Biometrika (2018), **103**, 1, *pp*. 1–16
Printed in Great Britain

Estimation of prediction error in time series

BY ALEXANDER AUE AND PRABIR BURMAN

Department of Statistics, University of California, One Shields Avenue, Davis, California 95616, U.S.A.

aaue@ucdavis.edu, pburman@ucdavis.edu

SUMMARY

The accurate estimation of prediction errors in time series is an important problem. It immediately affects the accuracy of prediction intervals but also the quality of a number of widely used time series model selection criteria such as AIC and others. Except for simple cases, however, it is difficult or even infeasible to obtain exact analytical expressions for one-step and multi-step predictions. This may be one of the reasons that, unlike in the independent case (see Efron, 2004), until today there has been no fully established methodology for time series prediction error estimation. Starting from an approximation to the bias-variance decomposition of the squared prediction error, this work is therefore concerned with the estimation of prediction errors in both univariate and multivariate stationary time series. In particular, several estimates are developed for a general class of predictors that includes most of the popular linear, nonlinear, parametric and nonparametric time series models used in practice, where causal invertible ARMA and nonparametric AR processes are discussed as lead examples. Simulation results indicate that the proposed estimators perform quite well in finite samples. The estimates may also be used for model selection when the purpose of modeling is prediction.

Some key words: Accumulated prediction error; ARMA models; Cross-validation; Multi-step ahead prediction; Multivariate time series; Nonparametric autoregressive processes; Univariate time series

1. INTRODUCTION

In this paper, the problem of estimating the prediction error of a univariate or multivariate stationary time series is considered in a general setup that can accommodate both linear or nonlinear and parametric or nonparametric specifications. In the literature, many models are available for estimating a series of dependent observations, for example through ARMA, hidden Markov, threshold, fractional ARMA, bilinear and nonparametric autoregressive processes. Most of these can be subsumed under the general model introduced in Priestley (1980). Corresponding multivariate methods are available if more than one feature is recorded over time. More details on many of these models and their respective estimation procedures can be found in the comprehensive books by Shumway and Stoffer (2010) and Fan and Yao (2005).

The estimation of the prediction error is important in itself, since it allows for the construction of prediction intervals. Prediction error estimates are also often used to obtain reasonable models for fitting the data. Rissanen Rissanen (1986) proposed an estimate, the accumulated prediction error (APE), that was designed as a model selection criterion. In addition, there are many other criteria for model selection such as AIC (Akaike, 1974), AICC (Hurvich and Tsai, 1989), BIC (Schwartz, 1978), HQIC (Hannan and Quinn, 1979) and MDL (Rissanen, 1984), among others. These model selection criteria involve estimating a measure of divergence, often the Kullback–Leibler distance or the prediction error, between the true model and the estimated model. If the purpose is to select the "correct" model assuming that a low dimensional "correct" model is available, then BIC or APE are suitable, while AIC and AICC tend to produce biased estimates of the divergence. If the purpose is prediction, AIC and AICC may be more suitable when the existence of a correct model of low dimension is uncertain because they seek to obtain

unbiased estimates of the divergence. Detailed information on model selection methods can be found in the monographs Claeskens and Hjort (2008) and Burnham and Anderson (2002).

In order to motivate the results put forward, the APE criterion for the univariate case is briefly discussed and related to the proposed methods. If $Y_1, \ldots, Y_n$ is the observed stationary time series data and $\hat{Y}_{n+1|n}$ the predictor of Y_{n+1} on the basis of this data, then the prediction error is given by $\mathrm{PE}_n = E[\{Y_{n+1} - \hat{Y}_{n+1|n}\}^2]$. Rissanen's APE seems to have been based on the idea of cross-validation. If $\hat{Y}_{t+1|t}$ is the predictor of Y_{t+1} based on the first t observations $Y_1, ..., Y_t$, then the APE estimate of the prediction error PE_n is given by $\widehat{\mathrm{PE}}_n^{\mathrm{R}} = (n-m)^{-1} \sum_{t=m+1}^{n-1} (Y_{t+1} - \hat{Y}_{t+1|t})^2$ where $m = \lfloor \delta n \rfloor$, $0 < \delta < 1$, is a fraction of the sample size and $\lfloor \cdot \rfloor$ denotes integer part. This has been called one-sided cross-validation by Hart and co-workers in a series of papers; see Hart and Lee (2005) and the references therein. Other cross-validation type approaches in time series may be found in Burman et al. (1994); Racine (2000) as well as in the references cited in these papers. APE is not necessarily a good estimate of PE_n: If m is small, $\widehat{\mathrm{PE}}_n^{\mathrm{R}}$ may have non-negligible bias in estimating PE_n, while if m is close to n, the bias may be low, but the estimate may have high variability instead. There are many papers dealing with prediction and model selection aspects of APE, for example Ing (2007), Hemerly and Davis (1989), Wei (1992), Speed and Yu (1993) and Findley (2005). In a fairly detailed analysis, it is pointed out in Ing (2007) that APE may not always select the best predictive model unless m is close to n.

This paper does not directly address the issue of model selection. Rather it is concerned with the estimation of the prediction error for univariate and multivariate stationary processes using the idea of cross-validation for time series. It is perhaps surprising that, unlike for independent, identically distributed (i.i.d.) observations (see Efron, 2004), there is not yet a fully developed methodology for the estimation of prediction errors for observations dependent in time. This work is seeking to fill in this gap. In the case of i.i.d. data for which the size of the learning set (on which the estimate is based) is considerably smaller that the sample size, correction terms are needed to produce approximately unbiased estimates of the prediction error (see Burman, 1989, 1990). Similar issues arise in the time series context. However, they require different methods because of the dependence inherent in the observations. Unbiased estimation of prediction errors requires correction as will be seen in Section 3. Establishing these corrections is the main contribution of this paper. The general results will be discussed more specifically for causal and invertible ARMA processes and nonparametric AR processes. It will be shown in a simulation study that the suggested methods work well in finite samples and particularly that it can provide significant improvements on Rissanen's APE. The proposed estimates may also be used for model selection if the purpose of modeling is prediction, but that issue is not investigated further in this paper.

2. NOTATIONS AND PRELIMINARIES

Suppose that observations $Y_1, \ldots, Y_n$ have been obtained from the univariate stationary time series

$$Y_t = \mu_t + \varepsilon_t, \tag{1}$$

where μ_t is the conditional mean of Y_t given the past, ε_t are mean zero i.i.d. and ε_t is independent of μ_t. Suppose that a model $\{\mu_t(\theta)\}$ has been employed to estimate $\{\mu_t\}$. It is important to point out that the means $\{\mu_t\}$ may or may not belong to the proposed class $\{\mu_t(\theta)\}$. No assumptions are made regarding the linearity or nonlinearity and parametric or nonparametric form of model (1). For instance, an AR(2) model could be fit to the data, in which case $\mu_t(\theta) = \theta_1 Y_{t-1} + \theta_2 Y_{t-2}$, but the data generating process may be an AR(∞) autoregression and then $\mu_t = \sum_{j=1}^{\infty} c_j Y_{t-j}$ instead. If $\hat{\theta}_s$ is the estimate of θ on the basis of the first s observations, then the estimate of μ_t is given by $\mu_t(\hat{\theta}_s)$ and the corresponding residual is denoted by $\varepsilon_t(\hat{\theta}_s) = Y_t - \mu_t(\hat{\theta}_s)$. In the AR(2) case, one gets

$$\mu_t(\hat{\theta}_s) = \hat{\theta}_{s1} Y_{t-1} + \hat{\theta}_{s2} Y_{t-2},$$

$$\varepsilon_t(\hat{\theta}_s) = (c_1 - \hat{\theta}_{s1}) Y_{t-1} + (c_2 - \hat{\theta}_{s2}) Y_{t-2} + \sum_{j=3}^{\infty} c_j Y_{t-j} + \varepsilon_t$$

and the bias-variance trade-off guiding the development of the proposed methods emerges. For more general time series the residuals $\varepsilon_t(\hat{\theta}_s)$ may have to be understood as approximations, since the infinite past is not available for the computations. In all standard cases, these approximations are justified by an exponentially decaying dependence on the past. Following standard practice in time series, no explicit distinctions are made here. Further examples follow in Section 2.3.

The variance for predicting Y_{n+1}, when the model has been estimated on the basis of the entire available data $Y_1, \ldots, Y_n$, is given by $\mathrm{PE}_n = E[\varepsilon^2_{n+1}(\hat{\theta}_n)]$. Suppose that $n-k$ residuals are available (this happens for ARMA fits when the first few estimates may not be available), then an empirical estimate of PE_n is

$$\widehat{\mathrm{PE}}_n^{\mathrm{emp}} = \frac{1}{n-k}\sum_{t=k}^{n-1} \varepsilon^2_{t+1}(\hat{\theta}_n).$$

It is well known that $\widehat{\mathrm{PE}}_n^{\mathrm{emp}}$ may not be a good estimate of PE_n (see Efron, 2004). If the model is estimated on the basis of the first s observations $Y_1, \ldots, Y_s$, then $\mathrm{PE}_s = E[\varepsilon^2_{s+1}(\hat{\theta}_s)]$ and $\widehat{\mathrm{PE}}_s^{\mathrm{emp}} = (s-k)^{-1}\sum_{t=k+1}^{s} \varepsilon^2_t(\hat{\theta}_s)$. For $m = \lfloor \delta n \rfloor$ with $0 < \delta < 1$, Rissanen's APE estimate of PE_n is defined as

$$\widehat{\mathrm{PE}}_n^{\mathrm{R}} = \frac{1}{n-m}\sum_{t=m}^{n-1} \varepsilon^2_{t+1}(\hat{\theta}_t).$$

When m is small (δ is close to 0), the bias in estimating PE_n by $\widehat{\mathrm{PE}}_n^{\mathrm{R}}$ may not be small. On the other hand, if m is close to n (δ is close to 1), $\widehat{\mathrm{PE}}_n^{\mathrm{R}}$ is almost unbiased for PE_n, but it may not be stable (that is, may have higher variability) since it is based on few residuals. A more detailed discussion on these issues is given below.

2.1. *Properties of* PE_n

At the outset, we state that mean and variance expressions used in the following are understood to be in the asymptotic sense; see Efron (1982) and (Rao, 1973, Chapter 6). For simplicity of exposition, the discussion is focused on the case when the method of estimation is least squares. However, the arguments given are valid for other methods of estimation. In order to obtain estimates of the conditional means μ_t, one often uses a parametric or nonparametric model $\mu_t(\theta)$ and then minimizes the sum of squares of deviations $\sum_{t=1}^{n}[Y_t - \mu_t(\theta)]^2$ to derive the estimates of θ. In the case of penalty methods (see, for example, the approaches via regularization methods in Gerencser, 1992; H. Wang and Tsai, 2007), one adds a penalty term to the sum of squared deviations and then carries out the minimization procedure. Whatever the procedure, there is a tuning parameter, say, p associated with it. The tuning parameter p may be the dimension of θ or a function of the penalty parameter. Typically, the larger the value of p, the smaller the bias in estimating μ_t. However, the larger the value of p, the larger the variance associated with estimating μ_t.

Let $\bar{\theta}$ be the minimizer of $E[\{\mu_t - \mu_t(\theta)\}^2]$, and denote $\beta_0 = E[\{\mu_{n+1} - \mu_{n+1}(\bar{\theta})\}^2]$. Note that β_0, which does not depend on n, is the square of the model bias and it decreases as p increases. Using the arguments given in Section 2 of the Supplementary Material to the paper one may write $\mathrm{PE}_n \approx \sigma^2_\varepsilon + \beta_0 + n^{-1}\beta_1$, or, more generally,

$$\mathrm{PE}_n \approx \sigma^2_\varepsilon + \beta_0 + \frac{\beta_1}{n} + \frac{\beta_2}{n^2}, \tag{2}$$

where the values of β_0, β_1 and β_2 depend on the tuning parameter p in such a way that β_0 decreases and β_1 increases as p increases. Underlying (2) is a common asymptotic expansion of mean square errors (for example of Hoffding or von Mises type) which holds under reasonable smoothness conditions on $\mu_t(\theta)$ as a function of θ. The interested reader is referred to the informative book by Taniguchi (1991).

2.2. *Properties of* $\widehat{\mathrm{PE}}_n^{\mathrm{emp}}$ *and* $\widehat{\mathrm{PE}}_n^{\mathrm{R}}$

First, the expected value of $\widehat{\mathrm{PE}}_n^{\mathrm{emp}}$ is approximated. From the arguments given in Section 2 of the Supplementary Material, it follows that $E[\widehat{\mathrm{PE}}_n^{\mathrm{emp}}] \approx \sigma_\varepsilon^2 + \beta_0 + n^{-1}\beta_3$, or, more generally,

$$E\Big[\widehat{\mathrm{PE}}_n^{\mathrm{emp}}\Big] \approx \sigma_\varepsilon^2 + \beta_0 + \frac{\beta_3}{n} + \frac{\beta_4}{n^2}, \tag{3}$$

noting that the constants associated with the n^{-1} and n^{-2} terms are not the same as in the case of PE_n. Oberve that PE_n and $E[\widehat{\mathrm{PE}}_n^{\mathrm{emp}}]$ have the same constant term $\sigma_\varepsilon^2 + \beta_0$. Therefore, from (2) and (3),

$$\mathrm{PE}_n - E\Big[\widehat{\mathrm{PE}}_n^{\mathrm{emp}}\Big] \approx \frac{\beta_5}{n} + \frac{\beta_6}{n^2}. \tag{4}$$

In the second part of this section, the expected value of the APE estimate $\widehat{\mathrm{PE}}_n^{\mathrm{R}}$ of PE_n is approximated. The following notation will be helpful for later discussions. For any real number z, let

$$\varrho_z(\ell) = \frac{1}{n-\ell-m+1}\sum_{t=m}^{n-\ell}\left(\frac{n}{t}\right)^z, \tag{5}$$

where ℓ is a positive integer. Note that $\varrho_z(\ell)$ depends also on m and n, but for notational simplicity this is suppressed. Since the case $\ell = 1$ will frequently appear, denote $\varrho_z(1)$ by ϱ_z. Arguing as in Section 2.1, it follows that

$$\begin{aligned} E\Big[\widehat{\mathrm{PE}}_n^{\mathrm{R}}\Big] &= \frac{1}{n-m}\sum_{t=m}^{n-1}\mathrm{PE}_t \\ &\approx \frac{1}{n-m}\sum_{t=m}^{n-1}\left(\sigma_\varepsilon^2 + \beta_0 + \frac{\beta_1}{t} + \frac{\beta_2}{t^2}\right) \\ &\approx \sigma_\varepsilon^2 + \beta_0 + \varrho_1\frac{\beta_1}{n} + \varrho_2\frac{\beta_2}{n^2}, \end{aligned}$$

with

$$\begin{aligned} \varrho_1 = \varrho_1(1) &= \frac{n}{n-m}\sum_{t=m}^{n-1}\frac{1}{t} \\ &= -\frac{1}{1-\delta}\log\delta + \frac{1}{2n} + O\left(\frac{1}{n^2}\right). \end{aligned}$$

Note that the constant ϱ_1 is larger than 1 for any $0 < \delta < 1$. It is about 1.72 when $\delta = 0.3$ and is equal to 1.39 when $\delta = 0.5$. Also, $\varrho_1 \approx 1$ when $\delta \approx 1$. So $\widehat{\mathrm{PE}}_n^{\mathrm{R}}$ is almost unbiased for PE_n when $\delta \approx 1$, but it is not expected to be stable (that is, it is likely to exhibit higher variability) since this estimate is averaged over few squared residuals. When δ is of order n^{-1}, then the factor ϱ_1 behaves like $\log n$ and this explains APE's ability to select the correct model when there is indeed a correct model of low dimension (see Hemerly and Davis, 1989). The preceding gives, in analogy to (4), that

$$\mathrm{PE}_n - E\Big[\widehat{\mathrm{PE}}_n^{R}\Big] \approx (1-\varrho_1)\frac{\beta_1}{n} + (1-\varrho_2)\frac{\beta_2}{n^2}. \tag{6}$$

The considerations in this section help set up improved estimators for the prediction error. Two specific proposals will be part of Section 3, where corrections based on (3) for $\widehat{\mathrm{PE}}_n^{\mathrm{emp}}$ and the corresponding expression for $\widehat{\mathrm{PE}}_n^{\mathrm{R}}$ in (6) are introduced.

2.3. *Examples*

Example 1 *(Linear prediction with AR processes).* Let $\{Y_t\}$ be an arbitrary stationary time series and suppose that AR(p) models

$$Y_t = \theta_1 Y_{t-1} + \cdots + \theta_p Y_{t-p} + \varepsilon_t$$

are fitted to the data. If $\{Y_t\}$ is invertible, it admits an AR(∞) representation from which one can deduce the form $\mu_t = \sum_{j=1}^{\infty} c_j Y_{t-j}$ for the conditional mean, the sequence $\{c_j\}$ being the invertibility weights. Signifying transposition by $'$ and denoting by $\hat{\theta}_n = (\hat{\theta}_{n1}, \ldots, \hat{\theta}_{np})'$ the LSE of $\theta = (\theta_1, \ldots, \theta_p)'$, it follows that AR fits aim to predict μ_{n+1} by $\mu_{n+1}(\hat{\theta}_n) = \sum_{j=1}^{p} \hat{\theta}_{nj} Y_{n+1-j}$. The corresponding prediction residual is given by

$$\begin{aligned}\varepsilon_{n+1}(\hat{\theta}_n) &= Y_{n+1} - \mu_{n+1}(\hat{\theta}_n) \\ &= \sum_{j=1}^{p} (c_j - \hat{\theta}_{nj}) Y_{n+1-j} + \sum_{j=p+1}^{\infty} c_j Y_{n+1-j} + \varepsilon_{n+1}.\end{aligned}$$

If the data generating process is causal and invertible, computations for the bias-variance decomposition of the prediction error PE_n are somewhat simpler. This will be part of the next example.

Example 2 *(Linear prediction with ARMA processes).* Let the data generating process $\{Y_t\}$ be specified as a stationary, causal and invertible time series in an i.i.d. sequence $\{\varepsilon_t\}$ with zero mean and variance σ_ε^2. Causality and invertibility ensure the existence of sequences $\{c_j\}$ and $\{d_j\}$ such that the AR(∞) representation $Y_t = \sum_{j=1}^{\infty} c_j Y_{t-j} + \varepsilon_t$ and the MA(∞) representation $Y_t = \sum_{j=0}^{\infty} d_j \varepsilon_{t-j}$ exist. Using the MA(∞) representation, it follows immediately that the autocovariance function is given by $\gamma(h) = E[Y_{t+h} Y_t] = \sigma_\varepsilon^2 \sum_{j=0}^{\infty} d_j d_{j+h}$. Using the AR($\infty$) representation, the conditional mean in (1) becomes $\mu_t = \sum_{j=1}^{\infty} c_j Y_{t-j}$.

Assume that causal and invertible ARMA(p, q) models

$$Y_t = a_1 Y_{t-1} + \cdots + a_p Y_{t-p} + \varepsilon_t + b_1 \varepsilon_{t-1} + \cdots + b_q \varepsilon_{t-q}, \tag{7}$$

are entertained as fits to the data. Necessary and sufficient conditions on the parameter $\theta = (a_1, \ldots, a_p, b_1, \ldots, b_q)'$ ensuring causality and invertibility of (7) can be found in Shumway and Stoffer (2010). These lead to respective coefficient sequences $\{c_j(\theta)\}$ and $\{d_j(\theta)\}$ which can be computed in an iterative fashion (see Box et al., 2008). With these, bias and variance of the prediction error can be analyzed.

To compute the model bias $\beta_0 = E[\{\mu_{n+1} - \mu_{n+1}(\bar{\theta})\}^2]$ if $q = 0$ and AR(p) models $Y_t = \theta_1 Y_{t-1} + \cdots + \theta_p Y_{t-p} + \varepsilon_t$ are fitted, let $c = (c_j)_{j=0}^{\infty}$ and $\Gamma = (\gamma(h - h'))_{h,h'=0}^{\infty}$. Let further Γ_p be the leading $p \times p$ principal submatrix of Γ and Γ_p^{∞} the $p \times \infty$ matrix consisting of the first p rows of Γ. It can now be shown that $\bar{\theta} = \Gamma_p^{-1} \Gamma_p^{\infty} c$ and

$$\beta_0 = c'(\Gamma - \Gamma_p^{\infty\prime} \Gamma_p^{-1} \Gamma_p^{\infty}) c.$$

The magnitude of the bias is determined by the decay of the AR(∞) weights $\{c_j\}$, the decay of the covariances $\gamma(h)$ (thus the MA(∞) weights $\{d_j\}$) and the order p of the fitted AR process. Clearly β_0 becomes smaller if the AR model order p gets larger. Further technical details on the expansion in (2) are offered in the Supplementary Material to the paper.

Example 3 *(Nonlinear prediction with nonparametric AR processes).* Suppose the goal is to model μ_t in (1) as a function $f(\boldsymbol{X}_t)$ of past observations $\boldsymbol{X}_t = (Y_{t-1}, \ldots, Y_{t-p})'$, where the function f is to be estimated from the data. This general model was considered by Lewis and Stevens (1991) and includes important subclasses, namely *additive models:* $f(\boldsymbol{X}_t) = f_1(Y_{t-1}) + \cdots + f_p(Y_{t-p})$ (see Chen and Liu, 2001); *threshold AR models:* $f(\boldsymbol{X}_t) = f_1(Y_{t-p}) Y_{t-1} + \cdots + f_p(Y_{t-p}) Y_{t-p}$ (see Tong, 1983); and *single index models:* $f(\boldsymbol{X}_t) = f_0(\beta_1 Y_{t-1} + \cdots + \beta_p Y_{t-p})$ (see Xia et al., 2002). A good account of additive models for independent data appears in Hastie and Tibshirani (1990). A discussion on many nonparametric

time series models is available in Fan and Yao (2005). Functional autoregressive and other models for vector time series as well as references to earlier work on many nonlinear time series models can be found in Jiang (2014).

The discussion here is carried out for additive models, even though all arguments are valid for the general nonparametric class with appropriate modifications. For the class of additive models, some restrictions on $f_2, \dots, f_p$ (such as $E[f_j(Y_{t-j})] = 0$ for $j = 2, \dots, p$) are needed for identifiability (see Stone, 1986; Burman, 1990). Suppose $f_j(\cdot)$ is being modeled as $\psi_j(\cdot)'\theta_j$, where the ψ_j are some k_j dimensional functions, often regression splines or their modifications for the purpose of identifiability (Burman, 1990). Let B be the $n \times p$ matrix whose t-th row is given by $\psi(\boldsymbol{X}_t)' = [\psi_1(Y_{t-1})', \dots, \psi_p(Y_{t-p})']$. Then μ_t is modeled as

$$\mu_t(\theta) = \psi_1(Y_{t-1})'\theta_1 + \cdots + \psi_p(Y_{t-p})'\theta_p = \psi(\boldsymbol{X}_t)'\theta,$$

where θ is the $k = k_1 + \cdots + k_p$ dimensional column vector of $\theta_1, \dots, \theta_p$. This leads to the linear model $Y = B\theta + \varepsilon$ with corresponding least squares estimate

$$\hat{\theta}_n = (BB')^{-1}B'Y,$$

assuming temporarily for the ease of discussion that Y_s, $s = -p+1, \dots, 0$, are available for the estimation. Hence the prediction of μ_{n+1} is $\mu_{n+1}(\hat{\theta}_n) = \psi(\boldsymbol{X}_{n+1})'\hat{\theta}_n$. The Supplementary Material to the paper gives arguments on why the prediction error expansion in (2) holds for these additive models.

3. PROPOSED ESTIMATES OF PE_n

3.1. *The modified empirical estimate*

Consider the empirical estimate $\widehat{\mathrm{PE}}_n^{\mathrm{emp}}$ of PE_n. Even though it may not be a good estimate of PE_n, equation (4) suggests that it may be used as initial estimate in a first step. In a second step, one then obtains an estimate of the bias in estimating PE_n by $\widehat{\mathrm{PE}}_n^{\mathrm{emp}}$ to adjust the empirical estimates. An estimate of the expected bias $E[\mathrm{PE}_n - \widehat{\mathrm{PE}}_n^{\mathrm{emp}}]$ can be set up as

$$C_n(w) = \frac{1}{n-m}\sum_{t=m}^{n-1} w_t\Big(\varepsilon_{t+1}^2(\hat{\theta}_t) - \widehat{\mathrm{PE}}_t^{\mathrm{emp}}\Big),$$

where the weights $\{w_t\}$ need to be chosen appropriately to estimate $E[\mathrm{PE}_n - \widehat{\mathrm{PE}}_n^{\mathrm{emp}}]$ well. The modified empirical estimate is then given by

$$\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w) = \widehat{\mathrm{PE}}_n^{\mathrm{emp}} + C_n(w). \tag{8}$$

It follows from equations (4) and (8) that

$$\mathrm{PE}_n - E\Big[\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w)\Big] \approx [1 - g_1(w)]\frac{\beta_5}{n} + [1 - g_2(w)]\frac{\beta_6}{n^2},$$

where

$$g_k(w) = \frac{1}{n-m}\sum_{t=m}^{n-1}\left(\frac{n}{t}\right)^k w_t, \qquad k = 1, 2. \tag{9}$$

The first-order bias correction requires that $g_1(w) = 1$, whereas the second-order bias correction requires $g_1(w) = 1$ and $g_2(w) = 1$. First-order bias correction can for example be achieved with the simple weights $w_{1t} = n^{-1}t$, since in this case $g_1(w_1) = 1$, and thus an approximate unbiased (first-order) estimate of $E[\mathrm{PE}_n - \widehat{\mathrm{PE}}_n^{\mathrm{emp}}]$ is given by $C_n(w_1)$.

Example 4 *(First-order bias correction for AR processes).* Suppose an AR(p) model is fitted to the data. Let $\hat{\theta}_t = (\hat{\theta}_{t1}, \dots, \hat{\theta}_{tp})$ be the estimates of the autoregressive parameters on the basis of observations

$Y_1, \ldots, Y_t$. In this case, $\varepsilon_s(\hat{\theta}_t) = Y_s - \sum_{j=1}^{p} \hat{\theta}_{tj} Y_{s-j}$, $s \geq p+1$, $\widehat{\mathrm{PE}}_t^{\mathrm{emp}} = (t-p)^{-1} \sum_{s=p}^{t-1} \varepsilon_{s+1}^2(\hat{\theta}_t)$ and consequently the bias is estimated as

$$C_n(w_1) = \frac{1}{n(n-m)} \sum_{t=m}^{n-1} t\left(\varepsilon_{t+1}^2(\hat{\theta}_t) - \frac{1}{t-p} \sum_{s=p}^{t-1} \varepsilon_{s+1}(\hat{\theta}_t)^2 \right).$$

where $m = \lfloor \delta n \rfloor$ can be chosen roughly as a quarter of the sample size, that is, $\delta = .25$.

A reasonable way would be to select those weights that minimize the variance of the estimate of PE_n. Unfortunately, this may not be feasible in general since the estimates are complicated quantities and their variances depend on the unknown (conditional) mean function $\{\mu_t\}$. However, one may aim at a less ambitious criterion for weight selection utilizing the following lemma whose proof is given in Section 2 of the Supplementary Material.

LEMMA 1. *Let $\{Y_t\}$ be a stationary time series according to model* (1) *in an i.i.d. sequence $\{\varepsilon_t\}$. Let $\sum \psi_t(w)\varepsilon_{t+1}^2$ be the collection of the squares of ε_t terms in the expression for $\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w)$. Let g_1 and g_2 be as in* (9)*. Then the following statements hold.*

(a) The minimum of $\sum \psi_t(w)^2$ with respect to the sequence $\{w_t\}$ subject to the constraint $g_1(w) = 1$ is attained at $w_t = \varrho_2^{-1} t^{-1} n$, where $\rho_2 \approx 1/\delta$.

(b) The minimum of $\sum \psi_t(w)^2$ with respect to the sequence $\{w_t\}$ subject to the constraints $g_1(w) =$ and $g_2(w) = 1$ is given by $w_t = \lambda_1 t^{-1} n + \lambda_2 t^{-2} n^2$, where $\lambda_2 = (\varrho_2 \varrho_4 - \varrho_2^2)^{-1}(\varrho_4 - \varrho_3)$ and $\lambda_1 = \varrho_2^{-1}(1 - \varrho_3 \lambda_2)$, and $\rho_3 \approx (1+\delta)/(2\delta^2)$ and $\rho_4 \approx (1+\delta+\delta^2)/\delta^3$.

Note that

$$\widehat{\mathrm{PE}}_n^{\mathrm{ME}} = \sum_{t=0}^{n-1} \psi_t(w)\varepsilon_{t+1}^2 + R,$$

where the term R involves the model bias terms $b_{t+1} = \mu_{t+1} - \mu_{t+1}(\bar{\theta})$ and terms of the form $d_{t+1}(\hat{\theta}_t) = \mu_{t+1}(\hat{\theta}_t) - \mu_{t+1}(\bar{\theta})$, $d_{t+1}(\hat{\theta}_n)$, recalling that $\bar{\theta}$ is the minimizer of $E[\{\mu_t - \mu_t(\theta)\}^2]$. If the remainder term R is ignored and the variance of $\sum_{t=0}^{n-1} \psi_t(w)\varepsilon_{t+1}^2$, which is proportional to $\sum_{t=0}^{n-1} \psi_t(w)^2$, is minimized, then the selection of weights can be based on Lemma 1. The following remark summarizes the main findings.

Remark 1. (i) The results of Lemma 1 do not depend on the type of stationary time series considered. This means in particular that the weights selected through minimization of $\sum_{t=0}^{n-1} \psi_t(w)\varepsilon_{t+1}^2$ do not depend on the underlying data generating process.

Elimination of first-order bias. The minimization of the criterion function $\sum_{t=0}^{n-1} \psi_t(w)^2$ subject to the constraint $g_1(w) = 1$ leads to the weight sequence $\{w_{2t}\}$ given by $w_{2t} = \varrho_2^{-1} t^{-1} n$;

(ii) *Elimination of second-order bias.* The minimization of the criterion function $\sum_{t=0}^{n-1} \psi_t(w)^2$ subject to the constraints $g_1(w) = 1$ and $g_2(w) = 1$ leads to the weight sequence $\{w_{3t}\}$ given by $w_{3t} = \lambda_1 t^{-1} n + \lambda_2 t^{-2} n^2$, where λ_1 and λ_2 are constants specified in Lemma 1.

Simulations have not revealed much of a difference in the estimates using w_1 and w_2. It can be shown that the use of w_2 leads to a slightly inflated second-order bias term. The weights given in (ii) lead to the elimination of first-order and second-order bias terms, but this seems to come at the cost of higher variability of the estimate. These findings are in line with those in Efron (1982), who recommends against correcting second-order bias for independent data. We reiterate Efron's recommendation and recommend only a first-order bias correction when estimating the prediction error for time series. For completeness, we also report the second-order bias correction and linear combinations of weights aiming for first- and second-order bias correction. Their properties might be considered in future research. For example, it is reasonable to consider linear combinations of estimates of the form $\widehat{\mathrm{PE}}^{\mathrm{ME}}(w_1)$ and $\widehat{\mathrm{PE}}^{\mathrm{ME}}(w_2)$ or linear combinations of $\widehat{\mathrm{PE}}^{\mathrm{ME}}(w_1)$ and $\widehat{\mathrm{PE}}^{\mathrm{ME}}(w_3)$, so that the second-order bias term is reduced but not

completely eliminated. Here a linear combination of $\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_1)$ and $\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_3)$ is given. First note that

$$\mathrm{PE}_n - E\Big[\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_1)\Big] \approx (1-\varrho_1)\frac{\beta_6}{n^2},$$

and

$$\mathrm{PE}_n - E\Big[\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_3)\Big] = O\bigg(\frac{1}{n^3}\bigg).$$

Let therefore $w_{13} = \alpha w_1 + (1-\alpha)w_3$ for some $0 < \alpha < 1$. Hence the estimate of the form

$$\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_{13}) = \alpha\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_1) + (1-\alpha)\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_3),$$

with $0 < \alpha < 1$, has bias

$$\mathrm{PE}_n - E\Big[\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_{13})\Big] \approx \alpha(1-\varrho_1)\frac{\beta_6}{n^2}.$$

It should be pointed out that as α increases the bias is reduced but the variance becomes bigger. Based on simulations, we recommend the use of $\alpha = 0.3$ balances the trade-offs between bias and variance reasonably well. A theoretical study of the properties of these estimators of prediction error is, however, beyond the scope of the present paper.

3.2. *The modified Rissanen estimate*

The original Rissanen estimate uses simple averages of $\varepsilon_{t+1}^2(\hat{\theta}_t)$ to compute the prediction error, but one may instead consider weighted averages as well. These give rise to the modified Rissanen estimate

$$\widehat{\mathrm{PE}}_n^{\mathrm{MR}}(v) = \frac{1}{n-m}\sum_{t=m}^{n-1} v_t\varepsilon_{t+1}^2(\hat{\theta}_t), \tag{10}$$

where the weights $\{v_t\}$ average to 1. In order to answer the question of how to choose the weights, notice that

$$\begin{aligned} E\Big[\widehat{\mathrm{PE}}_n^{\mathrm{MR}}(v)\Big] &\approx \frac{1}{n-m}\sum_{t=m}^{n-1}\bigg(\sigma_\varepsilon^2 + \beta_0 + \frac{\beta_1}{t} + \frac{\beta_2}{t^2}\bigg)v_t \\ &= f_0(v)(\sigma_\varepsilon^2+\beta_0) + f_1(v)\frac{\beta_1}{n} + f_2(v)\frac{\beta_2}{n^2}, \end{aligned}$$

where

$$f_k(v) = \frac{1}{n-m}\sum_{t=m}^{n-1}\bigg(\frac{n}{t}\bigg)^k v_t, \qquad k = 0,1,2. \tag{11}$$

Approximate (first-order) unbiasedness requires that $f_0(v) = f_1(v) = 1$. There are many sequences $\{v_t\}$ that satisfy these conditions and this issue will be discussed further below. First, the focus will be briefly on the particular weights $v_{1t} = \lambda_0 + \lambda_1 n^{-1}t$. Some calculations show that first-order unbiasedness can be achieved by choosing

$$\lambda_1 = \frac{\varrho_1 - 1}{\varrho_{-1}\varrho_1 - 1}, \qquad \lambda_0 = 1 - \varrho_{-1}\lambda_1,$$

with constants ϱ_{-1} and ϱ_1. Simulations revealed that the modified Rissanen estimates tend to have small bias. However, they also seem to have higher variability than the corresponding modified empirical estimates $\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w)$; see Section 5. Turning to the general weight selection, the following lemma is the counterpart to Lemma 1. Its proof is given in the Supplementary Material.

LEMMA 2. *Let $\{Y_t\}$ be a stationary time series according to model (1) in an i.i.d. sequence $\{\varepsilon_t\}$. Let $f_0(v), f_1(v)$ and $f_2(v)$ be as in (11). Then the following statements hold.*

(a) The minimum of $\sum_{m=t}^{n-1} v_t^2$ with respect to $\{v_t\}$ subject to the constraints $f_0(v) = 1$ and $f_1(v) = 1$ is attained at $v_t = \lambda_0 + \lambda_1 t^{-1} n$, where $\lambda_1 = (\varrho_2 - \varrho_1^2)^{-1}(1 - \varrho_1)$ and $\lambda_0 = 1 - \varrho_1 \lambda_1$.

(b) The minimum of $\sum_{m=t}^{n-1} v_t^2$ with respect to $\{v_t\}$ subject to the constraints $f_0(v) = 1$, $f_1(v) = 1$ and $f_2(v) = 1$ is attained at $v_t = \lambda_0 + \lambda_1 t^{-1} n + \lambda_2 t^{-2} n^2$, where λ_0, λ_1 and λ_2 are solutions of the equation

$$\begin{pmatrix} 1 & \varrho_1 & \varrho_2 \\ \varrho_1 & \varrho_2 & \varrho_3 \\ \varrho_2 & \varrho_3 & \varrho_4 \end{pmatrix} \begin{pmatrix} \lambda_0 \\ \lambda_1 \\ \lambda_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

Similar arguments as in the previous section applied to the modified Rissanen estimate show that the leading term of $\widehat{\mathrm{PE}}_n^{\mathrm{MR}}(v)$ is equal to $(n-m)^{-1} \sum_{t=m}^{n-1} v_t \varepsilon_{t+1}^2$ and the variance of this term is proportional to $\sum_{t=m}^{n-1} v_t^2$. Let $f_0(v)$, $f_1(v)$ and $f_2(v)$ be as in display (11). The following remark summarizes the findings of Lemma 2.

Remark 2. (iv) The results of Lemma 2 do not depend on the type of stationary time series considered. This means in particular that the weights selected through minimization of $\sum_{t=0}^{n-1} v_t^2$ do not depend on the underlying data generating process.

(v) *Elimination of first-order bias:* The minimization of the criterion function $\sum_{t=m}^{n-1} v_t^2$ subject to the constraints $f_0(v) = 1$ and $f_1(v) = 1$ leads to the weight sequence $\{v_{2t}\}$ given by $v_{2t} = \lambda_0 + \lambda_1 t^{-1} n$, where λ_0 and λ_1 are constants specified in Lemma 2.

(vi) *Elimination of second-order bias:* The minimization of the criterion function $\sum_{t=m}^{n-1} v_t^2$ subject to the constraints $f_0(v) = 1$, $f_1(v) = 1$ and $f_2(v) = 1$ leads to the weight sequence $\{v_{3t}\}$ given by $v_{3t} = \lambda_0 + \lambda_1 t^{-1} n + \lambda_2 t^{-2} n^2$, where λ_0, λ_1 and λ_2 are constants specified in Lemma 2.

Simulations indicate that there are only minor differences between the estimates using v_1 and v_2. The estimate using v_3 gives almost unbiased estimates, but it comes with the cost of higher variability.

4. EXTENSION TO MULTIVARIATE SETTINGS AND MULTI-STEP PREDICTIONS

4.1. *The multivariate setting*

The model in (1) is now understood to be d-dimensional. While in the univariate case PE_n (and also its estimate) is a real number, it has to be replaced by a variance-covariance matrix for predicting the vector Y_{n+1} in the multivariate setting. If the model has been estimated on the basis of the entire available data $Y_1, \ldots, Y_n$, the prediction error matrix is given by

$$\mathrm{PE}_n = E[\varepsilon_{n+1}(\hat{\theta}_n)\varepsilon_{n+1}(\hat{\theta}_n)'].$$

Now following along the lines of the arguments introduced in Section 2.1 and supposing that $n - k$ residuals are available (e.g., VARMA fits), the empirical estimate of PE_n becomes

$$\widehat{\mathrm{PE}}_n^{\mathrm{emp}} = \frac{1}{n-k} \sum_{t=k}^{n-1} \varepsilon_{t+1}(\hat{\theta}_n)\varepsilon_{t+1}(\hat{\theta}_n)'.$$

In order to introduce Rissanen's APE estimate in d dimensions, write similarly, if the model is estimated on the basis of the first s observations, that $\mathrm{PE}_s = E[\varepsilon_{s+1}(\hat{\theta}_s)\varepsilon_{s+1}(\hat{\theta}_s)']$ and $\widehat{\mathrm{PE}}_s^{\mathrm{emp}} = (s-k)^{-1} \sum_{t=k+1}^{s} \varepsilon_t(\hat{\theta}_s)\varepsilon_t(\hat{\theta}_s)'$. With $m = \lfloor \delta n \rfloor$, $0 < \delta < 1$, Rissanen's APE estimate of PE_n is then given by

$$\widehat{\mathrm{PE}}_n^{\mathrm{R}} = \frac{1}{n-m} \sum_{t=m}^{n-1} \varepsilon_{t+1}(\hat{\theta}_t)\varepsilon_{t+1}(\hat{\theta}_t)'.$$

If one wishes to obtain a numerical measure of the prediction error in the multivariate case, then one may use the trace of $\mathrm{PE}_n H$ (that is,$E[\hat{\varepsilon}'_{n+1|n} H \hat{\varepsilon}_{n+1|n}]$) when H is some $d \times d$ positive definite matrix. Choices for H include the inverse of the variance-covariance matrix V_ε of ε_t and the diagonal matrix $\mathrm{diag}(V_\varepsilon)^{-1}$. One may also take H to be the identity matrix if all the component series are equally scaled, and the determinant of the matrix PE_n as a numerical measure. The performance of $\widehat{\mathrm{PE}}_n^{\mathrm{emp}}$ and $\widehat{\mathrm{PE}}_n^{\mathrm{R}}$ is worse when a numerical measure is taken as can be seen from the simulation results in Section 5.

The modification of $\widehat{\mathrm{PE}}_n^{\mathrm{emp}}$ and $\widehat{\mathrm{PE}}_n^{\mathrm{R}}$ to $\widehat{\mathrm{PE}}_n^{\mathrm{ME}}$ and $\widehat{\mathrm{PE}}_n^{\mathrm{MR}}$ in the multivariate case is exactly the same as in the univariate case except that, for any residual vector e, one needs to write ee' instead of e^2. The details are therefore skipped to conserve space.

4.2. *Multiple-step predictions*

It is often of interest to analyze the performance of longer-term predictions, say, h-steps ahead. Observe that the optimal forecast of Y_{n+h} on the basis of observations Y_s, $s \le n$, is given by the conditional mean $\mu_{n+h}^{(h)} = E[Y_{n+h}|Y_s,\, s \le n]$. Let $\hat{\mu}_t^{(h)}(\hat{\theta}_s)$ be the estimated value of $\mu_t^{(h)}$ when the model has been fitted on the basis of the observations $Y_1, \ldots, Y_s$ and denote the h-step ahead residual $Y_t - \hat{\mu}_t^{(h)}(\hat{\theta}_s)$ by $\varepsilon_t^{(h)}(\hat{\theta}_s)$. The prediction variance-covariance matrix, its empirical and Rissanen estimates are then

$$\begin{aligned}
\mathrm{PE}_n(h) &= E\Big[\varepsilon_{n+h}^{(h)}(\hat{\theta}_n)\varepsilon_{n+h}^{(h)}(\hat{\theta}_n)'\Big],\\
\widehat{\mathrm{PE}}_n^{\mathrm{emp}}(h) &= \frac{1}{n-h-k+1}\sum_{t=k}^{n-h} \varepsilon_{t+h}^{(h)}(\hat{\theta}_n)\varepsilon_{t+h}^{(h)}(\hat{\theta}_n)',\\
\widehat{\mathrm{PE}}_n^{\mathrm{R}}(h) &= \frac{1}{n-h-m+1}\sum_{t=m}^{n-h} \varepsilon_{t+h}^{(h)}(\hat{\theta}_t)\varepsilon_{t+h}^{(h)}(\hat{\theta}_t)',
\end{aligned}$$

where the formulation is given in the multivariate form. All discussions and issues addressed in this work are valid also for h-step ahead predictions and do not require any special treatment. Arguments for obtaining modified estimates of the h-step ahead prediction error matrix $\mathrm{PE}_n(h)$ mirror those used for the one-step ahead case and are briefly discussed here.

The modified h-step ahead empirical estimate is defined as

$$\widehat{\mathrm{PE}}_n^{ME}(w;h) = \widehat{\mathrm{PE}}_n^{\mathrm{emp}}(h) + C_n(w;h),$$

where the form of the correction term is similar to the one-step ahead case. The correction term has the form

$$C_n(w;h) = \frac{1}{n-h-m+1}\sum_{t=k}^{n-h} w_t\Big(\hat{\varepsilon}_{t+h}(\hat{\theta}_t)^2 - \widehat{\mathrm{PE}}_t^{\mathrm{emp}}(h)\Big),$$

where the weights $\{w_t\}$ satisfy the condition $g_1(w;h) = (n-h-k+1)^{-1}\sum_{t=k}^{n-h} t^{-1} n w_t = 1$ in order to correct for first-order bias. To correct also for second-order bias, the weights must additionally satisfy $g_2(w;h) = (n-h-k+1)^{-1}\sum_{t=k}^{n-h} t^{-2} n^2 w_t = 1$. The rest of the discussion in Section **??** on choosing different weights applies here, too.

The modified h-step ahead Rissanen estimate is defined as

$$\widehat{\mathrm{PE}}_n^{\mathrm{MR}}(h) = \frac{1}{n-h-m+1}\sum_{t=m}^{n-h} v_t e_{t+h}^{(h)}(t) e_{t+h}^{(h)}(t)',$$

with weights $\{v_t\}$ satisfying $f_0(v) = 1$ and $f_1(v) = 1$ to achieve first-order bias correction and $f_0(v) = 1$, $f_1(v) = 1$ and $f_2(v) = 1$ to achieve second-order bias correction, where the functions $f_\ell(v)$, $\ell = 0, 1, 2$, are defined in display (11). All the discussion on choosing weights are similar to the one-step ahead case.

5. NUMERICAL RESULTS

Simulations were carried out in order to investigate the performances of Rissanen's APE and the proposed biased-corrected estimators for one-step prediction errors. Suppose that observations $Y_1, \ldots, Y_n$ are obtained from a d-dimensional stationary time series following (1). As numerical measure for the variance-covariance matrix the trace

$$\mathrm{tr}(\mathrm{PE}_n V_\varepsilon^{-1}) = E[e_{n+1}(n)' V_\varepsilon^{-1} e_{n+1}(n)]$$

is used, where V_ε is the variance-covariance matrix of the centered innovation sequence $\{\varepsilon_t\}$. It should be noted that $\mathrm{tr}(\mathrm{PE}_n V_\varepsilon^{-1}) = d + D^2$, where $D^2 = E[(\hat{\mu}_{n+1}(n) - \mu_{n+1})' V_\varepsilon^{-1} (\hat{\mu}_{n+1}(n) - \mu_{n+1})]$. For any generic estimate $\widehat{\mathrm{PE}}_n$ of PE_n, mean and standard deviation of $\widehat{\mathrm{PE}}_n - \mathrm{PE}_n$ were obtained. More specifically, they were computed based on the following estimates.

- *Rissanen's APE:* $\widehat{\mathrm{PE}}_n^{\mathrm{R}} - \mathrm{PE}_n$;
- *Modified empirical estimates:* $\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w) - \mathrm{PE}_n$ with three choices of weights: (i) $w_{1t} = n^{-1}t$, (ii) w_{3t} as defined in Section 3.1, and (iii) the linear combination $w_{13,t} = 0.3w_{1t} + 0.7w_{3t}$;
- *Modified Rissanen estimate:* $\widehat{\mathrm{PE}}_n^{\mathrm{MR}}(v_1) - \mathrm{PE}_n$, where $v_{1t} = \lambda_0 + \lambda_1 n^{-1}t$ as given in Section 3.2.

All results presented here have been based on 10,000 simulation runs. However, PE_n itself was estimated based on 25,000 simulation runs to achieve greater accuracy.

The following time series models were considered.

Univariate setting: For $d = 1$, the ARMA(1,1) model

$$Y_t = aY_{t-1} + \varepsilon_t + b\varepsilon_{t-1}, \qquad t = 1, \ldots, n,$$

was simulated with $\{\varepsilon_t\}$ i.i.d. $N(0,1)$ innovations and $n = 100$. In each case, the fitted models were AR(p) using Yule-Walker estimation with $p = 2, 12$. As in Ing (2007), various combinations of parameters (a, b) were considered, namely $(a, b) = (0.5, 0.5)$ (model 1), $(0.95, 0.5)$ (model 2) and $(0.5, 0.95)$ (model 3). Model 1 is rather simple, since the dependence dies rather rapidly, but the other two are extreme cases. In model 2, the AR root is close to the unit circle and thus it is close to non-stationarity. For this case, the variance part of PE_n can often be large. The third case is close to non-invertibilty of the MA part and predicting future values requires a large number of past observations. Therefore the bias may not be small when fitting an AR(p) model.

Multivariate setting: For $d = 4$ four independent copies of the univariate ARMA(1,1) model were simulated with $n = 100$ and VAR(p) models of order $p = 2, 7$ were fitted to the data. In practice one would perhaps not use a VAR(7) model for this four-dimensional time series, since there are other methods available for such cases; for example, reduced rank regression. The purpose here, however, is to investigate how easy or difficult it is to estimate the prediction error and how the proposed estimates perform under different modeling conditions.

The simulation results are summarized in Figures 1–4 and also given in Tables 1–4. The results show that APE is not a good estimate of PE_n except when δ is close to one, which is expected. All the proposed estimates seem to have small biases for any δ. Among the four proposed estimates, the bias of $\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_1)$ appears to be the highest. As expected, the standard deviation of all estimates increases with δ. It should also be noted that $\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_1)$ (followed by APE) seems to have the smallest standard deviation for any δ. The variability of $\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_3)$ and $\widehat{\mathrm{PE}}_n^{\mathrm{MR}}(v_1)$ are higher than the others. The hybrid estimate $\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_3)$ appears to have a lower bias than $\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_1)$, but it has a higher variability. For the four-dimensional case, the inadequacy of APE as an estimate of the prediction error becomes evident. The performances of $\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_1)$ and $\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_3)$ are the most satisfactory of the proposed estimates.

The simulation results seem to indicate that one should operate with a small δ such as 0.3 and use either $\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_1)$ or $\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_3)$ as an estimate of PE_n. Overall, the proposed estimates offer a significant improvement over Rissanen's classical APE method.

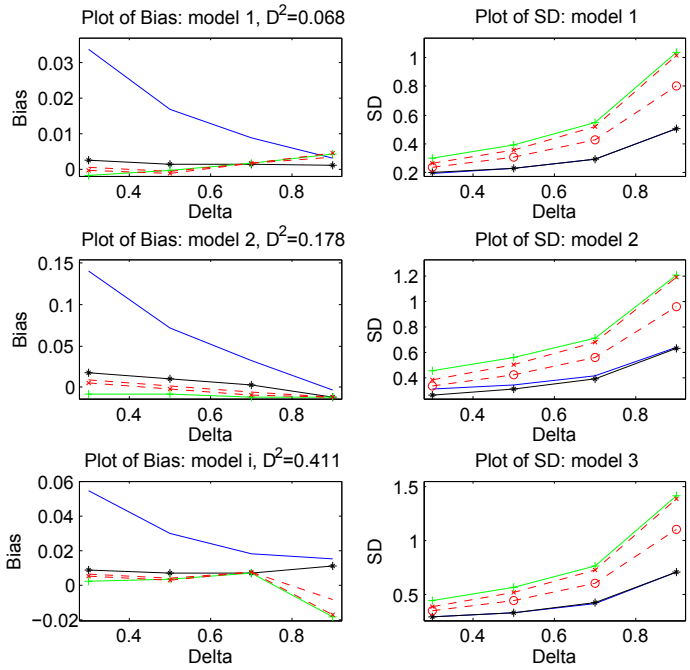

Fig. 1. Bias (left) and standard deviation (right) of estimates of PE_n for the univariate ARMA(1,1) data generating process $Y_t = aY_{t-1} + \varepsilon_t + b\varepsilon_{t-1}$ and fitted AR(2) processes: Rissanen's APE $\widehat{\mathrm{PE}}_n^{\mathrm{R}}$ (——); modified empirical estimates $\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_1)$ (–∗–∗–), $\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_{13})$ (–○–○–) and $\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_3)$ (–×–×–); modified Rissanen estimate $\widehat{\mathrm{PE}}_n^{\mathrm{MR}}(v_1)$ (–+–+–).

6. DISCUSSION

Two novel methods are introduced to tackle the difficult problem of estimating the (h-step ahead) prediction error of a stationary time series, which may be univariate or multivariate, linear or nonlinear and parametric or nonparametric. The model in (1) allows explicitly for the case of model misspecification often encountered in practice. The proposed method utilizes a bias-variance decomposition of the prediction error to suggest correction terms to modify an empirical estimate and Rissanen's classical APE, including guidelines for weight selection. Two examples, linear causal and invertible ARMA processes and nonlinear AR processes, highlight the usefulness of the methodology. Simulations confirm bias reductions of about an order of magnitude compared to APE and, for certain weight choices, an overall reduction in prediction error.

REFERENCES

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723.

Box, G. E. P., Jenkins, G. M. and Reinsel, G. C. (2008) *Time Series Analysis: Forecasting and Control (4th ed.)*. New York: Wiley.

Burman, P. (1989) A comparative study of ordinary cross-validation, v-fold cross-validation and repeated learning-testing methods. *Biometrika*, **76**, 503–514.

— (1990) Estimation of generalized additive models. *Journal of Multivariate Analysis*, **32**, 230–255.

Burman, P., Chow, E. and Nolan, D. (1994) A cross validatory method for dependent data. *Biometrika*, **81**, 351–358.

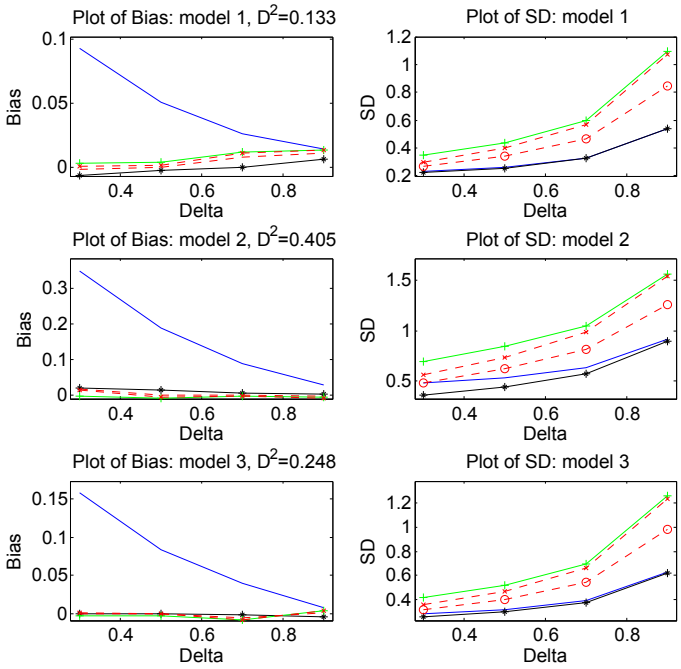

Fig. 2. Same as in Figure 1 but with fitted AR(12) processes.

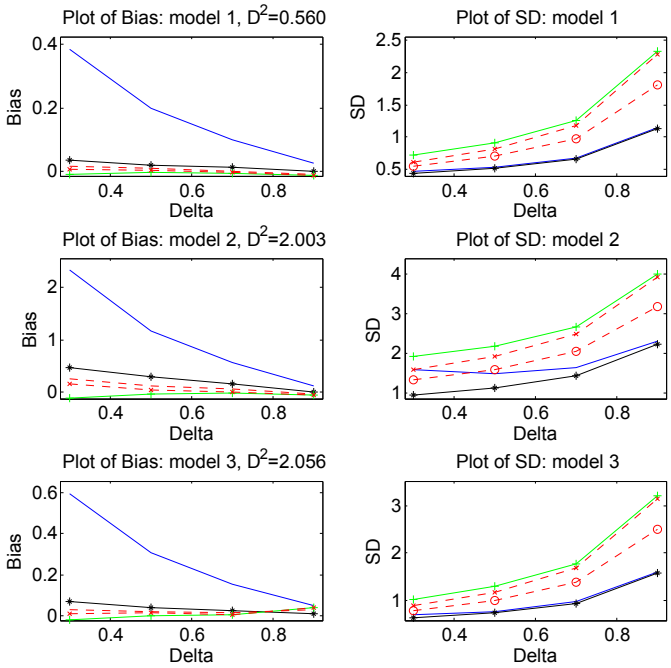

Fig. 3. Same as in Figure 1 but for the multivariate VARMA(1,1) model with fitted VAR(2) processes.

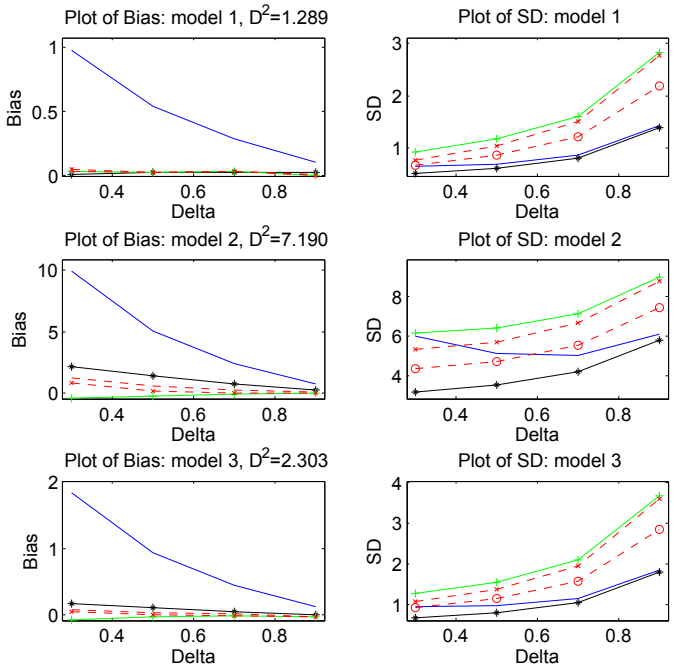

Fig. 4. Same as in Figure 1 but for the multivariate VARMA(1,1) model with fitted VAR(7) processes.

Burnham, K. and Anderson, D. R. (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach (2nd ed)*. New York: Springer-Verlag.

Chen, R. and Liu, L.-M. (2001) Functional coefficient autoregressive models: Estimation and tests of hypotheses. *Journal of Time Series Analysis*, **22**, 151–160.

Claeskens, G. and Hjort, N. (2008) *Model Selection and Model Averaging*. Cambridge University Press.

Efron, B. (1982) *The Jacknife, the Bootstrap and Other Resampling Plans*. Philadelphia: SIAM.

— (2004) The estimation of prediction error: Covariance penalties and cross-validation. *Journal of the American Statistical Association*, **99**, 619–642.

Fan, J. and Yao, Q. (2005) *Nonlinear Time Series: Nonparametric and Parametric Methods*. New York: Springer-Verlag.

Findley, D. F. (2005) Asymptotic second moment properties of out-of-sample forecast errors of misspecified regarima models and the optimality of gls. *Statistica Sinica*, **15**, 447–476.

Gerencser, L. (1992) Ar(∞) estimation and nonparametric stochastic complexity. *IEEE Transactions on Information Theory*, **38**, 1768–1778.

H. Wang, G. L. and Tsai, C.-L. (2007) Regression coefficient and autoregressive order shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, **69**, 63–78.

Hannan, E. J. and Quinn, B. G. (1979) The determination of the order of an autoregression. *Journal of the Royal Statistical Society, Series B*, **41**, 190–195.

Hart, J. D. and Lee, C.-L. (2005) Robustness of one-sided cross-validation to autocorrelation. *Journal of Multivariate Analysis*, **92**, 77–96.

Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models*. Boca Raton, FL: Chapman and Hall.

Hemerly, E. M. and Davis, M. H. A. (1989) Strong consistency of the PLS criterion for order determination of autoregressive processes. *The Annals of Statistics*, **17**, 941–946.

Hurvich, C. and Tsai, C.-L. (1989) Regression and time series model selection in small samples. *Biometrika*, **76**, 297–307.

Ing, C.-K. (2007) Accumulated prediction errors, information criteria and optimal forecasting for autoregressive time series. *The Annals of Statistics*, **36**, 1239–1277.

Jiang, J. (2014) Multivariate functional-coefficient regression models for nonlinear vector time series data. *Biometrika*, **101**, 689–702.

Lewis, P. A. W. and Stevens, J. G. (1991) Nonlinear modeling of time series using multivariate adaptive regression splines (mars). *Journal of the American Statistical Association*, **86**, 864–877.

(a, b)	estimate	$\delta = .3$	.5	.7	.9
	$\widehat{\mathrm{PE}}_n^{\mathrm{R}}$	0.033	0.017	0.009	0.003
		(0.20)	(0.23)	(0.29)	(0.50)
	$\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_1)$	0.003	0.001	0.002	0.001
		(0.20)	(0.23)	(0.29)	(0.50)
(.5,.5)	$\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_{13})$	0.001	−0.000	0.002	0.004
		(0.24)	(0.31)	(0.43)	(0.80)
	$\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_3)$	−0.000	−0.001	0.002	0.005
		(0.26)	(0.36)	(0.52)	(1.01)
	$\widehat{\mathrm{PE}}_n^{\mathrm{MR}}(v_1)$	−0.002	−0.000	0.002	0.004
		(0.30)	(0.39)	(0.55)	(1.03)
	$\widehat{\mathrm{PE}}_n^{\mathrm{R}}$	0.141	0.072	0.032	−0.003
		(0.31)	(0.35)	(0.42)	(0.64)
	$\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_1)$	0.017	0.010	0.003	−0.011
		(0.26)	(0.31)	(0.40)	(0.63)
(.95,.5)	$\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_{13})$	0.009	0.002	−0.006	−0.012
		(0.34)	(0.43)	(0.56)	(0.96)
	$\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_3)$	0.005	−0.002	−0.009	−0.012
		(0.38)	(0.50)	(0.68)	(1.19)
	$\widehat{\mathrm{PE}}_n^{\mathrm{MR}}(v_1)$	−0.008	−0.008	−0.011	−0.013
		(0.46)	(0.56)	(0.72)	(1.21)
	$\widehat{\mathrm{PE}}_n^{\mathrm{R}}$	0.055	0.030	0.018	0.015
		(0.29)	(0.33)	(0.42)	(0.71)
	$\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_1)$	0.009	0.007	0.007	0.011
		(0.29)	(0.33)	(0.42)	(0.71)
(.5,.95)	$\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_{13})$	0.007	0.004	0.007	−0.008
		(0.35)	(0.44)	(0.60)	(1.10)
	$\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_3)$	0.005	0.003	0.008	−0.017
		(0.38)	(0.51)	(0.73)	(1.39)
	$\widehat{\mathrm{PE}}_n^{\mathrm{MR}}(v_1)$	0.002	0.004	0.007	−0.018
		(0.44)	(0.56)	(0.77)	(1.41)

Table 1. *Bias (variance) of estimates of* $\widehat{\mathrm{PE}}_n - \mathrm{PE}_n$ *for the univariate ARMA(1,1) data generating process* $Y_t = aY_{t-1} + \varepsilon_t + b\varepsilon_{t-1}$ *and fitted AR(2) processes.*

Priestley, M. B. (1980) State-dependent models: A general approach to non-linear time series analysis. *Journal of Time Series Analysis*, **1**, 47–71.

Racine, J. (2000) Consistent cross-validatory model-selection for dependent data: hv-block cross-validation. *Journal of Econometrics*, **99**, 39–61.

Rao, C. R. (1973) *Linear Statistical Inference and Its Applications (3rd ed.)*. New York: Wiley.

Rissanen, J. (1984) Universal coding, information, prediction, and estimation. *IEEE Transactions on Information Theory*, **30**, 629–636.

— (1986) Order estimation by accumulated prediction errors. *Journal of Applied Probability*, **23A**, 55–61.

Schwartz, G. (1978) Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.

Shumway, R. H. and Stoffer, D. S. (2010) *Time Series Analysis and Its Applications: With R Examples (3rd ed.)*. New York: Springer-Verlag.

(a,b)	estimate	$\delta = .3$	.5	.7	.9
	$\widehat{\mathrm{PE}}_n^{\mathrm{R}}$	0.092	0.051	0.027	0.014
		(0.23)	(0.26)	(0.33)	(0.54)
	$\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_1)$	-0.006	-0.002	0.000	0.006
		(0.22)	(0.26)	(0.32)	(0.54)
(.5,.5)	$\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_{13})$	-0.001	0.001	0.008	0.011
		(0.27)	(0.34)	(0.47)	(0.85)
	$\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_3)$	0.001	0.002	0.011	0.013
		(0.30)	(0.40)	(0.57)	(1.07)
	$\widehat{\mathrm{PE}}_n^{\mathrm{MR}}(v_1)$	0.004	0.004	0.012	0.014
		(0.35)	(0.44)	(0.60)	(1.09)
	$\widehat{\mathrm{PE}}_n^{\mathrm{R}}$	0.349	0.188	0.089	0.030
		(0.48)	(0.53)	(0.63)	(0.91)
	$\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_1)$	0.020	0.014	0.007	0.005
		(0.36)	(0.44)	(0.57)	(0.89)
(.95,.5)	$\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_{13})$	0.017	0.001	0.001	-0.003
		(0.49)	(0.62)	(0.82)	(1.26)
	$\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_3)$	0.016	-0.004	-0.001	-0.006
		(0.56)	(0.74)	(0.99)	(1.54)
	$\widehat{\mathrm{PE}}_n^{\mathrm{MR}}(v_1)$	-0.001	-0.008	-0.001	-0.005
		(0.70)	(0.84)	(1.05)	(1.57)
	$\widehat{\mathrm{PE}}_n^{\mathrm{R}}$	0.158	0.084	0.040	0.008
		(0.28)	(0.32)	(0.39)	(0.63)
	$\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_1)$	0.000	-0.000	-0.002	-0.004
		(0.26)	(0.30)	(0.38)	(0.62)
(.5,.95)	$\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_{13})$	0.000	-0.000	-0.006	0.001
		(0.32)	(0.40)	(0.54)	(0.98)
	$\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_3)$	0.001	0.007	-0.008	0.004
		(0.36)	(0.47)	(0.66)	(1.23)
	$\widehat{\mathrm{PE}}_n^{\mathrm{MR}}(v_1)$	-0.002	-0.003	-0.008	0.004
		(0.42)	(0.52)	(0.70)	(1.25)

Table 2. *Same as Table 1 but with fitted AR(12) processes.*

Speed, T. and Yu, B. (1993) Model selection and prediction: Normal regression. *The Annals of the Institute of Statistical Mathematics*, **45**, 35–54.

Stone, C. (1986) Additive regression and other nonparametric models. *The Annals of Statistics*, **13**, 689–705.

Taniguchi, M. (1991) *Higher Order Asymptotic Theory for Time Series Analysis*. New York: Springer-Verlag.

Tong, H. (1983) *Threshold Models in Non-linear Time Series Analysis*. New York: Springer-Verlag.

Wei, C. Z. (1992) On predictive least squares principles. *The Annals of Statistics*, **20**, 1–42.

Xia, Y., Tong, H., Li, W. K. and Zhu, L. (2002) An adaptive estimation of dimension reduction space (with discussion). *Journal of the Royal Statistical Society, Series B*, **64**, 363–410.

(a,b)	estimate	$\delta = .3$	.5	.7	.9
	$\widehat{\mathrm{PE}}_n^{\mathrm{R}}$	0.39	0.20	0.10	0.03
		(0.47)	(0.54)	(0.68)	(1.15)
	$\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_1)$	0.04	0.02	0.02	0.00
		(0.45)	(0.52)	(0.67)	(1.14)
(.5,.5)	$\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_{13})$	0.02	0.01	0.01	-0.01
		(0.55)	(0.71)	(0.97)	(1.80)
	$\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_3)$	0.01	0.01	0.00	-0.01
		(0.61)	(0.82)	(1.18)	(2.28)
	$\widehat{\mathrm{PE}}_n^{\mathrm{MR}}(v_1)$	-0.01	0.00	-0.00	-0.01
		(0.72)	(0.91)	(1.25)	(2.32)
	$\widehat{\mathrm{PE}}_n^{\mathrm{R}}$	2.33	1.18	0.56	0.12
		(1.61)	(1.51)	(1.66)	(2.32)
	$\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_1)$	0.46	0.29	0.15	0.01
		(0.97)	(1.14)	(1.44)	(2.24)
(.95,.5)	$\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_{13})$	0.24	0.11	0.05	-0.05
		(1.34)	(1.59)	(2.05)	(3.19)
	$\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_3)$	0.15	0.04	0.00	-0.07
		(1.61)	(1.92)	(2.50)	(3.91)
	$\widehat{\mathrm{PE}}_n^{\mathrm{MR}}(v_1)$	-0.12	-0.05	-0.03	-0.07
		(1.93)	(2.18)	(2.67)	(3.99)
	$\widehat{\mathrm{PE}}_n^{\mathrm{R}}$	0.60	0.31	0.15	0.05
		(0.68)	(0.76)	(0.96)	(1.60)
	$\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_1)$	0.07	0.04	0.02	0.01
		(0.63)	(0.73)	(0.94)	(1.58)
(.5,.95)	$\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_{13})$	0.03	0.02	0.01	0.03
		(0.78)	(1.00)	(1.38)	(2.50)
	$\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_3)$	0.01	0.01	0.01	0.04
		(0.88)	(1.17)	(1.68)	(3.15)
	$\widehat{\mathrm{PE}}_n^{\mathrm{MR}}(v_1)$	-0.02	0.00	0.00	0.04
		(1.02)	(1.29)	(1.77)	(3.21)

Table 3. *Same as in Table 1 but for the VARMA(1,1) model with fitted VAR(2) processes.*

(a,b)	estimate	$\delta = .3$	.5	.7	.9
	$\widehat{\mathrm{PE}}_n^{\mathrm{R}}$	0.98	0.54	0.28	0.10
		(0.66)	(0.70)	(0.87)	(1.43)
	$\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_1)$	0.00	0.02	0.03	0.03
		(0.52)	(0.62)	(0.81)	(1.40)
(.5,.5)	$\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_{13})$	0.03	0.02	0.03	0.01
		(0.67)	(0.87)	(1.22)	(2.19)
	$\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_3)$	0.05	0.02	0.03	-0.00
		(0.77)	(1.04)	(1.50)	(2.77)
	$\widehat{\mathrm{PE}}_n^{\mathrm{MR}}(v_1)$	0.03	0.02	0.03	-0.00
		(0.93)	(1.17)	(1.60)	(2.82)
	$\widehat{\mathrm{PE}}_n^{\mathrm{R}}$	9.9	5.07	2.37	0.71
		1 (6.01)	(5.14)	(5.04)	(6.10)
	$\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_1)$	2.16	1.40	0.75	0.26
		(3.19)	(3.57)	(4.20)	(5.79)
(.95,.5)	$\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_{13})$	1.25	0.53	0.24	0.07
		(4.39)	(4.71)	(5.53)	(7.42)
	$\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_3)$	0.86	0.16	0.03	-0.02
		(5.32)	(5.69)	(6.65)	(8.77)
	$\widehat{\mathrm{PE}}_n^{\mathrm{MR}}(v_1)$	-0.42	-0.24	-0.07	-0.03
		(6.16)	(6.40)	(7.13)	(8.95)
	$\widehat{\mathrm{PE}}_n^{\mathrm{R}}$	1.83	0.94	0.45	0.13
		(0.95)	(0.98)	(1.16)	(1.87)
	$\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_1)$	0.17	0.11	0.06	0.01
		(0.69)	(0.82)	(1.06)	(1.82)
(.5,.95)	$\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_{13})$	0.09	0.04	0.01	-0.02
		(0.93)	(1.16)	(1.59)	(2.85)
	$\widehat{\mathrm{PE}}_n^{\mathrm{ME}}(w_3)$	0.05	0.01	-0.00	-0.03
		(1.10)	(1.39)	(1.96)	(3.59)
	$\widehat{\mathrm{PE}}_n^{\mathrm{MR}}(v_1)$	-0.07	-0.03	-0.01	-0.03
		(1.28)	(1.56)	(2.10)	(3.67)

Table 4. *Same as in Table 1 but for the VARMA(1,1) model with fitted VAR(7) processes.*