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 $E(\cdot \,|\, \boldsymbol{y})$ 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 $\boldsymbol{z}_t' = (\boldsymbol{x}_t', 1, 0, \ldots, 0)$,

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

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

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

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

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

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

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

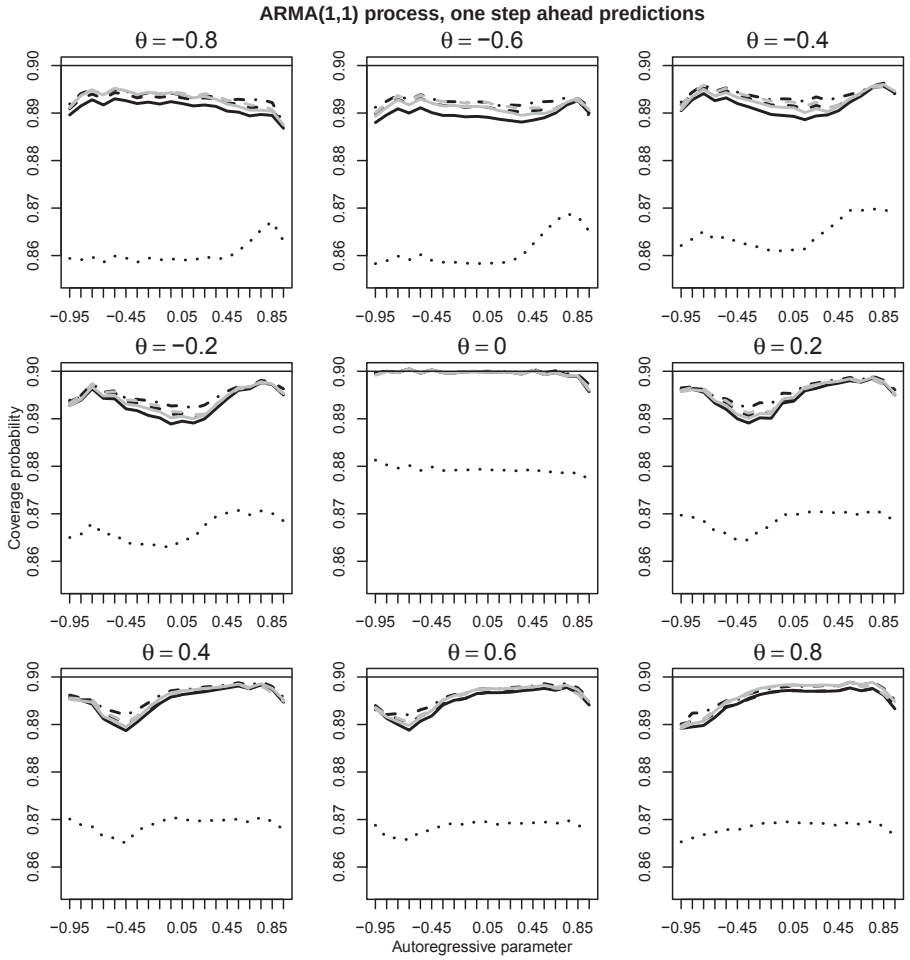

Figure 1. Coverage probabilities of one step ahead prediction intervals for ARMA(1,1) processes. The lines are: black dotted line = plug-in method, the solid black line = approximate joint Jeffreys's prior, the solid gray line = exact joint Jeffreys's prior, the dashed black line = approximate marginal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

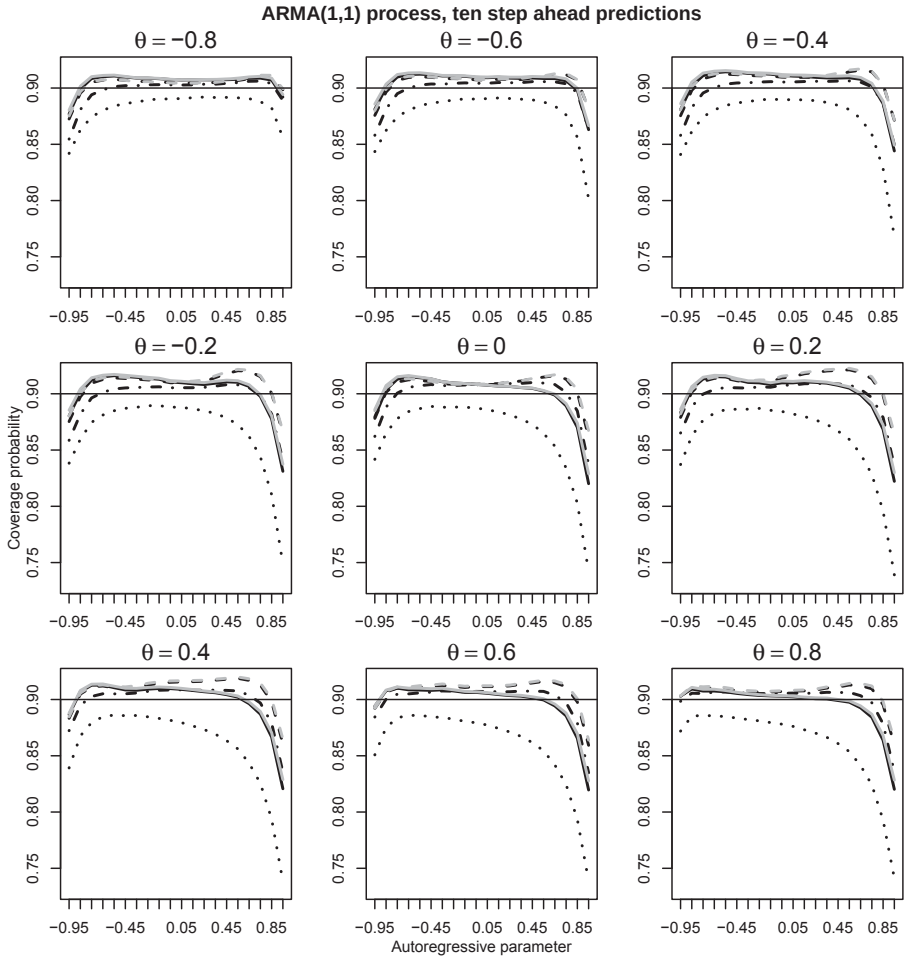

Figure 2. Coverage probabilities of ten step ahead prediction intervals for ARMA(1,1) processes. The lines are: black dotted line = plug-in method, the solid black line = approximate joint Jeffreys's prior, the solid gray line = exact joint Jeffreys's prior, the dashed black line = approximate marginal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

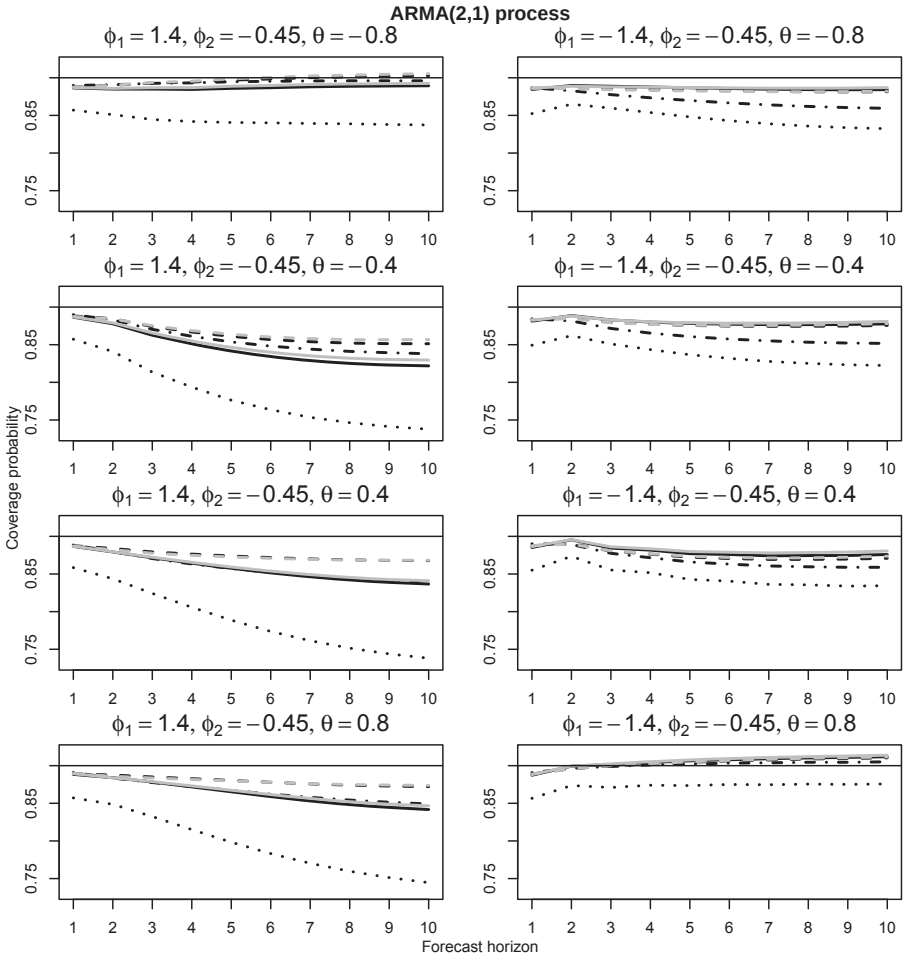

Figure 3. Coverage probabilities of the prediction intervals of varying step sizes for ARMA(1,1) processes. The lines are: black dotted line = plug-in method, the solid black line = approximate joint Jeffreys's prior, the solid gray line = exact joint Jeffreys's prior, the dashed black line = approximate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

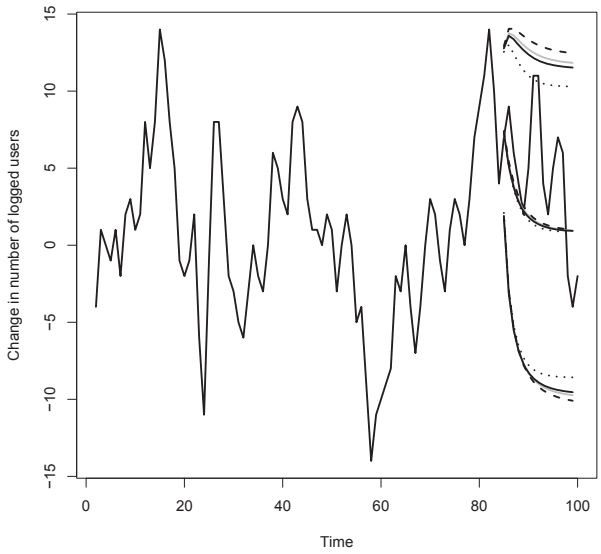

Figure 4. The prediction bands for the change of the number of users logged on to the Internet during the last 15 minutes. The lines are the black dotted line = the traditional plug-in method, the solid black line = approximate joint Jeffreys's prior, the dashed black line = approximate marginal Jeffreys's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