DISCUSSION PAPER SERIES

No. 3809

ON THE SELECTION OF FORECASTING MODELS

Atsushi Inoue and Lutz Kilian

INTERNATIONAL MACROECONOMICS

Centre for Economic Policy Research

www.cepr.org

Available online at: **www.cepr.org/pubs/dps/DP3809.asp**

ISSN 0265-8003

ON THE SELECTION OF FORECASTING MODELS

Atsushi Inoue, North Carolina State University
Lutz Kilian, European Central Bank (ECB), University of Michigan and CEPR

Discussion Paper No. 3809
March 2003

Centre for Economic Policy Research
90–98 Goswell Rd, London EC1V 7RR, UK
Tel: (44 20) 7878 2900, Fax: (44 20) 7878 2999
Email: cepr@cepr.org, Website: www.cepr.org

CEPR Discussion Paper No. 3809

March 2003

ABSTRACT

On the Selection of Forecasting Models*

It is standard in applied work to select forecasting models by ranking candidate models by their prediction mean squared error (PMSE) in simulated out-of-sample (SOOS) forecasts. Alternatively, forecast models may be selected using information criteria (IC). We compare the asymptotic and finite-sample properties of these methods in terms of their ability to mimimize the true out-of-sample PMSE, allowing for possible misspecification of the forecast models under consideration. We first study a covariance stationary environment. We show that under suitable conditions the IC method will be consistent for the best approximating model among the candidate models. In contrast, under standard assumptions the SOOS method will select over-parameterized models with positive probability, resulting in excessive finite-sample PMSEs. We also show that in the presence of unmodelled structural change both methods will be inadmissible in the sense that they may select a model with strictly higher PMSE than the best approximating model among the candidate models.

JEL Classification: C22, C52 and C53
Keywords: forecast accuracy, information criteria, model selection, simulated out-of-sample method and structural change

Atsushi Inoue
Dept of Agricultural &Resource Economics
Box 8109
North Carolina State University
Raleigh, NC 27695 8109
USA
Email: atsushi@unity.ncsu.edu

Lutz Kilian
European Central Bank
Eurotower
Kaiserstr 29
D-60311 Frankfurt a.M
GERMANY
Tel: (11 69) 1344 8668
Fax: (11 69) 1344 8553
Email: lutz.kilian@ecb.int

*We thank seminar participants at the European Central Bank, the EUI Florence, and the 2003 Econometric Society Meeting in Washington, DC, and especially Todd Clark, Valentina Corradi, Kirstin Hubrich, Michael McCracken, Helmut Lütkepohl, Barbara Rossi, Jim Stock, and Norman Swanson for helpful discussions. The views expressed in this Paper do not necessarily reflect the opinion of the ECB or its staff.

Submitted 06 February 2003

1 Introduction

We are interested in generating predictions for a scalar variable y_t. We have in mind a number of candidate models involving alternative sets of predictors. Data are available for period $1, ..., T$. The problem is how to select among the candidate models the forecast model that will generate the most accurate forecast for period $T + 1$ in terms of the prediction mean squared error (PMSE).[1] Although other metrics could be considered such as the ability of models to forecast signs or turning points, the use of PMSEs is by far the most common evaluation criterion in macroeconometrics and we will follow that tradition.

Forecast model selection problems such as this one arise frequently in macroeconometrics and finance. For example, we may be interested in forecasting exchange returns based on alternative sets of macroeconomic fundamentals or we may be interested in forecasting inflation based on alternative formulations of the Phillips curve. Similar issues arise in lag order selection for autoregressive models or in the selection of the number of factors for dynamic factor models.

In studying the theoretical properties of forecast model selection methods, it will be useful to distinguish between nested and nonnested forecast model comparisons. An example of a nonnested model comparison would be the choice between alternative specifications of the Phillips curve relationship based on alternative measures of the output gap. Specifically, we may be interested in choosing between two regression models designed to forecast changes in inflation ($\Delta\pi_t$):

$$\begin{aligned} \widehat{\Delta\pi_{t+1}} &= \widehat{\alpha}\Delta\pi_t + \widehat{\beta}gap_t \\ \widehat{\Delta\pi_{t+1}} &= \widehat{\alpha}\Delta\pi_t + \widehat{\beta}u_t \end{aligned}$$

where gap_t and u_t refer to detrended output and the demeaned unemployment rate, respectively. An example of a nested forecast model comparison would be the comparison of a random walk model for the exchange rate (e_t) with a regression model based on deviations of the exchange rate from economic fundamentals ($e_t - f_t$):

$$\begin{aligned} \widehat{\Delta e_{t+1}} &= \widehat{\alpha} \\ \widehat{\Delta e_{t+1}} &= \widehat{\alpha} + \widehat{\beta}(e_t - f_t) \end{aligned}$$

Here the random walk model is nested in the model based on economic fundamentals. In applied work, it is common to consider several forecast models, involving typically some nested and some nonnested pairs of models. Our analysis will cover all these cases.

The tools used by practitioners for selecting among competing models are the same whether the forecasting models are nested or not. It is standard practice in applied work to select the best forecasting model by ranking candidate models, $i = 1, ..., M$, by their root PMSE in

[1] Our explicit objective is to select one forecast model among the candidate models. An alternative approach would be to search for optimal combinations of forecasts from several models. We do not address this problem in this paper. See Swanson and Zeng (2001) for a review of the literature on how to combine forecasts from competing models.

simulated recursive out-of-sample forecasts (see e.g., Meese and Rogoff 1983; Stock 1999; Stock and Watson 1999a,b, 2000). This simulated out-of-sample (SOOS) procedure involves fitting each candidate model on the first S observations and evaluating the mean squared error of the forecast of observation $S+1$, for $S = R, R+1, R+2, ..., T-1$.[2]

Although the SOOS methodology is most common in practice, alternatively, forecast models may be selected by evaluating their in-sample PMSE subject to a penalty term. This approach has been used for example by Bossaerts and Hillion (1999) to select among alternative predictors of stock returns. The penalty term is necessary because minimizing in-sample PMSEs alone simply amounts to maximizing R^2 and leads to overfitting and poor predictive accuracy when comparing models of different dimension. This idea is the guiding principle underlying the use of information criteria (IC). Forecast model selection based on information criteria involves estimating all M candidate models on the sample period $1, ..., T$ and selecting the forecast model that minimizes a suitably chosen criterion function, denoted by $IC(.)$. When we are interested in one-step ahead forecasts, this function takes the form:

$$IC(i) = \ln(\widehat{\sigma}_i^2) + n_i * c_T, \tag{1}$$

where n_i is the number of parameters used in the estimation of forecast model i, $i = 1, ..., M$, and $\widehat{\sigma}_i^2 = SSR_i/T$. The term SSR_i denotes the sum of the squared residuals of the forecast model i evaluated over the sample period $1, ..., T$. The term c_T is a function of the sample size T, subject to suitable restrictions. We select the forecast model that yields the smallest criterion value. When all candidate models involve the same number of regressors, this procedure amounts to ranking forecast models by their in-sample fit, measured by $\ln(\widehat{\sigma}_i^2)$. Since the term $n_i * c_T$ is increasing in the number of parameters, models with additional regressors receive a penalty to avoid overfitting.

In this paper, we compare the asymptotic and finite-sample properties of these two forecast model selection methods. For simplicity attention is restricted to one-step ahead forecasts. The focus is on linear forecast models. It is assumed that the regressand and regressors have been transformed to make them stationary. Extensions to nearly integrated environments or nonlinear forecast models are not covered in this paper. Our theory allows for some or all forecast models under consideration to be misspecified due to omitted variables. It even allows the data generating process (DGP) to be nonlinear, provided the functional central limit theorem (FCLT) holds. This fact is important because in practice the true model is unlikely to be known. Throughout the paper we will postulate that the number of candidate forecast models is fixed

[2]It is sometimes suggested that we select forecast models based on the outcome of tests of equal predictive accuracy. It is important to note that from the point of view of forecast model selection nothing would be gained from conducting such a test. To see this consider two forecasting models and consider the implications first of a failure to reject the null and then of rejection of the null. Suppose that Model 1 has a lower estimated recursive PMSE than Model 2. In this case, the standard procedure of ranking the models by their estimated recursive PMSE suggests that we select Model 1. This conclusion would also be supported by a test. If we cannot reject the null of equal predictive accuracy at a given significance level, we conclude that the recursive PMSEs are equal and hence there is no loss from having selected Model 1 for forecasting. Alternatively, if we reject the null in favor of the alternative that the recursive PMSE of Model 1 is lower, we conclude that Model 1 is preferred. Thus, the forecast model selection decision does not depend on the outcome of the test.

as the sample size increases. This assumption seems reasonable in many contexts, in which the number of predictors increases slowly over time, if at all. For example, the set of economic fundamentals used for exchange rate prediction has changed little since Meese and Rogoff's (1983) comprehensive empirical study.

In section 2, we study the ability of the IC and SOOS methods to select the model with the smallest possible true out-of-sample PMSE, among the candidate models. Our analysis breaks new ground in that we analyze the asymptotic properties of the SOOS method for prediction under standard assumptions on the choice of R similar to those used in the closely related literature on simulated out-of-sample inference. We obtain results that in part differ markedly from the predictive least squares results obtained by Rissanen (1986) and Wei (1992) under the assumption that R is fixed with respect to T. We will show that our asymptotic results are useful in understanding the finite-sample behavior of the SOOS method. In addition, we extend the existing literature on the IC method. Although the IC approach to model selection has been studied extensively in the context of providing good in-sample fits, little attention has been given to its implications for out-of-sample prediction.

We compare the IC method and the SOOS method along two dimensions. First, we study their ability to select the model with the lowest population PMSE among a set of candidate models. Second, we study whether they satisfy the principle of parsimony. The principle of parsimony comes into play when we have the choice between two or more forecast models with the same population PMSE. The principle of parsimony states that of any two models of different dimension, but with the same PMSE in population, we always prefer the more parsimonious model on the grounds that this model is likely to have smaller PMSE in finite samples. This idea is closely related to the principle of parsimony as discussed by Granger and Newbold (1986) and Box and Jenkins (1970) in the context of ARMA model selection.

When the true model is included among the forecast models under consideration, the principle of parsimony reduces to the standard notion of consistency. Consistency refers to the ability of a model selection method to detect the true model or DGP with probability 1 asymptotically, when this model is included among the forecast models under consideration. This notion of consistency must be modified when all forecast models under consideration are misspecified. In the latter case, the principle of parsimony dictates that the best approximating model is the most parsimonious model among the models with minimum population PMSE at $T+1$. Thus, the relevant notion of consistency under misspecification will refer to the ability of a forecast model selection method to select this best approximating model with probability 1 in the limit.

In section 2.3. we focus on the covariance stationary environment. Our first main theoretical result is that under suitable conditions both the IC method and the SOOS method are equally asymptotically valid in the sense of selecting models with smallest possible PMSE. This result is important because it provides a formal justification for the use of these methods, even when the forecast models are misspecified. Our second main result is that under standard assumptions the SOOS method will tend to be inconsistent for the best approximating model in the sense that it asymptotically will select an overparameterized model with positive probability. This tendency to overparameterize will result in excessive finite-sample PMSEs. In contrast, under suitable conditions on the penalty term, the IC method will always select the true model with probability one asymptotically among a set of candidate models that includes the true model, or, alternatively, will select the best approximating model with probability one asymptotically

if all candidate models are misspecified. This theoretical result applies equally to nested and to nonnested model comparisons. As a result, we would expect the models selected by the IC method to have lower finite-sample PMSEs in out-of-sample forecasting. We show, however, that not all information criteria have this attractive theoretical property and that existing criteria may have to be modified.

In section 3, we investigate by simulation the extent, to which this inconsistency of the SOOS method affects the finite-sample root PMSE at $T+1$. For pairwise nested model comparisons the resulting increase in root PMSE is quantitatively small, but clearly observable. As suggested by theory, the effect of model overparameterization on the root PMSE tends to become negligible as the sample size gets large. We also investigate the relative finite-sample performance of the IC method and the SOOS method in comparisons involving more than two models. We find that in many empirically plausible situations the IC method may offer substantial gains in accuracy relative to the SOOS method. The largest gains occur in small and moderately large samples. The gains in accuracy decline as the sample size increases, as suggested by theory.

There are two main exceptions to this result. Both occur in small samples only. One exception is a nonnested comparison involving persistent data. In this case, spurious correlations may arise in small samples that favor the SOOS method. The other exception involves nested forecast model comparisons in which the more parsimonious model is false. In this case, for sufficiently small samples, the misspecification bias from underparameterizing the forecast model will be more than offset by the reduction in parameter uncertainty. This trade-off between misspecification bias and estimation variance allows misspecified models to have a lower out-of-sample PMSE in small samples than even the true model. Thus, the PMSE criterion in small samples may favor the method with the greater tendency to underfit relative to the true model, and methods that detect the true model (or the best approximating model) with high probability in small samples may be inferior forecasting tools. We conclude that there can be no unambiguous ranking of the IC method and the SOOS method in small samples. As the sample size increases, however, in all cases, the bias-variance trade-off favors the consistent forecast model selection method, as suggested by theory, and the spurious correlations vanish.

Finally, in section 3, we also investigate the sensitivity of our results to the choice of R/T, and we show that our asymptotic thought experiment provides a better approximation than the Rissanen (1986) asymptotics when R is large relative to T.

A common concern is that information criteria may be misleading when the forecasting environment changes over time (e.g., Stock (1999), p. 24). In section 2.4., we relax the assumption of covariance stationarity and allow for the presence of unmodeled structural change in the paramters of the DGP. Our analysis shows that in the presence of unmodeled structural change indeed the IC method is inadmissible in the sense that it may select models with strictly higher out-of-sample PMSE than the best approximating model among the candidate models. We show, however, that the same result also holds for the SOOS method. Our proof of inadmissibility is based on a counterexample for nonnested model comparisons. For nested model comparisons, we are able to establish the inconsistency of both the IC and SOOS method by counterexample. Thus, the theoretical advantages of the IC method do not extend to situations, in which there is unmodeled structural change. In the presence of unmodeled change, both methods can be shown to overparameterize relative to the best approximating model. Although we were unable to prove inadmissibility in the nested case, we note that there is no presumption of admissibility.

Unfortunately, there is no ready alternative to the IC and SOOS methods of forecast model selection. An interesting topic for future research will be an investigation of the extent to which our inadmissibility and inconsistency results in section 2.4. are practically relevant. Some preliminary simulation evidence in section 3 suggests that deterministic structural breaks have no systematic effect on the relative performance in finite samples of the IC method and SOOS method. There is no evidence that the SOOS method of forecast model selection is more reliable than the IC method in the presence of unmodeled parameter shifts. This result casts doubt on the perception that the SOOS method provides at least partial protection against parameter instability, but much more work is needed for a final verdict.

Our theoretical analysis clearly establishes that under standard assumptions there is no presumption that the SOOS method is more reliable than the IC method on a priori grounds, but that in fact the IC method can be expected in some cases to deliver higher finite-sample accuracy in out-of-sample forecasting. Thus, there is no strong support for the current practice of selecting models by the SOOS methodology. We also conclude that the presence of unmodeled structural change is a serious challenge to current methods of forecast model selection.

2 Asymptotic Theory

2.1 Preliminaries

Our objective in this paper is to select among M candidate models the forecast model that will generate the most accurate forecast for period $T+1$ in terms of the prediction mean squared error (PMSE), given data for period $1, ..., T$. The number of forecast models under consideration, M, is presumed to be fixed with respect to the sample size. This model selection problem may be approached using the IC method or the SOOS method. In this section, we will compare these methods along two dimensions. First, we study their ability to select asymptotically the model with the lowest population PMSE at $T+1$ among a set of candidate models. Methods of forecast model selection that asymptotically may select models with strictly higher PMSE than the best approximating model will be referred to as *inadmissible* model selection rules.

Second, we study whether the IC and SOOS methods satisfy the *principle of parsimony*. The principle of parsimony comes into play when we have the choice between two or more forecast models with the same population PMSE. The principle of parsimony states that of any two models of different dimension, but with the same population PMSE, we always prefer the more parsimonious model on the grounds that this model is likely to have smaller PMSE in finite samples.

When the true model is included among the forecast models under consideration, the principle of parsimony reduces to the standard notion of consistency. Consistency refers to the ability of a model selection method to detect the true model or DGP with probability 1 asymptotically, when this model is included among the forecast models under consideration. This notion of consistency must be modified when all forecast models under consideration are misspecified. In the latter case, the principle of parsimony dictates that the best forecast model is the most parsimonious model among the models with minimum population PMSE. Thus, the relevant notion of consistency under misspecification will refer to the ability of a forecast model selection

method to select this best forecast model with probability 1 in the limit. We illustrate these concepts by three examples:

Example 1. Suppose the DGP is $y_t = \varepsilon_t$ where $\varepsilon_t \sim NID(0, \sigma^2)$. The two forecast models under consideration include the no-change forecast of y_{T+1} and the forecast generated by an AR(1) regression model. In this nested model comparison, both forecast models have a population of σ^2, but only a method that selects the no-change forecast model will satisfy the principle of parsimony. Any method that assigns positive probability to the AR(1) model, in contrast, will be inconsistent. A similar situation might arise in nonnested model comparisons when two models have the same population PMSE, but the more parsimonious model is the true model.

Alternatively, consider an example in which all forecast models under consideration are misspecified.

Example 2. Suppose that the DGP is an $MA(1)$ model, i.e., $y_t = \varepsilon_t - \theta\varepsilon_{t-1}$, where $\varepsilon_t \sim NID(0, \sigma^2)$,and we compare a no-change forecast of y_{T+1} to the forecast generated by an AR(1) regression model. In this case, both forecast models are misspecified. The population PMSE of the no-change forecast model is $(1+\theta^2)\sigma^2$, whereas the PMSE of the AR(1) model is $[(1+\theta^2) - \theta^2/(1+\theta^2)]\sigma^2$. For $\theta \neq 0$, the latter PMSE is lower by construction and hence the AR(1) model is the best forecast model among the two candidate models. Forecast model selection methods that select the best approximating model with probability 1 asymptotically, will be considered consistent. In this example, the principle of parsimony does not come into play because the best model is uniquely determined by the PMSE rank.

Example 3. Finally, suppose again the DGP is $y_t = \varepsilon_t$ where $\varepsilon_t \sim NID(0, \sigma^2)$. We compare the $MA(1)$ forecast model and the $AR(1)$ forecast model. In this case, both models have the same population PMSE of σ^2 and both models satisfy the principle of parsimony because they involve the same number of parameters. Hence, we are unable to rank these models from the point of view of our objective function. Both models must be considered consistent for the best approximating model.

2.2 Notation

We first present some theory for pairwise comparisons of forecast models. This results may be extended to multiple comparisons by repeated application of the pairwise comparison, as we will subsequently show in Corollary 1. Let x_t and z_t denote l and k dimensional vectors of deterministic or predetermined regressors (possibly including lagged dependent variables). We are interested in comparing two regression models:

$$\text{Model 1:} \quad y_t = \alpha' x_t + u_t \tag{2}$$

$$\text{Model 2:} \quad y_t = \beta' z_t + v_t \tag{3}$$

where u_t and v_t may be serially correlated. In the special case when Model 1 is nested in Model 2, x_t is a subvector of z_t, $\beta = [\alpha' \ \gamma']'$ and $u_t \equiv v_t$ such that the unrestricted model with $\gamma \neq 0$ can be written as

$$y_t = \beta' z_t + v_t = \alpha' x_t + \gamma' w_t + v_t \tag{4}$$

and the restricted model with $\gamma = 0$ can be written as

$$y_t = \alpha' x_t + v_t. \tag{5}$$

In this context, the SOOS method is implemented as follows. We fit each model by ordinary least squares (OLS) on the first S observations and evaluate its mean squared forecast error on observation $S+1$, for $S = R, R+1, R+2, ..., T-1$. The recursive OLS estimators are defined by $\hat{\alpha}_t = (\sum_{s=1}^{t} x_s x_s')^{-1} \sum_{s=1}^{t} x_s y_s$, and $\widehat{\beta}_t = (\sum_{s=t}^{t} z_s z_s')^{-1} \sum_{s=1}^{t} z_s y_s$. The resulting recursive PMSEs are given by

$$\begin{aligned} \widehat{\sigma}_{1R}^2 &= (1/(T-R)) \sum_{t=R+1}^{T} \bar{u}_t^2 \\ \widehat{\sigma}_{2R}^2 &= (1/(T-R)) \sum_{t=R+1}^{T} \bar{v}_t^2, \end{aligned}$$

where $\bar{u}_t = y_t - \hat{\alpha}_{t-1}' x_t$, $\bar{v}_t = y_t - \hat{\beta}_{t-1}' z_t$. The model with the smallest recursive PMSE is re-estimated on the data up to T and used for forecasting y_{T+1}.

In contrast, the IC method is implemented as follows. We fit each model by OLS on the first T observations. The OLS estimators are defined by $\widehat{\alpha} = (\sum_{t=1}^{T} x_t x_t')^{-1} \sum_{t=1}^{T} x_t y_t$, and $\widehat{\beta} = (\sum_{t=1}^{T} z_t z_t')^{-1} \sum_{t=1}^{T} z_t y_t$. We evaluate the criterion function:

$$\begin{aligned} IC(1) &= \ln(\sum_{t=1}^{T} \widehat{u}_t^2/T) + \dim(\alpha) * c_T \\ IC(2) &= \ln(\sum_{t=1}^{T} \widehat{v}_t^2/T) + \dim(\beta) * c_T, \end{aligned}$$

where $\widehat{u}_t = y_t - \hat{\alpha}' x_t$ and $\widehat{v}_t = y_t - \hat{\beta}' z_t$. The form of c_T depends on the criterion chosen. For the Schwarz Information Criterion (SIC) due to Schwarz (1978), for example, we have $c_T = \ln(T)/T$. The model with the smallest IC value is chosen for forecasting y_{T+1}.

We conclude this subsection by defining the population PMSEs for the $(T+1)$th observation associated with each forecast model. Let

$$\begin{aligned} \alpha_{T+1} &= \text{argmin}_\alpha E[(y_{T+1} - \alpha' x_{T+1})^2], \\ \beta_{T+1} &= \text{argmin}_\beta E[(y_{T+1} - \beta' z_{T+1})^2], \\ \sigma_{1,T+1}^2 &= E[(y_{T+1} - \alpha_{T+1}' x_{T+1})^2], \\ \sigma_{2,T+1}^2 &= E[(y_{T+1} - \beta_{T+1}' z_{T+1})^2]. \end{aligned}$$

Here α_{T+1} and β_{T+1} correspond to the parameter values that result in the minimum PMSE forecast for model 1 and 2, respectively. The minimum PMSE of Model 1 and Model 2 is denoted as $\sigma^2_{1,T+1}$ and $\sigma^2_{2,T+1}$, respectively, and corresponds to the population PMSE of the model. When the process is covariance stationary, $\sigma^2_{1,T+1}$ and $\sigma^2_{2,T+1}$ are independent of T, that is, $\sigma^2_{1,T+1} = \sigma^2_1$ and $\sigma^2_{2,T+1} = \sigma^2_2$ where

$$\begin{aligned}
\alpha_0 &= \text{argmin}_\alpha \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} E[(y_t - \alpha' x_t)^2], \\
\beta_0 &= \text{argmin}_\beta \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} E[(y_t - \beta' z_t)^2], \\
\sigma^2_1 &= \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} E[(y_t - \alpha_0' x_t)^2], \\
\sigma^2_2 &= \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} E[(y_t - \beta_0' z_t)^2].
\end{aligned}$$

By analogy, the population PMSEs obtained by recursive estimation can be obtained as follows. Let

$$\begin{aligned}
\alpha_r &= \text{argmin}_\alpha \lim_{T\to\infty} \frac{1}{[rT]}\sum_{t=1}^{[rT]} E[(y_t - \alpha' x_t)^2], \\
\beta_r &= \text{argmin}_\beta \lim_{T\to\infty} \frac{1}{[rT]}\sum_{t=1}^{[rT]} E[(y_t - \beta' z_t)^2], \\
\sigma^2_{1R} &= \lim_{T\to\infty} \frac{1}{T-[\pi T]}\sum_{t=[\pi T]+1}^{T} E[(y_t - \alpha_{(t-1)/T}' x_t)^2], \\
\sigma^2_{2R} &= \lim_{T\to\infty} \frac{1}{T-[\pi T]}\sum_{t=[\pi T]+1}^{T} E[(y_t - \beta_{(t-1)/T}' z_t)^2],
\end{aligned}$$

where $R/T \to \pi \in (0,1)$ as $T \to \infty$ and $r \in [\pi, 1]$. Here α_r and βr correspond to the parameter values that result in the minimum recursive PMSE forecast for Model 1 and Model 2, respectively. The minimum recursive PMSE of Model 1 and Model 2 is denoted as σ^2_{1R} and σ^2_{2R}, respectively, and corresponds to the population recursive PMSE. Note that under covariance stationarity $\sigma^2_1 = \sigma^2_{1R}$ and $\sigma^2_2 = \sigma^2_{2R}$.

2.3 Results for the Covariance Stationary Case

This section collects our assumptions and theoretical results for the covariance stationary case. The proofs can be found in the appendix.

Assumption 1.

(a) x_t, y_t and z_t satisfy the uniform law of large numbers in second moments, that is, $\sup_{\pi \le r \le 1} \|(1/T)\sum_{t=1}^{[rT]} x_t x_t' - E(x_t x_t')\| = o_p(1)$, $\sup_{\pi \le r \le 1} \|(1/T)\sum_{t=1}^{[rT]} x_t y_t - E(z_t z_t')\| = o_p(1)$, $\sup_{\pi \le r \le 1} \|(1/T)\sum_{t=1}^{[rT]} z_t z_t' - E(x_t y_t')\| = o_p(1)$, $\sup_{\pi \ge r \le 1} \|(1/T)\sum_{t=1}^{[rT]} z_t y_t - E(z_t y_t')\| = o_p(1)$ where $\lim_{T\to\infty}(1/T)\sum_{t=1}^{[rT]} E(x_t x_t')$ and $\lim_{T\to\infty}(1/T)\sum_{t=1}^{[rT]} E(z_t z_t')$ are positive definite uniformly in $r \in [\pi, 1]$. $\text{vec}([x_t'\ z_t'\ y_t]'[x_t'\ z_t'\ y_t])$ satisfies the functional central limit theorem (FCLT).

(b) $\lim_{T\to\infty}(1/T)\sum_{t=1}^{[rT]} E[x_t x_t']$ and $\lim_{T\to\infty}(1/T)\sum_{t=1}^{[rT]} E[z_t z_t']$ are positive definite for $r \in [\pi, 1]$.

(c) Let $f_t = E(y_t|\mathcal{F}_t)$ and $\varepsilon_t = y_t - f_t$ where $\mathcal{F}_t$ is the σ-field generated by x_s and z_s for $s \le t$. Then $\sup_t E[f_t^2] < \infty$, $\sigma^2 = E(\varepsilon_t^2|\mathcal{F}_{t-1})$ a.s. and $\sup_t E(\varepsilon_t^k|\mathcal{F}_{t-1}) < \infty$ a.s. for some $k > 2$.

(d) There are positive definite matrices Γ_1 and Γ_2 and positive semi-definite matrices Σ_1 and Σ_2 such that $\Gamma_1 = \text{plim}(1/T)\sum_{t=1}^{T} x_t x_t'(f_t) - \alpha_{t-1}' x_t)^2$, $\Gamma_2 = \text{plim}(1/T)\sum_{t=1}^{T} z_t z_t'(f_t - \beta_{t-1}' z_t)^2$, $\Sigma_1 = \text{plim}(1/T)\sum_{t=1}^{T} x_t x_t'(f_t - \alpha_{t-1}' x_t)^2$ and $\Sigma_2 = \text{plim}(1/T)\sum_{t=1}^{T} z_t z_t'(f_t - \beta_{t-1}' z_t)^2$.

Assumption 2.

(a) If the forecast models are nested, then $c_T = o(1)$ and $Tc_T \to \infty$ as $T \to \infty$.

(b) If the forecast models are nonnested, then $c_T = o(1)$ and $T^{1/2}c_T \to \infty$ as $T \to \infty$.

Assumption 3. $R/T \to \pi \in (0, 1)$ as $T \to \infty$.

Assumption 4. $[x_t', z_t', y_t]'$ are covariance stationary.

Assumption 1(a) is satisfied, for example, if x_t, y_t and z_t are near-epoch dependent (NED) with respect to mixing processes (see Wooldridge and White 1988). Thus Assumption 1 includes both nested and nonnested pairs of models and it covers both correctly specified and misspecified models. Assumption 1 even allows for the forecast model to be misspecified due to nonlinearities in the DGP, provided the FCLT holds (see Davidson 2002). Assumption 1 does not include unit root processes, however. Note that for the IC method a slightly weaker version of Assumption 1(a) involving the CLT in place of the FCLT would suffice. Assumption 3 is a typical condition for simulated out-of-sample inference in that R increases with T, as $T \to \infty$.[3] At the end of this subsection we will discuss an alternative assumption due to Rissanen (1986) and Wei (1992) that involves fixing R, as $T \to \infty$. Assumption 4 is required to establish that the recursive PMSE for model i, σ_{iR}^2, and the population PMSE, $\sigma_{i,T+1}^2$, are the same in the limit.

Assumption 2 imposes conditions on the penalty term in the information criterion (see Sin and White 1996). In the nested case, these conditions are met for example by the SIC, which

[3]For example, McCracken (2000) postulates $\pi \in (0, 1)$ as $T \to \infty$. Clark and McCracken (2001) and Corradi et al. (2001) allow for $\pi \in (0, 1]$ as $T \to \infty$. West (1996) and West and McCracken (1998) require R and $T - R$ to diverge to infinity such that $\pi \in [0, 1]$ as $T \to \infty$.

sets $c_T = \ln(T)/T$, but they are not met by the Akaike Information Criterion (AIC). Assumption 2(b) for the nonnested case rules out even the SIC. Specifically, the SIC will be inconsistent for the best approximating model when two nonnested models have the same asymptotic PMSE at $T+1$. The problem is that the SIC will tend to overfit relative to the best approximating model with positive probability. In all other cases of interest, the SIC will be consistent. Notwithstanding this violation of the principle of parsimony, however, the SIC remains an admissible method of forecast model selection because asymptotically it will correctly select a model with minimum out-of-sample PMSE.

More generally, note that the in-sample PMSEs (and their differences) converge in probability to their population counterparts at rate $T^{1/2}$, provided that one model has a smaller population PMSE than the other. This rate holds whether the models are nested or not. Thus, as long as the penalty term is of order smaller than $O(T^{-1/2})$, any information criterion will be admissible. This result implies that, for example, the AIC, the SIC and even R^2 are all admissible model selection criteria.

The partial inconsistency of the SIC illustrates the need for Assumption 2(b). Interestingly, none of the commonly used information criteria (such as AIC, SIC or the PIC of Phillips and Ploberger 1996) satisfies the conditions of Assumption 2(b). One modified information criterion that meets these conditions has been proposed by Sin and White (1996). They suggest $c_T = (\ln(T)/T)^{1/2}$. Note that this new criterion is not the only criterion one could propose. Since there are many alternative criteria that would be equally asymptotically valid, further study will be required to arrive at a criterion with satisfactory small-sample properties. We will not pursue this question in this paper.

We now formally establish the consistency of model selection based on modified information criteria that meet Assumption 2:

Theorem 1. [Consistency of Forecast Model Selection based on the Modified IC Method]
Suppose that Assumptions 1, 2 and 4 hold.

(a) If $\sigma^2_{1,T+1} = \sigma^2_{2,T+1}$ and if $\dim(\alpha_0) < \dim(\beta_0)$, then

$$\lim_{T\to\infty} P(IC(1) < IC(2)) \;=\; 1.$$

(b) If $\sigma^2_{1,T+1} = \sigma^2_{2,T+1}$ and if $\dim(\alpha_0) > \dim(\beta_0)$, then

$$\lim_{T\to\infty} P(IC(1) > IC(2)) \;=\; 1.$$

(c) If $\sigma^2_{1,T+1} < \sigma^2_{2,T+1}$ then

$$\lim_{T\to\infty} P(IC(1) < IC(2)) \;=\; 1.$$

(d) If $\sigma^2_{1,T+1} > \sigma^2_{2,T+1}$ then

$$\lim_{T\to\infty} P(IC(1) > IC(2)) \;=\; 1.$$

Parts (a) and (b) of Theorem 1 cover situations in which two models have the same PMSE in population. This situation arises for example for nested model comparisons when the restricted model is the DGP. Parts (a) and (b) imply that suitably modified information criteria will be consistent in that case in that they will select the more parsimonious of the two nested models with probability 1 asymptotically. A similar situation arises when two nonnested models of different dimensions have the same population PMSE. In that case, part (a) and (b) of Theorem 1 imply that information criteria will select the more parsimonious model of two nonnested models that have the same population PMSE. The latter case is primarily of academic interest, however, since in practice two nonnested models are unlikely to have the same population PMSE. In that sense, for most practical purposes there is little to choose between the SIC and the modified IC.

Parts (c) and (d) of Theorem 1 cover the case in which one model has lower population PMSE than the other. This situation arises for nested model comparison when the unrestricted model is the DGP. It is also likely to arise in nonnested model comparisons. In this case, model selection based on information criteria is consistent in the sense that it will select the model with the lower population PMSE with probability 1 in the limit.

We now turn to the analysis of the SOOS method. The following theorem shows that the SOOS method is not necessarily consistent for the best approximating model:

Theorem 2. [Possible Inconsistency of Model Selection based on the SOOS Method]
Suppose that Assumptions 1, 3 and 4 hold.

(a) If $\sigma^2_{1,T+1} = \sigma^2_{2,T+1}$ and if $\dim(\alpha_0) < \dim(\beta_0)$, then

$$\lim_{T\to\infty} P(\widehat{\sigma}^2_{1R} < \widehat{\sigma}^2_{2R}) < 1.$$

(b) If $\sigma^2_{1,T+1} = \sigma^2_{2,T+1}$ and if $\dim(\alpha_0) > \dim(\beta_0)$, then

$$\lim_{T\to\infty} P(\widehat{\sigma}^2_{1R} > \widehat{\sigma}^2_{2R}) < 1.$$

(c) If $\sigma^2_{1,T+1} < \sigma^2_{2,T+1}$, then

$$\lim_{T\to\infty} P(\widehat{\sigma}^2_{1R} < \widehat{\sigma}^2_{2R}) = 1.$$

(d) If $\sigma^2_{1,T+1} > \sigma^2_{2,T+1}$, then

$$\lim_{T\to\infty} P(\widehat{\sigma}^2_{1R} > \widehat{\sigma}^2_{2R}) = 1.$$

Parts (a) and (b) of Theorem 2 cover situations in which two models have the same PMSE in population. This situation will arise for example when the more parsimonious of two nested models is the DGP or when two nonnested models have the same population PMSE. Parts (a) and (b) show that the SOOS method is inconsistent in this case in the sense that it will select an

overparameterized model with positive probability. Although we had mentioned earlier that the SIC suffers from a similar inconsistency, when two nonnested models have the same asymptotic PMSE, the failure of the SOOS method is much more serious because it also extends to nested models with the same asymptotic PMSE. The latter case is highly empirically relevant. Part (c) and (d) of Theorem 2 cover the case in which one model has lower population PMSE than the other, as would be the case for nested model comparison if the unrestricted model is the DGP or for nonnested model comparisons in the likely case that the population PMSEs differ. Part (c) and (d) of Theorem 2 imply that the SOOS method is consistent in this case.

We conclude that, as a matter of theory, only modified information criteria are guaranteed to be consistent forecast model selection criteria. This conclusion applies equally to nonnested and to nested forecast model comparisons, but is especially relevant for the latter situation. Since we do not know a priori whether the data were generated by the restricted model or the unrestricted model, there always is the risk that the SOOS method will select an overparameterized model. This inconsistency does not matter asymptotically because the inconsistent model selected by the SOOS method will still have the same PMSE in the limit. In finite samples, however, overparameterized models will have excessive PMSEs. Thus, we would expect that in practice the IC method of forecast model selection will result in more accurate out-of-sample forecasts than the SOOS method. We will explore this question in a simulation study in section 3.

Theorems 1 and 2 cover pairwise model comparisons. In many applications, however, we are interested in comparing more than two forecast models. The following corollary generalizes our results to such multiple forecast model comparisons.

Suppose that there are M models. Let $\sigma^2_{i,T+1}$ denote the population PMSE for the ith model, $i = 1, ..., M$, and let $\sigma^2_{1,T+1} = \min_i \sigma^2_{i,T+1}$. As defined earlier, n_i denotes the number of parameters of model i and n_1 the number of parameters of model 1.

Corollary 1. [Asymptotic Properties of the Modified IC and SOOS Methods in Multiple Forecast Model Comparisons]

Suppose that Assumptions 1, 2, 3 and 4 hold for each model.

(a) Suppose there is some $i = 2, ..., M$ such that $\sigma^2_{i,T+1} = \sigma^2_{1,T+1}$ and $n_i > n_1$, then the IC method is consistent in the sense of selecting Model 1 with probability 1 in the limit.

(b) Suppose that $\sigma^2_{i,T+1} > \sigma^2_{1,T+1}$ for all $i = 2, ..., M$, then the IC method is consistent in the sense of selecting Model 1 with probability 1 in the limit.

(c) Suppose there is some $i = 2, ..., M$, such that $\sigma^2_{i,T+1} = \sigma^2_{1,T+1}$ and $n_i > n_1$, then the SOOS method is inconsistent in the sense of not selecting Model with probability 1 in the limit.

(d) Suppose that $\sigma^2_{i,T+1} > \sigma^2_{1,T+1}$ for all $i = 2, ..., M$, then the SOOS method is consistent in the sense of selecting Model 1 with probability 1 in the limit.

Corollary 1 follows from repeated application of Theorems 1 and 2. Note that these results also extend to the SIC, provided the comparison involves no nonnested model pairs with $\sigma^2_{i,T+1} = \sigma^2_{j,T+1}$.

It is important to note that our results differ from the predictive least squares results obtained by Rissanen (1986) and Wei (1992). Rissanen (1986) studied the asymptotic properties of the SOOS method under the assumption that R is fixed with respect to T. He referred to this method as predictive least squares. When the true model is included among the forecast models, Rissanen (1996) shows that predictive least squares is asymptotically equivalent to the SIC. This result immediately implies that predictive least squares is consistent for nested forecast model comparisons. More generally, when the best model differs from the true model, Wei (1992) shows that the implicit penalty term of predictive least squares converges at the same rate as the SIC. Although the two criteria are not equivalent, both are consistent for the best approximating model among two nested models. In contrast, for nonnested model comparisons neither the SIC nor predictive least squares will be consistent. This result follows immediately from the inconsistency of the SIC for nonnested model comparisons.

In sharp contrast, we showed that the SOOS method of forecast model selection will be inconsistent for both nested and nonnested model comparisons. This difference in results can be traced to different assumptions about the choice of R as $T \to \infty$. Whereas Rissanen and Wei treat R as fixed with respect to the sample size, we presume that the researcher will gradually increase R as more observations become available. To appreciate the difference that this assumption makes, it is instructive to express the selection criterion of the predictive least squares method in the same form as the information criterion. To facilitate this comparison we will make some simplifying assumptions. For details the reader is referred to Appendix B.

Suppose that the model is correctly specified i.e.,$\Gamma_1^{-1}\Sigma_1 = 0$, and assume that covariance stationary holds. Then it follows from the derivations in the Appendix, that the SOOS model selection criterion can be decomposed into the sum of a term describing the fit of the model and a remainder term which may be viewed as an implicit penalty term not unlike the penalty term of an information criterion. For example, for model 1 we obtain the approximation:

$$\ln(\hat{\sigma}_{1R}^2) = \ln\left(\frac{\sum_{t=1}^{T}\hat{u}_t^2(T) - \sum_{t=1}^{R}\hat{u}_t^2(R)}{T-R}\right) + \frac{\ln(T/R)}{T-R}\dim(\alpha) + o_p(1), \tag{6}$$

where $\hat{u}_t(S) = y_t - \hat{\alpha}_S' x_t$. Given this result, we may consider two alternative thought experiments: (i) R is fixed with respect to T and (ii) R is fixed as a fraction of T. For fixed R as in Rissanen (1986), it follows that:

$$\ln(\widehat{\sigma}_{1R}^2) \;=\; \ln(\widehat{\sigma}_1^2) + \frac{\ln T}{T}\dim(\alpha) + o_p(1). \tag{7}$$

From this expression, predictive least squares is seen to be asymptotically equivalent to the SIC (see Rissanen, 1986 and Wei, 1992) and hence consistent for nested forecast model comparisons, but not for nonnested comparisons, as we have shown. Inspection of equation (7) also makes it clear that the predictive least squares method is admissible for both nested and nonnested comparisons for the same reasons that information criteria are admissible.

Now let $R = [\pi T] + o(T)$ instead. This assumption corresponds to the standard thought experiment underlying the SOOS method described in this paper. In that case we obtain:

$$\ln(\widehat{\sigma}_{1R}^2) \;=\; \ln(\widehat{\sigma}_1^2) - \frac{\ln(\pi)}{(1-\pi)T}\dim(\alpha) + o_p(1). \tag{8}$$

Because the implicit penalty term converges to zero at the same rate as the penalty term of AIC, it can be seen immediately that the SOOS method is generally inconsistent, but admissible.

In practice, of course, a researcher faces a given finite sample size. Which of these two alternative thought experiments provides a better approximation to the finite-sample behavior of the SOOS method can be answered only by a simulation study. Heuristically, one would expect our asymptotic thought experiment to provide a better approximation when R is large relative to T. We will address this question in section 3. For the time being we only note that the thought experiment underlying our analysis is consistent with the assumptions of the rapidly growing literature on simulated out-of-sample inference (see footnote 3).

2.4 Results in the Presence of Unmodeled Structural Change

The main conclusion of the preceding analysis was that under covariance stationarity both the IC method and the SOOS method are asymptotically valid (in the sense of selecting models with minimum $\sigma_{i,T+1}^2$, $i = 1, ..., M$), but that only the IC method will be consistent for the best approximating model (in the sense of selecting the most parsimonious among the minimum PMSE models). We now turn to the analysis of the effects of unmodeled structural change on the properties of forecast model selection methods.[4] This problem is of paramount importance for practitioners. Allowing for structural change in the parameters of the DGP in our context amounts to dropping Assumption 4.

In related work, Sin and White (1996) study the ability of the IC method to select the best approximating model among a number of candidate models. They define the best approximating model in terms of the in-sample PMSE, σ_i^2. They show that the IC approach remains consistent for the best approximating model in terms of σ_i^2 under general conditions that also allow for certain forms of structural change. This result may seem to suggest that the IC method may also be able to select the best approximating model in terms of out-of-sample PMSE, $\sigma_{i,T+1}^2$. As we will show in this subsection, this is not the case. In Theorem 3 we will show that indeed the IC method will pick the model with minimum σ_i^2 under our assumptions, but - in the absence of covariance stationarity - nothing ensures that $\min[\sigma_1^2, \sigma_2^2] = \min[\sigma_{1,T+1}^2, \sigma_{2,T+1}^2]$. This fact may be proved formally by counterexample. We will provide one example, in which the IC method is inconsistent for the best approximating model. We will provide another example, in which the IC method even asymptotically will select models with higher out-of-sample PMSE, σ_{T+1}^2, than the best approximating model. The latter counterexample establishes that the IC method of forecast model selection is inadmissible (in the sense defined in section 2.1) in the presence of unmodeled structural change.

What about the use of the SOOS method? It is sometimes believed that the SOOS method - unlike the IC method- offers at least partial protection against parameter instability. We there-

[4]We do not consider the case when the date and form of structural change are known or can be consistently estimated because in that case the forecast model may be modified to incorporate the structural change and the results of the previous section will continue to apply.

fore investigate whether the SOOS method can be relied upon to select the $\min[\sigma^2_{1,T+1}, \sigma^2_{2,T+1}]$ model even in the presence of structural change. In Theorem 4 we will show that the SOOS method continues to select the $\min[\sigma^2_{1R}, \sigma^2_{2R}]$ model in the presence of unmodeled structural change. This result is not enough, however, to establish the theoretical validity of the SOOS method for forecast model selection because in the absence of covariance stationarity nothing ensures that the $\min[\sigma^2_{1R}, \sigma^2_{2R}] = \min[\sigma^2_{1,T+1}, \sigma^2_{2,T+1}]$. We will show by counterexample that the SOOS method not only may be inconsistent for the best approximating model at $T+1$, but that it may select models with strictly higher out-of-sample PMSE than the best approximating model. As in the case of the IC method the root cause of the failure of the SOOS method is that it uses a loss function that differs from the loss function of interest to the forecaster. Thus, neither the IC method nor the SOOS method has any theoretical justification in the presence of unmodeled structural change.

We now formally illustrate the consequences of dropping the assumption of covariance stationarity for pairwise model comparisons. Theorems 3 and 4 evaluate each model selection method in terms of its underlying loss function, which in the absence of covariance stationarity may differ from the loss function of interest to the forecaster.

Theorem 3. [Asymptotic Properties of the Modified IC Method in the Presence of Unmodeled Structural Change]

Suppose that Assumptions 1, 2 and 3 hold.

(a) If $\sigma^2_1 = \sigma^2_2$ and if $\dim(\alpha_0) < \dim(\beta_0)$, then $\lim_{T\to\infty} P(IC(1) < IC(2)) = 1$.

(b) If $\sigma^2_1 = \sigma^2_2$ and if $\dim(\alpha_0) > \dim(\beta_0)$, then $\lim_{T\to\infty} P(IC(1) > IC(2)) = 1$.

(c) If $\sigma^2_1 < \sigma^2_2$, then $\lim_{T\to\infty} P(IC(1) < IC(2)) = 1$.

(d) If $\sigma^2_1 > \sigma^2_2$, then $\lim_{T\to\infty} P(IC(1) > IC(2)) = 1$.

Theorem 4. [Asymptotic Properties of the SOOS Method in the Presence of Unmodeled Structural Change]

Suppose that Assumptions 1 and 3 hold.

(a) If $\sigma^2_{1R} = \sigma^2_{2R}$ and if $\dim(\alpha_0) < \dim(\beta_0)$, then $\lim_{T\to\infty} P(\widehat{\sigma}^2_{1R} < \widehat{\sigma}^2_{2R}) < 1$.

(b) If $\sigma^2_{1R} = \sigma^2_{2R}$ and if $\dim(\alpha_0) > \dim(\beta_0)$, then $\lim_{T\to\infty} P(\widehat{\sigma}^2_{1R} > \widehat{\sigma}^2_{2R}) < 1$.

(c) If $\sigma^2_{1R} < \sigma^2_{2R}$, then $\lim_{T\to\infty} P(\widehat{\sigma}^2_{1R} < \widehat{\sigma}^2_{2R}) = 1$.

(d) If $\sigma^2_{1R} > \sigma^2_{2R}$, then $\lim_{T\to\infty} P(\widehat{\sigma}^2_{1R} > \widehat{\sigma}^2_{2R}) = 1$.

Theorems 3 and 4 establish that the IC method will select the model with minimum in-sample PMSE and that the SOOS method will select the model with minimum recursive PMSE, even in the presence of unmodeled structural change. This result is not sufficient, however, to establish the asymptotic validity of these methods. The reason is that the loss function implicit in IC and SOOS model selection differs from that of the applied forecaster. Our presumption is that the forecaster is concerned with minimizing the PMSE for period $T+1$. It would not be

helpful to the forecaster, if a procedure succeeded in finding the model with minimum in-sample PMSE or minimum recursive PMSE, unless this model also happens to be the model with minimum out-of-sample PMSE. Under the assumptions of Theorems 3 and 4 nothing ensures that this will be the case. Recall that given covariance stationarity the recursive PMSE and the population PMSE of a given model will coincide in the limit. In contrast, in the presence of structural change, the in-sample and recursive PMSE in population will in general differ from the population PMSE of the forecast model at $T+1$. Thus, the criteria used by the IC and SOOS methods for ranking forecast models no longer coincide with the criterion of interest to the forecaster, even asymptotically. This fact will be immaterial, as long as the minimum under the in-sample and recursive PMSE criterion still coincides with the minimum under the PMSE criterion. Unfortunately, this need not be the case in general. The following counterexamples show that - if there is structural change in the DGP - the IC and SOOS methods of forecast model selection will not necessarily select the model with smallest out-of-sample PMSE. Example 4 focuses on a nested model comparison in the presence of an unmodeled one-time shift in the mean. Example 5 focuses on a nonnested model comparison in the presence of a time-varying mean.

Example 4. [Inconsistency of the IC and SOOS Methods for Nested Model Comparisons in the Presence of Unmodeled Structural Change]

Consider the DGP

$$y_t = I(t > [T/2]) + \varepsilon_t,$$

where ε_t is iid with zero mean and variance σ^2. The structural change consists of a one-time shift in the mean. The econometrician compares two forecast models:

$$y_t = \mu + u_t, \tag{9}$$

$$y_t = c + \rho y_{t-1} + v_t. \tag{10}$$

We will refer to the models in (9) and (10) as Model 1 and Model 2. For Model 1 the minimum value of the loss function σ^2_{T+1} is obtained at $\mu = 1$. At that value we have $\sigma^2_{1,T+1} = \sigma^2$. For Model 2 the loss function $\sigma^2_{2,T+1}$is minimized at the same value of σ^2 for the parameter choices $c = 1$ and $\rho = 0$. Although asymptotically both models have the same out-of-sample forecast accuracy, by the principle of parsimony the econometrician prefers Model 1 to Model 2.

We now turn to the question of which model the IC method would select. Asymptotically, minimizing the information criterion function amounts to minimizing the in-sample PMSE, $\sigma^2_i, i = 1, 2$, because the penalty term vanishes as $T \to \infty$ and the log function is a monotonic transformation. The in-sample PMSE of Model 1 is minimized at $\mu = 1/2$ with a minimum value of

$$\sigma^2_1 = \sigma^2 + 1/4.$$

The in-sample PMSE of Model 2 is minimized at $c = 2\sigma^2/(1 + 4\sigma^2)$ and $\rho = 1/(1 + 4\sigma^2)$ with a minimum value of

$$\sigma^2_2 = \sigma^2 + \frac{4\sigma^4 + \sigma^2}{16\sigma^4 + 8\sigma^2 + 1}$$

which is strictly less than $\sigma^2 + 1/4$. Thus, the IC method would select Model 2.

Now consider the SOOS method. For simplicity, suppose that the fraction of the sample used for the initial estimate is $\pi = 1/2$. Then using the first $[rT]$ observations for estimation, the asymptotic recursive PMSE for Model 1 and Model 2 is minimized at

$$\begin{aligned} c_r &= \begin{cases} 0, & \text{for } 0 \leq r < 1/2, \\ 1 - 1/(2r) & \text{for } 1/2 \leq r \leq 1. \end{cases} \\ (c_r, \rho_r) &= \begin{cases} (0,0) & \text{for } 0 \leq r < 1/2, \\ (\frac{2r-1}{4\sigma^2 r^2 + 2r - 1}, -\frac{2\sigma^2 r(2r-1)}{4\sigma^2 r^2 + 2r - 1}, & \text{for } 1/2 \leq r \leq 1, \end{cases} \end{aligned}$$

respectively. The resulting values of the SOOS loss function of Model 1 and of Model 2 are.

$$\begin{aligned} \sigma^2_{1R} &= \frac{1}{2}\sigma^2 + \frac{7}{8} \\ \sigma^2_{2R} &= \int_{1/2}^{1} (1 + \rho_r^2)\sigma^2 + \int_{1/2}^{1} (1 - c_r - \rho_r)^2 dr \\ &= O(\sigma^2), \end{aligned}$$

Thus $\sigma^2_{2R} < \sigma^2_{1R}$ for sufficiently small σ^2. In that case, the SOOS method also selects Model 2.

Since the IC and the SOOS method both fail to select Model 1, which satisfies the principle of parsimony, we conclude that they are both inconsistent for the best approximating model. Thus, the theoretical advantages of the IC method do not in general extend to situations, in which there is unmodeled structural change. As the next example shows, there may be even more serious consequences when structural change is not explicitly modelled by the forecaster.

Example 5. [Inadmissibility of the IC and SOOS Methods for Nonnested Model Comparisons in the Presence of Unmodeled Structural Change]

Suppose that Assumptions 1 and 3 hold. Further suppose that the dependent variable is generated by:

$$y_t = \frac{t}{T} + \varepsilon_t$$

where ε_t is iid with mean zero and variance σ^2. This model is nonstationary in that the mean is subject to smooth structural change. Consider two regression models. Model 1 only involves estimating the mean:

$$y_t = \mu + u_t \tag{11}$$

Model 2 is a fitted autoregression of the form:

$$y_t = \rho y_{t-1} + v_t. \tag{12}$$

The values of the $(T+1)$ loss function of Model 1 and of Model 2 are:

$$\sigma^2_{1,T+1} = \sigma^2$$

$$\sigma^2_{2,T+1} = \sigma^2 + \frac{\sigma^4(T+1)^2 + \sigma^2(T+1)^2}{(1+\sigma^2)^2T^2}.$$

Thus, Model 1 is preferable from the point of view of a forecaster.

The asymptotic PMSE of the Model 1 is minimized at $\mu = 1/2$:

$$\sigma^2_1 \;=\; \sigma^2 + \frac{1}{24}$$

and that of Model 2 is minimized at $\rho = 1/(1+3\sigma^2)$:

$$\sigma^2_2 \;=\; \frac{18\sigma^6 + 15\sigma^4 + 4\sigma^2}{18\sigma^4 + 12\sigma^2 + 2}.$$

For $\pi = 1/2$, the asymptotic recursive PMSE of Model 1 is minimized at $\mu_r = r/2$ for $0 \leq r \leq 1$:

$$\sigma^2_{1R} = \frac{1}{2}\sigma^2 + \frac{7}{96}.$$

The corresponding asymptotic recursive PMSE of Model 2 is minimized at $\rho_r = r^2/(r^2+3\sigma^2)$:

$$\sigma^2_{2R} \;=\; \frac{1}{2}\sigma^2 + \sigma^2 \int_0^1 \frac{r^6 + 3\sigma^2 r^4 + 9\sigma^2 r^2}{r^4 + 6\sigma^2 r^2 + 9\sigma^4} dr.$$

By inspection, when σ^2 is sufficiently small, e.g., $\sigma^2 = 1/12$, we have $\sigma^2_2 < \sigma^2_1$ and $\sigma^2_{2R} < \sigma^2_{1R}$. Thus, the IC and SOOS methods select Model 2, which has higher out-of-sample PMSE than Model 1. We conclude that both criteria are inadmissible for forecast model selection in the sense defined in section 2.1.

Example 5 proves that - in the presence of unmodeled structural change - neither the IC nor the SOOS method are admissible for nonnested model comparisons. Although we were unable to come up with a counterexample that would disprove the admissibility of the IC and SOOS methods for nested model comparisons, there is no presumption that these methods are admissible in the nested case. It is important to note that at present there is no theoretical basis for the application of either method in that context.

An interesting topic for future research will be an investigation of the extent to which our inadmissibility and inconsistency results in this subsection are practically relevant. If it could be shown that for most practical purposes at least the inadmissibility result is unlikely to arise, we might regain some confidence in model rankings based on the IC or SOOS method. We will provide some preliminary simulation evidence in section 3 that addresses this point.

3 Simulation Evidence

In this section we illustrate the implications of our theoretical results for applied work. We quantify both the probabilities of selecting the true (or best approximating) model and the

implications of model selection for the root PMSE of forecast models. A full-fledged simulation study is beyond the scope of this paper. Rather we focus on some stylized examples motivated by problems frequently encountered in empirical research. We evaluate the relative merits of the IC approach to forecast model selection and the SOOS methodology in a controlled environment. We distinguish between nested and nonnested forecast model comparisons (as well as situations that involve both nested and nonnested forecast models) and we differentiate between environments in which the true model is contained among the forecast models considered and environments in which all forecast models are misspecified. Subsection 3.1 abstracts from the possibility of structural change. The issue of structural change will be dealt with in subsection 3.2.

3.1 Baseline Setting without Structural Change

For expository purposes, we focus on the problem of predicting a scalar variable y_t. We postulate a DGP for y_t and generate 5000 trials of data of length $T+1$, where $T\epsilon\{50, 100, 150, 200, 300, 400, 500, 1000\}$. For each sample of length $T+1$, we use the first T observations to select the best forecast model among a set of candidate forecast models. The forecast model selection is carried out alternatively using the Schwarz Information Criterion (SIC) and using the simulated recursive out-of-sample methodology (abbreviated as SOOS). Since none of our DGPs involves comparisons of nonnested models with exactly the same PMSE, all consistency results we derived for the modified IC approach will carry over to the SIC. The best forecast model selected by each method is re-estimated on the first T observations and used to predict y_{T+1}. This procedure allows us to evaluate the prediction error of these forecast models for period $T+1$ in a genuine out-of-sample setting. The performance of each procedure is evaluated by averaging the squared prediction errors for each sample size across the 5000 trials. We follow the common practice of presenting the square root of this prediction mean squared error (abbreviated as RPMSE). In addition, we compute the probabilities that each procedure selects the true model (or the best approximating model when the true model is not included among the forecast models under consideration).

In this subsection we will study two prototypical applications of our theory. First, we will consider the problem of forecasting variables like asset returns, inflation or economic growth based on economic fundamentals. The second application will involve the selection of lag orders in univariate autoregression. Although the SOOS method is not commonly used for the purpose of selecting autoregressive lag orders, this application fits naturally into our framework. The discussion is organized as follows. We begin with pairwise forecast model comparisons. Such pairs of models may be either nested or nonnested. The subsequent analysis is devoted to multiple model comparisons. Comparisons of more than two models may be viewed as a collection of pairwise comparisons. That collection of pairs may consist of nonnested pairs of models only or it may consist of some nested and some nonnested pairs. Initially, we will implement the SOOS method with $R = 0.9T$. Alternative assumptions will be considered in subsections 3.1.3. and 3.1.4.

3.1.1 Application 1: Forecasting based on Economic Fundamentals

Case 1: Pairwise Comparison of Nested Models

The first example of a nested forecast model comparison is motivated by the literature on exchange rate return prediction. In this literature interest centers on the random walk model of nominal exchange rates, which implies that exchange rate returns should be unpredictable. Thus, a natural benchmark is the no-change forecast of returns. The alternative forecast models under consideration involve a number of predictors based on economic fundamentals (see Meese and Rogoff 1983). We study a stylized version of this forecasting problem. Our DGP for returns is $y_t = u_t$ where $u_t \sim NID(0, \sigma^2)$ for simplicity. Thus the true process is unpredictable white noise. We are interested in predicting y_{T+1}. We can select from two competing forecast models. The first competitor is the no-change forecast $y_{T+1|T} = 0$. The other competitor takes the form $y_{T+1|T} = \widehat{\alpha} + \widehat{\beta} w_T$, where $w_t = \rho w_{t-1} + \eta_t$. We postulate that $\eta_t \sim NID(0, diag(\sigma_\eta^2)$. We set $\rho = 0.9$, $\sigma_\eta^2 = 0.005$ and $\sigma^2 = 0.05$. These values are close to values obtained in empirical research on quarterly exchange rate returns and monetary fundamentals (see e.g., Mark 1995).

Since the no-change forecast model is nested in the model based on fundamentals, by Theorems 1 and 3 only the SIC method can be expected to select the true model with probability one asymptotically. Since the restricted model is the true model, the SOOS method will tend to select overparameterized models with positive probability, even asymptotically. The simulation results in Table 1a are consistent with this asymptotic result. Even for samples as small as $T = 50$ the SIC selects the true model with a probability of about 98%. The probability of selecting the true model steadily increases with T and exceeds 99.5% for $T = 300$. In contrast, the probability that the simulated out-of-sample procedure selects the true model never exceeds about 66% for any sample size and remains stable as T increases. Put differently, the simulated out-of-sample methodology selects an overparameterized model with about 34% probability.

A second implication of the theory is that the RPMSE of the model selected by the SIC method should be no lower than that of the true model asymptotically, but higher than that of the model selected by the simulated out-of-sample procedure in finite samples. It is not immediately obvious how quantitatively important this effect is because, as T increases, the parameters of the redundant regressors will be estimated with increasing precision and their effect on the RPMSE will become negligible. This consideration suggests that we would expect the biggest RPMSE differences in moderately large samples, rather than in large samples. The simulation results in Table 1 are consistent with this intuition. The last column shows the percent loss in RPMSE caused by relying on the simulated out-of-sample methodology instead of the SIC. For $T = 50$ the loss is 0.9 percentage points. The percent loss steadily declines as T increases, but it becomes negligible only for $T = 150$. Although the losses shown in Table 1a may appear small, it is worth pointing out that they may be eliminated at no cost.

The results in Table 1a are based on the presumption that the true model is among the forecast models under consideration. This assumption may not be realistic. We therefore repeated the exercise with the important difference that the true model is not white noise, but follows an MA(1) process of the form $y_t = 0.1u_{t-1} + u_t$ where $u_t \sim NID(0, \sigma^2/(1 + 0.1^2))$. Thus the unconditional variance of y_t remains unchanged. The set of forecast models under consideration also remains unchanged. Note that the no-change forecast now is no longer the

true model, but still the best forecast model asymptotically (abbreviated as the "best model"). Table 1b shows that for the same sample size, the probability of selecting the best forecast model tends to be slightly lower than that of selecting the true model in Table 1a. This conclusion holds for both forecast model selection methods. It has no effect on the ranking of the methods in terms of the RPMSE, however. The SOOS method still implies small losses that are diminishing with T. Thus, the qualitative results are unchanged.

For the exercise in Table 1 we postulated that the true model (or best model) corresponds to the restricted model. In Table 2a we consider the alternative setting, in which the data were generated under the unrestricted model. Thus, the DGP is $y_t = 0.9y_{t-1} + u_t$,where $u_t \sim NID(0, 0.005)$. Otherwise the setting is identical to that in Table 1a. This situation is fundamentally different from the previous table. First, both the SIC and SOOS method will select the true model with probability 1 asymptotically and in the limit will be equally accurate. Second, it is possible for the model selection procedures to underfit relative to the asymptotically best forecast model. Thus, parsimony does not necessarily favor the best forecast model. This fact makes no difference asymptotically, of course, but it may have important implications for the small-sample behavior of forecast model selection procedures. Underfitting causes misspecification bias, but it also reduces the estimation error. Thus, the net effect on the RPMSE depends on the relative size of these effects, which in turn will depend on the DGP, the set of forecast models under consideration and the sample size.

As Table 2a shows, the probability of selecting the true model indeed reaches 100% for both methods for $T \geq 500$, as expected, but the SIC has much higher probability of selecting the true model in small samples. Even for $T = 50$ it selects the true model with probability 100%. In contrast, the SOOS method selects the true model only with a probability of 84% for that sample size. At the same time, use of the SOOS method implies RPMSE losses of up to 5% in small samples. As expected, however, these losses decline with T and are negligible for $T \geq 200$.

In Table 2b we relax the assumption that the true model is included in the forecast model comparison. Instead we postulate that the best predictor available is merely positively correlated with the true predictor. Recall the DGP $y_t = 0.9y_{t-1} + u_t$,where $u_t \sim NID(0, 0.005)$. The model based on economic fundamentals now takes the form $y_{T+1|T} = \widehat{\alpha} + \widehat{\beta} w_T$, where $w_t = \rho w_{t-1} + \eta_t$ and $corr(\eta_t\ ,\ u_t) = 0.95$. This assumption implies that $y_{T+1|T} = \widehat{\alpha} + \widehat{\beta} w_t$ is merely the best approximating model for the DGP among the two models under consideration. Theorems 1 and 3 imply that this approximation error does not affect the consistency of the forecast model selection methods. Both methods will select the best model with probability 1 asymptotically. This prediction is borne out by the results in Table 2b. For $T = 1000$, both methods select the best model with probability 1. Although both methods are less likely to select the best model in small samples than in Table 2a, the SIC has much higher probability of selecting the best model in small samples. It also results in more accurate predictions in small samples. The RPMSE losses of the SOOS method may be larger than without model misspecification and in the example persist even for $T = 500$. We conclude that as long as the best predictor is strongly correlated with the true predictor the results will be qualitatively similar to Table 2a.

Table 2c repeats the exercise under the assumption that the best approximating model is only weakly correlated with the true predictor. Specifically, we postulate that $corr(\eta_t\ ,\ u_t) = 0.6$. As

before, we find that the SIC has higher probability of selecting the best model in small samples. In fact, even though the probability that the SOOS method will select the best model steadily increases with T, as suggested by theory, even for $T = 1000$ this probability has reached only about 84%. In small samples, the differences are even more striking. For $T = 50$, for example, the SIC selects the best model with probability 90% compared with only 59% for the SOOS method. Thus, in terms of the probability of selecting the true model, the SIC becomes relatively even more attractive than in Tables 2a and 2b. Nevertheless, as the last column shows, in small samples the SOOS method implies more accurate predictions. These RPMSE gains may be as high as 8% for $T = 50$. This tendency is only reversed in favor of the SIC for $T \geq 200$. As before, there is little to choose between the SIC and the SOOS method for very large samples.

Why are the small-sample results in Table 2c so different from the earlier findings, which tended to favor the SIC? Although they may seem odd at first, the results in Table 2c have a straightforward explanation. Unlike in Table 1, where the more parsimonious of the two models coincided with the best model, in Table 2 the less parsimonious model is the best model. This feature causes a trade-off between misspecification bias and estimation variance reduction that was absent in Table 1. The presence of such a trade-off means that the objectives of selecting the true (or best) model and the objective of generating a forecast with low RPMSE are no longer identical in small samples. For very small samples, underfitting relative to the best model will tend to make sense because the effect of the variance reduction on the RPMSE will outweigh the bias caused by the omission of relevant predictors. In fact, in this setting, underfitting will be advantageous even relative to the (asymptotically) best forecast model. Note that in Table 2c the RPMSE of the SOOS method is actually smaller than the RPMSE of the best model. Thus, a method that has high probability of selecting the true model will be disadvantageous for such small T.

As the sample size increases, however, the benefits from variance reduction decline and the costs of misspecification bias increase. A method that continues to underfit in that range will actually result in higher RPMSE, as seen in Table 2c, where for $T \geq 300$ the RPMSE of the SOOS method exceeds that of the best model. How large these respective ranges are, will depend on how strong the signal-to-noise ratio is in the data. When the best model is quite different from the alternative model, the misspecification bias will dominate even in small samples. This is what can be observed in Tables 2a and 2b. Only in a noisy environment, as in Table 2c, where the best model is not as clearly distinguished from the alternative model in small samples, the variance reduction effect dominates.

We conclude that the simulation evidence is fully consistent with the theoretical results of section 2 that asymptotically the SIC and SOOS method both will select a model with minimum out-of-sample PMSE. The simulation results also are consistent with the theoretical result that when the data are generated from the restricted model, the SIC method can be expected to have higher accuracy in finite samples than the SOOS method. In practice, of course, we do not know whether the data were generated under the restricted or the unrestricted model. We showed that the bias-variance trade-off that arises when the data are generated under the unrestricted model causes the ranking of the SIC and SOOS method to be ambiguous in small samples. In general, it is not clear which method will be favored by this bias-variance trade-off. For sufficiently large sample sizes, however, we found that the SIC performed at least as well as the SOOS method even when the data are generated from the unrestricted model.

Case 2: Pairwise Comparison of Nonnested Models

We now turn to pairwise comparison of nonnested models. In the nonnested case, both SIC and SOOS forecast model selection is consistent, provided that no two models have identical PMSE at $T+1$ in population. Thus, there is little to choose between them asymptotically. What is not clear is how these methods compare in small samples. The pervasiveness of the use of the simulated out-of-sample methodology in applied work may seem to suggest that this method must enjoy the better small-sample properties. The following example illustrates that this is not necessarily the case.

This example is motivated by empirical work on predicting variables like inflation, changes in inflation or economic growth based on economic fundamentals. The objective is to select among alternative, mutually exclusive sets of predictors. We postulate that the DGP is $y_t = 0.9y_{t-1} + u_t$,where $u_t \sim NID(0, \sigma^2)$. Again we are interested in predicting y_{T+1}. There are two alternative forecast models under consideration. The first forecast model is the true model forecast $y_{T+1|T} = \widehat{\alpha}_1 + \widehat{\beta}_1 z_{1T} = \widehat{\alpha}_1 + \widehat{\beta}_1 y_T$. The alternative model takes the form $y_{T+1|T} = \widehat{\alpha}_2 + \widehat{\beta}_2 z_{2T}$, where $z_{2t} = \rho z_{2t-1} + \eta_{2t}$. We postulate that $\eta_{2t} \sim NID(0, \sigma_\eta^2)$ and that η_{2t} is independent of u_t. As before, we set $\rho = 0.9$, $\sigma_\eta^2 = 0.005$ and $\sigma^2 = 0.005$. As each forecast model involves a different set of regressors, the models are nonnested.

Table 3a provides evidence that indeed the probabilities of selecting the true model approach one for larger sample sizes, as suggested by theory. This tendency holds whether we use the SIC or the SOOS method for forecast model selection. The SIC, however, has a much higher probability of selecting the true model in small samples. For example, for T =50 the SIC selects the true model with almost 100% probability, compared to only 86% for the simulated out-of-sample methodology. Although the reliability of the SOOS method improves with increases in the sample size, only for $T \geq 500$ the two methods are equally reliable. The differences in probabilities are also reflected in substantial differences in the RPMSE. The last column of Table 3a shows that for $T = 50$ the SOOS method involves a 6% increase in RPMSE relative to the SIC. Even for $T = 200$, the increase is almost 2%. As in the nested case, these RPMSE losses decline, as the sample size increases and become negligible for $T = 400$.

Tables 3b and Table 3c show the corresponding results under the assumption that the true model is not included among the candidate models, but is replaced by a predictor that is merely correlated with the true predictor. Specifically, we postulate that $z_{1t} = \rho z_{1t-1} + \eta_{1t}$ where $corr(\eta_{1t}\ ,\ u_t) = 0.95$ and $corr(\eta_{1t}\ ,\ u_t) = 0.6$, respectively. Whereas the results in Table 3b are qualitatively similar to those in Table 3a, in Table 3c the SOOS method in small samples actually has lower RPMSE than the SIC method, even though the SIC has higher probability of selecting the best approximating model for all T. This result is intriguing because - unlike in Table 2 - all models under consideration are of exactly the same dimension, so there is no bias-variance trade-off that could explain these results. The explanation is that in small samples by mere chance data generated from an independent process may be correlated with the variable to be forecast. Since both the variable to be forecast and the predictors under consideration in this exercise are persistent, a predictor that by chance exhibited high correlation in the past, it also likely to forecast well in the near future. Put differently, from an RPMSE point of view it may pay to exploit entirely spurious correlations in the data. This may be inferred from comparing the respective RPMSE columns of the SOOS method and of the best model in Table

3c. This example is a warning to users of the SOOS method that apparent predictability in small samples may be entirely spurious when the data are persistent. In contrast, the SIC does not suffer from this tendency. This example also reinforces the point made earlier that the objective of forecasting well is fundamentally different from the objective of identifying genuine relationships in the data that hold in population.

There are two immediate implications of this explanation of the results in Table 3c that can be verified empirically. One implication is that the RPMSE advantages of the SOOS method in small samples should vanish when the data are not persistent. This is exactly what happens when we repeat the exercise with an autoregressive root of 0.2 instead of 0.9. A second implication is that one would expect even greater advantages for the SOOS method in small samples, if we consider - all else equal - a larger number of nonnested candidate models because the chances of spurious correlations should increase with the number of independent models. This conjecture is borne out by multiple comparisons involving 10 nonnested models, as shown in Table 4 next.

Case 3: Multiple Comparisons of Pairs of Nonnested Models

The analysis underlying Table 4 is substantively identical to the simulation design for Table 3, except that now we consider a total of 9 independent nonnested predictors in addition to the true model (or the best approximating model). As before, the objective is to select among alternative, mutually exclusive sets of predictors. We postulate that the DGP is $y_t = 0.9y_{t-1} + u_t$,where $u_t \sim NID(0, \sigma^2)$. Again we are interested in predicting y_{T+1}. The first forecast model under consideration is the true model forecast $y_{T+1|T} = \widehat{\alpha}_1 + \widehat{\beta}_1 z_{1T} = \widehat{\alpha}_1 + \widehat{\beta}_1 y_T$. The other nine competitors take the form $y_{T+1|T} = \widehat{\alpha}_j + \widehat{\beta}_j z_{jT}$, $j = 2, ..., 10$, where $z_{jt} = \rho_j z_{jt-1} + \eta_{jt}$ $\forall$ j. For Table 4a we postulate that $\eta_{jt} \sim NID(0, diag(\sigma^2_\eta)$ and that η_{jt} is independent of u_t. As before, we set $\rho_j = 0.9$ $\forall$ j, $\sigma^2_\eta = 0.005$ and $\sigma^2 = 0.005$. For Tables 4b and 4c we let $\eta_{jt} \sim NID(0, diag(\sigma^2_\eta)$ with all but one η_{jt} independent of u_t. For that η_{jt}, say $j = 1$, we postulate $corr(\eta_{1t}\ ,\ u_t) = 0.95$ and $corr(\eta_{1t}\ ,\ u_t) = 0.9$, respectively. This assumption implies that $y_{T+1|T} = \widehat{\alpha}_1 + \widehat{\beta}_1 z_{1T}$ is the best approximating model for the DGP.

The simulation results in Table 4 are qualitatively similar to the results in Table 3, so we can be brief. The addition of more candidate models tends to lower the ability of the SIC to select the true model, but it has a disproportionately negative effect on the ability of the SOOS method to select the true model. It also appears to reinforce the advantages of the SIC in terms of RPMSE, provided the best predictor is at least strongly correlated with the true predictor. In Table 4a, the RPMSE losses implied by the SOOS method increase to 14% for $T = 50$ and 2% for $T = 200$ (compared to 6% and 1%, respectively in Table 3a). In Table 4b the implied RPMSE losses for the SOOS method may be as high as 8% (compared with a maximum of 4% in Table 3b). In Table 4c, however, the advantages of the SOOS method in small samples are reinforced. In small samples, the SOOS method now may result in RPMSE gains as high as 14% (compared with only 6% in table 3c). This result supports the earlier conjecture that adding more irrelevant predictors increases the chances of spurious correlations if the data are persistent. As in Table 3c, however, this small sample phenomenon vanishes for moderate and large sample sizes. For sufficiently large sample sizes, we find that the SIC always implies a somewhat lower RPMSE. This example illustrates that it is not obvious at all that the common

practice of relying on the simulated out-of-sample methodology for nonnested forecast model selection is the best available procedure.

Case 4: Multiple Comparisons of Some Nested and Some Nonnested Pairs of Models

The previous analysis maintained the assumption that all models under consideration are either nested or nonnested. In applied work it is common to consider a collection of forecast models that may involve nested pairs as well as nonnested pairs of models. In that case, the theoretical results of section 2 continue to apply, but it becomes harder to predict the relative performance of the SIC and SOOS method in small samples.

The DGP underlying Table 5 is the same as that for Table 1. The difference is that now we can select from ten possible predictive relationships. The first competitor is the no-change forecast $y_{T+1|T} = 0$. The other nine competitors take the form $y_{T+1|T} = \widehat{\alpha}_j + \widehat{\beta}_j w_{jT}$, $j = 1, ..., 9$, where $w_{jt} = \rho_j w_{jt-1} + \eta_{jt}$.$\forall$ j. We postulate that $\eta_{jt} \sim NID(0, diag(\sigma_\eta^2)$. We set $\rho_j = 0.9$ $\forall$ j and $\sigma_\eta^2 = 0.005$. Thus, the no change forecast model is nested in the other nine models, which in turn are pairwise nonnested. Since all models but the no-change forecast model are fairly distant from the true model, the results - not surprisingly - are qualitatively similar to Table 1. The main difference is that the probabilities of selecting the true model are lower in small samples and the RPMSE losses associated with the SOOS method are greater and persist for larger sample sizes. Table 5 shows losses as high as 3% in small samples. These results are fully consistent with the theoretical results of Theorems 1 and 3.

If Table 5 may be viewed as the counterpart to Table 1, Table 6 is the counterpart to Table 2. The difference again is the inclusion of additional nonnested predictors among the set of candidate models. We postulate the DGP $y_t = 0.9y_{t-1} + u_t$,where $u_t \sim NID(0, \sigma^2)$. Again we are interested in predicting y_{T+1}. There are ten alternative forecast models under consideration. The first forecast model is the no-change forecast model forecast $y_{T+1|T} = 0$. The other nine competitors take the form $y_{T+1|T} = \widehat{\alpha}_j + \widehat{\beta}_j z_{jT}$, $j = 2, ..., 10$, where $z_{jt} = \rho_j z_{jt-1} + \eta_{jt}$ $\forall$ j. We postulate that $\eta_{jt} \sim NID(0, diag(\sigma_\eta^2)$ with all but one η_{jt} independent of u_t. For that η_{jt}, say $j = 2$, we postulate $corr(\eta_{2t}\ ,\ u_t) > 0$. This assumption implies that $y_{T+1|T} = \widehat{\alpha}_2 + \widehat{\beta}_2 z_{2T}$ is the best approximating model for the DGP. As before, we set $\rho_j = 0.9$ $\forall$ j, $\sigma_\eta^2 = 0.005$ and $\sigma^2 = 0.005$.

What makes this application interesting is that the no-change forecast model is the most parsimonious model without being the best approximating model. Clearly, in this environment misspecification may improve forecast accuracy in small samples. As in Table 2, the best forecast model in small samples may be more parsimonious than the model that we know to be the best approximation asymptotically. At the same time, too much parsimony will undermine forecast accuracy even in small samples. Thus, the optimal degree of parsimony will change with the sample size. In addition, the results for Table 4c suggest that in small samples there may be a tendency for spurious predictive relationships, which should favor the SOOS method. Since the true model is clearly different from the no-change model, we would expect the results to be qualitatively similar to Table 4. This is indeed what Table 6 shows.

3.1.2 Application 2: Lag Order Selection for Autoregression

Another application of our theoretical results involves the selection of the best forecasting model among AR(p) models of order $p\epsilon\{0, 1, 2, ..., \overline{p}\}$. Although the SOOS method is not normally used for autoregressive lag order selection, there is no a priori reason why it could not be used, and contrasting the performance of the IC and the SOOS method in this context will be instructive. As before, we postulate that the number of models under consideration is fixed, as the sample size increases. An alternative thought experiment could be that $\overline{p}(T) \to \infty$ at a suitable rate as $T \to \infty$. We do not allow for that possibility here. Let p_0 denote the true lag order (or the lag order of the best approximating autoregressive model).

We begin by postulating an AR(1) DGP of the form $y_t = 0.9y_{t-1}+u_t$,where $u_t \sim NID(0, 0.005)$. The parameter settings are motivated by Mark (1995). Again we are interested in predicting y_{T+1}. The nine AR(p) forecast models under consideration are indexed by $p\epsilon\{0, 1, 2, ..., 8\}$. All regression models include an intercept. Note that the true AR(1) model is nested in AR(p) models with $p > 1$, but is not nested in the model with $p = 0$. Our theoretical results suggest that for large enough samples the SOOS method will overfit relative to p_0 with positive probability, but that both methods of selecting the forecast model will select a model with minimum PMSE asymptotically.

The simulation evidence is consistent with that result. First we focus on the probabilities that each procedure selects the true AR(1) model. Table 7a shows that the probability that the SOOS method will select the true model rises with the sample size from 28% for $T = 50$ to 35% for $T = 150$, but then fails to improve much further. We know from theory that it will not reach 100% in the limit. In contrast, the probability that the SIC will select the true model will reach 100% in the limit. The observed probability in Table 7a is 92% for $T = 50$ and approaches 99% for $T = 1000$. As Table 7a shows, the SIC has much higher probability of selecting the true model than the SOOS method in small samples. Even for $T = 1000$ the SOOS method selects the true model only with 36% compared with 99% for the SIC.

Given the possibility of underfitting, there are no theoretical predictions for the relative RPMSE of the SIC and SOOS method in small samples. Table 7a shows that the SOOS method implies RPMSE losses as high as 6% in small samples. although, as suggested by theory, these losses diminish as the sample size increases. This result is not surprising since the true slope parameter of 0.9 is fairly high and an AR(0) model clearly a poor predictor. It can be shown that the poor small-sample accuracy of the SOOS method in this example is driven by overfitting relative to p_0.

Note that even stronger results in favor of the SIC would be obtained for $p_0 = 0$ because in this case the true model is also the most parsimonious model and the bias-variance trade-off is eliminated. This case, however, seems of minor practical relevance and will not be considered here. Instead, we focus on higher-order DGPs. The results in Table 7b are based on an AR(4) DGP of the form $y_t = 1.2y_{t-1} - 0.3y_{t-2} + 0.4y_{t-3} - 0.4y_{t-4} + u_t$,where $u_t \sim NID(0, 0.005)$. The dominant autoregressive root is the same as for the AR(1). The parameter settings are again based on fitted values for the quarterly data in Mark (1995). This AR(4) model is nested in AR(p) models with $p > 4$, but is not nested in AR(p) models with $p < 4$. Thus, there is greater scope than in Table 7a for underfitting to improve the out-of-sample accuracy in small samples. Not surprisingly, the probability that the SIC selects the true model drops from 92% for $T = 50$

in Table 7a to only 47% in Table 7b. A similar drop occurs for the SOOS method. Qualitatively, however, the results are unchanged. The SIC is much more suited for detecting the true model, and there are still RPMSE losses for the SOOS method as high as 4% that diminish only slowly with increases in the sample size.

For Tables 7a and 7b we postulated that $0 < p_0 < \overline{p}$, where $\overline{p} = 8$. This ensures that the true model is contained among the candidate models. We now turn to a case, in which none of the AR(p) models under consideration is correctly specified. This might happen, for example, if the true model is a finite-order process with $p_0 > \overline{p}$. It also might happen if the true model is an invertible ARMA(p,q) process that can be represented only as an AR(∞) model. In this case, the asymptotically best approximating model will be the AR($\overline{p}$) forecast model, as the following proposition shows. We focus on the example of an MA(1) DGP without loss of generality. The result generalizes to any stationary invertible ARMA(p,q) process.

Proposition 1. [Best Approximating Model for an MA(1) DGP]
Suppose that the DGP is $y_t = \varepsilon_t - \theta\varepsilon_{t-1}$. where $\theta \neq 0$ and ε_t is white noise. Then the best approximating model in terms of the population out-of-sample PMSE, $\sigma^2_{p,T+1}$, among the set of AR(p) models of order $p\epsilon\{0, 1, 2, ..., \overline{p}\}$ will be the AR($\overline{p}$) model.

The proof of this proposition is in the appendix. Although the AR($\overline{p}$) model is the best approximating model asymptotically, estimating this AR($\overline{p}$) model in small samples involves considerable parameter uncertainty, and there is reason to believe that in small samples more parsimonious models will have lower RPMSE, provided the moving average component is small. The extent to which underfitting will be beneficial will depend on the magnitude of the moving average parameters. We explore these issues for the DGP is $y_t = 0.9y_{t-1} + 0.2u_{t-1} + u_t$ where $u_t \sim NID(0, 0.005)$. The dominant autoregressive root is the same as in Tables 7a and 7b. The best approximating model is the AR(8) model.

Note that in this case there is no scope for overfitting since $p_0 = \overline{p}$. This is similar to a situation, in which the researcher by accident selects a $\overline{p}$ that equals the lag order p_0 of the true model. Thus, both the SIC and the SOOS method must select the best model with probability 1 asymptotically. Nevertheless, as shown in Table 7c, even for $T = 1000$ the AR(8) is rarely selected by either method. Interestingly, in this case the SIC is the more parsimonious method in small samples (unlike in Table 2 for example), and as a result produces more accurate forecasts. The RPMSE loss of the SOOS method may be as high as 5% in small samples, but diminishes, as expected, as the sample size increases. The SIC works so well, despite selecting the wrong model with probability 1, because it tends to select models that are even more parsimonious than the best model and hence tend to have lower population RPMSE. This can be seen by comparing the respective RPMSE columns in Table 7c. In fact, the SOOS method also tends to underfit and hence improves upon the AR(8) model, but it does so less frequently than the SIC method.

This result is not surprising, given the small magnitude of the moving average component. In fact, this specific ARMA DGP is not very different from the AR(1) process in Table 7a. It can be shown, however, that the probability of selecting the AR(8) model increases in T as well as in the size of the MA parameter. As the sample size increases, the SIC selects the true model

with increasing probability and the asymptotics of Theorem 1 start taking effect, but this occurs only for sample sizes much larger than shown in Table 7c.

3.1.3 How Practically Relevant Are the Rissanen Asymptotics?

So far we have maintained the assumption that R is a fixed fraction of the sample size. Our results in Tables 1-7 were based on the assumption that $R = 0.9T$. We now investigate the performance of the SOOS method under the alternative assumption that R is fixed for all T, as postulated by Rissanen (1986) and Wei (1992). Specifically, we postulate that $R = 0.9 * 50$ for all T and recompute the results in Tables 1-7. Under this alternative thought experiment, we would expect the probability that the SOOS method will select the best approximating model to increase with the sample size. Given the space constraints these results are not shown. They are available upon request. For $T = 50$, by construction, the results are the same as the results already shown, but for larger T the choice of R matters. We find that indeed the ability of the SOOS method to select the best approximating model improves with the sample size, as suggested by asymptotic theory, but the improvement is not quite as rapid as for the SIC.

Which of these two sets of results is more relevant for applied work? Given that a researcher typically works with given T and R, the question is which thought experiment will provide a better approximation to the finite-sample behavior of the SOOS method for given T and R. Heuristically, one would expect that the Rissanen asymptotics to work best when R is small relative to T, and our standard asymptotics to work best when R is large relative to T. The latter case motivated our choice of $R/T = 0.9$ in the simulation study. We now complete the analysis by considering the other extreme case of $R/T = 0.1$. For the forecast model comparisons based on economic fundamentals we find that the ability of the SOOS method to select the best forecast model tends to be higher than when $R/T = 0.9$, but typically somewhat lower than that of the SIC. Moreover, in some cases, the ability of the SOOS method to select the best forecast model declines for larger T. In those cases, the SIC is much better suited for detecting the best forecast model. In one example the SOOS method selects the best forecasting model even for $T = 1000$ only with probability 73.5% compared with 98.5% for the SIC. Thus it appears that even when R is small relative to T our asymptotics provide a better approximation than the Rissanen asymptotics for fixed R. Qualitatively similar results are also obtained for the autoregressive application in Table 7. In that application the number of parameters to be estimated and invertibility problems in computing the OLS estimator dictate that $R \geq 0.4T$.

3.1.4 How Robust Are the Simulation Results to the Choice of R/T?

A second and unrelated question is how robust our out-of-sample PMSE rankings of the SIC and of the SOOS method are to the choice of R/T. Our main results are based on $R/T = 0.9$. How do the rankings change when R/T is small? To answer this question we recomputed the results in Tables 1-6 for $R/T = 0.1$. We found that for $T = 50$ the relative performance of the SIC and SOOS methods is mixed with little to choose between the SIC and SOOS method. For $T \geq 100$, the SIC virtually always has lower PMSE than the SOOS method, although the advantages of the SIC tend to be small. Further simulation experiments suggest that for $R/T = 0.25$ the SIC clearly dominates the SOOS method in terms of forecast accuracy, irrespective of the

degree of model misspecification, although the gains in accuracy rarely exceed one percentage point. Similar results are also obtained for the autoregressive application with $R = 0.4T$. These additional simulation results suggest that for small R relative to T the choice of forecast model selection method matters little. The largest differences between the IC and SOOS methods can be expected when R is large relative to T, as in the main simulation study.

3.2 Modified Setting with Unmodeled Structural Change

An important consideration for applied forecasters is the possibility of unmodeled structural change along the sample path. Common examples of structural change include intercept shifts, variance shifts and changes in the slope parameters of the DGP. Often this structural change cannot be detected reliably, given the small sample, and its precise form is unclear. Thus, it is of practical interest to study the implications of possible structural change for the problem of forecast model selection. Moreover, as we showed in section 2, the theoretical justification for the IC and SOOS methods of forecast model selection breaks down in the presence of unmodeled structural change, so we are unsure about the performance of these methods even asymptotically.

We illustrate the consequences of deterministic structural change for forecast models based on economic fundamentals. For expository purposes we focus on one-time structural shifts. In the simulation study we postulate that the structural shift occurs in period λT where $\lambda = 0.5$. Intercept shifts are normalized to be of the same size as one innovation standard deviation. Innovation variance shifts are normalized to correspond to a doubling of the innovation standard deviation. Slope parameter shifts correspond to a 10 percent decrease in the value of the slope parameter. Otherwise the simulation design remains unchanged. Notably, $R/T = 0.9$. To conserve space we only present results for the pairwise model comparisons corresponding to Tables 1-3 and we focus on the RPMSE comparisons. The first column in Tables 8, 9, and 10 shows the benchmark results for the model without structural change. The other columns show the results for the model with intercept shift, variance shift and slope shift, respectively. Note that the forecast models that were the best approximating models in Tables 1-3 may or may not be the best approximating models in the presence of structural change. Determining the best model in each case would require an analytic comparison of their asymptotic PMSEs at $T + 1$. We do not pursue this question here since our ultimate interest is the RPMSE comparison.

Table 8 focuses on pairwise nested comparisons where the restricted model is the DGP. The second column of Table 8 shows that the presence of structural change - far from favoring the SOOS method - may actually favor the SIC. In the presence of an intercept change, the RPMSE losses for the SOOS method, which were comparatively minor in the benchmark case, may be as high as 6% in small samples. In contrast, variance shifts and slope shifts (where applicable) have little effect on the benchmark results. In all cases, the SIC method results in more accurate forecasts.

Table 9 shows the corresponding results for pairwise nested comparisons where the unrestricted model is the DGP. In this case, an intercept shift may all but wipe out the RPMSE losses associated with the SOOS method in the benchmark case, without affecting the overall ranking. In contrast, variance and slope shifts may increase or decrease the RPMSE losses without systematically affecting the results. Overall, the qualitative results are unchanged.

Table 10 deals with the pairwise comparison of nonnested forecast models. Although the

presence of variance shifts or slope shifts may change the results somewhat relative to the benchmark, there is no systematic pattern and the qualitative results are preserved. The same is true for intercept shifts with the important difference that intercept shifts may affect the ranking for $T = 500$ and $T = 1000$ (but not for smaller T) when the best predictor is only weakly correlated with the true predictor. This anomaly is a finite-sample phenomenon only that vanishes as the sample size is increased further.

The simulation results in Tables 8-10 suggest that there is no reason to revise our assessment of the relative merits of the IC and the SOOS method in practice. For the specific examples investigated here, we find that the qualitative results that we obtained under covariance stationarity are not much affected by unmodeled structural change, notwithstanding the inadmissibility and inconsistency of the model selection rules in general. Moreover, there is no evidence that the SOOS method of forecast model selection is more reliable than the IC method in the presence unmodeled structural change. More often than not, the IC method is more reliable. This result casts doubt on the perception that the SOOS method provides at least partial protection against parameter instability. That conclusion of course may be specific to our choice of DGPs, and much more work is needed for a final verdict.

4 Concluding Remarks

We studied the problem of how to select the best linear forecast model from a number of candidate models. Our objective has been to select the model with the smallest out-of-sample prediction mean squared error (PMSE). We have implicitly assumed that the number of models under consideration is fixed with respect to the sample size. The two most common methods of forecast model selection are information criteria (IC) and the simulated out-of-sample (SOOS) forecast comparison method. The aim of this paper has been to shed light on the asymptotic and finite-sample properties of these procedures.

We established the asymptotic validity of both methods for covariance stationary data under fairly general conditions that allow for forecast model misspecification. Our results cover many applications of interest in macroeconomics and finance. We showed that there is no strong support for the common current practice of selecting models by the SOOS methodology. Although both the IC and the SOOS method have asymptotic justification under standard assumptions, we showed that only the IC method is consistent for the asymptotically best forecast model. Notably, in forecast model comparisons that involve nested pairs of models, when the best forecast model is also the most parsimonious model, the SOOS method will select an overparameterized model with positive probability. A similar - if less practically relevant - inconsistency result is obtained when comparing nonnested models of differing dimensions that have the same PMSE.

Forecast model selection involves a tradeoff between finding a model with good fit and overfitting the data. In the case of the IC method, this tradeoff is solved by imposing an explicit penalty term. Under suitable assumption on this penalty term, which have been derived in this paper, the IC method will always select the less parsimonious model of two forecast models with the same asymptotic out-of-sample PMSE. These conditions rule out for example the AIC, but include criteria of the type proposed by Sin and White (1996). In contrast, in the case of the SOOS method of ranking forecast models by their simulated recursive PMSE the penalty term

is implicit. We showed that this implicit penalty term under standard assumptions converges to zero at the same rate as the AIC. This result implies that the SOOS method will assign positive probability to forecast models that are overparameterized relative to the true model or - in the absence of the true model - the best approximating model.

This inconsistency of SOOS model selection does not affect the out-of-sample forecast accuracy in the limit, because the parameter estimation uncertainty vanishes asymptotically. It does, however, inflate the finite-sample accuracy of out-of-sample forecasts. We quantified this finite-sample effect by simulation for some empirically plausible forecast settings and showed that indeed the SOOS method tends to select less accurate forecast models than the IC method. We also found that even in cases when the best forecast model is not the most parsimonious model and both methods are consistent as a result, the SIC method typically implies a lower out-of-sample PMSE in finite samples.

Although the IC method compares favorably to the SOOS method in many finite-sample situations, this result cannot be generalized. We characterized conditions under which the SOOS method may select more accurate forecast models in small samples. One possible reason for a reversal of the ranking is that, whenever it is possible to underfit relative to the asymptotically best approximating model, there will be a bias-variance trade-off in small samples. Since the nature of the bias-variance trade-off will depend on the set of models under consideration, on the DGP and on the sample size, there can be no generally applicable ranking of the IC method and SOOS method. A second reason is that even in the absence of a bias-variance trade-off, when the data are persistent, spurious correlation may arise in small samples that allow more accurate predictions than using the true model. These anomalies can be shown to disappear, as predicted by theory, as the sample size increases.

We also compared the finite-sample performance of the IC and the SOOS method under an alternative asymptotic thought experiment due to Rissanen (1986) and Wei (1992). Under their assumptions the SOOS method remains consistent for nested model comparisons, whereas the SOOS method is inconsistent for both nested and nonnested model comparisons under our assumptions. Although it appears impossible to choose between these alternative asymptotic results on a priori grounds, we showed that our asymptotic results tend to provide better finite-sample approximations than the asymptotic results of Rissanen.

A common concern is that information criteria may be misleading when the forecasting environment changes over time (e.g., Stock (1999), p. 24). Our theoretical analysis revealed that in the presence of unmodeled structural change the IC method is indeed inadmissible, in the sense that it may select models with strictly higher out-of-sample PMSE than the best approximating model. We showed, however, that the same result also holds for the SOOS method. The failure of both methods is driven by the fact that - in the presence of parameter instability - the loss function of the forecaster differs from the loss functions implicit in the choice of model selection method. Moreover, we demonstrated that even in cases, in which both methods remain asymptotically justified, they have a tendency to overparameterize relative to the best approximating model. Our analysis also shows that there is no presumption that the SOOS method is more reliable than the IC method in the presence of unmodeled structural change.

Notwithstanding these theoretical problems, there is no obvious alternative to the use of the IC and SOOS methods for forecast model selection. An important task for future research will

be the systematic investigation of the extent to which inadmissible methods of forecast model selection are robust to unmodeled structural change in practice. What remains to be seen is just how pervasive situations are, in which the inadmissibility of the IC and SOOS method arises, and to what extent the performance of the two criteria is affected by these problems in practice. Some preliminary simulation evidence in section 3 suggested that deterministic structural change has no systematic effect on the relative performance in finite samples of the IC method and the SOOS method. This result casts doubt on the perception that the SOOS method provides at least partial protection against parameter instability, but much more work is needed for a final verdict.

There are several other interesting extensions of the analysis presented in this paper. One extension would involve a study of environments in which the number of forecast models under consideration increases with the sample size. Another useful extension would be to allow for nearly integrated predictors. A third extension would focus on multi-step ahead prediction. Finally, we note that our results have narrowly focused on the evaluation of point forecasts under quadratic loss. Although quadratic loss functions are widely used in empirical work, it may be of interest to explore other loss functions for point forecasts. In addition, the use of prediction intervals or predictive densities has recently gained in popularity (see, e.g., Christoffersen 1998, Diebold, Gunther and Tay 1998). Thus, alternative metrics may be employed in judging methods of forecast model selection.

Appendix: Proofs

Proof of Theorem 1. It follows from Assumption 1 that

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}(y_t-\hat{\alpha}'x_t)^2 &= \frac{1}{T}\sum_{t=1}^{T}(y_t-\alpha_0'x_t)^2+\frac{2}{T}\sum_{t=1}^{T}(y_t-\alpha_0'x_t)x_t(\hat{\alpha}-\alpha_0)\\ &\quad+(\hat{\alpha}-\alpha_0)'\frac{1}{T}\sum_{t=1}^{T}x_tx_t'(\hat{\alpha}-\alpha_0)\\ &= \sigma_{1,0}^2+\frac{1}{T}\sum_{t=1}^{T}(y_t-\alpha_0'x_t)^2-\sigma_{1,0}^2+O_p(T^{-1}). \qquad (13)\end{aligned}$$

and

$$\frac{1}{T}\sum_{t=1}^{T}(y_t-\hat{\beta}'z_t)^2 = \sigma_{2,0}^2+\frac{1}{T}\sum_{t=1}^{T}(y_t-\beta_0'z_t)^2-\sigma_{2,0}^2+O_p(T^{-1}). \qquad (14)$$

Thus the desired results follow from (13), (14) and Assumption 2.

Proof of Theorem 2. Under Assumption 4, $\sigma_1^2=\sigma_{1R}^2$ and $\sigma_2^2=\sigma_{2R}^2$. Thus, Theorem 2 follows from the proof of Theorem 4 below with suitable changes of notation.

Proof of Theorem 3. Parts (a)–(d) follow from the proof of Theorem 1.

Proof of Theorem 4. Parts (c) and (d) follow from the consistency of $\hat{\sigma}_{1R}^2$ and $\hat{\sigma}_{2R}^2$. To prove parts (a) and (b), we will use a modified version of Theorem 2.3 and Lemma 2.1 of Wei (1992).

Theorem 5. In addition to Assumption 3, suppose that (i) $\varepsilon_t=h_t+\eta_t$; (ii) $\lim_{T\to\infty}(1/T)\sum_{t=1}^{T}x_tx_t'=\Gamma$ where Γ is a positive definite matrix; (iii) $\lim_{T\to\infty}(1/T)\sum_{t=1}^{T}x_tx_t'h_t^2=\Sigma$ where Σ is a positive semi-definite matrix; (iv) $\lim_{T\to\infty}(1/T)\sum_{t=1}^{T}x_tx_t'\eta_t^2=\sigma^2\Gamma$; and (v) $\sum_{t=R+1}^{T}x_t'\hat{\Gamma}_t^{-1}x_th_t\eta_t/t=O(\sum_{t=R+1}^{T}(x_t'\hat{\Gamma}_t^{-1}x_t/t)^2h_t^2)$ where $\hat{\Gamma}_t=(1/t)\sum_{s=1}^{t}x_sx_s'$. Then

$$\lim_{T\to\infty}\sum_{t=R+1}^{T}x_t'\hat{\Gamma}_t^{-1}x_t\varepsilon_t^2/t=\sigma^2\dim(x_t)+\mathrm{tr}(\Gamma^{-1}\Sigma). \qquad (15)$$

If we assume further that (vi) $\lim_{T\to\infty}x_T'(\hat{\beta}_{T-1}-\beta)=0$, then

$$\sum_{t=R+1}^{T}\bar{u}_t=\sum_{t=1}^{T}\hat{u}_t+[\ln(T/R)][\sigma^2\dim(x_t)+\mathrm{tr}(\Gamma^{-1}\Sigma)](1+o(1)). \qquad (16)$$

The proof of Theorem 5 is analogous to the proof of Theorem 2.3 of Wei (1992) with his Lemma 2.1 replaced by Lemma 1:

Lemma 1. In addition to Assumption 3, suppose that (i) $\varepsilon_t=h_t+\eta_t$; (ii) $\lim_{T\to\infty}(1/T)\sum_{t=1}^{T}x_tx_t'=$

Γ where Γ is a positive definite matrix; and (iii) $\lim_{T\to\infty}(1/T)\sum_{t=1}^{T} x_t x_t' h_t^2 = \Sigma$ where Σ is a positive semi-definite matrix. Then

$$\lim_{T\to\infty} \sum_{t=R+1}^{T} x_t' \hat{\Gamma}_t^{-1} x_t \varepsilon_t^2 = -\ln(\pi)\mathrm{tr}(\Gamma^{-1}\Sigma). \tag{17}$$

To conserve space, we omit the proof of Theorem 5, but present the proof of our Lemma 1.

Proof of Lemma 1. Let ϵ be an arbitrary positive number. Since

$$(1-\epsilon)x'\Gamma^{-1}x \leq x'\hat{\Gamma}_t^{-1}x \leq (1-\epsilon)x'\Gamma^{-1}x \quad \text{for } \forall x \tag{18}$$

for sufficiently large T (see equation 2.18 of Wei, 1992), it follows that

$$(1-\epsilon)\sum_{t=R+1}^{T} x_t'\Gamma^{-1}x_t\varepsilon_t^2/t \leq \sum_{t=R+1}^{T} x_t'\hat{\Gamma}_t^{-1}x_t\varepsilon_t^2/t \leq (1-\epsilon)\sum_{t=R+1}^{T} x_t'\Gamma^{-1}x_t\varepsilon_t^2/t \tag{19}$$

Since ϵ can be made arbitrarily small, we obtain

$$\lim_{T\to\infty}\sum_{t=R+1}^{T} x_t'\hat{\Gamma}_t^{-1}x_t\varepsilon_t^2/t = \lim_{T\to\infty}\sum_{t=R+1}^{T} x_t'\Gamma^{-1}x_t\varepsilon_t^2/t. \tag{20}$$

Let $S_t = (1/t)\sum_{s=1}^{t} x_t x_t'\varepsilon_t^2$. Then $\lim_{t\to\infty} S_t = \Sigma$. It follows from applications of summation by parts, the Toeplitz lemma (Hall and Heyde, 1980, p.31) and the definition of Euler's constant that

$$\begin{aligned}\sum_{t=R+1}^{T} x_t'\Gamma^{-1}x_t/t &= \mathrm{tr}(\Gamma^{-1}S_T) - \mathrm{tr}(\Gamma^{-1}S_R) + \sum_{t=R+1}^{T}\mathrm{tr}(\Gamma^{-1}S_t)/(1+t) \\ &= (\ln(T)-\ln(R))\mathrm{tr}(\Gamma^{-1}\Sigma) + o(1). \end{aligned}\tag{21}$$

(see equation 2.20 of Wei (1992) for the first equality). Thus the desired result follows from (20) and (21).

Following the proof of Theorem 4.1.1 of Wei (1992) with his Theorem 2.3 replaced by our Theorem 5, one can show that, for example, for model 1:

$$\sum_{t=R+1}^{T}\bar{u}_t^2 = \sum_{t=1}^{T}\hat{u}_t^2(T) - \sum_{t=1}^{R}\hat{u}_t^2(R) + \ln(T/R)[\dim(\alpha)\sigma_1^2 + \mathrm{tr}(\Gamma_1^{-1}\Sigma_1)] + o_p(1). \tag{22}$$

where $\hat{u}_t(S)$ are the OLS residuals from regressing $y_t - \alpha_{t-1}'x_t$ on x_t, based on the first S observations. After dividing both sides by $(T-R)$ and taking logs, it follows that

$$\ln(\hat{\sigma}_{1R}^2) = \ln\left(\frac{\sum_{t=1}^{T}\hat{u}_t^2(T) - \sum_{t=1}^{R}\hat{u}_t^2(R)}{T-R}\right) + \frac{\ln(T/R)}{T-R}[\dim(\alpha) + \mathrm{tr}(\Gamma_1^{-1}\Sigma_1)] + o_p(1). \tag{23}$$

When the models are nested and are correctly specified, we obtain the expression

$$\ln(\hat{\sigma}_{1R}^2) = \ln\left(\frac{\sum_{t=1}^{T}\hat{u}_t^2(T) - \sum_{t=1}^{R}\hat{u}_t^2(R)}{T-R}\right) + \frac{\ln(T/R)}{T-R}\dim(\alpha) + o_p(1), \tag{24}$$

from which the results for parts (a) and (b) follow.

Proof of Proposition 1.

Let

$$y_t \;=\; \varepsilon_t - \theta\varepsilon_{t-1}$$

where $\theta \neq 0$. Then the autoregressive coefficients of an $AR(p)$ model in population are given by

$$\begin{bmatrix} 1+\theta^2 & -\theta & & 0 \\ -\theta & 1+\theta^2 & & 0 \\ 0 & -\theta & & 0 \\ \vdots & & & \vdots \\ 0 & & -\theta & 1+\theta^2 \end{bmatrix}^{-1} \begin{bmatrix} -\theta \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{25}$$

All we need to show is that the last component of (25) is different from 0. The last component of (25) is different from 0 if and only if the $(p,1)$ component of the inverse of

$$A = \begin{bmatrix} 1+\theta^2 & -\theta & & 0 \\ -\theta & 1+\theta^2 & & 0 \\ 0 & -\theta & & 0 \\ \vdots & & & \vdots \\ 0 & & -\theta & 1+\theta^2 \end{bmatrix}$$

is different from zero. Let $B = A^{-1}$ and suppose that $b_{p,1} = 0$. Let $A = [a_1' \; a_2' \; \cdots \; a_p']'$ and $B = [b_1 \; b_2 \; \cdots \; b_p]$. By definition $AB = I$. Thus, it follows from $a_p b_1 = 0$ that $b_{p-1,1} = 0$. Then it follows from $a_{p-1}b_1 = 0$ that $b_{p-2,1} = 0$. Repeating this procedure, we obtain $b_{i,1} = 0$ for $i = 1, 2, ..., p$. However, $a_1 b_1 = 0$. This is a contradiction because $a_1 b_1$ must be one. Thus $b_{p,1}$ cannot be zero, which means that the last component of (25) must be different from zero. Therefore, if $p > q$ then the $AR(p)$ model is a better-approximating model than the $AR(q)$ model.

Table 1. Nested Forecast Model Comparison: Two Models Overfitting Relative to Best Model Possible

Application: Forecasting Based on Economic Fundamentals

(a) True Model Included Among Competitors

	SIC		*SOOS*		True Model	SOOS/SIC
T	P(True Model) *in Percent*	$RPMSE$	P(True Model) *in Percent*	$RPMSE$	$RPMSE$	$RPMSE$-Loss *in Percent*
50	97.6	0.221	65.4	0.223	0.221	0.9
100	98.8	0.221	66.0	0.221	0.221	0.3
150	99.3	0.221	66.4	0.221	0.221	0.1
200	99.4	0.221	65.3	0.221	0.221	0.1
300	99.6	0.221	65.5	0.221	0.221	0.1
400	99.9	0.221	64.9	0.221	0.221	0.0
500	99.9	0.221	64.9	0.221	0.221	0.1
1000	99.9	0.221	65.3	0.221	0.221	0.1

(b) True Model Not Included Among Competitors

	SIC		*SOOS*		Best Model	SOOS/SIC
T	P(Best Model) *in Percent*	$RPMSE$	P(Best Model) *in Percent*	$RPMSE$	$RPMSE$	$RPMSE$-Loss *in Percent*
50	95.8	0.222	63.7	0.222	0.221	0.3
100	97.7	0.221	64.2	0.221	0.221	0.1
150	98.4	0.221	64.6	0.221	0.221	0.0
200	98.7	0.221	63.4	0.221	0.221	0.1
300	99.1	0.221	63.1	0.221	0.221	0.1
400	99.5	0.221	62.8	0.221	0.221	0.0
500	99.5	0.221	63.0	0.221	0.221	0.1
1000	99.7	0.221	63.1	0.221	0.221	0.1

SOURCE: Based on 5000 trials. R=0.9T.

Table 2. Nested Forecast Model Comparison: Two Models Underfitting Relative to Best Model Possible

Application: Forecasting Based on Economic Fundamentals

(a) True Model Included Among Competitors

	SIC		*SOOS*		True Model	SOOS/SIC
T	P(True Model) *in Percent*	$RPMSE$	P(True Model) *in Percent*	$RPMSE$	$RPMSE$	$RPMSE$-Loss *in Percent*
50	100.0	0.072	83.6	0.076	0.072	5.0
100	100.0	0.071	92.3	0.073	0.071	3.4
150	100.0	0.070	95.7	0.072	0.070	2.3
200	100.0	0.070	97.9	0.071	0.070	1.0
300	100.0	0.070	99.6	0.070	0.070	0.1
400	100.0	0.070	99.9	0.070	0.070	0.1
500	100.0	0.070	100.0	0.070	0.070	0.1
1000	100.0	0.070	100.0	0.070	0.070	0

(b) Best Model Strongly Correlated with True Model

	SIC		*SOOS*		Best Model	SOOS/SIC
T	P(Best Model) *in Percent*	$RPMSE$	P(Best Model) *in Percent*	$RPMSE$	$RPMSE$	$RPMSE$-Loss *in Percent*
50	99.9	0.084	76.9	0.087	0.084	4.0
100	100.0	0.084	84.7	0.087	0.084	3.4
150	100.0	0.084	89.8	0.087	0.084	2.9
200	100.0	0.084	92.7	0.086	0.084	2.0
300	100.0	0.084	96.8	0.085	0.084	1.4
400	100.0	0.084	98.4	0.085	0.084	0.9
500	100.0	0.084	99.2	0.084	0.084	0.3
1000	100.0	0.084	100.0	0.084	0.084	0.0

(c) Best Model Weakly Correlated with True Model

	SIC		*SOOS*		Best Model	SOOS/SIC
T	P(Best Model) *in Percent*	$RPMSE$	P(Best Model) *in Percent*	$RPMSE$	$RPMSE$	$RPMSE$-Loss *in Percent*
50	90.2	0.131	59.1	0.121	0.131	-7.7
100	94.5	0.137	61.4	0.130	0.136	-5.1
150	97.5	0.137	64.3	0.134	0.137	-2.5
200	99.1	0.138	67.3	0.137	0.138	-0.7
300	99.6	0.138	70.7	0.138	0.138	0.4
400	99.9	0.138	73.7	0.139	0.138	1.1
500	100.0	0.138	75.7	0.140	0.138	1.5
1000	100.0	0.138	84.4	0.140	0.138	1.2

SOURCE: Based on 5000 trials. R=0.9T.

Table 3. Nonnested Forecast Model Comparison: Two Models

Application: Forecasting Based on Economic Fundamentals

(a) True Model Included Among Competitors

	SIC		*SOOS*		True Model	SOOS/SIC
T	P(True Model) *in Percent*	$RPMSE$	P(True Model) *in Percent*	$RPMSE$	$RPMSE$	$RPMSE$-Loss *in Percent*
50	99.9	0.072	85.7	0.076	0.072	5.6
100	100.0	0.071	94.1	0.073	0.071	3.2
150	100.0	0.070	97.4	0.071	0.070	1.0
200	100.0	0.070	98.5	0.071	0.070	0.6
300	100.0	0.070	99.6	0.070	0.070	0.2
400	100.0	0.070	99.8	0.070	0.070	0.1
500	100.0	0.070	100.0	0.070	0.070	0.0
1000	100.0	0.070	100.0	0.070	0.070	0

(b) Best Model Strongly Correlated with True Model

	SIC		*SOOS*		Best Model	SOOS/SIC
T	P(Best Model) *in Percent*	$RPMSE$	P(Best Model) *in Percent*	$RPMSE$	$RPMSE$	$RPMSE$-Loss *in Percent*
50	98.9	0.084	78.9	0.088	0.084	4.3
100	100.0	0.084	87.4	0.087	0.084	3.3
150	100.0	0.084	91.1	0.086	0.084	2.6
200	100.0	0.084	93.6	0.086	0.084	2.2
300	100.0	0.084	97.0	0.085	0.084	1.4
400	100.0	0.084	98.3	0.085	0.084	0.9
500	100.0	0.084	99.1	0.084	0.084	0.4
1000	100.0	0.084	99.9	0.084	0.084	0.0

(c) Best Model Weakly Correlated with True Model

	SIC		*SOOS*		Best Model	SOOS/SIC
T	P(Best Model) *in Percent*	$RPMSE$	P(Best Model) *in Percent*	$RPMSE$	$RPMSE$	$RPMSE$-Loss *in Percent*
50	74.9	0.131	59.8	0.124	0.131	-5.6
100	86.7	0.138	63.9	0.134	0.136	-3.0
150	92.5	0.138	67.1	0.136	0.137	-1.5
200	96.3	0.138	69.6	0.137	0.138	-0.6
300	99.0	0.138	72.6	0.139	0.138	0.8
400	99.7	0.138	75.2	0.140	0.138	1.2
500	99.9	0.138	77.1	0.140	0.138	1.2
1000	100.0	0.138	84.6	0.140	0.138	1.1

SOURCE: Based on 5000 trials. R=0.9T.

Table 4. Nonnested Forecast Model Comparison: Ten Models

Application: Forecasting Based on Economic Fundamentals

(a) True Model Included Among Competitors

	SIC		*SOOS*		True Model	SOOS/SIC
T	P(True Model) *in Percent*	$RPMSE$	P(True Model) *in Percent*	$RPMSE$	$RPMSE$	$RPMSE$-Loss *in Percent*
50	99.1	0.073	57.4	0.083	0.072	13.7
100	100.0	0.072	79.4	0.078	0.072	8.7
150	100.0	0.071	89.5	0.074	0.071	4.0
200	100.0	0.071	94.7	0.073	0.071	2.1
300	100.0	0.071	98.7	0.071	0.071	0.5
400	100.0	0.071	99.6	0.071	0.071	0.2
500	100.0	0.071	99.8	0.071	0.071	0.0
1000	100.0	0.071	100.0	0.071	0.071	0

(b) Best Model Strongly Correlated with True Model

	SIC		*SOOS*		Best Model	SOOS/SIC
T	P(Best Model) *in Percent*	$RPMSE$	P(Best Model) *in Percent*	$RPMSE$	$RPMSE$	$RPMSE$-Loss *in Percent*
50	94.1	0.086	48.0	0.092	0.085	6.0
100	99.8	0.085	66.6	0.092	0.085	7.7
150	100.0	0.085	78.3	0.090	0.085	5.6
200	100.0	0.085	85.1	0.089	0.085	4.7
300	100.0	0.085	93.0	0.087	0.085	2.3
400	100.0	0.085	96.5	0.086	0.085	1.2
500	100.0	0.085	97.8	0.086	0.085	1.0
1000	100.0	0.085	99.8	0.085	0.085	0.1

(c) Best Model Weakly Correlated with True Model

	SIC		*SOOS*		Best Model	SOOS/SIC
T	P(Best Model) *in Percent*	$RPMSE$	P(Best Model) *in Percent*	$RPMSE$	$RPMSE$	$RPMSE$-Loss *in Percent*
50	40.8	0.132	22.2	0.113	0.133	–14.4
100	60.0	0.141	31.8	0.126	0.137	-10.9
150	74.4	0.140	39.9	0.131	0.137	-6.3
200	83.0	0.140	46.0	0.134	0.137	-4.0
300	93.9	0.138	55.2	0.137	0.137	-0.6
400	97.5	0.138	61.3	0.139	0.137	0.8
500	99.2	0.138	66.1	0.139	0.138	0.7
1000	100.0	0.137	79.0	0.140	0.137	1.8

SOURCE: Based on 5000 trials. R=0.9T.

Table 5. Both Nested and Nonnested Forecast Model Comparisons: Ten Models
Overfitting Relative to Best Model Possible
Application: Forecasting Based on Economic Fundamentals

(a) True Model Included Among Competitors

	SIC		*SOOS*		True Model	SOOS/SIC
T	P(True Model) *in Percent*	$RPMSE$	P(True Model) *in Percent*	$RPMSE$	$RPMSE$	$RPMSE$-Loss *in Percent*
50	88.2	0.228	20.7	0.235	0.223	3.0
100	93.6	0.225	20.3	0.228	0.223	1.7
150	96.1	0.224	18.0	0.226	0.223	1.2
200	96.6	0.223	18.1	0.226	0.223	1.0
300	98.0	0.223	18.0	0.225	0.223	0.6
400	98.5	0.236	17.7	0.224	0.223	0.5
500	98.9	0.223	17.9	0.224	0.223	0.5
1000	99.5	0.223	18.4	0.224	0.223	0.2

(b) True Model Not Included Among Competitors

	SIC		*SOOS*		Best Model	SOOS/SIC
T	P(Best Model) *in Percent*	$RPMSE$	P(Best Model) *in Percent*	$RPMSE$	$RPMSE$	$RPMSE$-Loss *in Percent*
50	81.4	0.229	19.0	0.233	0.223	1.9
100	88.7	0.225	18.4	0.228	0.223	1.1
150	92.7	0.224	16.5	0.226	0.223	0.8
200	93.5	0.223	16.6	0.225	0.223	0.8
300	95.4	0.223	16.5	0.225	0.223	0.5
400	96.6	0.223	16.7	0.224	0.223	0.3
500	97.5	0.223	16.8	0.224	0.223	0.4
1000	98.5	0.223	17.1	0.224	0.223	0.2

SOURCE: Based on 5000 trials. R=0.9T.

Table 6. Both Nested and Nonnested Forecast Models: Ten Models Underfitting Relative to Best Model Possible

Application: Forecasting Based on Economic Fundamentals

(a) True Model Included Among Competitors

	SIC		*SOOS*		True Model	SOOS/SIC
T	P(True Model) *in Percent*	$RPMSE$	P(True Model) *in Percent*	$RPMSE$	$RPMSE$	$RPMSE$-Loss *in Percent*
50	99.3	0.073	53.7	0.082	0.072	13.3
100	100.0	0.072	77.6	0.078	0.072	8.7
150	100.0	0.071	88.4	0.074	0.071	4.4
200	100.0	0.071	94.3	0.073	0.071	2.2
300	100.0	0.071	98.6	0.071	0.071	0.4
400	100.0	0.071	99.5	0.071	0.071	0.2
500	100.0	0.071	99.8	0.071	0.071	0.0
1000	100.0	0.071	100.0	0.071	0.071	0

(b) Best Model Strongly Correlated with True Model

	SIC		*SOOS*		Best Model	SOOS/SIC
T	P(Best Model) *in Percent*	$RPMSE$	P(Best Model) *in Percent*	$RPMSE$	$RPMSE$	$RPMSE$-Loss *in Percent*
50	94.6	0.086	44.2	0.091	0.085	5.2
100	99.8	0.085	64.2	0.092	0.085	7.8
150	100.0	0.085	76.7	0.090	0.085	6.0
200	100.0	0.085	84.3	0.090	0.085	5.5
300	100.0	0.085	92.6	0.087	0.085	2.4
400	100.0	0.085	96.4	0.086	0.085	1.4
500	100.0	0.085	97.7	0.086	0.085	1.0
1000	100.0	0.085	99.8	0.085	0.085	0.1

(c) Best Model Weakly Correlated with True Model

	SIC		*SOOS*		Best Model	SOOS/SIC
T	P(Best Model) *in Percent*	$RPMSE$	P(Best Model) *in Percent*	$RPMSE$	$RPMSE$	$RPMSE$-Loss *in Percent*
50	42.3	0.132	20.4	0.110	0.133	-16.3
100	61.4	0.141	29.8	0.125	0.137	-11.5
150	75.9	0.140	37.8	0.130	0.137	-6.9
200	84.0	0.140	44.0	0.133	0.137	-4.5
300	94.3	0.138	53.4	0.137	0.137	-0.7
400	97.7	0.137	59.7	0.139	0.137	1.2
500	99.3	0.138	64.8	0.139	0.138	0.7
1000	100.0	0.137	78.0	0.140	0.137	1.9

SOURCE: Based on 5000 trials. R=0.9T.

Table 7. Both Nested and Nonnested Forecast Models: Nine Models
Application: Lag Order Selection for Autoregression

(a) AR(1) DGP: True Model Included Among Competitors

	SIC		*SOOS*		True Model	SOOS/SIC
T	P(True Model) *in Percent*	$RPMSE$	P(True Model) *in Percent*	$RPMSE$	$RPMSE$	$RPMSE$-Loss *in Percent*
50	91.9	0.071	27.5	0.076	0.071	6.2
100	95.6	0.070	32.7	0.072	0.070	3.6
150	97.2	0.070	34.9	0.071	0.070	2.4
200	97.6	0.070	35.7	0.071	0.070	1.8
300	98.2	0.070	36.7	0.070	0.070	0.7
400	98.5	0.070	36.4	0.070	0.070	0.6
500	98.6	0.070	37.1	0.070	0.069	0.4
1000	99.0	0.069	36.0	0.069	0.069	0.0

(b) AR(4) DGP: True Model Included Among Competitors

	SIC		*SOOS*		True Model	SOOS/SIC
T	P(True Model) *in Percent*	$RPMSE$	P(True Model) *in Percent*	$RPMSE$	$RPMSE$	$RPMSE$-Loss *in Percent*
50	46.9	0.078	19.5	0.080	0.075	2.2
100	82.1	0.074	24.3	0.076	0.073	3.5
150	94.5	0.072	27.3	0.074	0.072	2.8
200	97.1	0.072	29.4	0.073	0.072	2.2
300	98.0	0.072	33.6	0.073	0.072	1.7
400	98.5	0.071	36.0	0.072	0.071	1.1
500	98.5	0.071	37.6	0.072	0.071	1.2
1000	99.4	0.071	41.8	0.071	0.071	0.6

(c) ARMA(1,1) DGP: True Model Not Included Among Competitors

	SIC		*SOOS*		Best Model	SOOS/SIC
T	P(Best Model) *in Percent*	$RPMSE$	P(Best Model) *in Percent*	$RPMSE$	$RPMSE$	$RPMSE$-Loss *in Percent*
50	0	0.074	11.1	0.078	0.080	5.3
100	0	0.072	10.9	0.074	0.074	2.3
150	0	0.072	11.2	0.073	0.073	1.0
200	0	0.072	10.5	0.072	0.072	0.6
300	0	0.071	10.4	0.071	0.072	0.4
400	0	0.071	10.3	0.071	0.071	0.1
500	0	0.071	10.3	0.071	0.071	0.3
1000	0	0.071	10.1	0.071	0.071	0.2

SOURCE: Based on 5000 trials. R=0.9T.

Table 8. Nested Forecast Model Comparison: Two Models
Restricted Model Subject to Structural Change
Application: Forecasting Based on Economic Fundamentals

(a) True Model Included Among Competitors

	SOOS/SIC *RPMSE*-Loss *in Percent*			
T	*No Shift*	*Intercept Shift*	*Variance Shift*	*Slope Shift*
50	0.9	6.3	0.9	-
100	0.3	2.9	0.3	-
150	0.1	1.6	0.1	-
200	0.1	0.7	0.1	-
300	0.1	0.1	0.1	-
400	0.0	0.1	0.0	-
500	0.1	0.0	0.1	-
1000	0.1	0.0	0.0	-

(b) True Model Not Included Among Competitors

	SOOS/SIC *RPMSE*-Loss *in Percent*			
T	*No Shift*	*Intercept Shift*	*Variance Shift*	*Slope Shift*
50	0.3	6.4	0.3	0.9
100	0.1	3.3	0.1	0.3
150	0.0	1.9	0.0	0.1
200	0.1	1.0	0.1	0.1
300	0.1	0.3	0.1	0.1
400	0.0	0.2	0.0	0.0
500	0.1	0.0	0.1	0.1
1000	0.1	0.0	0.0	0.1

SOURCE: Based on 5000 trials. R=0.9T.

Table 9. Nested Forecast Model Comparison: Two Models
Unrestricted Model Subject to Structural Change
Application: Forecasting Based on Economic Fundamentals

(a) True Model Included Among Competitors

	SOOS/SIC *RPMSE*-Loss *in Percent*			
T	*No Shift*	*Intercept Shift*	*Variance Shift*	*Slope Shift*
50	5.0	0	4.9	5.7
100	3.4	0	2.9	3.9
150	2.3	0	1.4	3.1
200	1.0	0	0.4	1.4
300	0.1	0	0.3	0.5
400	0.1	0	0.1	0.2
500	0.1	0	0.0	0.1
1000	0	0	0.0	0.0

(b) Best Model Strongly Correlated with True Model

	SOOS/SIC *RPMSE*-Loss *in Percent*			
T	*No Shift*	*Intercept Shift*	*Variance Shift*	*Slope Shift*
50	4.0	0	4.2	2.7
100	3.4	0	3.1	3.6
150	2.9	0	1.9	2.9
200	2.0	0	1.7	2.3
300	1.4	0	0.7	1.3
400	0.9	0	0.3	0.7
500	0.3	0	0.0	0.4
1000	0.0	0	0.0	0.4

(c) Best Model Weakly Correlated with True Model

	SOOS/SIC *RPMSE*-Loss *in Percent*			
T	*No Shift*	*Intercept Shift*	*Variance Shift*	*Slope Shift*
50	-7.7	0.2	-8.9	-5.6
100	-5.1	0.1	-6.1	-2.5
150	-2.5	0.2	-4.1	-0.5
200	-0.7	0.3	-2.5	0.2
300	0.4	0.0	-0.8	0.1
400	1.1	0.0	-0.2	1.1
500	1.5	0.0	0.2	0.7
1000	1.2	0.0	0.3	0.6

SOURCE: Based on 5000 trials. R=0.9T.

Table 10. Nonnested Forecast Model Comparison: Two Models

True Model Subject to Structural Change

Application: Forecasting Based on Economic Fundamentals

(a) True Model Included Among Competitors

	SOOS/SIC *RPMSE*-Loss *in Percent*			
T	*No Shift*	*Intercept Shift*	*Variance Shift*	*Slope Shift*
50	5.6	5.2	4.6	5.6
100	3.2	2.6	2.6	3.2
150	1.0	1.4	1.3	1.8
200	0.6	1.0	0.9	1.2
300	0.2	0.2	0.4	0.5
400	0.1	0.1	0.1	0.2
500	0.0	0.0	0.0	0.1
1000	0	0.0	0.0	0.0

(b) Best Model Strongly Correlated with True Model

	SOOS/SIC *RPMSE*-Loss *in Percent*			
T	*No Shift*	*Intercept Shift*	*Variance Shift*	*Slope Shift*
50	4.3	4.3	3.6	3.1
100	3.3	3.3	3.2	3.4
150	2.6	2.6	1.9	2.6
200	2.2	2.2	1.7	2.1
300	1.4	1.4	0.7	1.3
400	0.9	1.1	0.3	0.7
500	0.4	-1.5	0.0	0.4
1000	0.0	-0.3	0.0	0.4

(c) Best Model Weakly Correlated with True Model

	SOOS/SIC *RPMSE*-Loss *in Percent*			
T	*No Shift*	*Intercept Shift*	*Variance Shift*	*Slope Shift*
50	-5.6	-5.6	-5.0	-4.4
100	-3.0	-3.0	-4.0	-1.8
150	-1.5	-1.5	-2.6	-0.1
200	-0.6	-0.6	-1.4	0.2
300	0.8	0.8	-0.1	0.7
400	1.2	0.9	0.3	0.9
500	1.2	-3.5	0.3	0.7
1000	1.1	-0.7	0.4	0.6

SOURCE: Based on 5000 trials. R=0.9T.

References

1. Bossaerts, P., and P. Hillion (1999), "*Implementing Statistical Criteria to Select Return Forecasting Models: What Do We Learn?" Review of Financial Studies*, 12, 405-428.
2. Box, G.E.P., and G.M. Jenkins (1970), *Time Series Analysis, Forecasting and Control*, Holden Day: San Francisco.
3. Clark, T.E., and M.W. McCracken (2000), "Not-for-Publication Appendix to "Tests of Equal Forecast Accuracy and Encompassing for Nested Models"," manuscript, Federal Reserve Bank of Kansas City and Louisiana State University.
4. Clark, T., and M.W. McCracken, M.W. (2001), "Tests of Equal Forecast Accuracy and Encompassing for Nested Models," *Journal of Econometrics*, 105, 85-110.
5. Christoffersen, P. (1998), "Evaluating Interval Forecasts," *International Economic Review*, 39, 841-862.
6. Corradi, V., N.R. Swanson, and C. Olivetti (2001), "Predictive Ability with Cointegrated Variables," *Journal of Econometrics*, 104, 315-358.
7. Davidson, J. (2002), "Establishing conditions for the functional central limit theorem in nonlinear and semiparametric time series processes," *Journal of Econometrics*, 106, 243-269.
8. Diebold, F.X., Gunther, T. and Tay, A. (1998), "Evaluating Density Forecasts, with Applications to Financial Risk Management," *International Economic Review*, 39, 863-883.
9. Granger, C.W.J., and P. Newbold (1986), *Forecasting Economic Time Series*, Academic Press: London.
10. Hall, P., and C.C. Heyde (1980), *Martingale Limit Theory and Its Application*, Academic Press: San Diego.
11. Inoue, A. and L. Kilian (2002), "In-Sample or Out-of-Sample Tests of Predictability: Which One Should We Use?" Working Paper No. 195, European Central Bank.
12. Magnus, J.R. and H. Neudecker (1999), *Matrix Differential Calculus with Applications in Statistics and Econometrics,* Revised Edition, John Wiley & Sons: Chichester, England.
13. Mark, N.C. (1995), "Exchange Rates and Fundamentals: Evidence on Long-Horizon Predictability," *American Economic Review*, 85, 201–218.
14. Meese, R.A., and K. Rogoff (1983), "Empirical Exchange Rate Models of the Seventies: Do They Fit Out-of-Sample?" *Journal of International Economics*, 14, 3–24.
15. McCracken, M.W. (1999), "Asymptotics for Out of Sample Tests of Causality," manuscript, Department of Economics, Louisiana State University.
16. McCracken, M.W. (2000), "Robust Out-of-Sample Inference," *Journal of Econometrics*, 99, 195-223.
17. Phillips, P.C.B., and W. Ploberger (1996), "Posterior Odds Testing for a Unit Root with Data-Based Model Selection," *Econometrica*, 64, 381-412.

18. Rissanen, J. (1986), "Stochastic Complexity and Modeling," *Annals of Statistics,* 14, 1080-1100.

19. Schwarz, G. (1978), "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461-464.

20. Sin, C.-Y., and H. White (1996), "Information Criteria for Selecting Possibly Misspecified Parametric Models," *Journal of Econometrics*, 71, 207-225.

21. Stock, J.H. (1999), "Forecasting Economic Time Series," forthcoming in: B. Baltagi (ed.), *Companion in Theoretical Econometrics*, Basil Blackwell.

22. Stock, J.H., and M.W. Watson (1999a), "A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series," in: R. Engle and H. White (eds.), *Festschrift for C.W.J. Granger*, Oxford University Press.

23. Stock, J.H., and M.W. Watson (1999b), "Forecasting Inflation," *Journal of Monetary Economics,* 44, 293–335.

24. Stock, J.H., and M.W. Watson (2001), "Forecasting Output and Inflation: The Role of Asset Prices," manuscript, Kennedy School of Government, Harvard University.

25. Swanson, N.R., and T. Zeng (2001), "Choosing Among Competing Econometric Forecasts: Regression-Based Forecast Combination Using Model Selection," *Journal of Forecasting,* 20, 425–440.

26. Wei, C.Z. (1992), "On Predictive Least Squares Principles," *Annals of Statistics* 20, 1–42.

27. West, K.D. (1986), "Asymptotic Inference About Predictive Ability," *Econometrica* 64, 1067–1084.

28. West, K.D., and M.W. McCracken. (1998), "Regression-Based Tests of Predictive Ability," *International Economic Review* 39, 817–840.

29. Wooldridge, J.M., and H. White (1988), "Some Invariance Principles and Central Limit Theorems for Dependent Heterogeneous Processes" *Econometric Theory* 4, 210–230.