,1) processes are shown, with varying parameter values and forecast horizon ranging from one to ten. Cases where $\phi_1 = -1.4$ correspond to alternating autocorrelation function, and in these cases coverage probabilities are usually higher than in non-alternating cases ($\phi_1 = 1.4$). Also, uniform stationary prior seems to perform slightly worse than Jeffreys's priors. Again in all cases the Bayesian methods are superior to the plug-in method. In non-alternating cases the marginal Jeffreys priors seem to give higher coverages than the joint versions, but in alternating cases the difference is negligible. Overall, Bayesian methods perform relatively well.

7 Predicting the number of Internet users

As an illustration, we apply our method to the series of the number of users logged on to an Internet server each minute over 100 minutes. The data is previously studied by Makridakis et al. (1998) and Durbin & Koopman (2001). The former authors fitted ARMA(3,0) to the differenced series, whereas the latter ones preferred ARMA(1,1) for the same series. We use here the first 84 differences for model fitting, and then compute the prediction intervals for the next 15 time points. The Akaike information criterion suggests ARMA(1,1) as the best model. The estimated ARMA coefficients are $\hat{\phi} = 0.65, \hat{\theta} = 0.49$. The additional two estimates are $\hat{\beta} = 0.84$, and $\hat{\sigma}^2 = 10.07$. The complete time series with the simulated 90% prediction intervals are shown in Figure 4, together with median estimates which are computed by setting $\alpha = 0.5$ in the Bayesian calculations. For the plug-in method, the mean is used. These simulations are based on 100,000 replicates. As the differences between exact and approximate versions of Jeffreys's prior turns out to

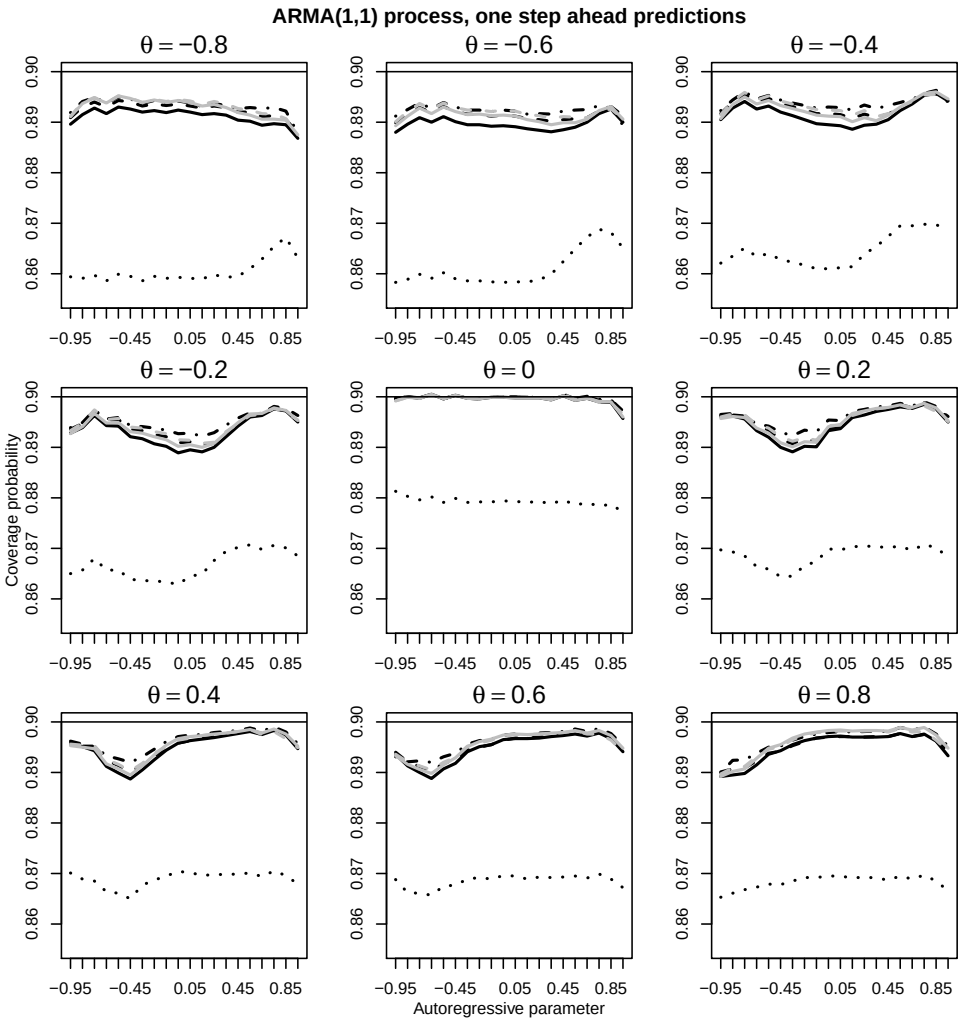

Figure 1. Coverage probabilities of one step ahead prediction intervals for ARMA(1,1) processes. The lines are: black dotted line = plug-in method, the solid black line = approximate joint Jeffreys's prior, the solid gray line = exact joint Jeffreys's prior, the dashed black line = approximate marginal Jeffreys's prior, the dashed gray line = exact marginal prior, the dot-and-dash line = uniform stationary prior.

be negligible, only approximate versions are shown. However, difference between joint and marginal priors is evident: marginal priors give substantially larger upper bounds for the prediction intervals. The upper bounds given by uniform prior is between the different Jeffreys priors, whereas the plug-in gives much smaller upper bounds than any of simulated intervals. On the lower bounds, differences are smaller.

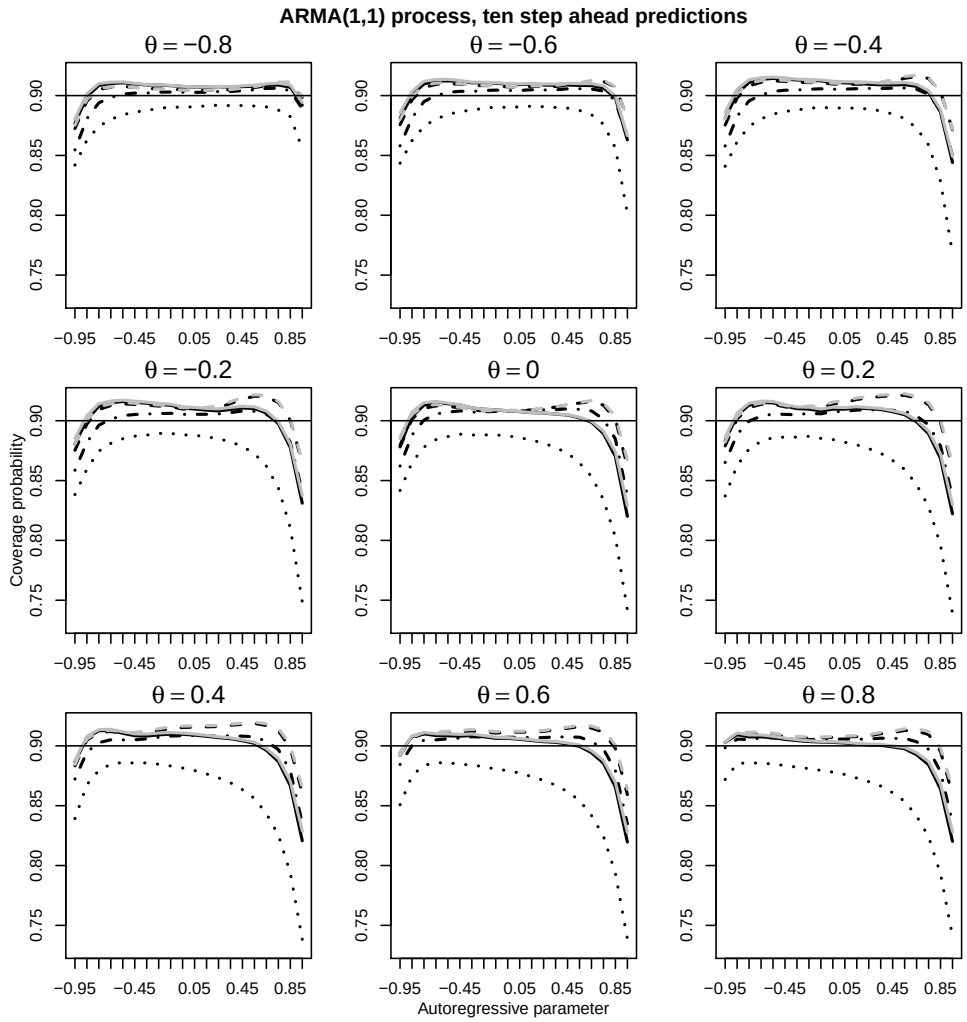

Figure 2. Coverage probabilities of ten step ahead prediction intervals for ARMA(1,1) processes. The lines are: black dotted line = plug-in method, the solid black line = approximate joint Jeffreys's prior, the solid gray line = exact joint Jeffreys's prior, the dashed black line = approximate marginal Jeffreys's prior, the dashed gray line = exact marginal prior, the dot-and-dash line = uniform stationary prior.

Given that the estimated model is correct, we can compute the average coverage probabilities of the intervals. These are given in Table 1 when the forecast horizon $h = 15$. The prediction limits and their standard errors are also given. The reported mean coverage probabilities are based on 10,000 series replicates. Within each replicate 100 values are used in (7) for the Bayesian prediction interval.

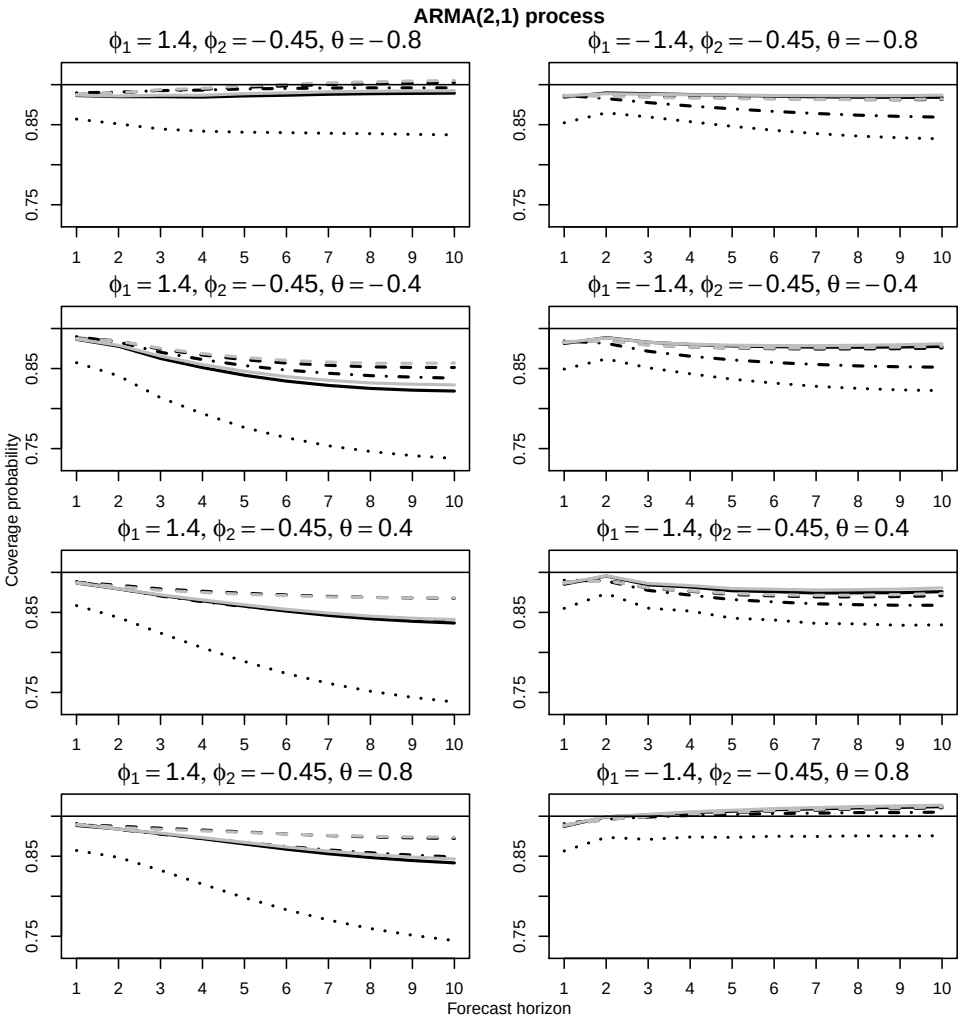

Figure 3. Coverage probabilities of the prediction intervals of varying step sizes for ARMA(1,1) processes. The lines are: black dotted line = plug-in method, the solid black line = approximate joint Jeffreys's prior, the solid gray line = exact joint Jeffreys's prior, the dashed black line = approximate marginal Jeffreys's prior, the dashed gray line = exact marginal prior, the dot-and-dash line = uniform stationary prior.

8 Discussion

In this paper we have extended the importance sampling approach presented in Helske & Nyblom (2013) from AR models to general ARMA models, and studied the effect of different prior choices on the coverage probabilities using simulated and real data. Extension of this approach to integrated ARMA models is straightforward. As may be inferred from sections 2 and 3, our method could be applied also to models outside the ARIMA framework. Compared to Markov Chain Monte

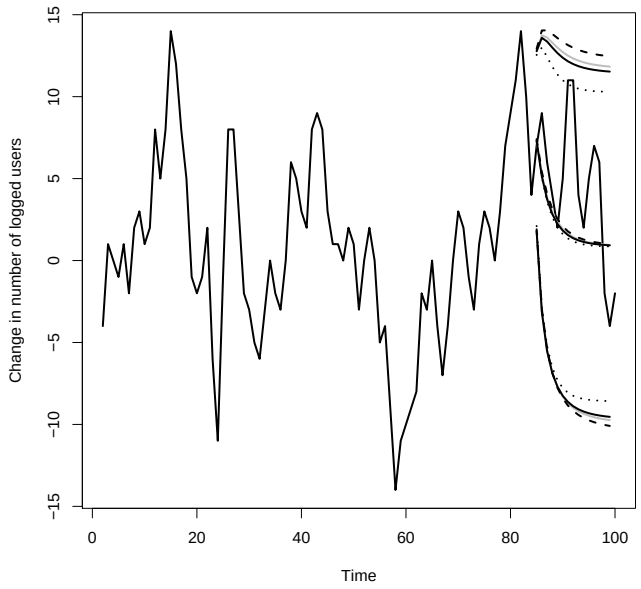

Figure 4. The prediction bands for the change of the number of users logged on to the Internet during the last 15 minutes. The lines are the black dotted line = the traditional plug-in method, the solid black line = approximate joint Jeffreys's prior, the dashed black line = approximate marginal Jeffreys's prior, the solid gray line = uniform stationary prior.

Table 1. Coverage probabilities and prediction limits for the Internet series with forecast horizon $h = 15$ and the nominal coverage probability of 0.9.

	Uniform	Joint	Marginal	Plug-in
Coverage	0.906	0.900	0.914	0.866
$\hat{b}_\alpha$	-9.73	-9.54	-10.09	-8.57
s.e.($\hat{b}_\alpha$)	0.02	0.02	0.06	–
$\hat{b}_{1-\alpha}$	11.83	11.53	12.46	10.29
s.e.($\hat{b}_{1-\alpha}$)	0.02	0.01	0.02	–

Carlo methods, we argue that method presented here is more straightforward to implement and understand, and it could also be computationally cheaper as we are only sampling the model parameters, not the future observations itself. Although we do not need to concern ourselves with the convergence problems of MCMC methods, careful checking of obtained importance weights is still needed. For example if the estimated model parameters are near the boundary of the stationary region with large variance, most of the weights can be zero due to the stationary constraint and there can be few simulated parameters with very large weights which

dominate the whole sample. On the other hand, this should also be visible in the standard errors of the prediction limits, which are easily obtained during prediction interval computation.

Our simulation studies show that a simple uniform prior with stationarity and invertibility constraints performs relatively well in most cases. As the uniform prior is computationally much cheaper than the different versions of Jeffreys's prior, we feel that it could be used as a default prior in practical cases. In addition, a similar check as in section 7 regarding the average coverage probabilities can give further information on the accuracy of the adopted prior.

References

Ansley, C.F. & Kohn, R. (1986). Prediction Mean Squared Error for State Space Models With Estimated Parameters. *Biometrika* 73, 467–473.

Barndorff-Nielsen, O.E. & Cox, D.R. (1996). Prediction and Asymptotics. *Bernoulli* 2, 319–340.

Beran, R. (1990). Calibrating Prediction Regions. *Journal of the American Statistical Association* 85, 715–723.

Box, G.E.P., Jenkins, G.M. & Reinsel, G.C. (2008). *Time Series Analysis: Forecasting and Control*. Fourth edition. Hoboken: Wiley.

Chatfield, C. (1993). Calculating Interval Forecasts. *Journal of Business & Economic Statistics* 11, 121–135.

Chatfield, C. (1996). Model Uncertainty and Forecast Accuracy. *Journal of Forecasting* 15, 495–508.

Clements, M.P. & Hendry, D.F. (1999). *Forecasting Non-stationary Economic Time Series*. Cambridge: The MIT Press.

Clements, M.P. & Kim, J.H. (2007). Bootstrap Prediction Intervals for Autoregressive Time Series. *Computational Statistics & Data Analysis* 51, 3580–3594.

Durbin, J. & Koopman, S.J. (2001). *Time Series Analysis by State Space Methods*. New York: Oxford University Press.

Fuller, W.A. & Hasza, D.P. (1981). Properties of Predictors for Autoregressive Time Series. *Journal of the American Statistical Association* 76, 155–161.

Grigoletto, M. (1998). Bootstrap Prediction Intervals for Autoregressions: Some Alternatives. *International Journal of Forecasting* 14, 447–456.

Helske, J. & Nyblom, J. (2013). Improved Frequentist Prediction Intervals for Autoregressive Models by Simulation. Submitted.

Kabaila, P. & Syuhada, K. (2008). Improved Prediction Limits for AR(p) and ARCH(p) processes. *Journal of Time Series Analysis* 29, 213–223.

Kim, J.H. (2004). Bootstrap Prediction Intervals for Autoregression Using Asymptotically Mean-Unbiased Estimators. *International Journal of Forecasting* 20, 85–97.

Lin, T.I. & Ho, H.J. (2008). A simplified approach to inverting the autocovariance matrix of a general ARMA(p, q) process. *Statistics and Probability Letters* 78, 36–41.

Makridakis, S., Wheelwright, S.C. & Hyndman, R.J. (1998). *Forecasting: Methods and Applications*. Third edition. New York: Wiley.

Masarotto, G. (1990). Bootstrap Prediction Intervals for Autoregressions. *International Journal of Forecasting* 6, 229–239.

Pascual, L., Romo, J. & Ruiz, E. (2004). Bootstrap Predictive Inference for ARIMA Processes. *Journal of Time Series Analysis* 25, 449–465.

Pfeffermann, D. & Tiller, R. (2005). Bootstrap Approximation to Prediction MSE for State-Space Models With Estimated Parameters. *Journal of Time Series Analysis* 26, 893–916.

Phillips, P.C.B. (1979). The Sampling Distribution of Forecasts From a First-Order Autoregression. *Journal of Econometrics* 9, 241–261.

Quenneville, B. & Singh, A.C. (2000). Bayesian Prediction Mean Squared Error for State Space Models With Estimated Parameters. *Journal of Time Series Analysis* 21, 219–236.

Rodriguez, A. & Ruiz, E. (2009). Bootstrap Prediction Intervals in State-Space Models. *Journal of Time Series Analysis* 30, 167–178.

Vidoni, P. (2004). Improved Prediction Intervals for Stochastic Process Models. *Journal of Time Series Analysis* 25, 137–154.

Vidoni, P. (2009). A Simple Procedure for Computing Improved Prediction Intervals for Autoregressive Models. *Journal of Time Series Analysis* 30, 577–590.

III

Helske, J., Nyblom, J., Ekholm, P., and Meissner, K. (2013). "Estimating aggregated nutrient fluxes in four Finnish rivers via Gaussian state space models". *Environmetrics*, 24(4):237– 247.

Research Article Environmetrics

Received: 21 May 2012, Revised: 4 February 2013, Accepted: 18 February 2013, Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/env.2204

Estimating aggregated nutrient fluxes in four Finnish rivers via Gaussian state space models

Jouni Helske[a]*, Jukka Nyblom[a], Petri Ekholm[b] and Kristian Meissner[b]

Reliable estimates of the nutrient fluxes carried by rivers from land-based sources to the sea are needed for efficient abatement of marine eutrophication. Although nutrient concentrations in rivers generally display large temporal variation, sampling and analysis for nutrients, unlike flow measurements, are rarely performed on a daily basis. The infrequent data calls for ways to reliably estimate the nutrient concentrations of the missing days. Here, we use the Gaussian state space models with daily water flow as a predictor variable to predict missing nutrient concentrations for four agriculturally impacted Finnish rivers. Via simulation of Gaussian state space models, we are able to estimate aggregated yearly phosphorus and nitrogen fluxes, and their confidence intervals.

The effect of model uncertainty is evaluated through a Monte Carlo experiment, where randomly selected sets of nutrient measurements are removed and then predicted by the remaining values together with re-estimated parameters. Results show that our model performs well for rivers with long-term records of flow. Finally, despite the drastic decreases in nutrient loads on the agricultural catchments of the rivers over the last 25 years, we observe no corresponding trends in riverine nutrient fluxes. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: simulation; sparse data; interpolation; Kalman filter; Kalman smoother

1. INTRODUCTION

Abatement of marine eutrophication calls for reliable estimates of the nutrient fluxes carried by rivers from land-based sources to the sea. Monitoring programs of many important rivers in Finland, and elsewhere, typically involves daily measurements of water flow, but due to the costs, much more infrequent sampling and analysis of phosphorus and nitrogen concentrations. Yet, the concentrations of nutrients often show large temporal variation, especially in rivers receiving loading from diffuse sources (Kauppila and Koskiaho, 2003). The more infrequent the water quality data are, the more sensitive the flux estimates are to the method used to estimate the concentrations for the unsampled days. Several interpolation and extrapolation methods have been suggested to estimate missing monitoring data (Young *et al.*, 1988; Rekolainen *et al.*, 1991; Kronvang and Bruhn, 1996; Quilbé *et al.*, 2006). Although many of the methods simply assume that the observation made on a specific day represents the concentration level for a longer period (e.g., between the midpoints of the preceding, current, and next observation), other approaches make use of the relationship between the concentration and some other variable, usually the flow.

Our aim is to develop a method for estimating fluxes of total phosphorus and total nitrogen for rivers mainly impacted by diffuse loading from agriculture for a given period, commonly a year. For prediction of the missing nutrient concentration measurements, we use a time varying regression model with an additional autoregressive component using the water flow measurements as predictor variables. Various simulation techniques are employed for evaluating our results. As a general framework, we use Gaussian state space models together with Kalman filter and smoother.

2. METHODS

2.1. Interpolation via state space models and simulation

Our approach to modeling nutrient concentrations and fluxes is based on state space modeling with Kalman filtering, smoothing, and interpolation. The form of the Gaussian state space model sufficient for our purposes is

* *Correspondence to: Jouni Helske, University of Jyväskylä Department of Mathematics and Statistics, P.O.Box 35 (MaD) Jyväskylä, FI 40014. E-mail: jouni.helske@jyu.fi*

a *University of Jyväskylä, Department of Mathematics and Statistics, Finland*

b *Finnish Environment Institute, Finland*

$$y_t = X_t\beta_t + \epsilon_t, \quad \epsilon_t \sim NID(0, H) \tag{1}$$

$$\beta_{t+1} = T\beta_t + \eta_t, \quad \eta_t \sim NID(0, Q), \quad t = 1, 2, \ldots, T \tag{2}$$

where NID stands for "normally and independently distributed." The first row (1) is called an observation equation and the second row (2) a state equation. The observed process $\{y_t\}$ may be a scalar or vector valued. The unobserved state process $\{\beta_t\}$ is often a vector process. The process starts with $\beta_1 \sim N(b_1, P_1)$ independently of error processes $\{\epsilon_t\}, \{\eta_t\}$. In our application the system matrix T is a time invariant diagonal matrix, whereas the system matrices X_t contain time varying predictor values. The state process $\{\beta_t\}$ is a latent process of time varying levels and regression coefficients. The model is defined in more detail in Sections 4.1 and 4.2. Further, the covariance matrices H and Q are time invariant.

In our application the interpolation problem arises because there are missing observations. Let Y comprise all the non-missing observations. If the value y_t at time t is missing, then the Kalman smoother provides its estimate as the conditional mean $\hat{y}_t = X_t\hat{\beta}_t$ together with $\hat{\beta}_t = E(\beta_t \mid Y)$ and the conditional covariance matrix $\mathrm{Var}(y_t \mid Y) = S_t$. The Gaussian assumption then yields

$$y_t \mid Y \sim N(\hat{y}_t, S_t) \tag{3}$$

which can be used for obtaining prediction error limits. Plainly, the interpolated value is unbiased in the sense that $E(y_t - \hat{y}_t) = 0$.

Formula (3) is useful for single missing values. However, our primary interest is a nonlinear compound measure over a time span $t+1, \ldots, t+s$ of length s (e.g., a calendar year), denoted by

$$m_{t,s} = \sum_{i=1}^{s} q_{t+i} e^{y_{t+i}}$$

where q_t is the water flow on the day t, and e^{y_t} is the daily nutrient concentration. If we had the values q_t and y_t measured on each day, then we would have correct nutrient fluxes. Admittedly, this is not exactly true due to the measurement errors, but it would satisfy the practical needs of evaluating the yearly fluxes. In the subsequent analysis, we focus on the effects of missing nutrient measurements compared to the ideal case of having all measurements.

In section 4, we define our model. Under the specified model, we replace the missing values with the estimates that are simply their conditional expectations. Furthermore, to assess their accuracy, we need the conditional variances as well. Formally, we need to determine

$$m_{t,s} = E\left[\sum_{i=1}^{s} q_{t+i} e^{y_{t+i}} \,\middle|\, Y\right] \tag{4}$$

$$V_{t,s} = \mathrm{Var}\left[\sum_{i=1}^{s} q_{t+i} e^{y_{t+i}} \,\middle|\, Y\right] \tag{5}$$

Although the conditional means are easily estimated by using known results of log-normal variables, the variances are more complicated because of correlations between the smoothed state variables (see Durbin and Koopman (2002, section 4.5)). Therefore, we rely on simulations (see Durbin and Koopman (2002)). Additionally, these simulations allow easy constructions for the prediction intervals, which are analytically intractable, because the distribution of the sum of the log-normal variables cannot be given in a closed form.

For simulating the missing observations conditionally on Y, we simulate realizations $(\tilde{\beta}, \tilde{\epsilon})$ from their joint conditional distribution $p(\beta, \epsilon|Y)$. Then simulated observations are obtained from $\tilde{y}_t = X_t\tilde{\beta}_t + \tilde{\epsilon}_t,\ t = 1, \ldots, n$. As we are simulating conditionally on Y, $\tilde{y}_t = y_t$ if y_t is observed, as y_t belongs to Y. The simulation from $p(\beta, \epsilon|Y)$ can be done by augmenting state vector β_t with disturbance ϵ_t, similarly as in Durbin and Koopman (2001, p. 131), and by using the simulation smoothing algorithm of Durbin and Koopman (2002) for the augmented state vector. With a large number of replications, the conditional mean (4) and variance (5) are computed naturally as averages. More specifically, let $\tilde{y}_{1j}, \ldots, \tilde{y}_{nj}$ be the j^{th} simulated series, and

$$\tilde{m}_{s,t,j} = \sum_{i=1}^{s} q_{t+i} e^{\tilde{y}_{t+i,j}}$$

Then, with N replicates, the conditional expectations and variances are obtained respectively as

$$m_{t,s} = \frac{1}{N}\sum_{j=1}^{N} \tilde{m}_{s,t,j}$$

$$V_{t,s} = \frac{1}{N}\sum_{j=1}^{N} (\tilde{m}_{s,t,j} - m_{t,s})^2$$

Assuming that the estimated model is true, the accuracy of the yearly total nutrient fluxes can be computed in terms of prediction intervals. The prediction interval with coverage probability $1-2\alpha$ is found by taking the r^{th} smallest and the r^{th} largest value among $\{\tilde{m}_{s,t,j}\}$, $j = 1,\ldots,N$ with $r = N\alpha$; denoted as $\tilde{m}_{s,t,\text{low}}$ and $\tilde{m}_{s,t,\text{up}}$. Assuming the estimated parameters true, the required prediction interval is

$$[\tilde{m}_{s,t,\text{low}}, \tilde{m}_{s,t,\text{up}}]$$

The other measure of accuracy is the coefficient of variation

$$\frac{\sqrt{V_{t,s}}}{m_{t,s}}$$

All the computations in this paper have been done in R (R Development Core Team, 2012), using the state space modeling package KFAS (Helske, 2012), where the simulation of the state vector is done by using the simulation smoothing with two antithetic variables to reduce the error due to the simulation (Durbin and Koopman, 2002).

2.2. Model fitting and evaluation

The unknown parameters of the nutrient concentration model can be estimated by maximum likelihood method, by using the Kalman filter for computing the log-likelihood of the model. The Kalman filter updating formulas yield us the predicted state $b_{t+1} = E(\beta_{t+1} \mid y_1,\ldots,y_t)$, the prediction $X_{t+1}b_{t+1}$ for y_{t+1}, the prediction error $v_{t+1} = y_{t+1} - X_{t+1}b_{t+1}$, and the prediction error variance (or the covariance matrix in multivariate case) $\text{Var}(v_t) = F_t$.

The log-likelihood of a linear Gaussian state space model can be written in terms of prediction errors and their covariance matrices, which in applications, depend on unknown parameters. Let us denote the parameter vector by ψ, and let $v_{t,\psi}$ and $F_{t,\psi}$ be the prediction errors and their covariance matrices under ψ. Then the likelihood is given by

$$\log L(\psi) = -\frac{np}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{n}\left(\log|F_{t,\psi}| + v'_{t,\psi}F^{-1}_{t,\psi}v_{t,\psi}\right)$$

where p is the dimension of y_t.

The non-stationary part of the state vector is initialized by the diffuse method suggested by (Durbin and Koopman, 2001), whereas the stationary components are assumed to have a stationary distribution at start. When the series $\{y_t\}$ is multivariate, we transform it into a univariate form as in Durbin and Koopman (2001). This enables us to treat totally and partially missing values automatically as well as automatically adjust the likelihood correctly.

The effect of model uncertainty, comprising parameter uncertainty and the uncertainty due to model choice, is evaluated by removing k nutrient measurement vectors from the dataset. The model is then fitted to the thinned data. Let f_k be the total nutrient flux of the removed days, and $\widehat{f}_k$ is the corresponding figure estimated using the thinned dataset. The relative error due to thinning is then $(\widehat{f}_k - f_k)/f_k$. Assuming the model is true and ignoring the parameter estimation error, the difference $e_k = \widehat{f}_k - f_k$ has mean zero. Furthermore, with larger k, the average error per day e_k/k tends to be smaller. The same is true also for the relative error $e_k/f_k = (e_k/k)/(f_k/k)$. Therefore, if we plot the absolute relative errors $|e_k|/f_k$ on the thinning size k, we expect to see a decreasing curve, given our model is true. However, if these values remain more or less constant or are increasing, then our model is severely biased.

The overall effect of thinning is assessed through a Monte Carlo experiment. We remove randomly k nutrient measurement vectors and compute the mean relative error

$$\text{MRE}_k = \frac{1}{B}\sum_{i=1}^{B}\frac{|\widehat{f}_{i,k} - f_{i,k}|}{f_{i,k}} \tag{6}$$

where B denotes the number of random replicates and i refers to ith replicate.

3. DATA

Our data consist of the concentrations of total phosphorus and total nitrogen, and daily water flow measurements from four rivers located in southern Finland, Paimionjoki, Aurajoki, Porvoonjoki, and Vantaanjoki, during 1985–2010. The nutrient data are taken from the databases of the Finnish Environment Institute. Total phosphorus and total nitrogen concentrations have been determined spectrometrically from water samples after digestion with peroxodisulphate.

The catchments of these rivers all have a high proportion of the agricultural land (24–43%, Table 1), and the soil is dominated by clay, which renders the water turbid. Much of the phosphorus in these rivers is transported in association with eroded soil particles. In addition, the catchments contain only few lakes (lake percentage 0.30–2.6), which results in high day to day variation in flow. In all the rivers, agriculture is the major source of nutrients.

Table 1. Catchment characteristics of the rivers studied

	Paimionjoki	Aurajoki	Porvoonjoki	Vantaanjoki
Catchment area, km^2	1088	874	1273	1686
Lakes, %	1.6	0.3	1.3	2.3
Agricultural land, %	42.8	36.8	31.2	23.8
Constructed area, %	2.5	4.8	4.1	9.2
Mean flow, $m^3 s^{-1}$	9.5	8.5	12.7	15.9
Wastewater load, % of total flux	0.5	0.7	12.3	6.3

At the beginning of our observation period, the Porvoonjoki has received substantial waste–water loading from the city of Lahti, but due to improved treatment the share of waste–water to total loading has decreased with time, to an average 12% of the anthropogenic loading. In the Vantaanjoki, the respective proportion of waste–water loading is 6.3%, whereas in the other two rivers, it is below 1%.

Daily measurements on nutrient concentrations are available for only 5–10% of the time, whereas flow measurements are usually available for each day. A few flow measurements are missing in the Paimionjoki and Aurajoki series. For the Paimionjoki, flow measurements are missing from mid-October to mid-November for 2004, whereas for the Aurajoki, flow values are missing on a single day in 1985 and on a total of 99 days between 2004–2010. The missing flow measurements in Paimionjoki and Aurajoki are estimated from an auxiliary four variate state space model defined as in (1) and (2) with all matrices X_t and T being identity matrices. The model is called a local level model, for example, see Harvey (1989). Amisigo and van de Giesen (2005) have used a similar model to patch gaps in daily riverflow series.

4. RESULTS

4.1. Relating nutrient concentration and river flow

It can be argued, as has been done by Wartiovaara (1975) and Rankinen *et al.* (2010), that the high water flow due to the precipitation has two opposite effects on the nutrient loadings. Precipitation increases the diffuse loading from the agriculture while simultaneously diluting waste–water loading. We have tried to take both these aspects into account. In Figure 1, we have plotted the concentrations on the flow, both in logarithms, but due to zero values, we have first added one to the flow values. To address both of the mutually opposing effects caused by precipitation-induced high flows, we have decided to regress the log-concentration y_t on both $\log(1+q_t)$, and $1/\log(2+q_t)$. In the latter, we have added two to ensure a finite value. Figure 1 includes also some regression curves: the loess curves of first degree (Cleveland and Devlin, 1988), and the ordinary least squares regression of y_t on $\beta_0 + \beta_1 \log(1+q_t)$ and on $\beta_0 + \beta_1 \log(1+q_t) + \beta_2/\log(2+q_t)$.

By visual inspection, the relation between the concentration and the flow seems to be linear or slightly curved in a log scale. Moreover, the loess curve and the regression curve from model with two predictor variables are quite close to each other, whereas the regression line from model with one predictor lies apart, especially for nitrogen measurements. Therefore, in some cases, it seems clearly beneficial to include both $x_{1,t} = \log(1+q_t)$ and $x_{2,t} = 1/\log(2+q_t) = 1/\log(1+e^{x_{1t}})$ as the predictor variables in the model. To treat all series equally, both predictors are present in each model. Note that this visual inspection with regression and loess curves is about finding the proper relationship between concentration and flow, and it ignores the time aspect of the problem which, as we will later see, is an important part of the modeling.

4.2. State space specification

As the phosphorus and nitrogen concentration measurements are correlated, we model them together but separately for each river. The model applied to each river is of the form

$$\begin{aligned} y_t^P &= \mu^P + \alpha_t^P + \beta_{1,t}^P x_{1,t} + \beta_{2,t}^P x_{2,t} + \epsilon_t^P \\ y_t^N &= \mu^N + \alpha_t^P + \beta_{1,t}^N x_{1,t} + \beta_{2,t}^N x_{2,t} + \epsilon_t^N \\ \alpha_{t+1} &= T\alpha_t + \xi_t \\ \beta_{t+1} &= \beta_t + \eta_t \end{aligned} \tag{7}$$

where $\left(y_t^P, y_t^N\right)$ is a bivariate process of the logarithms of phosphorus and nitrogen concentrations, respectively; β_t consists of all coefficients $\beta_{j,t}^i$, $i = P, N$, $j = 1, 2$, and α_t consists of zero-mean first-order autoregressive processes α_t^P and α_t^N with $T = \text{diag}[\phi^P, \phi^N]$ containing the corresponding autoregressive parameters. The disturbance processes $\epsilon_t \sim N(0, \Sigma_\epsilon)$, $\eta_t \sim N(0, \Sigma_\eta)$, and $\xi_t \sim N(0, \Sigma_\xi)$

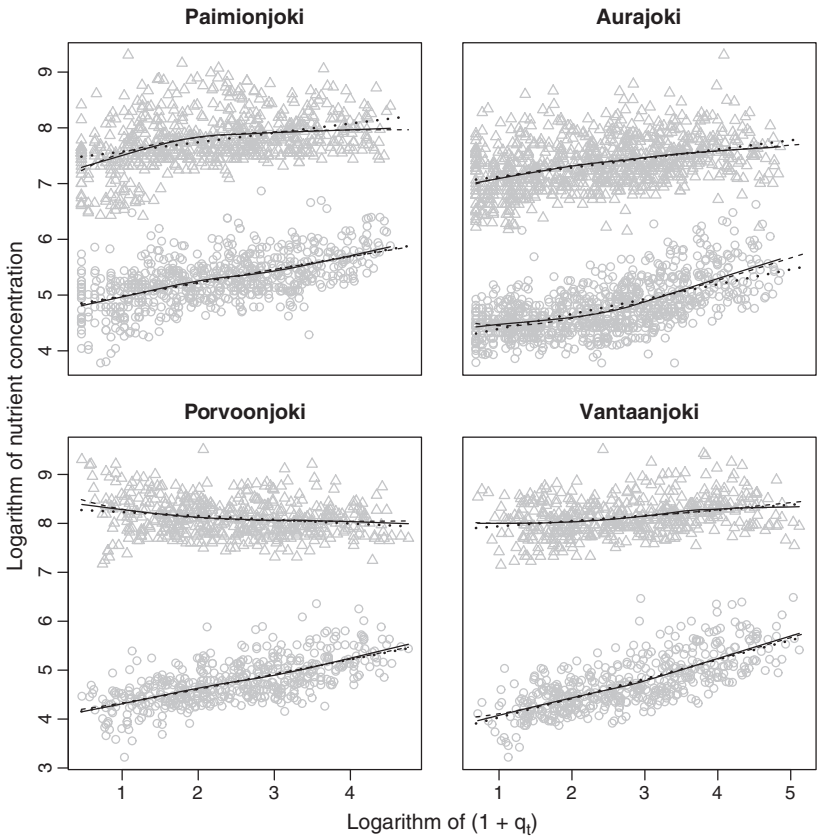

Figure 1. Scatter plots of log-concentrations of nutrients and $\log(1 + q_t)$, with loess curves of first degree (solid line), the regression curves with one explanatory variable (dotted line) and with two explanatory variables (dashed line). The circles correspond to the phosphorus measurements and triangles to the nitrogen measurements

are independent of each other. For simplicity, Σ_η is assumed to be a diagonal matrix. When the diagonal elements are positive, the regression coefficients vary according to a random walk allowing the dependence between the flow and the nutrient concentration to change in time.

Note that the model collapses to an ordinary regression model when $\Sigma_\eta = 0$ and $T = 0$ (i.e., $\phi^P = \phi^N = 0$). The first restriction means that the regression coefficients β_t are constants. The second one implies that level processes α_t^P, α_t^N are white noise processes merged into the errors ϵ_t^P and ϵ_t^N, respectively.

Zero variances for the components of coefficient process β_t are sometimes obtained. The state space modeling automatically handles the zero variances in the covariance matrices so that the time invariant regression coefficients coincide with the appropriate generalized least squares estimates. Also, the simulation algorithm is capable of handling the constant states without modifications.

The long-term seasonal weather conditions such as the starting times of snowmelt and autumn rains, as well as the short-term weather conditions such as daily temperature or precipitation also affect concentrations. We assume here that their effects come mainly through flow. In addition, we assume that other environmental effects are mostly captured by the latent autoregressive level processes and coefficient processes of the flow series. We deliberately aim at a parsimonious model with practical formulas for the interpolation of the nutrient fluxes, although the true phenomena behind the variation of nutrient concentrations are obviously more complicated than our model suggests.

4.3. Estimated nutrient fluxes and model parameters

The yearly estimates of the nutrient fluxes obtained by simulating the model are given in Table A.1 in the Appendix. Yearly estimates of nutrient fluxes with their simulated 95% prediction intervals are also shown in Figure 2. Each river exhibits a similar fluctuating patterns without a clear trend. Especially yearly phosphorus fluxes, but also nitrogen fluxes clearly peak in 2008, followed by an even larger drop in 2009. Overall, fewer nutrient measurements result in somewhat wider prediction intervals for Porvoonjoki and Vantaanjoki than for Paimionjoki and Aurajoki.

The estimated values of the unknown variance and autoregressive parameters are shown in Table 2.

Occasionally, the estimation process yields the variance estimates close to zero (i.e., values less than 10^{-8}). In such cases, these are replaced with fixed zeros, and estimation process for other parameters is repeated. In all cases, the likelihood remained practically unchanged.

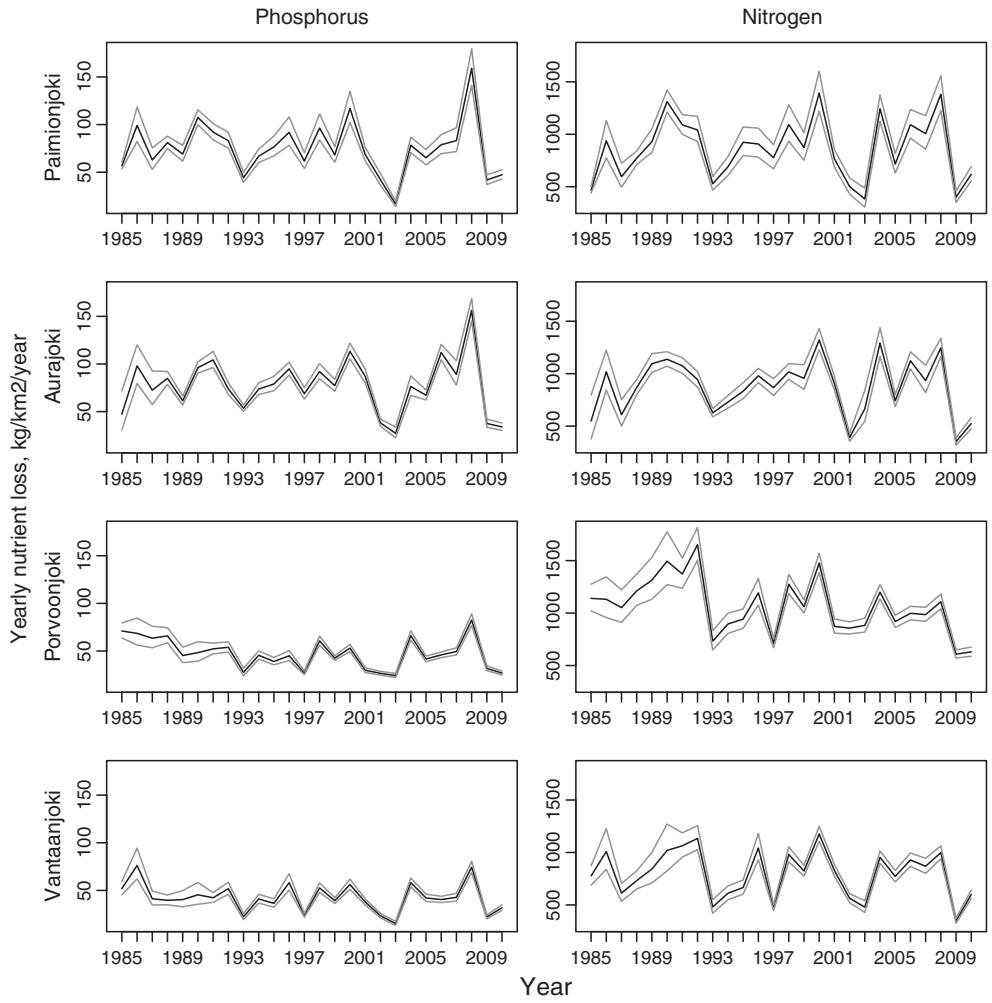

Figure 2. The estimated values and the simulated 95% prediction intervals of the yearly phosphorus and nitrogen fluxes

Standard errors of the estimates are computed by inverting the Hessian matrix given by the optimization function optim in R. The variance parameters are estimated in logarithmic scale. By using their standard errors, the confidence intervals for the log-variances can be obtained. Then the confidence intervals for the variances themselves are easily derived.

In all models, the values of autoregressive parameters are close to one ($0.95 - 0.98$), and therefore the standard errors might not be very useful, as the sampling distributions are far from normal distribution. The correlations ρ_{ξ^P,ξ^N} between the disturbances of the autoregressive processes are around 0.5 for all rivers. This indicates moderate long-term correlation between the underlying phosphorus and nitrogen concentration processes at a given flow level. The instantaneous correlations $\rho_{\epsilon^P,\epsilon^N}$, again given the flow, are smaller and more variable: 0.2 or slightly higher in the Paimionjoki and Porvoonjoki, and negligible in the Aurajoki and Vantaanjoki.

The coefficient processes are shown in Figure 3. Somewhat larger regression coefficients of the reciprocal log-flow of the Porvoonjoki and Vantaanjoki compared with those of Paimionjoki and Aurajoki are in concordance with the fact that the former rivers are subject to higher waste–water loads. Otherwise, the interpretation of the regression coefficient processes is difficult. Nevertheless, as predictive tools, individual river-specific models appear to be highly useful.

4.4. Model criticism

We have also tested models where the autoregressive processes have been replaced with random walks (i.e., $\phi^P = \phi^N = 1$) and a multivariate local level model without regressors, but where the concentration processes are augmented with water flow. In addition, we also tested the ordinary multivariate regression model. All these models yield large autocorrelations of the standardized residuals, and in case of time-varying models, there is a clear inverse relationship between the size of residuals and observed concentration. These apparent violations are avoided by using the model (7). However, even despite obvious violations of model assumptions, yearly estimates of the nutrient fluxes from different time varying models have very similar coefficients of variation with deviations being usually less than one percentage point.

Table 2. Estimates of the unknown parameters and their standard errors in parenthesis

	Paimionjoki	Aurajoki	Porvoonjoki	Vantaanjoki
$\sigma^2_{\eta_1^P}$	1.3×10^{-7}	0	6.4×10^{-7}	9.0×10^{-8}
$\log\left(\sigma^2_{\eta_{11}}\right)$	−15.83(2.07)	—	−14.27(1.05)	−16.22(3.75)
$\sigma^2_{\eta_2^P}$	1.4×10^{-6}	0	7.8×10^{-5}	5.5×10^{-5}
$\log\left(\sigma^2_{\eta_{12}}\right)$	−13.46(3.39)	—	−9.46(0.89)	−9.81(0.88)
$\sigma^2_{\eta_1^N}$	0	0	2.1×10^{-5}	7.7×10^{-6}
$\log\left(\sigma^2_{\eta_{21}}\right)$	—	—	−10.78(0.64)	−11.78(0.98)
$\sigma^2_{\eta_2^N}$	2.9×10^{-6}	5.8×10^{-6}	3.1×10^{-4}	3.9×10^{-5}
$\log\left(\sigma^2_{\eta_{22}}\right)$	−12.74(1.52)	−12.06(1.00)	−8.08(1.03)	−10.15(0.82)
$\sigma^2_{\xi^P}$	5.4×10^{-3}	1.0×10^{-2}	1.1×10^{-2}	7.0×10^{-3}
$\log\left(\sigma^2_{\xi_1}\right)$	−5.23(0.16)	−4.58(0.14)	−4.51(0.19)	−4.96(0.25)
$\sigma^2_{\xi^N}$	8.3×10^{-3}	1.1×10^{-2}	4.0×10^{-3}	4.0×10^{-3}
$\log\left(\sigma^2_{\xi_2}\right)$	−4.80(0.13)	−4.47(0.13)	−5.51(0.24)	−5.51(0.25)
$\sigma^2_{\epsilon^P}$	3.1×10^{-2}	1.8×10^{-2}	6.5×10^{-3}	2.2×10^{-2}
$\log\left(\sigma^2_{\epsilon_1}\right)$	−3.47(0.13)	−4.02(0.19)	−5.04(0.89)	−3.81(0.33)
$\sigma^2_{\epsilon^N}$	3.0×10^{-2}	1.7×10^{-2}	1.5×10^{-2}	1.3×10^{-2}
$\log\left(\sigma^2_{\epsilon_2}\right)$	−3.49(0.14)	−4.09(0.22)	−4.21(0.30)	−4.32(0.35)
ρ_{ξ^P,ξ^N}	0.58 (0.04)	0.46(0.04)	0.53(0.06)	0.48(0.06)
$\rho_{\epsilon^P,\epsilon^N}$	0.26(0.08)	0.07(0.12)	0.20(0.29)	0.02(0.24)
μ^P	4.47(0.13)	3.83(0.10)	3.11(0.25)	2.72(0.29)
μ^N	7.63(0.15)	7.62(0.11)	7.04(0.23)	6.83(0.24)
ϕ^P	0.98(0.004)	0.95(0.006)	0.95(0.009)	0.96(0.007)
ϕ^N	0.98(0.003)	0.96(0.005)	0.98(0.006)	0.98(0.005)

In the case of the ordinary regression model, the coefficients of variation are often substantially smaller. In Figure 4, the coefficients of variation are plotted against the yearly sample sizes of the concentration measurements. The coefficients of variation from the model (7) depend on the yearly sample sizes, whereas results from the ordinary regression model are overoptimistic and counterintuitive: uncertainty in the yearly flux estimate is independent from the amount of measurements in a given year. Both models use the daily water flow for the prediction of the missing concentration measurements, but the ordinary regression is immune to the time order of the measurements, and only the total number of measurements is important. However, we acknowledge that because yearly flux estimates are always conditioned on the model, all models underestimate the true errors of yearly flux estimates, and none of the models considered is "true".

The quantile-to-quantile plots of the standardized residuals of the models reveal heavier tails compared with the normal distribution. This would be problematic if the interest is on the daily values, but because we are interested in yearly values, we believe that the possible non-normality is not critical here. This is because the yearly measure of nutrient fluxes is a sum, which tends to be more normal than its components by the central limit theorem. For evaluating the effects of non-normality, we have made a simulation experiment where the errors ϵ_t are a random sample from a heavy-tailed bivariate t-distribution with three degrees of freedom scaled to have $\mathrm{Var}(\epsilon_t) = \Sigma_\epsilon$. New values representing the concentration measurements, on the same days as the true ones, are then simulated from the model with the estimated parameters. By using these simulated measurements, we fit our proposed model (under Gaussian assumptions), and we computed the coefficients of variation for the yearly fluxes. The coefficients of variation from the simulation are, on the average, within one percentage point of those obtained from the actual dataset thus displaying the negligible effect of non-normality.

The main purpose of our model is to estimate the yearly nutrient flux. To this end, we developed the thinning experiment explained at the end of section 2.2. We have made five experiments by randomly removing 10%, 20%, 30%, 40%, and 50% from the concentration values. The resulting relative errors (6) are reported in Table A.2 in the Appendix. The number of simulations is $B = 2000$, and each time, the parameters are re-estimated. If the model is correct, we expect a decreasing trend, and this is mostly what we observe. The loss of relative accuracy with 30% thinning is about 5% or less. However, the mean absolute errors $\mathrm{MAE}_k = \sum_{i=1}^{B} |\widehat{f}_{i,k} - f_{i,k}|/B$ increase rapidly as expected when thinning is increased (Table A.3 in the Appendix). When measuring bias using average errors $\mathrm{AE}_k = \sum_{i=1}^{B} (\widehat{f}_{i,k} - f_{i,k})/B$ (Table A.4 in the Appendix), the total phosphorus flux is underestimated in all rivers, whereas the total nitrogen flux is usually overestimated, except for Aurajoki, where the nitrogen flux is underestimated. Overall, the results suggest that our model performs well enough for practical

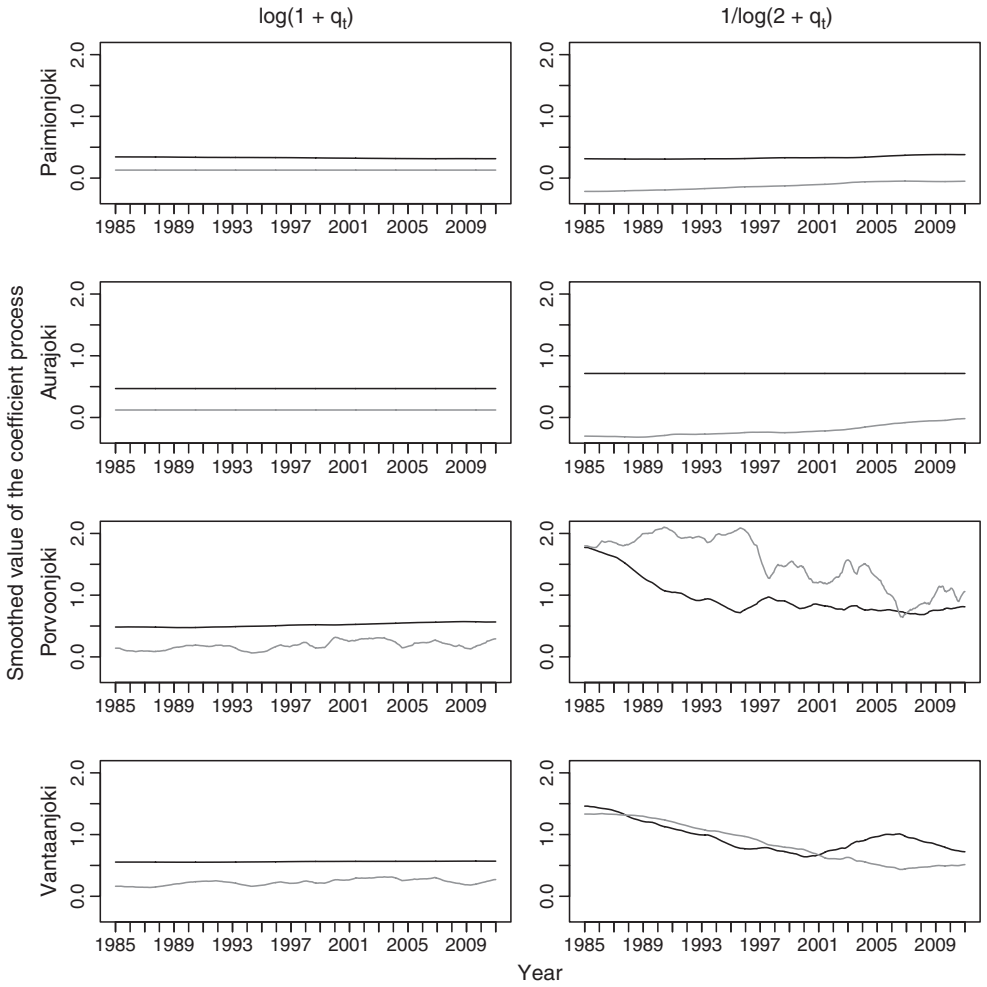

Figure 3. The smoothed coefficient processes corresponding to the predictor variables $\log(1 + q_t)$ (left) and $1/\log(2 + q_t)$ (right) for all four rivers. The black lines represent the processes corresponding to phosphorus observations, and the gray lines correspond to the nitrogen observations. Constant horizontal line corresponds to the null variance of the coefficient process

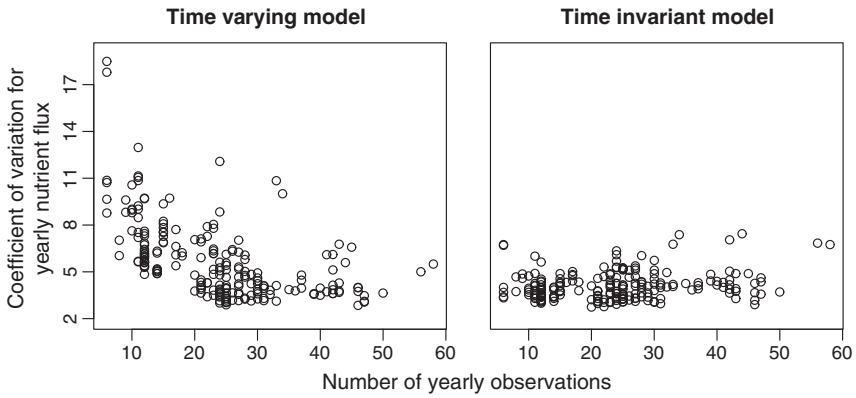

Figure 4. The relationship between coefficients of variation for yearly nutrient flux and the number of yearly nutrient concentration observations. The figure on left corresponds to final model (7), and the figure on right to the ordinary regression model

purposes. For the ordinary regression model, the mean relative and absolute errors are always larger, and prominently so for nitrogen fluxes. The average errors show that the ordinary regression model overestimates the nitrogen fluxes more than our model, whereas the bias of phosphorus fluxes is slightly smaller. Finally, we note that removing predictor $1/\log(2 + q_t)$ from the final model (7) always worsens the model performance compared with including it.

5. DISCUSSION

We have used Gaussian state space models with partially sparse data for modeling the yearly nutrient fluxes of four rivers running through catchments dominated by agricultural land use. The large proportion of "missing" daily nutrient concentration measurements for corresponding daily flow measurements increased the uncertainty regarding the model selection, parameter estimation, and prediction, thus encouraging the use of models with simple structure and large flexibility.

During the observational period covered by this study, Finnish agricultural farmlands experienced a substantial decrease in phosphorus and nitrogen balance OECD (2012). Despite this drastic decrease in nutrient balance, we did not observe any corresponding trends in nutrient fluxes over the last 25 years for any of the four rivers examined here. Greatly reduced nutrient balances do not always lead to concurrent reduction in riverine nutrient fluxes, for example, due to high nutrient reserves in soil and groundwater (e.g., Stålnacke *et al.* (2004)). Moreover, although nutrient balances form a crucial indicator of the risk of nutrient losses from agriculture, changes in other agricultural practices or in climate may have had an opposite effect on the load (Ekholm *et al.*, 2007).

While we have reported results when the daily water flow is only predictor variable, we have also augmented the model with locally important variables such as daily air temperature, precipitation, and several functions of these. To examine the possible effect of large scale climate patterns, we have also used the North Atlantic Oscillation and Arctic Oscillation indices in combination with flow. Additions of variables operating at either small (temperature and precipitation) or large scales (North Atlantic Oscillation or Arctic Oscillation) did not improve results for any of the models we used.

Many studies examining nutrient dynamics of rivers have stated the need for extensive datasets to be able to make precise statements on the nutrient flux (e.g., Rekolainen *et al.* (1991)). Although we are conscious that the thinning of an originally sparse data by half can include possible computational caveats and thus may lead to artifacts, our results seem to indicate that when daily flow data are available, relatively sparse data on nutrient concentrations can be used to estimate yearly fluxes. If the aim of monitoring is to assess yearly fluxes of principal nutrients from agriculturally dominated watersheds to receiving downstream locations (e.g., the sea), our findings imply the potential to lower the frequency of water quality (i.e., nutrient) sampling intensities for rivers with permanent gauging stations and long-term records of flow. It should be noted that these concentration measurements could be used for other types of analysis as well, where the number of samples cannot not be reduced.

Acknowledgements

The authors thank two anonymous referees for their comments, which have led to considerable improvements. Support by grant 110045 from Emil Aaltonen Foundation for J. Helske is gratefully acknowledged.

REFERENCES

OECD. 2012. Follow-up study of the impacts of agri-environmental measures in Finland. In *OECD, Evaluation of Agri-environmental Policies: Selected Methodological Issues and Case Studies*. OECD Publishing. DOI:10.1787/9789264179332-8-en.

Amisigo BA, van de Giesen NC. 2005. Using a spatio-temporal dynamic state–space model with the EM algorithm to patch gaps in daily riverflow series. *Hydrology and Earth System Sciences* **9**(3): 209–224, DOI: 10.5194/hess-9-209-2005. http://www.hydrol-earth-syst-sci.net/9/209/2005/.

Cleveland WS, Devlin SJ. 1988. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association* **83**(403): 596–610, DOI: 10.2307/2289282.

Durbin J, Koopman SJ. 2001. Time Series Analysis by State Space Methods. Oxford University Press: New York.

Durbin J, Koopman SJ. 2002. A simple and efficient simulation smoother for state space time series analysis. *Biometrika* **89**: 603–615, DOI: 10.1093/biomet/89.3.603.

Ekholm P, Granlund K, Kauppila P, Mitikka S, Niemi J, Rankinen K, Räike A, Räsänen J. 2007. Influence of EU policy on agricultural nutrient losses and the state of receiving surface waters in Finland. *Agricultural and Food Science* **16**(4): 282–300.

Harvey AC. 1989. Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press: Cambridge.

Helske J. 2012. KFAS: Kalman filter and smoother for exponential family state space models. R package version 0.9.11.

Kauppila P, Koskiaho J. 2003. Evaluation of annual loads of nutrients and suspended solids in Baltic rivers. *Nordic Hydrology* **34**(3): 203–220, DOI: 10.2166/nh.2003.013.

Kronvang B, Bruhn A. 1996. Choice of sampling strategy and estimation method for calculating nitrogen and phosphorus transport in small lowland streams. *Hydrological Processes* **10**: 1483–1501, DOI: 10.1002/(SICI)1099-1085(199611)10:11<1483::AID-HYP386>3.0.CO;2-Y.

Quilbé R, Rousseau AN, Duchemin M, Poulin A, Gangbazo G, Villeneuve JP. 2006. Selecting a calculation method to estimate sediment and nutrient loads in streams: application to the Beaurivage river (Québec, Canada). *Journal of Hydrology* **326**: 295–310, DOI: 10.1016/j.jhydrol.2005.11.008.

R Development Core Team. 2012. R: a language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.ISBN3-900051-07-0.

Rankinen K, Ekholm P, Sjöblom H, Rita H. 2010. Nutrient losses from catchments and the governing factors. In *Follow-up Study on the Impacts of Agri-Environment Measures (MYTVAS3) - Midterm Report*, Aakkula J, Manninen T, Nurro M (eds), Reports of Finland Ministry of Agriculture and Forestry 1; 122–131. In Finnish.

Rekolainen S, Posch M, Kämäri J, Ekholm P. 1991. Evaluation of the accuracy and precision of annual phosphorus load estimates from two agricultural basins in Finland. *Journal of Hydrology* **128**: 237–255, DOI: 10.1016/0022-1694(91)90140-D.

Stålnacke P, Vandsemb S, Grimvall A, Jolankai G. 2004. Changes in nutrient levels in some Eastern European rivers in response to large-scale changes in agriculture. *Water Science & Technology* **49**(3): 29–36.

Wartiovaara J. 1975. Amounts of substances discharged by rivers off the coast of Finland. *Publications of the Water Research Institute* **13**: 1–54.

Young TC, DePinto JV, Heidtke TM. 1988. Factors affecting the efficiency of some estimators of fluvial total phosphorus load. *Water Resources Research* **24**: 1535–1540, DOI: 10.1029/WR024i009p01535.

APPENDIX

Table A.1. Annual fluxes (kg/km^2/year) and the coefficient of variation in percentages for each river and nutrient

	Paimionjoki		Aurajoki		Porvoonjoki		Vantaanjoki	
	P	N	P	N	P	N	P	N
1985	57 (2.8)	469 (3.5)	53 (17.8)	547 (18.5)	71 (5.7)	1140 (5.6)	52 (7.0)	776 (6.0)
1986	99 (9.2)	939 (9.7)	100 (8.8)	1020 (9.6)	68 (10.7)	1132 (8.8)	76 (10.9)	1009 (9.7)
1987	63 (9.4)	596 (9.7)	72 (11.1)	609 (9.7)	63 (9.0)	1052 (7.5)	41 (9.0)	613 (7.2)
1988	81 (4.0)	772 (4.3)	85 (3.8)	867 (4.1)	66 (6.0)	1211 (6.3)	40 (6.7)	732 (6.0)
1989	69 (6.1)	929 (6.1)	62 (3.8)	1094 (4.1)	45 (9.0)	1314 (7.6)	40 (10.6)	842 (8.8)
1990	107 (3.6)	1312 (4.0)	96 (3.1)	1137 (3.1)	48 (10.9)	1494 (8.5)	45 (13.0)	1020 (11.0)
1991	92 (4.8)	1090 (4.5)	101 (3.6)	1076 (3.6)	52 (5.6)	1372 (5.3)	42 (6.2)	1064 (5.6)
1992	83 (5.2)	1041 (5.8)	72 (3.9)	945 (4.1)	54 (5.0)	1651 (4.9)	52 (6.2)	1134 (5.1)
1993	44 (6.0)	528 (6.2)	54 (3.1)	627 (3.3)	28 (7.4)	733 (6.4)	22 (7.3)	481 (6.8)
1994	67 (5.4)	690 (6.6)	73 (4.2)	729 (4.2)	46 (4.8)	896 (5.6)	41 (5.7)	611 (5.4)
1995	77 (6.9)	924 (7.6)	78 (4.6)	832 (4.7)	39 (5.1)	943 (4.9)	36 (6.3)	666 (5.2)
1996	92 (8.1)	907 (7.8)	94 (3.5)	980 (3.6)	45 (5.9)	1193 (5.3)	58 (7.6)	1042 (6.1)
1997	62 (6.9)	777 (7.5)	68 (4.4)	867 (4.8)	27 (3.7)	711 (3.1)	24 (4.1)	473 (3.0)
1998	96 (7.3)	1092 (8.2)	92 (4.0)	1017 (3.9)	61 (4.1)	1274 (3.6)	53 (4.8)	980 (3.8)
1999	68 (6.1)	872 (7.7)	79 (4.1)	950 (6.1)	43 (2.9)	1060 (3.1)	39 (4.0)	823 (3.2)
2000	117 (6.9)	1392 (7.1)	115 (3.7)	1317 (3.8)	53 (3.7)	1476 (3.2)	56 (4.8)	1176 (3.1)
2001	69 (5.1)	766 (5.6)	88 (4.0)	905 (3.6)	30 (4.3)	873 (4.0)	38 (4.3)	828 (3.4)
2002	42 (7.3)	501 (7.9)	39 (4.9)	391 (5.0)	27 (3.7)	856 (3.5)	23 (5.1)	563 (3.9)
2003	17 (8.8)	383 (12.1)	30 (10.0)	662 (10.8)	24 (4.5)	883 (3.9)	15 (7.1)	479 (5.9)
2004	78 (5.2)	1243 (5.2)	76 (6.1)	1251 (5.3)	66 (3.6)	1200 (3.0)	58 (4.1)	953 (3.1)
2005	65 (6.3)	717 (6.5)	65 (3.8)	722 (3.9)	42 (3.6)	919 (3.2)	42 (4.7)	772 (3.7)
2006	79 (6.4)	1091 (6.3)	106 (3.5)	1075 (3.6)	46 (3.2)	998 (3.3)	40 (4.2)	928 (3.5)
2007	83 (7.8)	1006 (8.0)	81 (6.6)	894 (6.8)	49 (3.8)	985 (3.5)	43 (4.8)	869 (4.3)
2008	159 (6.2)	1380 (6.2)	153 (3.7)	1245 (3.7)	83 (3.7)	1108 (3.3)	74 (4.1)	998 (3.2)
2009	42 (6.3)	400 (7.0)	35 (5.5)	335 (5.0)	32 (3.9)	609 (3.3)	22 (4.7)	348 (3.4)
2010	47 (5.4)	619 (5.6)	33 (5.6)	501 (5.1)	27 (3.8)	630 (3.6)	32 (4.8)	600 (3.4)

Table A.2. The mean relative error percentages and their standard errors for the final model (7) and for the ordinary least squares regression model with two predictors (marked by †)

	Paimionjoki		Aurajoki		Porvoonjoki		Vantaanjoki	
	P	N	P	N	P	N	P	N
MRE_{10}	5.8 (0.09)	4.8 (0.08)	6.5 (0.11)	5.3 (0.11)	6.4 (0.11)	3.9 (0.07)	7.1 (0.13)	4.7 (0.08)
$MRE^{\dagger}_{10}$	7.0 (0.12)	7.4 (0.13)	8.9 (0.15)	8.0 (0.14)	7.6 (0.13)	6.7 (0.11)	8.9 (0.15)	7.6 (0.13)
MRE_{20}	4.5 (0.07)	3.6 (0.06)	5.4 (0.09)	4.6 (0.08)	5.0 (0.08)	3.0 (0.05)	5.8 (0.10)	3.7 (0.06)
$MRE^{\dagger}_{20}$	5.0 (0.08)	5.9 (0.10)	6.8 (0.11)	6.3 (0.10)	6.0 (0.10)	5.1 (0.09)	7.0 (0.11)	6.0 (0.10)
MRE_{30}	4.0 (0.06)	3.3 (0.06)	5.1 (0.08)	4.1 (0.07)	4.5 (0.07)	2.7 (0.05)	5.3 (0.09)	3.3 (0.06)
$MRE^{\dagger}_{30}$	4.5 (0.07)	5.1 (0.09)	6.2 (0.10)	5.3 (0.09)	4.9 (0.08)	4.5 (0.07)	5.9 (0.10)	5.1 (0.09)
MRE_{40}	3.6 (0.06)	3.2 (0.05)	4.9 (0.08)	3.8 (0.06)	4.3 (0.07)	2.7 (0.05)	5.2 (0.08)	3.1 (0.06)
$MRE^{\dagger}_{40}$	4.1 (0.07)	5.0 (0.08)	5.7 (0.09)	4.9 (0.08)	4.7 (0.08)	4.2 (0.07)	5.7 (0.09)	4.7 (0.09)
MRE_{50}	3.5 (0.06)	3.1 (0.05)	4.9 (0.08)	3.9 (0.06)	4.3 (0.07)	2.8 (0.05)	4.9 (0.08)	3.3 (0.06)
$MRE^{\dagger}_{50}$	3.9 (0.07)	4.6 (0.08)	5.5 (0.09)	4.9 (0.08)	4.6 (0.07)	4.1 (0.07)	5.4 (0.09)	4.8 (0.08)

Table A.3. The mean absolute errors (metric tons) and their standard errors for the final model (7) and for the ordinary least squares regression model with two predictors (marked by †)

	Paimionjoki		Aurajoki		Porvoonjoki		Vantaanjoki	
	P	N	P	N	P	N	P	N
MAE_{10}	0.9 (0.02)	7.8 (0.14)	1.7 (0.03)	14.0 (0.32)	0.8 (0.02)	9.9 (0.17)	1.3 (0.03)	13.5 (0.24)
$MAE_{10}^{\dagger}$	1.1 (0.02)	11.8 (0.20)	2.2 (0.04)	20.3 (0.35)	1.0 (0.02)	16.7 (0.28)	1.5 (0.03)	21.7 (0.38)
MAE_{20}	1.4 (0.02)	11.8 (0.21)	2.7 (0.05)	23.7 (0.46)	1.3 (0.02)	15.1 (0.26)	2.0 (0.04)	20.9 (0.37)
$MAE_{20}^{\dagger}$	1.6 (0.03)	19.0 (0.31)	3.4 (0.06)	31.4 (0.52)	1.6 (0.03)	25.4 (0.43)	2.4 (0.04)	33.8 (0.59)
MAE_{30}	1.9 (0.03)	16.0 (0.28)	3.9 (0.07)	31.7 (0.55)	1.8 (0.03)	20.0 (0.35)	2.8 (0.05)	28.1 (0.50)
$MAE_{30}^{\dagger}$	2.1 (0.04)	24.8 (0.42)	4.7 (0.08)	40.3 (0.66)	1.9 (0.03)	33.6 (0.55)	3.1 (0.05)	43.6 (0.77)
MAE_{40}	2.3 (0.04)	20.9 (0.36)	5.0 (0.08)	39.1 (0.66)	2.3 (0.04)	26.9 (0.50)	3.6 (0.06)	35.3 (0.64)
$MAE_{40}^{\dagger}$	2.6 (0.04)	32.2 (0.53)	5.8 (0.10)	48.9 (0.83)	2.5 (0.04)	41.5 (0.71)	3.9 (0.07)	53.4 (0.96)
MAE_{50}	2.7 (0.05)	25.7 (0.44)	6.1 (0.10)	49.9 (0.79)	2.8 (0.05)	35.8 (0.65)	4.3 (0.07)	46.9 (0.82)
$MAE_{50}^{\dagger}$	3.0 (0.05)	37.8 (0.64)	6.9 (0.11)	61.2 (1.05)	3.0 (0.05)	51.3 (0.88)	4.7 (0.08)	68.7 (1.20)

Table A.4. The mean errors (metric tons) and their standard errors for the final model (7) and for the ordinary least squares regression model with two predictors (marked by †)

	Paimionjoki		Aurajoki		Porvoonjoki		Vantaanjoki	
	P	N	P	N	P	N	P	N
ME_{10}	−0.3 (0.02)	−0.1 (0.22)	−0.9 (0.05)	−4.6 (0.43)	−0.3 (0.02)	−0.1 (0.28)	−0.4 (0.04)	0.7 (0.39)
$ME_{10}^{\dagger}$	−0.2 (0.03)	5.0 (0.31)	−0.8 (0.06)	4.7 (0.57)	−0.2 (0.03)	3.7 (0.46)	−0.3 (0.04)	5.0 (0.61)
ME_{20}	−0.6 (0.04)	0.6 (0.33)	−1.8 (0.07)	−9.0 (0.67)	−0.6 (0.04)	−0.5 (0.43)	−0.9 (0.06)	1.2 (0.59)
$ME_{20}^{\dagger}$	−0.2 (0.04)	10.7 (0.47)	−1.5 (0.09)	9.4 (0.85)	−0.4 (0.04)	8.8 (0.68)	−0.6 (0.07)	11.5 (0.92)
ME_{30}	−0.8 (0.05)	6.1 (0.43)	−2.8 (0.09)	−9.9 (0.87)	−0.9 (0.05)	1.8 (0.57)	−1.3 (0.07)	3.8 (0.80)
$ME_{30}^{\dagger}$	−0.3 (0.06)	15.2 (0.60)	−2.6 (0.12)	12.2 (1.09)	−0.4 (0.05)	13.8 (0.88)	−1.0 (0.08)	16.8 (1.19)
ME_{40}	−0.9 (0.06)	9.1 (0.55)	−3.7 (0.11)	−11.4 (1.07)	−1.3 (0.06)	2.6 (0.78)	−1.9 (0.09)	5.5 (1.01)
$ME_{40}^{\dagger}$	−0.2 (0.07)	22.2 (0.74)	−3.2 (0.14)	18.4 (1.31)	−0.7 (0.07)	19.0 (1.09)	−1.5 (0.11)	22.4 (1.45)
ME_{50}	−0.9 (0.07)	11.6 (0.68)	−4.6 (0.13)	−11.2 (1.34)	−1.5 (0.07)	3.9 (1.03)	−2.3 (0.11)	6.3 (1.33)
$ME_{50}^{\dagger}$	−0.2 (0.09)	25.4 (0.90)	−4.0 (0.17)	23.3 (1.64)	−0.9 (0.08)	21.8 (1.36)	−1.7 (0.13)	28.5 (1.84)

