

Consistent Model Selection: Over Rolling Windows

Atsushi Inoue, Barbara Rossi and Lu Jin

Abstract In this chapter we analyze asymptotic properties of the simulated out-of-sample predictive mean squared error (PMSE) criterion based on a rolling window when selecting among nested forecasting models. When the window size is a fixed fraction of the sample size, Inoue and Kilian (J Econ 130: 273–306, 2006) show that the PMSE criterion is inconsistent. We consider alternative schemes under which the rolling PMSE criterion is consistent. When the window size diverges slower than the sample size at a suitable rate, we show that the rolling PMSE criterion selects the correct model with probability approaching one when parameters are constant or when they are time varying. We provide Monte Carlo evidence and illustrate the usefulness of the proposed methods in forecasting inflation.

1 Introduction

It is a common practice to compare models by out-of-sample predictive mean squared error (PMSE). For example, Meese and Rogoff (1983a,b) and Swanson and White (1997) compare models according to their PMSE calculated in rolling windows. Another common practice is to use a consistent information criterion such as the

A. Inoue (✉)

Department of Agricultural and Resource Economics, North Carolina State University,
Campus Box 8109, Raleigh, NC 27695-8109, USA
e-mail: atsushi@unity.ncsu.edu

B. Rossi

ICREA-UPF, Barcelona GSE and CR, calle Ramon Trias Fargas 25-27,
Barcelona 08005, SPAIN
e-mail: barbara.rossi@upf.edu

L. Jin

Department of Economics, North Carolina State University,
Campus Box 8110, Raleigh, NC 27695-8110, USA
e-mail: ljin@unity.ncsu.edu

Schwarz Information Criterion (SIC), used for example in Swanson and White (1997). Information criteria and the out-of-sample PMSE criteria deal with the issue of overfitting inherent in the in-sample PMSE criterion. Information criteria penalizes overparameterized models via penalty terms and are easy to compute. The out-of-sample PMSE criteria simulate out-of-sample forecasts and are very intuitive.¹

In a recent chapter, Inoue and Kilian (2006) show that the recursive and rolling PMSE criteria are inconsistent and recommend that consistent in-sample information criteria, such as the SIC, be used in model selection. They also show that even when there is structural change these out-of-sample PMSE criteria are not necessarily consistent. Their results are based on the assumption that the window size is proportional to the sample size.

In this chapter we consider an alternative framework in which the window size goes to infinity at a slower rate than the sample size. Under this assumption we show that the rolling-window PMSE criterion is consistent for selecting nesting linear forecasting models. When the nesting model is the truth, the criterion selects the nesting model with probability approaching one because the parameters and thus the PMSE are consistently estimated as the window size diverges. When the nested model is generating the data, the quadratic term in the quadratic expansion of the loss difference becomes dominant when the window size is small. Because the quadratic form is always positive, the criterion will select the nested model with probability approaching one. When the window size is large, however, the linear term and the quadratic term are of the same order and the sign cannot be determined. By letting the window size diverge slowly, the rolling PMSE criterion is consistent under a variety of environments, when parameters are constant or when they are time varying.

When the window size diverges at a slower rate than the sample size, the rolling regression estimator can be viewed as a nonparametric estimator (Giraitis et al. 2011) and time-varying parameters are consistently estimated. We show that our rolling-window PMSE criterion remains consistent even when parameters are time varying. When the window size is large, that is, when it is assumed to go to infinity at the same rate as the total sample size, the criterion is not consistent because the rolling regression estimator is oversmoothed. In the time-varying parameter case, the conventional information criterion is not consistent in general.

This chapter is related to, and different from, the works by West (1996); Clark and McCracken (2001); Giacomini and White (2006); Giacomini and Rossi (2010), and Rossi and Inoue (2011) in several ways. West (1996) and Clark and McCracken (2001) focus on comparing models' relative to forecasting performance when the window size is a fixed fraction of the total sample size, whereas Giacomini and

¹ The out-of-sample PMSE criteria are based on simulated out-of-sample predictions where parameters are estimated from a subsample to predict an observation outside the subsample. When subsamples always start with the first observation and use consecutive observations whose number is increasing, we call the simulated quadratic loss the recursive PMSE criterion. When subsamples are based on the same number of observations and are moving, we call the simulated quadratic loss the rolling PMSE criterion and the number of observations in the subsamples is the window size. See Inoue and Kilian (2006) for more technical definitions of these criteria.

White (2006) focus on the case where the window size is constant; this chapter focuses instead on the case where the window size goes to infinity but at a slower rate than the total sample size. Giacomini and Rossi (2010) argue that, in the presence of instabilities, traditional tests of predictive ability may be invalid, since they focus on the forecasting performance of the models on average over the out-of-sample portion of the data. To avoid the problem, they propose to compare models' relative predictive ability in the presence of instabilities by using a rolling window approach over the out-of-sample portion of the data. The latter helps them to follow the relative performance of the models as it evolves over time. In this chapter we focus on consistent model selection procedures, instead, rather than testing; furthermore, our focus is not to compare models' predictive performance over time, rather to select the best forecasting model asymptotically. Rossi and Inoue (2011) focus on the problem of performing inference on predictive ability that is robust to the choice of the window size. In this chapter, instead, we take as given the choice of the window size and our objective is not to perform tests; we focus instead on understanding whether it is possible to consistently select the true model depending on the size of the window relative to the total sample size.

The rest of this chapter is organized as follows: In Sect. 2 we establish the consistency of the rolling PMSE criterion under the standard stationary environment as well as under the time-varying parameter environment. In Sect. 3 we investigate the finite-sample properties of the rolling-window PMSE criterion. Section 4 demonstrates the usefulness of our criteria in forecasting inflation. Section 5 concludes.

2 Asymptotic Theory

Consider two nesting linear forecasting models, models 1 and 2, to generate h -steps ahead direct forecasts (where h is finite):

$$\text{Model 1 : } y_{t+h} = \alpha^* x_t + u_{t+h}, \tag{1}$$

$$\text{Model 2 : } y_{t+h} = \beta' z_t + v_{t+h} = \alpha' x_t + \gamma' w_t + v_{t+h}, \tag{2}$$

where $\dim(\alpha) = k$ and $\dim(\beta) = l$. The first terms on the right-hand sides of Eqs. (1) and (2), $\alpha^* x_t$ and $\beta' z_t$ are the population linear projections of y_{t+h} on x_t and z_t , respectively. Thus, z_t is uncorrelated with v_{t+h} , $\alpha^* = [E(x_t x_t')]^{-1} E(x_t y_{t+h})$ and $\beta = [E(z_t z_t')]^{-1} E(z_t y_{t+h})$.

Define the population quadratic loss of each model by

$$\sigma_1^2 = \lim_{T \rightarrow \infty} \frac{1}{T-h} \sum_{t=1}^{T-h} E[(y_{t+h} - \alpha' x_t)^2] = \lim_{T \rightarrow \infty} \frac{1}{T-h} \sum_{t=1}^{T-h} E(u_{t+h}^2),$$

$$\sigma_2^2 = \lim_{T \rightarrow \infty} \frac{1}{T-h} \sum_{t=1}^{T-h} E[(y_{t+h} - \beta' z_t)^2] = \lim_{T \rightarrow \infty} \frac{1}{T-h} \sum_{t=1}^{T-h} E(v_{t+h}^2).$$

Our goal is to select the model with smallest quadratic loss.

Let the window size used for parameter estimation be denoted by W for some $W > h$. Define the rolling ordinary least squares (OLS) estimators as follows, for $t = W + 1, \dots, T$:

$$\hat{\alpha}_{t,W} = \left(\sum_{s=t-W}^{t-h} x_s x'_s \right)^{-1} \sum_{s=t-W}^{t-h} x_s y_{s+h}, \tag{3}$$

$$\hat{\beta}_{t,W} = \left(\sum_{s=t-W}^{t-h} z_s z'_s \right)^{-1} \sum_{s=t-W}^{t-h} z_s y_{s+h}, \tag{4}$$

and the associated rolling PMSEs by:

$$\hat{\sigma}_{1,W}^2 = \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} \hat{u}_{t+h}^2, \tag{5}$$

$$\hat{\sigma}_{2,W}^2 = \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} \hat{v}_{t+h}^2, \tag{6}$$

where $\hat{u}_{t+h} = y_{t+h} - \hat{\alpha}'_{t,W} x_t$, $\hat{v}_{t+h} = y_{t+h} - \hat{\beta}'_{t,W} z_t$. We say that the rolling PMSE criterion is consistent if

- $\hat{\sigma}_{1,W}^2 < \hat{\sigma}_{2,W}^2$ with probability approaching one if $\sigma_1^2 = \sigma_2^2$; and
- $\hat{\sigma}_{1,W}^2 > \hat{\sigma}_{2,W}^2$ with probability approaching one if $\sigma_1^2 > \sigma_2^2$.

Under what conditions on the window size is the rolling PMSE criterion consistent? The existing results are not positive. When the window size is *large relative to the sample size* (i.e., $\exists \lambda \in (0, 1)$ s.t. $W = \lambda T + o(T)$), Inoue and Kilian (2005) show that the criterion is not consistent. Specifically, when $\sigma_1^2 = \sigma_2^2$, they show that the criterion selects model 2 with a positive probability resulting in the overparameterized model. We will discuss this result in more detail in the next section, where we will compare it with the theoretical results proposed in this chapter.

When the window size is *very small* (i.e., W is a fixed constant), it is straightforward to show that the criterion may not be consistent. For example, compare the zero-forecast model ($x_t = \emptyset$) and the constant-forecast model ($w_t = 1$) with $W = h = 1$. Suppose that $y_{t+1} = c + u_{t+1}$, where $u_t \sim iid(c, \sigma^2)$. Note that $\sigma_1^2 = c^2 + \sigma^2$ and $\sigma_2^2 = \sigma^2$. Since

$$\hat{\sigma}_{1,1}^2 = \frac{1}{T-1} \sum_{t=1}^{T-1} y_{t+1}^2 \xrightarrow{P} c^2 + \sigma^2,$$

$$\hat{\sigma}_{2,1}^2 = \frac{1}{T-1} \sum_{t=1}^{T-1} (y_{t+1} - y_t)^2 \xrightarrow{P} 2\sigma^2,$$

however, $\hat{\sigma}_{1,1}^2 < \hat{\sigma}_{2,1}^2$ with probability approaching one whenever $c^2 < \sigma^2$. This is because parameter estimation uncertainty never vanishes even asymptotically, when the window size is fixed.

The goal of the next section is to show that the criterion is consistent if the window size is small, but not too small, relative to the sample size in the following sense: $W \rightarrow \infty$ and $W/T \rightarrow 0$ as $T \rightarrow \infty$. Following Clark and McCracken (2000), we use the following notation: Let $q_{2,t} = z_t z_t'$, $q_{1,t} = x_t x_t'$, $B_i = [E(q_{it})]^{-1}$,

$$B_i(t) = \left[\frac{1}{W_h} \sum_{s=t-W}^{t-h} q_{i,s} \right]^{-1}, \quad H_1(t) = \frac{1}{W_h} \sum_{s=t-W}^{t-h} x_s (y_{s+h} - \alpha^* x_s), \quad H_2(t) = \frac{1}{W_h} \sum_{s=t-W}^{t-h} z_s v_{s+h},$$

where i is either 1 or 2 and $W_h = W - h + 1$.

2.1 Consistency of the Rolling-Window PMSE Criterion When Parameters are Constant

First, consider the case where the parameters are constant.

Assumption 1 As $T \rightarrow \infty$, $T^{1/2}/W = O(1)$ and $W/T \rightarrow 0$.

Assumption 2 (a) $\{[x_t' z_t' y_{t+h}]\}$ is covariance stationary and has finite 10 moments with $E(z_t z_t')$ positive definite and $B_2(t)$ positive definite for all t almost surely.

(b) $W^{1/2}(B_i(t) - B_i)$ and $W^{1/2}H_i(t)$ have finite fourth moments uniformly in t for $i = 1, 2$.

(c) $E(v_{t+h}|\mathcal{F}_t) = 0$ with probability one for $1, 2, \dots$, where \mathcal{F}_t is the σ field generated by $\{(y_{s+h}, z_s)\}_{s=1}^{t-h}$.

(d) $E[H_1'(t)B_1(x_t x_t' - E(x_t x_t'))B_1 H_1(t)] = o(W^{-1})$ and $E[H_2'(t)B_2(z_t z_t' - E(z_t z_t'))B_2 H_2(t)] = o(W^{-1})$ uniformly in t .

(e)

$$\begin{aligned} & \text{Cov} \left[\text{vech} \left(\sum_{t=W+1}^{T-h} H_i'(t)(B_i(t) - B_i)q_{i,t}(B_i(t) - B_i)H_i(t) \right) \right] \\ &= O \left(\sum_{t=W+1}^{T-h} \text{Cov} [\text{vech} (H_i'(t)(B_i(t) - B_i)q_{i,t}(B_i(t) - B_i)H_i(t))] \right), \\ & \text{Cov} \left[\text{vec} \left(\sum_{t=W+1}^{T-h} H_i'(t)B_i q_{i,t}(B_i(t) - B_i)H_i(t) \right) \right] \\ &= O \left(\sum_{t=W+1}^{T-h} \text{Cov} [\text{vec} (H_i'(t)B_i q_{i,t}(B_i(t) - B_i)H_i(t))] \right), \end{aligned}$$

$$\begin{aligned} & \text{Cov} \left[\text{vech} \left(\sum_{t=W+1}^{T-h} H_i'(t) B_i q_{i,t} B_i H_i(t) \right) \right] \\ &= O \left(\sum_{t=W+1}^{T-h} \text{Cov} [\text{vech} (H_i'(t) B_i q_{i,t} B_i H_i(t))] \right), \end{aligned}$$

for $i = 1, 2$.

Remark When the window size is assumed to be proportional to the sample size, $W = [rT]$ for $r \in [0, 1]$, the functional central limit theorem (FCLT) is often used to find the asymptotic properties of the recursive and rolling regression estimators (e.g., Clark and McCracken 2001). For example, if $h = 1$,

$$\sqrt{T}(\hat{\beta}_{t,W} - \beta) = \left(\frac{1}{T} \sum_{s=t-W}^{t-1} z_s z_s' \right)^{-1} \frac{1}{\sqrt{T}} \sum_{s=t-W}^{t-1} z_s v_{s+1}$$

and if $\text{vech}(z_t z_t')$ and $z_t v_{t+1}$ satisfy the FCLT, we obtain

$$\sqrt{T}(\hat{\beta}_{[rT]} - \beta) \Rightarrow \frac{\sigma}{r} [E(z_t z_t')]^{-1/2} B_l(r)$$

where $B_l(r)$ is the l -dimensional standard Brownian motion, provided $[z_t' v_{t+1}]'$ is covariance stationary. Thus, we have $\hat{\beta}_{t,W} - \beta = O_p(T^{-1/2})$ uniformly in t . When the window size diverges slower than the sample size it is tempting to use the same analogy and claims $\hat{\beta}_{t,W} - \beta = O_p(W^{-1/2})$ uniformly in t . This result does not follow from the FCLT, however, even though $\hat{\beta}_{t,W} - \beta = O_p(W^{-1/2})$ pointwise in t . To see why, let $z_t = 1$. Then

$$\begin{aligned} \hat{\beta}_{t,W} - \beta &= \frac{1}{W} \sum_{s=1}^{t-1} v_{s+1} - \frac{1}{W} \sum_{s=1}^{t-W-1} v_{s+1} \\ &= \frac{\sqrt{T}}{W} \frac{1}{\sqrt{T}} \sum_{s=1}^{t-1} v_{s+1} - \frac{\sqrt{T}}{W} \frac{1}{\sqrt{T}} \sum_{s=1}^{t-W-1} v_{s+1} \\ &= o_p \left(\frac{\sqrt{T}}{W} \right) \end{aligned}$$

uniformly in t , where the last equality follows from $\frac{1}{\sqrt{T}} \sum_{s=1}^{t-1} v_{s+1} - \frac{1}{\sqrt{T}} \sum_{s=1}^{t-W-1} v_{s+1} = o_p(1)$ by the FCLT and $W = o(T)$. Thus, the FCLT alone does not imply $\hat{\beta}_{t,W} - \beta = O_p(W^{-1/2})$ uniformly in t in general. This is why we need some high-level assumption, such as Assumptions 2(b)(d)(e).

Assumption 1 requires that W diverges slower than T . This assumption makes the convergence rates of terms in the expansion of the PMSE differential uneven which helps to establish the consistency of this criterion when the nested model is generating the data. Assumption 2(c) requires that the nesting model is (dynamically) correctly specified. Assumption 2(d) is trivially satisfied if z_t is strictly exogenous and allows for weak correlations between z_t and v_s . Assumption 2(e) is a high-level assumption and imposes that the variance of the sum is in the same order of the sum of variances. In other words, the summands are only weakly serially correlated so that their autocovariances decay fast enough. This assumption is somewhat related to the concept of essential stationarity of Wooldridge (1994, pp. 2643–2644). Assumptions somewhat similar to this condition are used in the central limit theorem for stationary and ergodic processes (e.g., Theorem 5.6 of Hall and Heyde 1980, p. 148) and the central limit theorem for near epoch-dependent processes (e.g., Theorem 5.3 of Gallant and White 1988, p. 76; Assumption C1 of Wooldridge and White 1988).

Theorem 1 *Under Assumptions 1 and 2, the rolling-window PMSE criterion is consistent.*

To compare our consistency result and the inconsistency result of Inoue and Kilian (2006), consider two simple competing models, $y_{t+h} = u_{t+h}$ (model 1) and $y_{t+h} = c + v_{t+h}$ (model 2) where v_{t+h} is i.i.d. with mean zero and variance σ_2^2 and $h = 1$. The difference of the out-of-sample PMSE can be written as

$$\hat{\sigma}_{2,W}^2 - \hat{\sigma}_{1,W}^2 = -\frac{2}{T - W - 1} \sum_{t=W+1}^{T-1} (\hat{c}_t - c)v_{t+1} + \frac{1}{T - W - 1} \sum_{t=W+1}^{T-1} (\hat{c}_t - c)^2$$

where $\hat{c}_t = (1/W) \sum_{s=t-W}^{t-1} y_{s+1}$. Assume that $c = 0$ in population.

When $W = [\lambda T]$ for some $\lambda \in (0, 1)$, it follows from Lemmas A6 and A7 of Clark and McCracken (2000) that

$$T \left(\hat{\sigma}_{2,W}^2 - \hat{\sigma}_{1,W}^2 \right) \xrightarrow{d} -\frac{2}{\lambda(1-\lambda)} \sigma_2^2 \int_{\lambda}^1 (B(r) - B(r-\lambda)) dB(r) + \frac{1}{\lambda^2(1-\lambda)} \sigma_2^2 \int_{\lambda}^1 (B(r) - B(r-\lambda))' (B(r) - B(r-\lambda)) dr$$

where $B(\cdot)$ is the standard Brownian motion. Because the probability that the right-hand side is negative is nonzero, the criterion is inconsistent when $c = 0$. This is the inconsistency result in Inoue and Kilian (2006).

When $W = o(T^{1/(1+2\varepsilon)})$ for some $\varepsilon \in (0, 1/2)$, the case considered in this chapter, we have:

$$\begin{aligned}
 W(\hat{\sigma}_{2,W}^2 - \hat{\sigma}_{1,W}^2) &= -\frac{2W^{\frac{1}{2}+\varepsilon}}{T-W-1} \sum_{t=W+1}^{T-1} \left(\frac{1}{W^{\frac{1}{2}+\varepsilon}} \sum_{s=t-W}^{t-1} v_{s+1} \right) v_{t+1} \\
 &\quad + \frac{1}{T-W-1} \sum_{t=W+1}^{T-1} \left(\frac{1}{W^{\frac{1}{2}}} \sum_{s=t-W}^{t-1} v_{s+1} \right)^2 \\
 &= \frac{1}{T-W-1} \sum_{t=W+1}^{T-1} \left(\frac{1}{W^{\frac{1}{2}}} \sum_{s=t-W}^{t-1} v_{s+1} \right)^2 + o_p(1)
 \end{aligned}$$

Because the right-hand side remains positive even asymptotically, the criterion will choose model 1 with probability approaching one. The key for the consistency result is that the last quadratic term in the expansion dominates the middle cross-term when the window size is small.

Lastly, it should be noted that our consistency result does not imply that the resulting forecast based on a slowly diverging window size is optimal. When parameters are constant, one would expect that the optimal forecast for the $T + 1$ st observation should be based on all T observations, not on the last W observations. Assumption 1 is merely a device to obtain the consistency of the rolling PMSE criterion.

2.2 Consistency of the Rolling-Window PMSE Criterion When Parameters are Time Varying

Sometimes it is claimed that out-of-sample PMSE comparisons are used to protect practitioners from parameter instability. As Inoue and Kilian (2006) show this is not always the case. In this section we show that the rolling PMSE criterion with small window sizes delivers consistent model selection even when parameters are time varying.

Suppose that the slope coefficients are time varying in the sense that

$$y_{T,t+h} = \beta \left(\frac{t}{T} \right)' z_{T,t} + v_{T,t+h} \tag{7}$$

where $\beta(r) = [\alpha(r)' \gamma(r)']'$ for $r \in [0, 1]$. When the slope coefficients are time varying, the second moments are also time varying. Let

$$\begin{aligned}
 \begin{bmatrix} \Gamma_{zz} \left(\frac{t}{T} \right) & \Gamma_{zy} \left(\frac{t}{T} \right) \\ \Gamma_{yz} \left(\frac{t}{T} \right) & \Gamma_{yy} \left(\frac{t}{T} \right) \end{bmatrix} &= \begin{bmatrix} \Gamma_{xx} \left(\frac{t}{T} \right) & \Gamma_{xw} \left(\frac{t}{T} \right) & \Gamma_{xy} \left(\frac{t}{T} \right) \\ \Gamma_{wx} \left(\frac{t}{T} \right) & \Gamma_{ww} \left(\frac{t}{T} \right) & \Gamma_{wy} \left(\frac{t}{T} \right) \\ \Gamma_{yx} \left(\frac{t}{T} \right) & \Gamma_{yw} \left(\frac{t}{T} \right) & \Gamma_{yy} \left(\frac{t}{T} \right) \end{bmatrix} \\
 &= \begin{bmatrix} E[x_{T,t}x'_{T,t}] & E[x_{T,t}w'_{T,t}] & E[x_{T,t}y_{T,t}] \\ E[w_{T,t}x'_{T,t}] & E[w_{T,t}w'_{T,t}] & E[w_{T,t}y_{T,t}] \\ E[y_{T,t}x'_{T,t}] & E[y_{T,t}w'_{T,t}] & E[y_{T,t}^2] \end{bmatrix},
 \end{aligned}$$

for $t = 1, 2, \dots, T$ and $T = 1, 2, \dots$. Let $\bar{B}_1\left(\frac{t}{T}\right) = [E(x_{T,t}x'_{T,t})]^{-1}$ and $\bar{B}_2\left(\frac{t}{T}\right) = [E(z_{T,t}z'_{T,t})]^{-1}$. Then $\beta(\cdot) = [\Gamma_{zz}(\cdot)]^{-1}\Gamma_{zy}(\cdot)$. We compare

$$y_{T,t+h} = \alpha\left(\frac{t}{T}\right)' x_{T,t} + u_{T,t+h} \quad (8)$$

and (7), where (7) simplifies to (8) if $\gamma(u) = 0$ for all $u \in [0, 1]$.

Assumption 3 As $T \rightarrow \infty$, $T^{1/2}/W = O(1)$ and $W = o(T^{2/3})$.

Assumption 4 (a)

$$\xi_t = \text{vech} \left\{ \begin{bmatrix} z_{T,t}z'_{T,t} & z_{T,t}y_{T,t+h} \\ y_{T,t+h}z'_{T,t} & y_{T,t+h}^2 \end{bmatrix} - \begin{bmatrix} \Gamma_{zz}\left(\frac{t}{T}\right) & \Gamma_{zy}\left(\frac{t}{T}\right) \\ \Gamma_{yz}\left(\frac{t}{T}\right) & \Gamma_{yy}\left(\frac{t}{T}\right) \end{bmatrix} \right\} \quad (9)$$

has finite fifth moments with $B_2(t)$ positive definite for all t almost surely.

(b) $W^{1/2}(B_i(t) - \bar{B}_i\left(\frac{t}{T}\right))$ and $W^{1/2}H_i(t)$ have finite fourth moments uniformly in t for $i = 1, 2$.

(c) $E(v_{T,t+h}|\mathcal{F}_{T,t}) = 0$ with probability one for $1, 2, \dots$, where $\mathcal{F}_{T,t}$ is the σ field generated by $\{(y_{T,s+h}, z_{T,s})\}_{s=1}^{t-h}$.

(d) $E[H'_i(t)\bar{B}_i\left(\frac{t}{T}\right)(q_{i,T,t} - E(q_{i,T,t}))\bar{B}_i\left(\frac{t}{T}\right)H_i(t)] = o(W^{-1})$ uniformly in t for $i = 1, 2$, where $q_{1,T,t} = x_{T,t}x'_{T,t}$ and $q_{2,T,t} = z_{T,t}z'_{T,t}$.

(e)

$$\begin{aligned} & \text{Cov} \left[\text{vech} \left(\sum_{t=W+1}^{T-h} H'_i(t) \left(B_i(t) - \bar{B}_i\left(\frac{t}{T}\right) \right) q_{i,T,t} \left(B_i(t) - \bar{B}_i\left(\frac{t}{T}\right) \right) H_i(t) \right) \right] \\ &= O \left(\sum_{t=W+1}^{T-h} \text{Cov} \left[\text{vech} \left(H'_i(t) \left(B_i(t) - \bar{B}_i\left(\frac{t}{T}\right) \right) q_{i,T,t} \left(B_i(t) - \bar{B}_i\left(\frac{t}{T}\right) \right) H_i(t) \right) \right] \right), \end{aligned}$$

$$\begin{aligned} & \text{Cov} \left[\text{vec} \left(\sum_{t=W+1}^{T-h} H'_i(t) \bar{B}_i\left(\frac{t}{T}\right) q_{i,T,t} \left(B_i(t) - \bar{B}_i\left(\frac{t}{T}\right) \right) H_i(t) \right) \right] \\ &= O \left(\sum_{t=W+1}^{T-h} \text{Cov} \left[\text{vec} \left(H'_i(t) \bar{B}_i\left(\frac{t}{T}\right) q_{i,T,t} \left(B_i(t) - \bar{B}_i\left(\frac{t}{T}\right) \right) H_i(t) \right) \right] \right), \end{aligned}$$

$$\begin{aligned} & \text{Cov} \left[\text{vech} \left(\sum_{t=W+1}^{T-h} H'_i(t) \bar{B}_i\left(\frac{t}{T}\right) q_{i,T,t} \bar{B}_i\left(\frac{t}{T}\right) H_i(t) \right) \right] \\ &= O \left(\sum_{t=W+1}^{T-h} \text{Cov} \left[\text{vech} \left(H'_i(t) \bar{B}_i\left(\frac{t}{T}\right) q_{i,T,t} \bar{B}_i\left(\frac{t}{T}\right) H_i(t) \right) \right] \right), \end{aligned}$$

where $i = 1, 2$.

- (f) $\Gamma_{zz}(u)$ is positive definite for all $u \in [0, 1]$, and $\alpha(\cdot) \equiv \Gamma_{xx}(\cdot)^{-1} \Gamma_{xy}(\cdot)$ and $\beta(\cdot) \equiv \Gamma_{zz}(\cdot)^{-1} \Gamma_{zy}(\cdot)$ satisfy a Lipschitz condition of order 1.

Remark Assumption 3 is more restrictive than Assumption 1 to keep the bias of the rolling regression estimator from interfering the consistency of the rolling PMSE estimator. Assumptions 4(a)(b) requires that ξ_t behaves like a stationary process with enough many moments. Assumptions 4(b)–(e) are analogs of Assumptions 2(b)–(e). Assumption 4(f) requires that the second moments change very smoothly.

Theorem 2 *Suppose Assumptions 3 and 4 hold. Then the rolling-window PMSE criterion is consistent.*

Remark The above consistency result is intuitive once it is recognized that the rolling regression estimator is a nonparametric regression estimator of parameters with a truncated kernel. For example, Cai (2007) establish the consistency and asymptotic normality of nonparametric estimators of time-varying parameters, and Giraitis et al. (2011) prove the consistency and asymptotic normality of nonparametric estimators for stochastic time-varying coefficient AR(1) models.

In general, the conventional information criteria, such as SIC, are not consistent when parameters are time varying. To show why that is the case consider comparing two competing models $y_{t+h} = u_{t+h}$ and $y_{t+h} = c + v_{t+h}$ for $h = 1$ when the data are generated from:

$$y_t = \frac{t}{T} - \frac{1}{2} + \varepsilon_t \tag{10}$$

where ε_t is i.i.d. with mean zero and variance σ^2 . Then the population in-sample PMSE of the zero forecast model is

$$\lim_{T \rightarrow \infty} E \left(\frac{1}{T-1} \sum_{t=1}^{T-1} y_{t+1}^2 \right) = \sigma^2 + \int_0^1 \left(r - \frac{1}{2} \right)^2 dr = \sigma^2 + \frac{1}{12}$$

The population in-sample PMSE of the forecast model that estimates the constant in rolling windows is also

$$\lim_{T \rightarrow \infty} \min_c E \left(\frac{1}{T-1} \sum_{t=1}^{T-1} (y_{t+1} - c)^2 \right) = \min_c \left(\sigma^2 + \int_0^1 (r - c)^2 dr \right) = \sigma^2 + \frac{1}{12}$$

Thus, the SIC would select the zero forecast model while the true DGP is a time-varying constant forecast model. Our criterion, by re-estimating the constant in rolling windows, is robust to time variation in the parameters and will select the second model with probability approaching unity asymptotically.

3 Monte Carlo Evidence

In this section we investigate the finite-sample performance of the rolling-window PMSE criterion in two Monte Carlo experiments. In the first experiment, we use the data generating process (DGP) of Clark and McCracken (2005) as it is similar to the empirical application that we will consider in the next section. In the second experiment, we use a simple DGP in which the dependent and independent variables both follow first-order autoregressive processes, and consider both constant parameter and time-varying parameter cases.

3.1 Simulation 1: DGP2 in Clark and McCracken (2005)

The second DGP of Clark and McCracken (2005) is based on estimates based on quarterly 1957:1–2004:3 data of inflation (Y) and the rate of capacity utilization in manufacturing (x). We consider restricted and unrestricted forecasting models as follows:

$$\text{Model 1 : } \Delta Y_{t+1} = \alpha_0 + \alpha_1 \Delta Y_t + \alpha_2 \Delta Y_{t-1} + u_{1,t+1} \tag{11}$$

$$\begin{aligned} \text{Model 2 : } \Delta Y_{t+1} = \alpha_0 + \alpha_1 \Delta Y_t + \alpha_2 \Delta Y_{t-1} + \gamma_1 x_{t-1} + \gamma_2 x_{t-2} + \gamma_3 x_{t-3} \\ + \gamma_4 x_{t-4} + u_{2,t+1} \end{aligned} \tag{12}$$

When the restricted model (11) is true, the DGP is parameterized using Eq. (7) in Clark and McCracken (2005):

$$\Delta Y_t = -0.316 \Delta Y_{t-1} - 0.214 \Delta Y_{t-2} + u_{y,t}, \tag{13}$$

$$\begin{aligned} x_t = -0.193 \Delta Y_{t-1} - 0.242 \Delta Y_{t-2} - 0.240 \Delta Y_{t-3} - 0.119 \Delta Y_{t-4} \\ + 1.427 x_{t-1} - 0.595 x_{t-2} + 0.294 x_{t-3} - 0.174 x_{t-4} + u_{x,t}, \end{aligned} \tag{14}$$

where

$$\begin{bmatrix} u_{y,t} \\ u_{x,t} \end{bmatrix} \overset{iid}{\sim} N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1.792 & 0.244 \\ 0.244 & 1.463 \end{bmatrix} \right). \tag{15}$$

When the unrestricted model (12) is the truth, the DGP is parameterized using Eq. (9) in Clark and McCracken (2005).

$$\begin{aligned} \Delta Y_t = -0.419 \Delta Y_{t-1} - 0.258 \Delta Y_{t-2} \\ + 0.331 x_{t-1} - 0.423 x_{t-2} + 0.309 x_{t-3} - 0.139 x_{t-4} + u_{y,t}, \end{aligned} \tag{16}$$

where x_t is defined as in Eq. (14) and

Table 1 Selection probabilities of the SIC

T	The restricted model is true	The unrestricted model is true
100	0.9901	0.6640
250	0.9977	0.9847
500	0.9997	1
1000	0.9997	1

Table 2 Selection probabilities of the PMSE criterion when the window size is a fixed fraction of the total sample size

π	T	The restricted model is true	The unrestricted model is true
0.2	100	0.9955	0.2326
	250	0.9914	0.9113
	500	0.9907	0.9997
	1000	0.9916	1
0.5	100	0.7459	0.8101
	250	0.7353	0.9845
	500	0.7383	0.9995
	1000	0.7427	1
0.8	100	0.3385	0.8476
	250	0.3682	0.9411
	500	0.3735	0.9841
	1000	0.3719	0.9985

$$\begin{bmatrix} u_{y,t} \\ u_{x,t} \end{bmatrix} \stackrel{iid}{\sim} N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1.517 & 0.244 \\ 0.244 & 1.463 \end{bmatrix} \right), \tag{17}$$

In both (15) and (17), the initial values of ΔY_t and x_t are generated with draws from the unconditional normal distribution. We compare the performance of the SIC and the rolling window PMSE criteria; the latter is implemented with a window size that is either (i) fixed relative to the sample size; (ii) proportional to the sample size; or (iii) diverging slower than the sample size. The number of Monte Carlo replications is set to 10,000. Tables 1, 2, 3, 4 report the empirical probabilities of selecting the correct model. If the procedure is correct, the corresponding probabilities in the tables should be unity.

Tables 1, 2 and 3 report the results for the SIC, the PMSE criterion with W proportional to T , and the PMSE criterion with fixed W , respectively. As expected, the SIC selects the correct model with probability approaching one as the sample size increases. The second last column of Table 2 shows that, when the window size is set to a fraction of the total sample size, $W = [\pi T]$, the PMSE criterion tends to overparameterize the model when π is not very small. When the window size is fixed to a small number ($W = 10$), the PMSE criterion tends to underparameterize the model. The results for $W = [0.2T]$, $W = 50$, and $W = 90$ seem to contradict our claim that these specifications of the window size should yield inconsistent

Table 3 Selection probabilities of the PMSE criterion when the window size is constant

W	T	The restricted model is true.	The unrestricted model is true.
10	100	1	0.0008
	250	1	0
	500	1	0
	1000	1	0
50	100	0.7459	0.8101
	250	0.9914	0.9113
	500	1	0.9729
	1000	1	0.9972
90	100	0.1937	0.8612
	250	0.8959	0.9840
	500	0.9954	0.9990
	1000	1	1

Table 4 Selection probabilities of the PMSE criterion when the window size is slowly diverging

W	T	The restricted model is true	The unrestricted model is true
$T^{1/3}$	100	N/A	N/A
	250	1	0
	500	1	0
	1000	1	0
$T^{1/2}$	100	1	0.0008
	250	1	0.0016
	500	1	0.0532
	1000	1	0.5512
$T^{3/4}$	100	0.9500	0.5947
	250	0.9749	0.9619
	500	0.9883	0.9998
	1000	0.9953	1

model selection; however, for reasonably large sample sizes, these specifications are observationally equivalent to the small window size specification we propose. Table 4 shows the results when the window size is small but diverging, $W = o(T)$. The results for $W = T^{3/4}$ support our consistency results. Although the window size $W = T^{1/3}$ and $W = T^{1/2}$ does not satisfy our sufficient condition (Assumption 1), the resulting criterion chooses the restricted model with probability approaching one when it is true. However, the PMSE criterion with $W = T^{1/3}$ fails to choose the unrestricted model when it is the truth.²

Overall, our results suggest that a window size that is a fixed fraction of the total sample size does not appear to give consistent results when Model 1 is the true data generating process. On the other hand, a constant window size $W = 10$ is not

² When $T = 100$, $W = T^{1/3}$ is too small to compute a rolling estimator.

consistent when Model 2 is true. The divergent window size, in general, consistently selects the correct model, asymptotically. When $W = T^{1/3}$, the consistency is not obvious due to the small window size, but unreported results show that the frequency of consistency will eventually converge to 1 when the total sample size becomes infinitely large.

The SIC does select the correct model asymptotically, and it appears to do so with an even higher probability that the PMSE criterion with a slowly diverging window size. However, as we will show in the next set of Monte Carlo simulations, the SIC will not select the correct model in the presence of time variation.

3.2 Simulation 2: Autoregressive DGP With/Without a Time-Varying Parameter

Next we consider two forecasting models

$$\text{Model 1: } y_t = \alpha y_{t-1} + u_{1,t}$$

$$\text{Model 2: } y_t = \alpha y_{t-1} + \gamma x_t + u_{2,t}$$

where the data are generated by

$$x_t = 0.5x_{t-1} + u_{x,t},$$

$$y_t = 0.5y_{t-1} + \gamma x_t + u_{y,t},$$

$u_{x,t} \sim iid N(0, 1)$ and $u_{y,t} \sim iid N(0, 1)$ are independent of each other. We consider four cases: $\gamma = 0$; $\gamma = 0.25$; $\gamma = 0.5$ and $\gamma = t/T - 0.5$. When $\gamma = 0$ Model 1 is true. Under the cases where $\gamma = 0.5$ or 0.25 , Model 2 is true. Even when $\gamma_{T,t} = t/T - 0.5$, Model 2 should be selected since the true data generating process does include a constant, although the constant is time varying. The number of Monte Carlo replications is set to 10,000.

Tables 5, 6, 7, and 8 report the empirical probabilities of selecting the right model for the SIC and the rolling-window PMSE criterion with $W = [\pi T]$, W being a constant, and $W = o(T)$, respectively, when γ is time invariant. As before, the SIC is consistent and the PMSE criterion tends to either overparameterize or underparameterize the model when W is a large fraction of T or when W is a small constant. The results when W is a small fraction of T ($\pi = 0.2$) or when W is 50 or 90 show that the PMSE criterion selects the correct model. This may be due to finite samples in which these window sizes are consistent with slowly diverging ones. The results in Table 8 show that the PMSE criterion selects the correct forecasting model with probability approaching one as the sample size increases when $W \rightarrow \infty$ and $T^{1/2}/W = O(1)$ as T grows.

The aforementioned results indicate that while the PMSE criterion with a slowly diverging window size is consistent the SIC tends to perform better. One advantage of the PMSE criterion over the SIC is that the PMSE criterion is robust to parameter

Table 5 Selection probabilities of the SIC

T	$\gamma = 0$	$\gamma = 0.25$	$\gamma = 0.5$
100	0.9645	0.7548	0.9989
250	0.9815	0.9826	1
500	0.9881	1	1
1000	0.9926	1	1

Table 6 Selection probabilities of the PMSE criterion when the window size is a fixed fraction of the sample

π	T	$\gamma = 0$	$\gamma = 0.25$	$\gamma = 0.5$
0.2	100	0.9364	0.5497	0.9795
	250	0.9411	0.9360	1
	500	0.9414	0.9981	1
	1000	0.9422	1	1
0.5	100	0.8075	0.7433	0.9759
	250	0.8100	0.9368	0.9998
	500	0.8089	0.9914	1
	1000	0.8182	0.9998	1
0.8	100	0.6724	0.6944	0.8784
	250	0.6787	0.8338	0.9753
	500	0.6882	0.9205	0.9971
	1000	0.6963	0.9800	0.9999

Table 7 Selection probabilities of the PMSE criterion when the window size is constant

W	T	$\gamma = 0$	$\gamma = 0.25$	$\gamma = 0.5$
10	100	0.9859	0.2170	0.8569
	250	0.9998	0.1118	0.9591
	500	1	0.0449	0.9945
	1000	1	0.0054	0.9996
50	100	0.8075	0.7433	0.9759
	250	0.9411	0.9360	1
	500	0.9856	0.9909	1
	1000	0.9982	1	1
90	100	0.6145	0.6421	0.7845
	250	0.8688	0.9568	1
	500	0.9479	0.9980	1
	1000	0.9885	1	1

instabilities. Table 9 reports the selection probabilities of the SIC and PMSE criterion when $\gamma_{T,t} = t/T - 0.5$. $\gamma_{T,t}$ is modeled so that the in-sample PMSE of Model 2 equals that of Model 1 while the out-of-sample PMSE of Model 2 is smaller than that

Table 8 Selection probabilities of the PMSE criterion when the window size is slowly diverging

W	T	$\gamma = 0$	$\gamma = 0.25$	$\gamma = 0.5$
$T^{1/3}$	100	0.9983	0.0092	0.0970
	250	0.9999	0.0040	0.4060
	500	1	0.0008	0.7201
	1000	1	0.0016	0.9959
$T^{1/2}$	100	0.9859	0.2170	0.8569
	250	0.9982	0.4115	0.9987
	500	0.9997	0.7848	1
	1000	1	0.9901	1
$T^{3/4}$	100	0.8909	0.6889	0.9858
	250	0.9213	0.9506	1
	500	0.9361	0.9980	1
	1000	0.9551	1	1

Table 9 Selection probabilities when a parameter is time varying

T	SIC	$W = T^{\frac{1}{3}}$	$W = T^{\frac{1}{2}}$	$W = T^{\frac{2}{3}}$
100	0.0489	0.0063	0.1943	0.4904
250	0.0313	0.0026	0.4567	0.8703
500	0.0215	0.0005	0.8664	0.9953
1000	0.0139	0.0015	0.9982	1.0000

of Model 1. Table 9 shows that the PMSE criterion selects Model 2 with empirical probability approaching one while the SIC selects Model 1.³

To summarize, the Monte Carlo results are consistent with our asymptotic theory and the PMSE criterion with a slowly diverging window size chooses the correct forecasting model with probability approaching one, no matter whether the parameters are time varying or not. On the other hand, although the SIC is consistent when the parameter is constant over time, it is inconsistent when the parameter is time varying.

4 Empirical Application

We consider forecasting quarterly inflation h -periods into the future. Let the regression model be:

$$y_{t+h}^h = \gamma_0 + \gamma_1(L) x_t + \gamma_2(L) y_t + u_{t+h}^h, t = 1, \dots, T \tag{18}$$

³ Technically, the window size $W = T^{2/3}$ does not satisfy our sufficient condition but yields good results.

where the dependent variable is $y_{t+h}^h = (400/h) \ln(P_{t+h}/P_t) - 400 \ln(P_t/P_{t-1})$ where P_t is the price level (CPI) at time t , h is the forecast horizon and equals four, so that the forecasts involve annual percent growth rates of inflation. $\gamma_1(L) = \sum_{j=0}^p \gamma_{1j} L^j$ and $\gamma_2(L) = \sum_{j=0}^q \gamma_{2j} L^j$, where L is the lag operator. Following Stock and Watson (2003), we consider several explanatory variables, x_t , one at a time. The explanatory variable, x_t , is either an interest rate or a measure of real output, unemployment, price, money, or earnings. The data are transformed to eliminate stochastic or deterministic trends and to quarterly frequencies. For a detailed description of the variables that we consider, see Table 10. We utilize quarterly, finally revised data available in January 2011. The earliest starting point of the sample that we consider is January 1959, although both M3 and the exchange rate series have a later starting date due to data availability constraints. Overall, this implies that the total sample size is about 240 observations. In the out-of-sample forecasting exercise, we estimate the number of lags (p and q) recursively by BIC; the estimation scheme is rolling with a window size of 40 observations. The benchmark model is an autoregressive model:

$$y_{t+h}^h = \gamma_0 + \gamma_2(L) y_t + u_{t+h}^h, \quad t = 1, \dots, T. \tag{19}$$

Results are reported in Fig. 1. The figure reports the ratio of the MSFE of the model, Eq. (18), relative to the MSFE of the autoregressive benchmark model, Eq. (19). According to the Monte Carlo simulations in the previous section, the most successful window sizes are between $T^{1/2}$ and $T^{2/3}$, which, given the available sample of data, implies between 16 and 39 observations.

Panel A reports results for predictors (x_t) that include real output measures. It is well known that such measures should be good predictors of inflation according to the Phillips curve. Several studies have documented the empirical success of Phillips curve models, see for example Stock et al. (1999a,b) and 2003, although the empirical results in Marcellino et al. (2003) suggests that the ability of such measures to forecast inflation in Europe is more limited than in the United States. The figure shows that capacity utilization, employment, and unemployment measures are very useful predictors for inflation. In fact, when the window size is less than about 80, the MSFE of the model is always smaller than that of the autoregressive benchmark, sometimes even substantially. Note that for larger window sizes the PMSE criterion would however suggest that the AR benchmark forecasts better than the economic model.

Earnings, instead, is not a successful predictor: in window sizes in the range between $T^{1/2}$ and $T^{2/3}$, it is significantly worse, and occasionally better, although only for larger window sizes. However, recall from the discussion in Sect. 2 that when the window size is large relative to the total sample size, Inoue and Kilian (2005) have shown that the PMSE criterion tends to select overparameterized models. When the window sizes are between $T^{1/2}$ and $T^{2/3}$, the previous sections showed that the PMSE criterion tends to select the correct model. This suggests that earnings are particularly unreliable for forecasting inflation.

Table 10 Series description

Label	Name	Period	Description	Source
<i>Asset Prices</i>				
rovnght@us	FEDFUNDS	1959 M1	Int rate: Fed funds (effective)	F
rtbill@us	TB3MS	1959 M1	Int rate: 3-month treasury bill, Sec Mkt rate	F
rbnds@us	GS1	1959 M1	Int rate: US treasury constant maturity, 1-Yr	F
rbndm@us	GS5	1959 M1	Int rate: US treasury constant maturity, 5-Yr	F
rbndl@us	GS10	1959 M1	Int rate: US treasury constant maturity, 10-Yr	F
stockp@us	SF500	1959 Q1	US share prices: S&P 500	F
exrate@us	I11..NELZF...	1960 M1	NEER from ULC	I
<i>Activity</i>				
rgdp@us	GDPC96	1959 Q1	Real GDP	F
ip@us	INDPRO	1959 M1	Industrial Prod. index, (sa)	F
capu@us	CAPUB04	1959 M1	Capacity utilization rate: manufacturing (sa)	F
emp@us	CE16OV	1959 M1	Civilian employment: thou. persons	F
unemp@us	UNRATE	1959 M1	Civilian Unemp rate (sa)	F
<i>Money</i>				
mon0@us	AMBSL	1959 M1	Monetary base: St. Louis Adj. (sa)	F
mon1@us	M1SL	1959 M1	Money: M1 (sa)	F
mon2@us	M2SL	1959 M1	Money: M2 (sa)	F
mon3@us	M3SL	1959 M1	Money: M3 (sa)	F
<i>Wages and Prices</i>				
ppi@us	PPIACO	1959 M1	Producer price index	F
earn@us	AHEMAN	1959 M1	Hourly earnings: Manufact. (msa)	F

Notes Sources are abbreviated as follows: D, datastream; F, federal reserve economic data; I, IMF international financial statistics; O, OECD main economic indicators; G, global insight

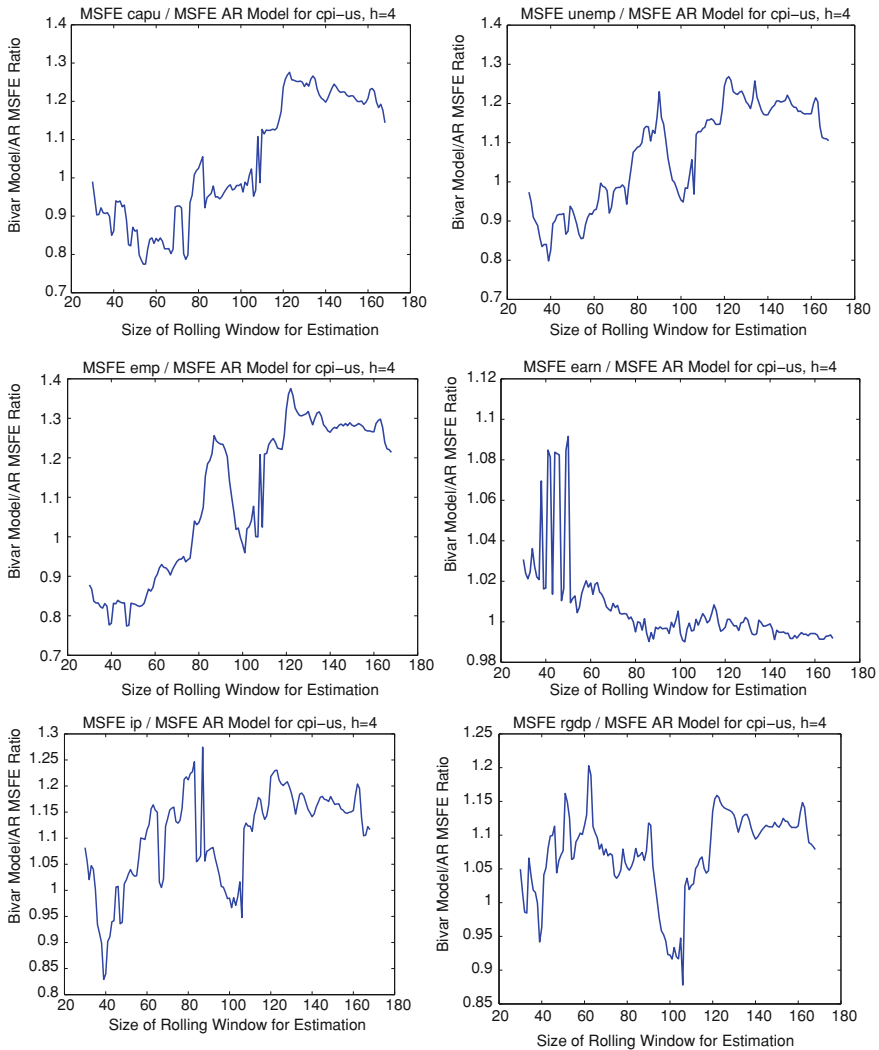


Fig. 1 QLR break test

The performance of industrial production and real GDP predictors, instead, is less clear: the ratio can be either above or below unity depending on the window size. Even for window sizes in the range between $T^{1/2}$ and $T^{2/3}$, the ratio can be either above or below unity. These results suggest instabilities in the forecasting performance of these predictors, and are consistent with the results in Rossi and Sekhposyan (2010), although the latter were interested in testing equal predictive ability rather than consistently selecting the correct model, as we do here. Rossi and Sekhposyan (2010) empirical evidence documented that the economic predictors have forecasting

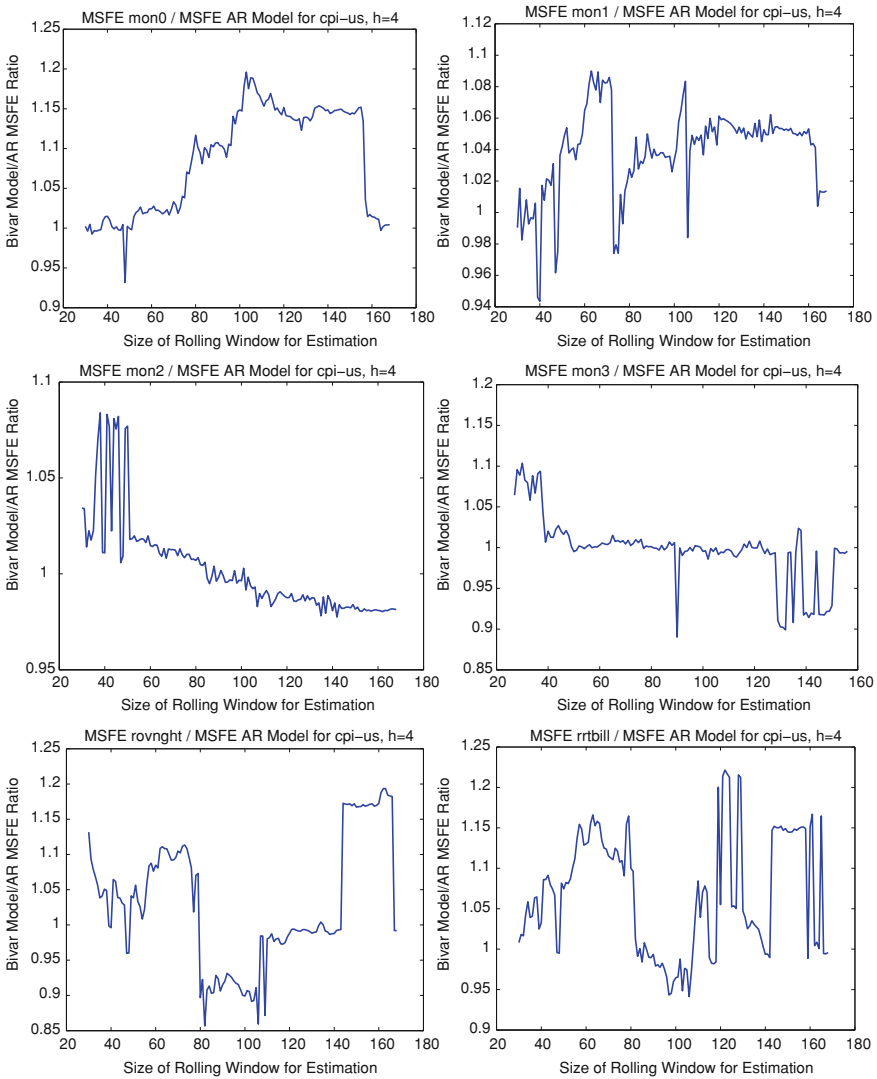


Fig. 1 continued

ability in the early part of their sample, but the predictive ability disappears in the later part of their sample. The reversals in predictive ability happened, according to their tests, around the time of the Great Moderation, which the literature dates back to 1983–1984 (see McConnell and Perez-Quiros 2000), similar to the results in D’Agostino et al. (2006).

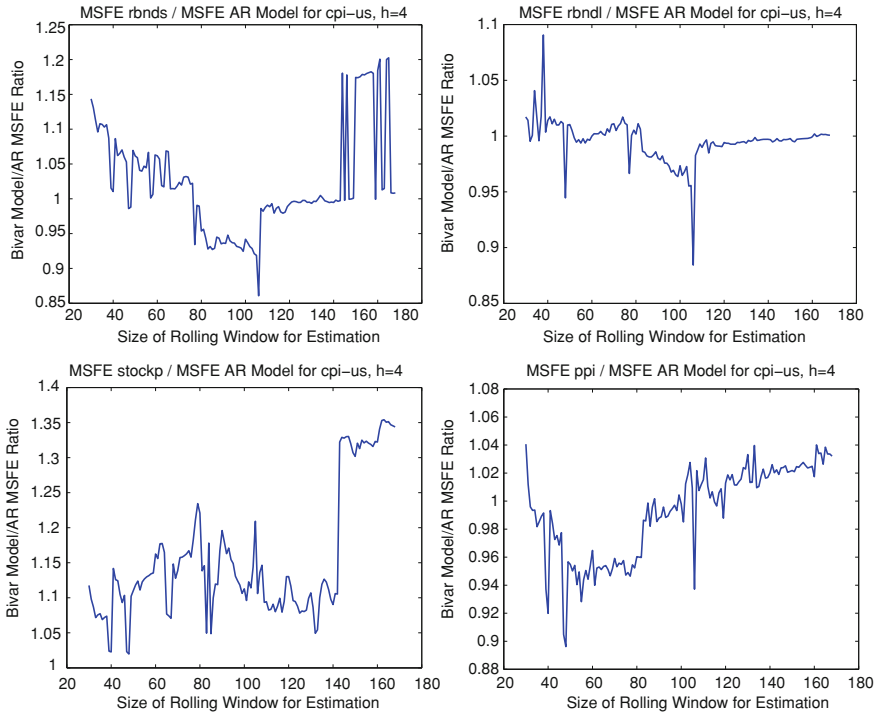


Fig. 1 continued

Panel B focuses on monetary measures. M1, M2, and M3 never have predictive ability except for some selected window sizes, again pointing to the presence of instabilities.

Panel C focuses on interest rates. The results are quite interesting. They show that interest rates (such as 1-year or 10-year bonds) appear to be very good predictors of inflation for medium window sizes, below 120–140 observations. Again, however, for very large window sizes the PMSE criterion would select the smaller model. Short-term interest rates tend to be useful predictors only when the window size is large, but again the ratio is below unity for some selected window sizes and above unity for others. Again, we conjecture that instabilities are important, as discussed in Rossi and Sekhposyan (2010).

Panel D focuses on other monetary variables. Stock prices are never useful for predicting inflation. Interestingly, the producer price index is a good predictor for inflation: the figure shows that for the relevant window sizes, the ratio of the MSFE of the model relative to that of the benchmark is always lower than unity, and it becomes higher than unity only for large window sizes.

Overall, our empirical results suggest that traditional Phillips curve predictors such as capacity utilization and unemployment are useful in forecasting inflation, as well as the producer price index. The empirical results for the other macroeconomic

Table 11 QLR break test P-values

Indicator	P-value
<i>A. Real output measures</i>	
Capacity utilization	0.00
Unemployment	0.00
Employment	0.00
Earnings	0.00
Industrial production	0.05
Real GDP	0.00
<i>B. Money measures</i>	
M0	0.00
M1	0.00
M2	0.00
M3	0.00
<i>C. Interest rates</i>	
Fed funds	0.00
Real 3-mo. Treasury bill	0.00
1-Year bond	0.04
10-Year bond	0.04
<i>D. Other nominal measures</i>	
Stock prices	0.03
Producer price index	0.00

Notes The table reports results for Andrews (1993) QLR test for structural breaks implemented with a HAC covariance estimator with a bandwidth equal to $(1/5)T$

predictors are not clearcut, and might signal the importance of instabilities in the data. In order to provide more information on the instability in the forecasting regressions we consider, we report joint tests for structural breaks in the parameters of Eq. (18) using Andrews (1993) test for structural breaks. Table 11 reports the p-values of the test, which confirm that instabilities are extremely important.

5 Concluding Remarks

There is a known break, forecasters tend to use post-break observations when they make forecasts. In other words, they base their forecasts on a “truncated window” instead of the full sample. This chapter shows that this type of ideas can deliver the consistency of the rolling PMSE criterion not only when parameters are time varying but also when they are constant over time.

In this chapter we focus on the rolling scheme. Inoue and Kilian (2006) show that the PMSE criterion based on the recursive scheme is inconsistent if the number of initial observations is large, i.e., a fixed fraction of the sample size, while Wei (1992) proves that it is consistent if the number of initial observations is very small, i.e., a fixed constant. One might be able to extend Wei (1992) result to the case in which the number of initial observations diverges at a rate slower than the sample

size. However, such a model selection criterion might not be robust to parameter instability.

It should be noted that our consistency results are based on correctly specified nested models. Although information criteria are not robust to parameter instabilities, they are robust to misspecification and nonnestedness (Sin and White 1996). We leave PMSE criterion-based model comparison of misspecified or non-nested models for future research.

The main object of forecasters is often to minimize PMSE rather than identify the true model. We are currently developing a data-dependent method for choosing the window size to achieve this goal in a separate chapter

Appendix

A.1 Lemmas

Next, we present a lemma similar to Lemma A2 of Clark and McCracken (2000).

Lemma 1 *Suppose that Assumptions 1 and 2 hold and that $\gamma = 0$. Then:*

- (a) $\frac{1}{T-h-W} \sum_{t=W+1}^{T-h} u_{t+h} x_t B_1(t) H_1(t) = o_p\left(\frac{1}{W}\right).$
- (b) $\frac{1}{T-h-W} \sum_{t=W+1}^{T-h} v_{t+h} z_t B_2(t) H_2(t) = o_p\left(\frac{1}{W}\right).$
- (c) $\frac{1}{T-h-W} \sum_{t=W+1}^{T-h} H_1'(t) B_1(t) x_t x_t' B_1(t) H_1(t) = \frac{1}{T-h-W} \sum_{t=W+h}^{T-1} H_1'(t) B_1 H_1(t) + o_p\left(\frac{1}{W}\right).$
- (d) $\frac{1}{T-h-W} \sum_{t=W+1}^{T-h} H_2'(t) B_2(t) z_t z_t' B_2(t) H_2(t) = \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} H_2'(t) B_2 H_2(t) + o_p\left(\frac{1}{W}\right).$

Proof of Lemma 1: The proofs for (a) and (c) are very similar to those for (b) and (d), respectively. For brevity, we only provide the proofs of (b) and (d). The results for (a) and (c) can be easily derived by replacing z_t and β by x_t and α , respectively.

Note that

$$\begin{aligned} \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} v_{t+h} z_t B_2(t) H_2(t) &= \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} v_{t+h} z_t B_2 H_2(t) \\ &\quad + \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} v_{t+h} z_t (B_2(t) - B_2) H_2(t) \end{aligned}$$

By Assumption 2(b) and Hölder's inequality, the second moments of the summands on the right-hand side are of order $O(W^{-1})$ and $O(W^{-2})$, respectively. Thus, it follows from Assumption 2(c) that the variance of the left-hand side is of order

$O(T^{-1}W^{-1})$. By the Chebyshev inequality and Assumption 1, the left-hand side is $o_p(W^{-1})$.

The proof of (d) is composed of two stages. In the first stage, we show that $B_2(t)$ in the equation can be approximated by its expectation B_2 , which is

$$\begin{aligned} & \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} H_2'(t) B_2(t) z_t z_t' B_2(t) H_2(t) \\ &= \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} H_2'(t) B_2 z_t z_t' B_2 H_2(t) + o_p\left(\frac{1}{W}\right) \end{aligned} \tag{A.1}$$

Since the left-hand side of Eq.(A.1) contains four terms,

$$\begin{aligned} & \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} H_2'(t) B_2(t) z_t z_t' B_2(t) H_2(t) \\ &= \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} H_2'(t) B_2 z_t z_t' B_2 H_2(t) \\ & \quad + \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} H_2'(t) (B_2(t) - B_2) z_t z_t' (B_2(t) - B_2) H_2(t) \\ & \quad + \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} H_2'(t) B_2 z_t z_t' (B_2(t) - B_2) H_2(t) \\ & \quad + \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} H_2'(t) (B_2(t) - B_2) z_t z_t' B_2 H_2(t), \end{aligned} \tag{A.2}$$

which include the first term in the right-hand side of Eq. (A.1).

By Assumption 2(b) and Hölder’s inequality, the second moments of the summands in the last three terms are of order $O(W^{-4})$, $O(W^{-3})$, and $O(W^{-3})$, respectively. Thus, their first moments are at most $O(W^{-3}) = o(W^{-1})$. By using these and Assumption 2(e), the second moments of the last three terms are thus of the order $O(T^{-1}W^{-4})$, $O(T^{-1}W^{-3})$ and $O(T^{-1}W^{-1})$, respectively. By the Chebyshev inequality and Assumption 1, these last three terms are of the order $o_p(W^{-1})$, proving (A.1).

The second stage of the proof of (d) is to show that we can further approximate $z_t z_t'$ in the first term in the right-hand side of Eq.(A.2) by its expectation $E(z_t z_t')$. Adding and subtracting $E(z_t z_t')$, we obtain

$$\frac{1}{T-h-W} \sum_{t=W+1}^{T-h} H_2'(t) B_2 z_t z_t' B_2 H_2(t)$$

$$\begin{aligned}
 &= \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} H_2'(t) B_2 E(z_t z_t') B_2 H_2(t) \\
 &\quad + \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} H_2'(t) B_2 (z_t z_t' - E(z_t z_t')) B_2 H_2(t) \tag{A.3}
 \end{aligned}$$

The mean of the second term is $o_p(W^{-1})$ by Assumption 2(d). The second moments of the summand in the second term is $O(W^{-2})$ by Assumption 2(b). Using these and Assumption 2(e), the second moment of the second term is of the order $o(W^{-2})$. By the Chebyshev inequality, (A.3) is $o_p(W^{-1})$.

Lemma 2 *Suppose that Assumptions 3 and 4 hold and that $\gamma(\cdot) = 0$.*

- (a) $\frac{1}{T-h-W} \sum_{t=W+1}^{T-h} u_{T,t+h} x_{T,t} B_1(t) H_1(t) = o_p\left(\frac{1}{W}\right)$.
- (b) $\frac{1}{T-h-W} \sum_{t=W+1}^{T-h} v_{T,t+h} z_{T,t} B_2(t) H_2(t) = o_p\left(\frac{1}{W}\right)$.
- (c) $\frac{1}{T-h-W} \sum_{t=W+1}^{T-h} H_1'(t) B_1(t) x_{T,t} x_{T,t}' B_1(t) H_1(t)$
 $= \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} H_1'(t) \bar{B}_1\left(\frac{t}{T}\right) H_1(t) + o_p\left(\frac{1}{W}\right)$.
- (d) $\frac{1}{T-h-W} \sum_{t=W+1}^{T-h} H_2'(t) B_2(t) z_{T,t} z_{T,t}' B_2(t) H_2(t)$
 $= \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} H_2'(t) \bar{B}_2\left(\frac{t}{T}\right) H_2(t) + o_p\left(\frac{1}{W}\right)$.

Proof of Lemma 2 Under Assumptions 3 and 4 the proof of Lemma 2 takes exactly the same steps as the proof of Lemma 1 except that B_i , u_t , and v_t are replaced by $\bar{B}_i\left(\frac{t}{T}\right)$, $u_{T,t}$, and $v_{T,t}$, respectively. This is because Lemma 2 is written in terms of $u_{T,t}$ and $v_{T,t}$ rather than in terms of $\hat{\alpha}_{t,W} - \alpha\left(\frac{t}{T}\right)$ and $\hat{\beta}_{t,W} - \beta\left(\frac{t}{T}\right)$ which we deal with in the proof of Theorem 2.

A.2 Proofs of Theorems

Proof of Theorem 1 Note that the PMSEs $\hat{\sigma}_{1,W}^2$ and $\hat{\sigma}_{2,W}^2$ can be expanded as

$$\begin{aligned}
 \hat{\sigma}_{1,W}^2 &= \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} (y_{t+h} - \hat{\alpha}'_t x_t)^2 \\
 &= \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} (y_{t+h} - \alpha^{*'} x_t - (\hat{\alpha}'_t x_t - \alpha^{*'} x_t))^2
 \end{aligned}$$

$$\begin{aligned}
 &= \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} (y_{t+h} - \alpha^{*'} x_t)^2 \\
 &\quad - \frac{2}{T-h-W} \sum_{t=W+1}^{T-h} (y_{t+h} - \alpha^{*'} x_t) x_t' (\hat{\alpha}_t - \alpha^*) \\
 &\quad + \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} (\hat{\alpha}_t' - \alpha^{*'}) x_t x_t' (\hat{\alpha}_t - \alpha^*) \tag{A.4}
 \end{aligned}$$

and

$$\begin{aligned}
 \hat{\sigma}_{2,W}^2 &= \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} (y_{t+h} - \hat{\beta}_t' z_t)^2 \\
 &= \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} (y_{t+h} - \beta' z_t - (\hat{\beta}_t' z_t - \beta' z_t))^2 \\
 &= \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} (y_{t+h} - \beta' z_t)^2 \\
 &\quad - \frac{2}{T-h-W} \sum_{t=W+1}^{T-h} (y_{t+h} - \beta' z_t) z_t' (\hat{\beta}_t - \beta) \\
 &\quad + \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} (\hat{\beta}_t' - \beta') z_t z_t' (\hat{\beta}_t - \beta), \tag{A.5}
 \end{aligned}$$

respectively, where $\alpha^* = [E(x_t x_t')]^{-1} E(x_t y_{t+h})$. There are two cases: the case in which the data are generated from model 1, i.e., $\gamma = 0$ (case 1) and the case in which the data are generated from model 2, i.e., $\gamma \neq 0$ (case 2).

In case 1, the actual model is $y_{t+h} = \alpha' x_t + v_{t+h}$. The first component of $\hat{\sigma}_{2,W}^2$ in Eq. (A.5) is numerically identical to the first component of $\hat{\sigma}_{1,W}^2$ in Eq. (A.4) because $\gamma = 0$ and $\alpha - \alpha^* = 0$. Note that all the other components converge to zero faster since all parameters are consistently estimated. Under the case where Model 1 is true, the difference between the probability limit of $\hat{\sigma}_{1,W}^2$ and $\hat{\sigma}_{2,W}^2$ is zero, which does not identify which model is the true model. Only comparing the probability limits of $\hat{\sigma}_{1,W}^2$ and $\hat{\sigma}_{2,W}^2$ as T and W go to infinity and W diverges slowly than T is not sufficient for the model selection to indicate that $\lim_{T \rightarrow \infty, W \rightarrow \infty} P(\hat{\sigma}_{1,W}^2 < \hat{\sigma}_{2,W}^2) = 1$. However, if we can tell whether $\hat{\sigma}_{1,W}^2$ is always smaller than $\hat{\sigma}_{2,W}^2$ along the path of convergence of T and W toward infinity, the true model can still be identified. Since the models are nested $u_{t+h} = v_{t+h}$, it follows from (A.4) and (A.5) that

$$\begin{aligned}
\hat{\sigma}_{2,W}^2 - \hat{\sigma}_{1,W}^2 &= \frac{2}{T-h-W} \sum_{t=W+1}^{T-h} [v_{t+h} z_t' (\hat{\beta}_t - \beta) - v_{t+h} x_t' (\hat{\alpha}_t - \alpha)] \\
&\quad + \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} [(\hat{\beta}_t' - \beta') z_t z_t' (\hat{\beta}_t - \beta) \\
&\quad \quad - (\hat{\alpha}_t' - \alpha') x_t x_t' (\hat{\alpha}_t - \alpha)] \\
&= \frac{2}{T-h-W} \sum_{t=W+1}^{T-h} [v_{t+h} z_t' B_2(t) H_2(t) - v_{t+h} x_t' B_1(t) H_1(t)] \\
&\quad + \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} [H_2(t)' B_2(t) z_t z_t' B_2(t) H_2(t) \\
&\quad \quad - H_1(t)' B_1(t) x_t x_t' B_1(t) H_1(t)] \\
&= \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} [H_2(t)' B_2 H_2(t) - H_1(t)' B_1 H_1(t)] + o_p\left(\frac{1}{W}\right)
\end{aligned} \tag{A.6}$$

where the last equality follows from Lemma 1(a)–(d).

To get the sign of Eq. (A.6), we first define Q by

$$Q = [E(z_t z_t')]^{\frac{1}{2}} \left\{ [E(z_t z_t')]^{-1} - \begin{bmatrix} [E(x_t x_t')]^{-1} & \mathbf{0}_{l \times (k-l)} \\ \mathbf{0}_{(k-l) \times l} & \mathbf{0}_{(k-l) \times (k-l)} \end{bmatrix} \right\} [E(z_t z_t')]^{\frac{1}{2}} \tag{A.7}$$

as in Lemma A.4 of Clark and McCracken (2000). Clark and McCracken (2000) show that the Q matrix is symmetric and idempotent. An idempotent matrix is positive semidefinite, which means for all $v \in \mathfrak{R}^k$, $v^T Q v \geq 0$. It implies that

$$\begin{aligned}
&\left[\frac{1}{W_h^{\frac{1}{2}}} \sum_{s=t-W}^{t-h} z_s v_{s+h} \right]' [E(z_t z_t')]^{-1} \left[\frac{1}{W_h^{\frac{1}{2}}} \sum_{s=t-W}^{t-h} z_s v_{s+h} \right] \\
&\quad - \left[\frac{1}{W_h^{\frac{1}{2}}} \sum_{s=t-W}^{t-h} x_s v_{s+h} \right]' [E(x_t x_t')]^{-1} \left[\frac{1}{W_h^{\frac{1}{2}}} \sum_{s=t-W}^{t-h} x_s v_{s+h} \right] \\
&= \left[\frac{1}{W_h^{\frac{1}{2}}} \sum_{s=t-W}^{t-h} z_s v_{s+h} \right]' \left\{ [E(z_t z_t')]^{-1} - \begin{bmatrix} [E(x_t x_t')]^{-1} & \mathbf{0}_{l \times (k-l)} \\ \mathbf{0}_{(k-l) \times l} & \mathbf{0}_{(k-l) \times (k-l)} \end{bmatrix} \right\} \\
&\quad \times \left[\frac{1}{W_h^{\frac{1}{2}}} \sum_{s=t-W}^{t-h} z_s v_{s+h} \right]
\end{aligned}$$

$$\begin{aligned}
 &= \left[\frac{1}{W_h^{\frac{1}{2}}} \sum_{s=t-W}^{t-h} z_s v_{s+h} \right]' [E(z_t z_t')]^{-\frac{1}{2}} \cdot Q \cdot [E(z_t z_t')]^{-\frac{1}{2}} \\
 &\quad \times \left[\frac{1}{W_h^{\frac{1}{2}}} \sum_{s=t-W}^{t-h} z_s v_{s+h} \right] \geq 0 \tag{A.8}
 \end{aligned}$$

Note that the probability that $[E(z_t z_t')]^{-1/2} W_h^{-1/2} \sum_{s=t-W}^{t-h} z_s v_{s+h}$ lies in the null space of Q for infinitely many t approaches zero because the dimension of the null space is $l < k$. Thus, the average of (A.8) over t is positive with probability approaching one. Combining the results in Eqs. (A.6) and (A.8), we find that the sign of $W(\hat{\sigma}_{2,W}^2 - \hat{\sigma}_{1,W}^2)$ is always positive with probability approaching one. Therefore, when $\gamma = 0$, $\hat{\sigma}_{1,W}^2 < \hat{\sigma}_{2,W}^2$ with probability approaching one.

In case 2, that is, when Model 2 is the true model, we have $y_{t+h} = \beta' z_t + v_{t+h} = \alpha' x_t + \gamma' w_t + v_{t+h}$. By Assumptions 2(a)(b), the second and third terms on the right-hand sides of (A.4) and (A.5) are both $o_p(T^{1/2}/W)$ and $o_p(T/W^2)$, respectively. Thus, they are $o_p(1)$ by Assumption 1. The first term on the right-hand side of Eq. (A.5) converges to the variance of v_{t+h} as the sample size T goes to infinity:

$$\frac{1}{T-h-W} \sum_{t=W+1}^{T-h} (y_{t+h} - \beta' z_t)^2 = \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} v_{t+h}^2 \xrightarrow{p} \sigma_2^2. \tag{A.9}$$

Similarly, the first term on the right-hand side of Eq. (A.4) converges in probability to the variance of $u_{t+h} \equiv y_{t+h} - \alpha^* x_t$:

$$\begin{aligned}
 \hat{\sigma}_{1,W}^2 &= \frac{1}{T-h-W} \sum_{t=W+h}^{T-1} (y_{t+h} - \alpha^* x_t)^2 + o_p(1) \\
 &\xrightarrow{p} E \left[(y_{t+h} - \alpha^* x_t)^2 \right] \\
 &= E \left[(\alpha' x_t + \gamma' w_t + v_{t+h} - \alpha^* x_t)^2 \right] \\
 &= E \left[(v_{t+h} + (\alpha' - \alpha^*) x_t + \gamma' w_t)^2 \right] \\
 &= \sigma_2^2 + \begin{bmatrix} \alpha - \alpha^* \\ \gamma \end{bmatrix}' \begin{bmatrix} E(x_t x_t') & E(x_t w_t') \\ E(w_t x_t') & E(w_t w_t') \end{bmatrix} \begin{bmatrix} \alpha - \alpha^* \\ \gamma \end{bmatrix} > \sigma_2^2. \tag{A.10}
 \end{aligned}$$

Therefore, when Model 2 is true, the PMSEs satisfy $P(\hat{\sigma}_{1,W}^2 > \hat{\sigma}_{2,W}^2) = 1$ as $T \rightarrow \infty$ and $W \rightarrow \infty$, where W diverges slower than T .

Proof of Theorem 2 Note that the PMSEs, $\hat{\sigma}_{1,W}^2$ and $\hat{\sigma}_{2,W}^2$ can be expanded as

$$\begin{aligned}
\hat{\sigma}_{1,W}^2 &= \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} \left(y_{T,t+h} - \alpha^* \left(\frac{t}{T} \right)' x_{T,t} \right)^2 \\
&\quad - \frac{2}{T-h-W} \sum_{t=W+1}^{T-h} \left(y_{T,t+h} - \alpha^* \left(\frac{t}{T} \right)' x_{T,t} \right) x'_{T,t} \left(\hat{\alpha}_t - \alpha^* \left(\frac{t}{T} \right) \right) \\
&\quad + \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} \left(\hat{\alpha}'_t - \alpha^* \left(\frac{t}{T} \right) \right)' x_{T,t} x'_{T,t} \left(\hat{\alpha}_t - \alpha^* \left(\frac{t}{T} \right) \right)
\end{aligned} \tag{A.11}$$

and

$$\begin{aligned}
\hat{\sigma}_{2,W}^2 &= \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} \left(y_{T,t+h} - \beta' \left(\frac{t}{T} \right) z_{T,t} \right)^2 \\
&\quad - \left(\frac{2}{T-h-W} \right) \sum_{t=W+1}^{T-h} \left(y_{T,t+h} - \beta' \left(\frac{t}{T} \right) z_{T,t} \right) z'_{T,t} \left(\hat{\beta}_t - \beta \left(\frac{t}{T} \right) \right) \\
&\quad + \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} \left(\hat{\beta}'_t - \beta' \left(\frac{t}{T} \right) \right)' z_t z'_{T,t} \left(\hat{\beta}_t - \beta \left(\frac{t}{T} \right) \right),
\end{aligned} \tag{A.12}$$

respectively. If we show that each of

$$\begin{aligned}
&\frac{1}{T-h-W} \sum_{t=W+1}^{T-h} \left(y_{T,t+h} - \alpha^* \left(\frac{t}{T} \right)' x_{T,t} \right) x'_{T,t} \left(\hat{\alpha}_t - \alpha^* \left(\frac{t}{T} \right) \right) \\
&\quad - \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} u_{T,t+h} x'_{T,t} B_1(t) H_1(t),
\end{aligned} \tag{A.13}$$

$$\begin{aligned}
&\frac{1}{T-h-W} \sum_{t=W+1}^{T-h} \left(\hat{\alpha}'_t - \alpha^* \left(\frac{t}{T} \right) \right)' x_{T,t} x'_{T,t} \left(\hat{\alpha}_t - \alpha^* \left(\frac{t}{T} \right) \right) \\
&\quad - \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} H_1(t)' B_1(t) z_{T,t} z'_{T,t} B_2(t) H_2(t),
\end{aligned} \tag{A.14}$$

$$\begin{aligned}
&\frac{1}{T-h-W} \sum_{t=W+1}^{T-h} \left(y_{T,t+h} - \beta \left(\frac{t}{T} \right)' z_{T,t} \right) z'_{T,t} \left(\hat{\beta}_t - \beta \left(\frac{t}{T} \right) \right) \\
&\quad - \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} v_{T,t+h} z'_{T,t} B_2(t) H_2(t),
\end{aligned} \tag{A.15}$$

$$\begin{aligned} & \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} \left(\hat{\beta}'_t - \beta \left(\frac{t}{T} \right) \right)' z_{T,t} z'_{T,t} \left(\hat{\beta}_t - \beta \left(\frac{t}{T} \right) \right) \\ & - \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} H_2(t)' B_2(t) z_{T,t} z'_{T,t} B_2(t) H_2(t), \end{aligned} \tag{A.16}$$

are $o_p(1/W)$ when the data are generated from model 1 (case 1) and are $o_p(1)$ when the data are generated from model 2 (case 2), the proof of Theorem 2 takes exactly the same steps as the proof of Theorem 1. Thus, it remains to show that (A.13)–(A.16) are $o_p(W^{-1})$ in case 1 and $o_p(1)$ in case 2. Note that the bias of the rolling regression estimator can be written as:

$$\begin{aligned} \hat{\beta}_{W,t} - \beta \left(\frac{t}{T} \right) &= B_2(t) \frac{1}{W_h} \sum_{s=t-W}^{t-h} z_s \left[v_{s+h} + z'_s \left(\beta \left(\frac{s}{T} \right) - \beta \left(\frac{t}{T} \right) \right) \right] \\ &= B_2(t) H_2(t) + \frac{B_2(t)}{W_h} \sum_{s=t-W}^{t-h} z_s z'_s \left(\beta \left(\frac{s}{T} \right) - \beta \left(\frac{t}{T} \right) \right) \end{aligned} \tag{A.17}$$

Thus, the difference (A.15) is

$$\begin{aligned} & \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} v_{T,t+h} z'_{T,t} B_2(t) \frac{1}{W_h} \sum_{s=t-W}^{t-h} z_s z'_s \left(\beta \left(\frac{s}{T} \right) - \beta \left(\frac{t}{T} \right) \right). \\ &= \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} v_{T,t+h} z'_{T,t} \bar{B}_2 \left(\frac{t}{T} \right) \frac{1}{W_h} \sum_{s=t-W}^{t-h} z_s z'_s \left(\beta \left(\frac{s}{T} \right) - \beta \left(\frac{t}{T} \right) \right) \\ &+ \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} v_{T,t+h} z'_{T,t} \left(B_2(t) - \bar{B}_2 \left(\frac{t}{T} \right) \right) \frac{1}{W_h} \\ &\times \sum_{s=t-W}^{t-h} z_s z'_s \left(\beta \left(\frac{s}{T} \right) - \beta \left(\frac{t}{T} \right) \right). \end{aligned} \tag{A.18}$$

By Assumption 4(c), the summands have zero mean. By Hölder’s inequality and Assumptions 4(b)(c)(e)(f), the second moments of the right-hand side terms are $O(W/T^2)$. By Chebyshev’s inequality, (A.15) is $O_p(W^{1/2}/T)$ which is $o_p(1/W)$ by Assumption 3. It can be shown that (A.13) is also $o_p(1/W)$ in a similar fashion.

The difference (A.16) is the sum of the following three terms:

$$\frac{1}{T-h-W} \sum_{t=W+1}^{T-h} v_{T,t+h} z_{T,t} z'_{T,t} B_2(t) \frac{1}{W_h} \sum_{s=t-W}^{t-h} z_s z'_s \left(\beta \left(\frac{s}{T} \right) - \beta \left(\frac{t}{T} \right) \right), \tag{A.19}$$

$$\frac{1}{T-h-W} \sum_{t=W+1}^{T-h} \left(\beta \left(\frac{s}{T} \right) - \beta \left(\frac{t}{T} \right) \right)' \frac{1}{W_h} \sum_{s=t-W}^{t-h} z_s z_s' B_2(t) z_{T,t} z_{T,t}' v_{T,t+h}, \tag{A.20}$$

$$\begin{aligned} & \frac{1}{T-h-W} \sum_{t=W+1}^{T-h} \left(\beta \left(\frac{s}{T} \right) - \beta \left(\frac{t}{T} \right) \right)' \frac{1}{W_h} \sum_{s=t-W}^{t-h} z_s z_s' B_2(t) z_{T,t} \\ & \times z_{T,t}' B_2(t) \frac{1}{W_h} \sum_{s=t-W}^{t-h} z_s z_s' \left(\beta \left(\frac{s}{T} \right) - \beta \left(\frac{t}{T} \right) \right), \end{aligned} \tag{A.21}$$

Using Chebyshev’s inequality, Hölder’s inequality, Assumptions 3 and 4(b)(c)(e)(f), it can be shown that (A.19), (A.20), and (A.21) are $O_p(W^{1/2}T^{-2})$, $O_p(W^{1/2}T^{-2})$ and $O_p(W^2T^{-2})$ all of which are $o_p(W^{-1})$. It can be shown that (A.14) is also $o_p(1/W)$ when $\gamma(\cdot) = 0$ in an analogous fashion.

The rest of the proof of Theorem 2 takes exactly the same steps as the proof of Theorem 1 except that α^* , β , B_i , u_t , v_t , x_t , y_t , z_t and Lemma 1 is replaced by $\alpha \left(\frac{t}{T} \right)$, $\beta \left(\frac{t}{T} \right)$, $\bar{B}_i \left(\frac{t}{T} \right)$, u_{Tt} , v_{Tt} , x_{Tt} , y_{Tt} , z_{Tt} and Lemma 2, respectively.

References

Andrews, D.W.K. (1993), “Tests for Parameter Instability and Structural Change with Unknown Change Point”, *Econometrica* 61, 821–856.

D’Agostino, A., D. Giannone, and P. Surico (2006), “(Un)Predictability and Macroeconomic Stability” ECB Working Paper 605.

Cai, Z., (2007), “Trending Time-Varying Coefficient Time Series Models with Serially Correlated Errors”, *Journal of Econometrics*, 136, 163–188.

Clark, T.E., and M.W. McCracken (2000), “Not-for-Publication Appendix to ” Tests of Equal Forecast Accuracy and Encompassing for Nested Models, unpublished manuscript, Federal Reserve Bank of Kansas City and Louisiana State University.

Clark, T.E., and M.W. McCracken (2001) “Tests of Equal Forecast Accuracy and Encompassing for Nested Models”, *Journal of Econometrics*, 105, 85–110.

Clark, T.E., and M.W. McCracken (2005), “Evaluating Direct Multistep Forecasts”, *Econometric Reviews*, 24, 369–404.

Gallant, A.R., and H. White (1988), *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*, Basil Blackwell: New York, NY.

Giacomini, R. and B. Rossi (2010), “Forecast Comparisons in Unstable Environments”, *Journal of Applied Econometrics* 25(4), 595–620.

Giacomini, R. and H. White (2006), “Tests of Conditional Predictive Ability”, *Econometrica* 74(6), 1545–1578.

Giraitis, L., G. Kapetanios and T. Yates (2011), “Inference on Stochastic Time-Varying Coefficient Models”, unpublished manuscript, Queen Mary, University of London, and the Bank of England.

Hall, P., and C.C. Heyde (1980), *Martingale Limit Theory and its Application*, Academic Press: San Diego CA.

Inoue, A., and L. Kilian (2006), “On the Selection of Forecasting Models”, *Journal of Econometrics*, 130, 273–306.

- Marcellino, Massimiliano, James H. Stock and Mark W. Watson (2003), "Macroeconomic Forecasting in the Euro Area: Country-Specific vs. Area-Wide Information", *European Economic Review*, 47(1), pages 1–18.
- McConnell, M.M., and G. Perez-Quiros (2000), "Output Fluctuations in the United States: What Has Changed Since the Early 1980" *American Economic Review*, 90(5), 1464–1476.
- Meese, R. and K.S. Rogoff (1983a), "Exchange Rate Models of the Seventies. Do They Fit Out of Sample?", *The Journal of International Economics* 14, 3–24.
- Meese, R. and K.S. Rogoff (1983b), "The Out of Sample Failure of Empirical Exchange Rate Models", in Jacob Frankel (ed.), *Exchange Rates and International Macroeconomics*, Chicago: University of Chicago Press for NBER.
- Rossi, B., and A. Inoue (2011), "Out-of-Sample Forecast Tests Robust to the Choice of Window Size", mimeo.
- Rossi, B., and T. Sekhposyan (2010), "Have ModelsTM Forecasting Performance Changed Over Time, and When?", *International Journal of Forecasting*, 26(4).
- Sin, C.-Y., and H. White (1996), "Information Criteria for Selecting Possibly Misspecified Parametric Models", *Journal of Econometrics*, 71, 207–225.
- Stock, James H. and Mark W. Watson (1999a), "Business Cycle Fluctuations in U.S. Macroeconomic Time Series", in *Handbook of Macroeconomics*, Vol. 1, J.B. Taylor and M. Woodford, eds, Elsevier, 3–64.
- Stock, James H. and Mark W. Watson (1999b), "Forecasting Inflation", *Journal of Monetary Economics*, 44, 293–335.
- Swanson, N.R., and H. White (1997), "A Model Selection Approach to Real-Time Macroeconomic Forecasting Using Linear Models and Artificial Neural Networks", *Review of Economics and Statistics*.
- Wei, C.Z. (1992), "On Predictive Least Squares Principles", *Annals of Statistics*, 20, 1–42.
- West, K.D. (1996), "Asymptotic Inference about Predictive Ability", *Econometrica* 64, 1067–1084.
- Wooldridge, J.M., and H. White (1988), "Some Invariance Principles and Central Limit Theorems for Dependent Heterogeneous Processes", *Econometric Theory* 4, 210–230.
- Wooldridge, J.M. (1994), "Estimation and Inference for Dependent Processes", in the *Handbook of Econometrics*, Volume IV, Edited by R.F. Engle and D.L. McFadden, Chapter 45, pp. 2639–2738.