Journal of Econometrics 158 (2010) 142–155

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Averaging estimators for autoregressions with a near unit root

Bruce E. Hansen *

Department of Economics, 1180 Observatory Drive, University of Wisconsin, Madison, WI 53706, United States

ARTICLE INFO

Article history:
Available online 12 March 2010

ABSTRACT

This paper uses local-to-unity theory to evaluate the asymptotic mean-squared error (AMSE) and forecast expected squared error from least-squares estimation of an autoregressive model with a root close to unity. We investigate unconstrained estimation, estimation imposing the unit root constraint, pre-test estimation, model selection estimation, and model average estimation. We find that the asymptotic risk depends only on the local-to-unity parameter, facilitating simple graphical comparisons. Our results strongly caution against pre-testing. Strong evidence supports averaging based on Mallows weights. In particular, our Mallows averaging method has uniformly and substantially smaller risk than the conventional unconstrained estimator, and this holds for autoregressive roots far from unity. Our averaging estimator is a new approach to forecast combination.

1. Introduction

This paper reopens the question of selection between unit root and stationary autoregressions. Rather than approaching the question from the vantage of hypothesis testing, we attack the question from the viewpoint of minimizing risk as measured by mean-squared error and out-of-sample expected squared forecast error. Our view is that if the purpose of autoregressive models is for estimation and forecasting, then model selection methods should be designed to minimize risk. As a general rule, hypothesis testing is inappropriate for this purpose, and we find that this rule remains true in the context of near non-stationary time series.

We consider an autoregressive model, and study asymptotic risk using a local-to-unity asymptotic framework. We study the asymptotic performance of the unconstrained least-squares estimator, the estimator imposing the unit root restriction, an optimal weighted average, the Dickey–Fuller pre-test estimator, the Mallows/AIC selection estimator, and finally the Mallows averaging estimator. We consider two measures of risk: asymptotic in-sample mean-squared error (AMSE), and asymptotic out-of-sample expected squared forecast error. In the local-to-unity framework, both risk measures depend exclusively on the local-to-unity parameter, facilitating graphical comparisons. The conclusions are clear. On the one hand, we find that the classic Dickey–Fuller pre-test estimator has very high risk. On the other hand, we find that our new Mallows averaging estimator has uniformly and substantially low risk. It is the preferred estimation method among those considered.

We now discuss some of the related literature.

There is a very large literature concerning test for autoregressive unit roots, starting with the seminal work of Dickey and Fuller (1979, 1981). The local-to-unity asymptotic framework was introduced by Chan and Wei (1987) and Phillips (1988a,b).

Many methods have been proposed for selecting the order of a stationary autoregression, including Akaike's final prediction error (Akaike, 1970), AIC (Akaike, 1973), Mallows' C_p (Mallows, 1973), BIC (Schwarz, 1978), $S_h(k)$ (Shibata, 1980), and predictive least squares (Rissanen, 1986). There is also a large literature exploring the asymptotic performance of these methods, including Wei (1992), Bhansali (1996), Lee and Karagrigoriou (2001), Ing (2003, 2004), Ing and Wei (2003, 2005) and Inoue and Kilian (2006). All of these papers focus on model selection for stationary observations, and none consider averaging estimators.

There is also a literature studying the effect on forecasting performance of whether or not to impose a unit root on an estimated autoregression and the role of unit root pre-testing. Franses and Kleibergen (1996) compare the empirical forecasting performance of the two models using the predictive least-squares criterion. Kemp (1999) studies forecast errors from a nearly integrated process at long horizons. Diebold and Kilian (2000) investigate the role of Dickey–Fuller pre-testing on long-horizon forecasting. Clements and Hendry (2001) study the impact of incorrect model choice on forecast mean-squared error. Kim et al. (2004) give asymptotic expressions for mean-squared forecast error in estimated models with a linear trend. Two papers which are close in method to ours are Stock (1996) and Elliott (2006). Both use local-to-unity asymptotics to evaluate the distribution of long-horizon forecasts based on pre-test estimators.

Autoregressive models with unit roots are a special case of cointegrated vector autoregressions (Engle and Granger, 1987).

* Tel.: +1 608 263 3880.
E-mail address: behansen@wisc.edu.
URL: http://www.ssc.wisc.edu/~bhansen.

0304-4076/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2010.03.022

There is a small literature on information-based methods for selection of cointegration rank. Gonzalo and Pitarakis (1998) and Aznar and Salvador (2002) discuss conditions for consistent model selection, and Kapetanios (2004) argues that the AIC is not a good selector of cointegration rank. Chao and Phillips (1999) analyze the problem using Bayes methods and propose the Posterior Information Criterion.

The averaging estimator discussed in this paper was introduced by Hansen (2007). It has also been applied to out-of-sample forecasting in stationary models by Hansen (2008), models with a structural break by Hansen (2009), and to heteroskedastic regressions by Hansen and Racine (2007). The idea of using a local-to-zero parameterization to study the asymptotic distribution of pre-test and model average estimators was developed by Hjort and Claeskens (2003).

Forecast combination was introduced in the seminal work of Bates and Granger (1969) and Granger and Ramanathan (1984) and spawned a large literature. Some excellent reviews include Granger (1989), Clemen (1989), Diebold and Lopez (1996), Hendry and Clements (2002), Timmermann (2006) and Stock and Watson (2006). Stock and Watson (1999, 2004, 2005) have provided detailed empirical evidence demonstrating the gains in forecast accuracy through forecast combination. A related paper is Pesaran and Timmermann (2007) which proposes forecast combination methods in regression models subject to structural breaks.

The paper is organized as follows. Section 2 presents the model and the base estimators. Section 3 presents the asymptotic analysis of mean-squared error. Section 4 presents asymptotic forecast risk. Section 5 covers Dickey–Fuller pre-testing. Section 6 presents Mallows selection. Section 7 introduces the Mallows averaging estimator. Section 8 evaluates the finite sample performance using simulation. Section 9 introduces a generalized Mallows averaging estimator. Section 10 concludes. Proofs of the theorems are presented in the Appendix. A Gauss program which calculates the MMA estimator is available on the author's webpage www.ssc.wisc.edu/~bhansen.

2. Model and estimation

Our model writes an observed series as a sum of its deterministic and stochastic components:

$$y_t = \beta_0 + \beta_1 t + \cdots + \beta_p t^p + S_t \tag{1}$$

where p is the order of the trend component. The leading case of interest is $p = 1$, a linear time trend. The stochastic component S_t is an AR$(k+1)$, written as

$$\Delta S_t = \alpha_0 S_{t-1} + \alpha_1 \Delta S_{t-1} + \cdots + \alpha_k \Delta S_{t-k} + e_t \tag{2}$$

where e_t is a homoskedastic martingale difference sequence (MDS) with variance σ^2. The Eq. (2) has a unit root when $\alpha_0 = 0$. We assume that all other roots of the Eq. (2) are stationary.

Differencing (1) and substituting (2) implies

$$\Delta y_t = \delta_t'\theta_0 + x_t'\theta_1 + z_t'\alpha + e_t \tag{3}$$

where

$$\delta_t = \begin{pmatrix} 1 \\ t \\ \vdots \\ t^{p-1} \end{pmatrix}, \qquad x_t = \begin{pmatrix} t^p \\ y_{t-1} \end{pmatrix}, \qquad z_t = \begin{pmatrix} \Delta y_{t-1} \\ \vdots \\ \Delta y_{t-k} \end{pmatrix},$$

$$\theta_1 = \begin{pmatrix} -\alpha_0\beta_p \\ \alpha_0 \end{pmatrix} \qquad \alpha = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_k \end{pmatrix},$$

and θ_0 is a function of the parameters in (1)–(2).

The optimal one-step-ahead predictor for Δy_t is the conditional mean

$$\mu_t = \delta_t'\theta_0 + x_t'\theta_1 + z_t'\alpha. \tag{4}$$

We consider three estimators of μ_t. Our baseline is unconstrained least-squares estimation of (3)

$$\Delta y_t = \delta_t'\hat\theta_0 + x_t'\hat\theta_1 + z_t'\hat\alpha + \hat e_t. \tag{5}$$

We set $\hat\mu_t = \delta_t'\hat\theta_0 + x_t'\hat\theta_1 + z_t'\hat\alpha$. This estimator has $p+2+k$ fitted coefficients.

Our second estimator imposes the unit root $\alpha_0 = 0$ which implies that $\theta_1 = 0$. The least-squares estimates under this restriction is

$$\Delta y_t = \delta_t'\tilde\theta_0 + z_t'\tilde\alpha + \tilde e_t. \tag{6}$$

We set $\tilde\mu_t = \delta_t'\tilde\theta_0 + z_t'\tilde\alpha$. This estimator has $p+k$ fitted coefficients, two fewer than the unconstrained estimator.

Our third estimator is obtained by taking a weighted average of $\hat\mu_t$ and $\tilde\mu_t$. Let $w \in [0,1]$ be the weight assigned to the unconstrained estimator. The averaging estimator is

$$\hat\mu_t(w) = w\hat\mu_t + (1-w)\,\tilde\mu_t.$$

3. Asymptotic mean-squared error

To evaluate the quality of our estimators, we use two measures of risk. In this section we consider the (asymptotic) in-sample mean-squared error, which measures the average fit. It is not a direct measure of forecasting performance because the estimates are constructed using the entire sample. Despite this qualification, we will see later that the in-sample AMSE is a convenient criterion because it is related to conventional information criterion.

To evaluate these measures, we use the local-to-unity asymptotic framework. Specifically, we let

$$\alpha_0 = \frac{ca}{n}$$

where

$$a = 1 - a_1 - \cdots a_k$$

and c is held fixed as $n \to \infty$. Let $W(r)$ denote a standard Brownian motion and define the diffusion process

$$\mathrm{d}W_c(r) = cW_c(r) + \mathrm{d}W(r) \tag{7}$$

which satisfies

$$W_c(r) = \int_0^r \exp\left(c\,(r-s)\right)\mathrm{d}W(s). \tag{8}$$

Also define the trend functions

$$\delta(r) = \begin{pmatrix} 1 \\ r \\ \vdots \\ r^{p-1} \end{pmatrix},$$

$$X_c(r) = \begin{pmatrix} r^p \\ W_c(r) \end{pmatrix}, \tag{9}$$

and the detrended processes

$$W_c^*(r) = W_c(r) - \int_0^1 W_c\delta'\left(\int_0^1 \delta\delta'\right)^{-1}\delta(r)$$

$$X_c^*(r) = X_c(r) - \int_0^1 X_c\delta'\left(\int_0^1 \delta\delta'\right)^{-1}\delta(r).$$

Theorem 1. *The AMSE of the constrained estimator is*

$$m_0(c,p,k) \equiv \lim_{n\to\infty}\frac{1}{\sigma^2}\sum_{t=1}^n E\left(\tilde\mu_t - \mu_t\right)^2 = m_0\left(c,p\right) + k \tag{10}$$

where

$$m_0(c,p) = EF_{0c} + p,$$

$$F_{0c} = c^2 \int_0^1 W_c^{*2}. \tag{11}$$

For $p = 0$ we can calculate

$$m_0(c,0) = -\frac{c}{2} - \left(\frac{1-\exp(2c)}{4}\right) \tag{12}$$

and for $p = 1$,

$$m_0(c,1) = -\frac{c}{2} - \left(\frac{1-\exp(2c)}{4}\right) - \left(\frac{\exp(2c)-1}{2c}\right) + 2\left(\frac{\exp(c)-1}{c}\right). \tag{13}$$

The AMSE of the unconstrained estimator is

$$m_1(c,p,k) \equiv \lim_{n\to\infty} \frac{1}{\sigma^2} \sum_{t=1}^{n} E\left(\hat{\mu}_t - \mu_t\right)^2 = m_1(c,p) + k \tag{14}$$

where

$$m_1(c,p) = EF_{1c} + p,$$

$$F_{1c} = \left(\int_0^1 \mathrm{d}W X_c^{*\prime}\right)\left(\int_0^1 X_c^* X_c^{*\prime}\right)^{-1}\left(\int_0^1 X_c^* \mathrm{d}W\right). \tag{15}$$

A closed-form expression for $m_1(c,p)$ is not available, but for all p,

$$\lim_{c\to-\infty} m_1(c,p) = 2 + p.$$

The AMSE of the averaging estimator is

$$m_w(c,p,k) \equiv \lim_{n\to\infty} \frac{1}{\sigma^2} \sum_{t=1}^{n} E\left(\hat{\mu}_t(w) - \mu_t\right)^2 = m_w(c,p) + k \tag{16}$$

where

$$m_w(c,p) = w^2 m_1(c,p) + (1-w)^2 m_0(c,p) + 2w(1-w) m_{01}(c,p),$$

$$m_{01}(c,p) = -E\left(c\int_0^1 W_c^* \mathrm{d}W\right) + p.$$

For all p, $m_{01}(0,p) = p$. When $p = 0$ then

$$m_{01}(c,0) = 0$$

and for $p = 1$

$$m_{01}(c,1) = \left(\frac{\exp(c)-1}{c}\right). \tag{17}$$

The weight w which minimizes $m_w(c,p,k)$ is

$$w_m(c,p) = \frac{m_0(c,p) - m_{01}(c,p)}{m_0(c,p) + m_1(c,p) - 2m_{01}(c,p)}$$

and the minimized AMSE is

$$m_{w_m}(c,p,k) = \frac{m_0(c,p)m_1(c,p) - m_{01}(c,p)^2}{m_0(c,p) + m_1(c,p) - 2m_{01}(c,p)} + k.$$

The AMSE of all estimators are the sum of k plus the additional component $m_0(c,p)$, $m_1(c,p)$, or $m_w(c,p)$. k is the normalized variance from the estimation of the coefficient α which is common across the three models and estimators. For the constrained estimator, $m_0(c,p)$ reflects variance from the estimation of θ_1 and the bias arising from the imposed unit root restriction. For the unconstrained estimator, $m_1(c,p)$ is the normalized variance from estimation of the coefficients on x_t and δ_t and is thus non-standard.

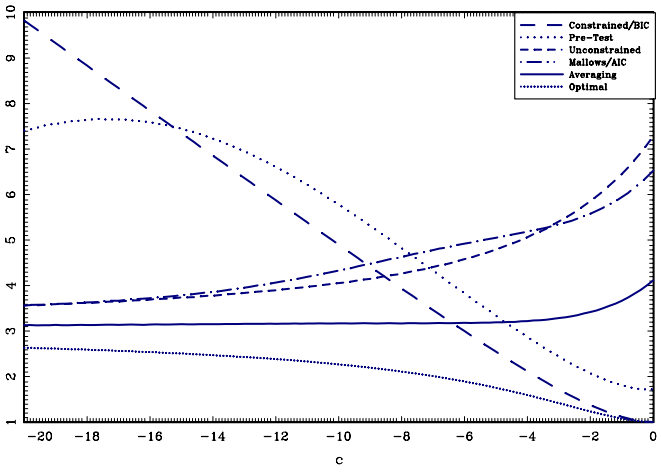

Fig. 1. Asymptotic mean-squared error.

For the averaging estimator, $m_w(c,p)$ is a convex weighted average of the constrained and unconstrained components, less an interaction term.

While $m_0(c,p)$ and $m_{01}(c,p)$ can be calculated analytically, in general the function $m_1(c,p)$ must be calculated by simulation.

The optimal weight $w_m(c,p)$ is independent of k and is strictly between 0 and 1 for $c < 0$. This means that the AMSE of the optimal averaging estimator is strictly less than both the unrestricted and restricted estimators.

The AMSE of the constrained, unconstrained, and optimal estimators for $p = 1$ and $k = 0$ (that is, the functions $m_0(c,1)$, $m_1(c,1)$ and $m_{w_m}(c,1)$) are displayed[1] in Fig. 1 for c ranging from -20 to 0. This corresponds to the model with a fitted intercept and linear time trend (the case with an intercept only ($p = 0$) is qualitatively similar). From the display, we can see that the AMSE of the constrained estimator is approximately linear in c, monotonically increasing as c moves away from zero. The AMSE of the unconstrained estimator is also monotonic, but with the opposite slope. The latter obtains its maximal value of 7.3 at $c = 0$, and asymptotically approaches 3 as $c \to -\infty$. The AMSE curves intersect at $c = -8.5$, meaning that for $c > -8.5$, the restricted (unit root) estimator has lower AMSE than the unconstrained estimator, while for $c < -8.5$ the unconstrained estimator has lower AMSE. For all c, the AMSE of the optimal averaging estimator is substantially below the AMSE of the other two estimators. The optimal estimator is infeasible, but its AMSE suggests that there are potentially large gains may be available from averaging.

4. Asymptotic forecast risk

In this section we consider an alternative measure of risk, the asymptotic one-step-ahead expected squared forecast error, which is a more direct measure of forecasting ability than AMSE.

Theorem 2. *The asymptotic forecast risk of the constrained estimator is*

$$f_0(c,p,k) \equiv \lim_{n\to\infty} \frac{n}{\sigma^2} E\left(\tilde{\mu}_{n+1} - \mu_{n+1}\right)^2 = f_0(c,p) + k \tag{18}$$

where

[1] Figs. 1 and 2 are computed on a grid of 101 evenly-spaced points from -20 to 0. Figs. 3–5 are computed on a grid of 21 points. Functions without analytic expressions were calculated by simulation. The asymptotic distributions were approximated by finite-sample counterparts with 1000 observations. 500,000 simulation replications were used.

$$f_0(c,p) = ET_{0c}^2,$$

$$T_{0c} = -cW_c^*(1) + \delta(1)'\left(\int_0^1 \delta\delta'\right)^{-1}\int_0^1 \delta \mathrm{d}W. \tag{19}$$

When $p = 0$ then $T_{0c} = -cW_c(1)$ and

$$f_0(c,0) = c\left(\frac{\exp(2c)-1}{2}\right).$$

When $p = 1$ then $T_{0c} = (1-c)W_c(1)$ and

$$f_0(c,1) = (1-c)^2\left(\frac{\exp(2c)-1}{2c}\right).$$

The asymptotic forecast risk of the unconstrained estimator is

$$f_1(c,p,k) \equiv \lim_{n\to\infty}\frac{n}{\sigma^2}E(\hat{\mu}_{n+1}-\mu_{n+1})^2 = f_1(c,p)+k \tag{20}$$

where

$$f_1(c,p) = E\left(T_{1c}^2\right)$$

$$\begin{aligned} T_{1c} &= \delta(1)'\left(\int_0^1 \delta\delta'\right)^{-1}\int_0^1 \delta \mathrm{d}W \\ &\quad + X_c^*(1)'\left(\int_0^1 X_c^*X_c^{*\prime}\right)^{-1}\int_0^1 X_c^*\mathrm{d}W. \end{aligned} \tag{21}$$

The asymptotic forecast risk of the averaging estimator is

$$\begin{aligned} f_w(c,p,k) &= f_w(c,p)w^2f_1(c,p) + (1-w)^2f_0(c,p) \\ &\quad + 2w(1-w)f_{01}(c,p) + k \end{aligned} \tag{22}$$

where

$$f_w(c,p) = w^2f_1(c,p) + (1-w)^2f_0(c,p) + 2w(1-w)f_{01}(c,p)$$
$$f_{01}(c,p) = E(T_{1c}T_{0c}).$$

The weight which minimizes $f_w(c,p,k)$ is

$$w_f(c,p) = \frac{f_0(c,p)-f_{01}(c,p)}{f_0(c,p)+f_1(c,p)-2f_{01}(c,p)}$$

and the minimized risk is

$$f_{w_f}(c,p,k) = \frac{f_0(c,p)f_1(c,p)-f_{01}(c,p)^2}{f_0(c,p)+f_1(c,p)-2f_{01}(c,p)} + k.$$

The functions $f_1(c,p)$ and $f_{01}(c,p)$ must be calculated by simulation.

The asymptotic forecast risk of the constrained, unconstrained, and optimal estimators for $p = 1$ and $k = 0$ is displayed in Fig. 2. The features are qualitatively similar those displayed in Fig. 1.

5. Pre-testing

The choice between the constrained estimator $\tilde{\mu}_t$ and the unconstrained estimator $\hat{\mu}_t$ may be determined by the data. A common practice is pre-testing using a unit root test. The pre-test estimate selects $\hat{\mu}_t$ if the test rejects the null of the unit root, otherwise it selects $\tilde{\mu}_t$. For concreteness, let us focus on the Dickey–Fuller t-test, which is based on the t-ratio

$$DF_n = \frac{\hat{\alpha}-1}{s(\hat{\alpha})}$$

where $s(\hat{\alpha})$ is the OLS standard error for $\hat{\alpha}$. Let r be a critical value. (For example, if $p = 1$ the asymptotic 5% critical value is $r = -3.41$.) The pre-test estimator is

$$\hat{\mu}_t^{\mathrm{df}} = \hat{\mu}_t 1(DF_n \le r) + \tilde{\mu}_t 1(DF_n > r).$$

We now present the AMSE and asymptotic forecast risk for this estimate.

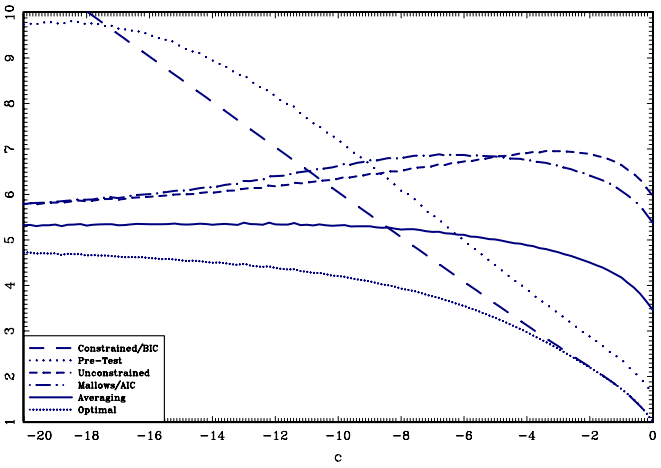

Fig. 2. Asymptotic forecast risk.

Theorem 3.

$$\begin{aligned} m_{\mathrm{df}}(c,p,k) &= \lim_{n\to\infty}\frac{1}{\sigma^2}\sum_{t=1}^n E\left(\hat{\mu}_t^{\mathrm{df}}-\mu_t\right)^2 \\ &= E(F_{1c}1(DF_c \le r)) + E(F_{0c}1(DF_c > r)) + p + k \end{aligned}$$

and

$$\begin{aligned} f_{\mathrm{df}}(c,p,k) &= \lim_{n\to\infty}\frac{n}{\sigma^2}E\left(\hat{\mu}_{n+1}^{\mathrm{df}}-\mu_{n+1}\right)^2 \\ &= E\left(T_{1c}^2 1(DF_c \le r)\right) + E\left(T_{0c}^2 1(DF_c > r)\right) + k \end{aligned}$$

where F_{0c}, F_{1c}, T_{0c}, and T_{1c} are defined in (11), (15), (19) and (21), respectively,

$$DF_c = \frac{\int_0^1 W_c^\tau \mathrm{d}W_c}{\left(\int_0^1 \left(W_c^\tau\right)^2\right)^{1/2}},$$

and

$$W_c^\tau(r) = W_c(r) - \int_0^1 W_c\tau'\left(\int_0^1 \tau\tau'\right)^{-1}\tau(r),$$

$$\tau(r) = \begin{pmatrix} 1 \\ \vdots \\ r^p \end{pmatrix}.$$

The AMSE and asymptotic forecast risk of the pre-test estimator are displayed in Figs. 1 and 2, respectively. Both measures of risk are low for small values of $-c$ but quite large for moderate to large values of $-c$. The goal of pre-testing is presumably to gain the benefits of both the constrained and unconstrained estimators, but examining the figures we see that this is not the case. The pre-test estimator has smaller risk than the unconstrained estimator only for very small values of $-c$, but for most of the parameter space the pre-test estimator has higher risk, and the discrepancy can be quite large. This is similar to the behavior of pre-test estimates in stationary models.

This conclusion appears to clash with the assertions in Stock (1996) and Diebold and Kilian (2000) that pre-testing can be useful for selection of forecasting models. However, the tables in Stock (1996) clearly show similar behavior to the results in Figs. 1 and 2, even for his DF-GLS pre-test estimator. Diebold and Kilian (2000) largely miss the difficulty by focusing exclusively on very small and very large values of $|c|$—both cases where pre-testing works well. These papers also focus on the long-horizon forecasting case, where they argue that pre-testing is more valuable, while in this paper we focus exclusively on one-step-ahead forecasting.

6. Mallows selection

As an alternative to pre-testing, the choice between $\hat{\mu}_t$ and $\tilde{\mu}_t$ can be made using an information criterion, such as the AIC, BIC, or Mallows. For convenience we focus on the Mallows criterion as its linear structure is easy to characterize. We discuss selection based on AIC and BIC following Theorem 5 below.

The Mallows (1973) criterion is a penalized sum of squared errors, designed to be approximately unbiased for the in-sample AMSE. In our local-to-unity model, the optimal Mallows criteria for the unrestricted and restricted models are

$$M_0(c) = n\tilde{\sigma}^2 + 2\hat{\sigma}^2 (m_{01}(c, p) + k)$$

and

$$M_1(c) = n\hat{\sigma}^2 + 2\hat{\sigma}^2 (m_1(c, p) + k),$$

respectively, where $\hat{\sigma}^2 = n^{-1}\sum_{t=1}^n \left(y_t - \hat{\mu}_t\right)^2$ and $\tilde{\sigma}^2 = n^{-1}\sum_{t=1}^n (y_t - \tilde{\mu}_t)^2$ are the estimates of σ^2 from the two models. This claim is justified by the following result.

Theorem 4.

$$\frac{EM_0(c)}{\sigma^2} - n \to m_0(c, p, k)$$
$$\frac{EM_1(c)}{\sigma^2} - n \to m_1(c, p, k).$$

Theorem 4 shows that the criteria $M_0(c)$ and $M_1(c)$ are (after normalization) asymptotically unbiased estimates of the AMSE. This demonstrates that these are appropriate Mallows criterion for model selection. The penalty terms in $M_0(c)$ and $M_1(c)$ are non-standard. This is analogous to the finding of Chao and Phillips (1999) in their study of Bayesian model selection in reduced rank VARs.

Unfortunately, the criteria $M_0(c)$ and $M_1(c)$ are infeasible since they depend on the unknown c. We suggest evaluating the criterion for the restricted model $M_0(c)$ at the restricted value $c = 0$, viz. $M_0 = M_0(0)$ and the criterion for the unrestricted model $M_1(c)$ at the opposite asymptote, viz. $M_1 = \lim_{c\to-\infty} M_1(c)$. Since $m_{01}(0, p) = p$ and $\lim_{c\to-\infty} m_1(c, p) = 2 + p$ these values are

$$M_0 = n\tilde{\sigma}^2 + 2\hat{\sigma}^2(p + k)$$

and

$$M_1 = n\hat{\sigma}^2 + 2\hat{\sigma}^2(2 + p + k)$$

respectively, which are the classic Mallows (1973) criterion for the restricted and unrestricted models (as $p + k$ is the number of fitted parameters in the unit root model and $2 + p + k$ is the number of parameters in the unrestricted model). That is, while $M_0(c)$ and $M_1(c)$ are optimal yet infeasible, the limits $M_0 = M_0(0)$ and $M_1 = \lim_{c\to-\infty} M_1(c)$ correspond to the conventional model selection criterion.

Mallows selection picks the model with the smallest criterion (M_0 or M_1). This is equivalent to selecting the unrestricted estimator when $F_n \geq 2((2 + p + k) - (p + k)) = 4$ where

$$F_n = n\left(\frac{\tilde{\sigma}^2 - \hat{\sigma}^2}{\hat{\sigma}^2}\right) \tag{23}$$

is the classic Wald statistic for the joint exclusion of y_{t-1} and the time trend from (3). The Mallows selected estimator is then

$$\hat{\mu}_t^m = \hat{\mu}_t 1(F_n \geq 4) + \tilde{\mu}_t 1(F_n < 4).$$

We now characterize the AMSE and asymptotic forecast risk of Mallows selection.

Theorem 5.

$$m_m(c, p, k) = \lim_{n\to\infty} \frac{1}{\sigma^2}\sum_{t=1}^n E\left(\hat{\mu}_t^m - \mu_t\right)^2$$
$$= E(F_{1c}1(F_c \geq 4)) + E(F_{0c}1(F_c < 4)) + p + k$$

and

$$f_m(c, p, k) = \lim_{n\to\infty} \frac{n}{\sigma^2} E\left(\hat{\mu}_t^m - \mu_{n+1}\right)^2$$
$$= E(T_{1c}^2 1(F_c \geq 4)) + E\left(T_{0c}^2 1(F_c < 4)\right) + k$$

where F_{0c}, F_{1c}, T_{0c}, and T_{1c} are defined in (11), (15), (19) and (21), and

$$F_c = \left(\int_0^1 dW_c X_c^{*\prime}\right)\left(\int_0^1 X_c^* X_c^{*\prime}\right)^{-1}\left(\int_0^1 X_c^* dW_c\right). \tag{24}$$

From Theorem 5 we can also deduce the asymptotic performance of AIC and BIC selection. In fact, Mallows and AIC selection are asymptotically equivalent, as AIC selects the unrestricted estimator when $n\log\left(\tilde{\sigma}^2/\hat{\sigma}^2\right) \geq 4$, and $n\log\left(\tilde{\sigma}^2/\hat{\sigma}^2\right) = F_n + o_p(1)$ in the local-to-unity model. It follows that AIC selection has the same asymptotic risk as Mallows selection. In contrast, BIC is asymptotically equivalent to restricted estimation, as BIC selects the unrestricted estimator when $n\log\left(\tilde{\sigma}^2/\hat{\sigma}^2\right) \geq 2\ln(n)$ which occurs with probability tending to zero as n diverges, since $n\log\left(\tilde{\sigma}^2/\hat{\sigma}^2\right) \approx F_n$ has a non-degenerate asymptotic distribution in the local-to-unity model. Since the unrestricted model is in a local neighborhood of the restricted model, BIC selects the restricted model with probability approaching one and thus BIC selection has asymptotic risk equal to the restricted estimator.

The AMSE of the Mallows/AIC selection estimator is displayed in Fig. 1 and its forecast risk is displayed in Fig. 2. (The BIC selection estimator has the same asymptotic risk as the constrained estimator.) The risk of the Mallows/AIC selection estimator is very close to that of the unconstrained estimator. The evidence suggests that there is little reason to consider the selection estimator rather than unconstrained estimation.

7. Mallows averaging

Just as for selection, the Mallows criterion for the averaging estimator (Hansen, 2007) is a penalized sum of squared errors, designed to be approximately unbiased for the in-sample AMSE. For any w let

$$\hat{\sigma}^2(w) = n^{-1}\sum_{t=1}^n \left(y_t - \hat{\mu}_t(w)\right)^2$$

be the variance estimate using the averaging estimator $\hat{\mu}_t(w)$. In the local-to-unity model the optimal Mallows criterion is

$$M_w(c) = n\hat{\sigma}^2(w) + 2\hat{\sigma}^2(w(m_1(c, p) + k) + (1 - w)(m_{01}(c, p) + k)).$$

This claim is justified by the following result.

Theorem 6.

$$\frac{EM_w(c)}{\sigma^2} - n \to m_w(c, p, k).$$

This shows that the criterion $M_w(c)$ is an asymptotically unbiased estimates of the AMSE for the averaging estimator. As in the case of selection this criterion is unfeasible, and as before we suggest replacing $m_1(c, p)$ with $\lim_{c\to-\infty} m_1(c, p) = 2 + p$ and $m_{01}(c, p)$ with $m_{01}(0, p) = 0$, leading to the feasible criterion

$$M_w = n\hat{\sigma}^2(w) + 2\hat{\sigma}^2(w(2 + p + k) + (1 - w)(p + k))$$
$$= n\hat{\sigma}^2(w) + 2\hat{\sigma}^2(2w + p + k)$$

which is identical to the Mallows averaging criterion proposed in Hansen (2007).

The Mallows selected weight $\hat{w}$ is the value which minimizes M_w over $w \in [0, 1]$. Since the criterion is quadratic in w there is an explicit solution.

Theorem 7. *The minimizer of M_w is*

$$\hat{w} = \begin{cases} 1 - \dfrac{2}{F_n} & \text{if } F_n > 2 \\ 0 & \text{otherwise} \end{cases}$$

where F_n is the classic Wald statistic (23).

The Mallows averaging estimator is the weighted average of the unrestricted and restricted least-squares estimators, using the Mallows weight $\hat{w}$.

$$\hat{\mu}_t^a = \hat{w}\hat{\mu}_t + \left(1 - \hat{w}\right)\tilde{\mu}_t = \begin{cases} \tilde{\mu}_t & \text{if } F_n \le 2 \\ \left(1 - \dfrac{2}{F_n}\right)\hat{\mu}_t + \left(\dfrac{2}{F_n}\right)\tilde{\mu}_t & \text{otherwise.} \end{cases}$$

Theorem 8. *The Mallows selected weight has the asymptotic distribution*

$$\hat{w} \xrightarrow{\mathrm{d}} \pi_c = \begin{cases} 1 - \dfrac{2}{F_c} & \text{if } F_c > 2 \\ 0 & \text{otherwise,} \end{cases}$$

where F_c is defined in (24). The AMSE of the Mallows averaging estimator is

$$\begin{aligned} m_a(c) &= \lim_{n\to\infty} \frac{1}{\sigma^2} \sum_{t=1}^{n} E\left(\hat{\mu}_t^a - \mu_t\right)^2 \\ &= E\left(\pi_c^2 F_{1c}\right) + E\left((1 - \pi_c)^2 F_{0c}\right) \\ &\quad - 2cE\left(\pi_c\left(1 - \pi_c\right)\int_0^1 \mathrm{d}WW_c^*\right) + 1 \end{aligned}$$

and the asymptotic forecast risk is

$$\begin{aligned} f_a(c) &= \lim_{n\to\infty} \frac{n}{\sigma^2} E\left(\hat{\mu}_{n+1}\left(\hat{w}\right) - \mu_{n+1}\right)^2 \\ &= E\left(\pi_c T_{1c} + \left(1 - \pi_c\right) T_{0c}\right)^2 \end{aligned}$$

where F_{0c}, F_{1c}, T_{0c}, and T_{1c} are defined in (11), (15), (19) and (21), respectively.

The AMSE and asymptotic forecast risk of the Mallows averaging estimator are displayed in Figs. 1 and 2. Its performance is stunning relative to the other estimators. In both figures, the risk of the averaging estimator is uniformly smaller than the unrestricted estimator, and for most values of c the improvement is quite significant. The risk reduction (relative to unrestrictive estimation) is significant even for large values of $|c|$. For example, at $c = -20$ the reduction in AMSE is about is about 15%. Overall, the averaging estimator has the impressively low risk, and is the best choice among the feasible estimators considered.

8. Finite sample MSE and forecast risk

The analysis of the previous sections has been asymptotic. We have found that the AMSE and forecast risk are only a function of the local-to-unity parameter c and the order of the trend function p, and is affected by the autoregressive order k only by an intercept shift. In particular, the asymptotic theory is invariant to the other model parameters, including those which determine the short-run dynamics. In this section, we investigate whether or not these features continue to hold in finite samples. For simplicity, we focus on forecast risk, as this is the primary criterion of interest.

Our finite sample investigation uses the AR($k+1$) model (1)–(2) with $p = 1$ and e_t i.i.d. $N(0, 1)$. We set the trend parameters $\beta_0 = \beta_1 = 0$. In this section, we assume that k is known, and consider the values $k = 0, 4, 8$. The sample sizes are $n = 50$ and $n = 200$.

For our first experiment, we set all the remaining autoregressive parameters to zero, $\alpha_1 = \cdots = \alpha_k = 0$. This allows us to investigate the effects of sample size and autoregressive order, holding the serial correlation properties constant. Setting $\alpha_0 = c/n$, we vary c on a grid from -20 to 0. This implies a range for α_0 of $[-0.4, 0]$ for $n = 50$ and a range of $[-0.1, 0]$ for $n = 200$.

For each parameter configuration, we calculate the forecast risk $nE(\hat{\mu}_{n+1} - \mu_{n+1})^2$ for three estimators: the unrestricted least-squares estimator $\hat{\mu}_{n+1}$, the Dickey–Fuller pre-test estimator $\hat{\mu}_{n+1}^{\mathrm{df}}$, and the Mallows averaging estimator $\hat{\mu}_{n+1}^a$. The risk was calculated by Monte Carlo simulation, taking the average of $n(\hat{\mu}_{n+1} - \mu_{n+1})^2$ across 500,000 simulation draws.

The results are presented in Fig. 3. There are six panels, one for each (n, k) pair. In each panel, the forecast risk is plotted as a function of c. These panels are finite sample analogs of the asymptotic risk as reported in Fig. 2. What is striking is that all of the panels in Fig. 3 are quite similar to Fig. 2. The scaled finite sample forecast risk is nearly identical to the asymptotic risk. The only exception can be seen in the lower-left panel, for $n = 50$ and $k = 8$, where the unrestricted estimator has relatively high forecast risk, and is noticeably dominated by the pre-test and averaging estimators for all values of c. What is most important, however, is that the risk of the Mallows averaging estimator is uniformly less than that of the unrestricted estimator, in all cases considered.

Our second experiment adds serial correlation. We do this by setting the autoregressive parameters as $\alpha_j = -(-\theta)^j$ for $j = 1, \ldots, k$, for $\theta = 0.6$ (the results are not sensitive to this choice). We then set $\alpha_0 = (1 - a_1 - \cdots a_k)c/n$ as indicated by the asymptotic theory.

We repeated the experiment as described above, for $k = 4$ and 8 (since $k = 0$ is redundant with the prior experiment). The results are presented in Fig. 4 and are very similar to Fig. 3. As predicted by the asymptotic theory, the forecast risk is relatively invariant to the autoregressive parameters.

9. General Mallows averaging

We now consider a more general setting where the number of autoregressive lags k is unknown. Let the set of models be indexed by both k and the possible unit root restriction. This is model (1)–(2) with $k \in \{0, 1, \ldots, K\}$. For each $k = 0, \ldots, K$, let $\hat{\mu}_t(k)$ and $\tilde{\mu}_t(k)$ denote the least-squares estimates of μ_t from the regressions (5) and (6).

The averaging estimator is a weighted average of these $2K + 2$ estimates. For each k, let w_{1k} be the weight assigned to $\hat{\mu}_t(k)$, and let w_{0k} be the weight assigned to $\tilde{\mu}_t(k)$. The weights are non-negative and sum to one: $w_{1k} \ge 0$, $w_{0k} \ge 0$, and $\sum_{k=0}^{K}(w_{0k} + w_{1k}) = 1$. The general averaging estimator of μ_t is

$$\hat{\mu}_t(W) = \sum_{k=0}^{K}\left(w_{0k}\tilde{\mu}_t(k) + w_{1k}\hat{\mu}_t(k)\right)$$

where

$$W = \begin{pmatrix} w_{0k} \\ \vdots \\ w_{0k} \\ w_{1k} \\ \vdots \\ w_{1k} \end{pmatrix}.$$

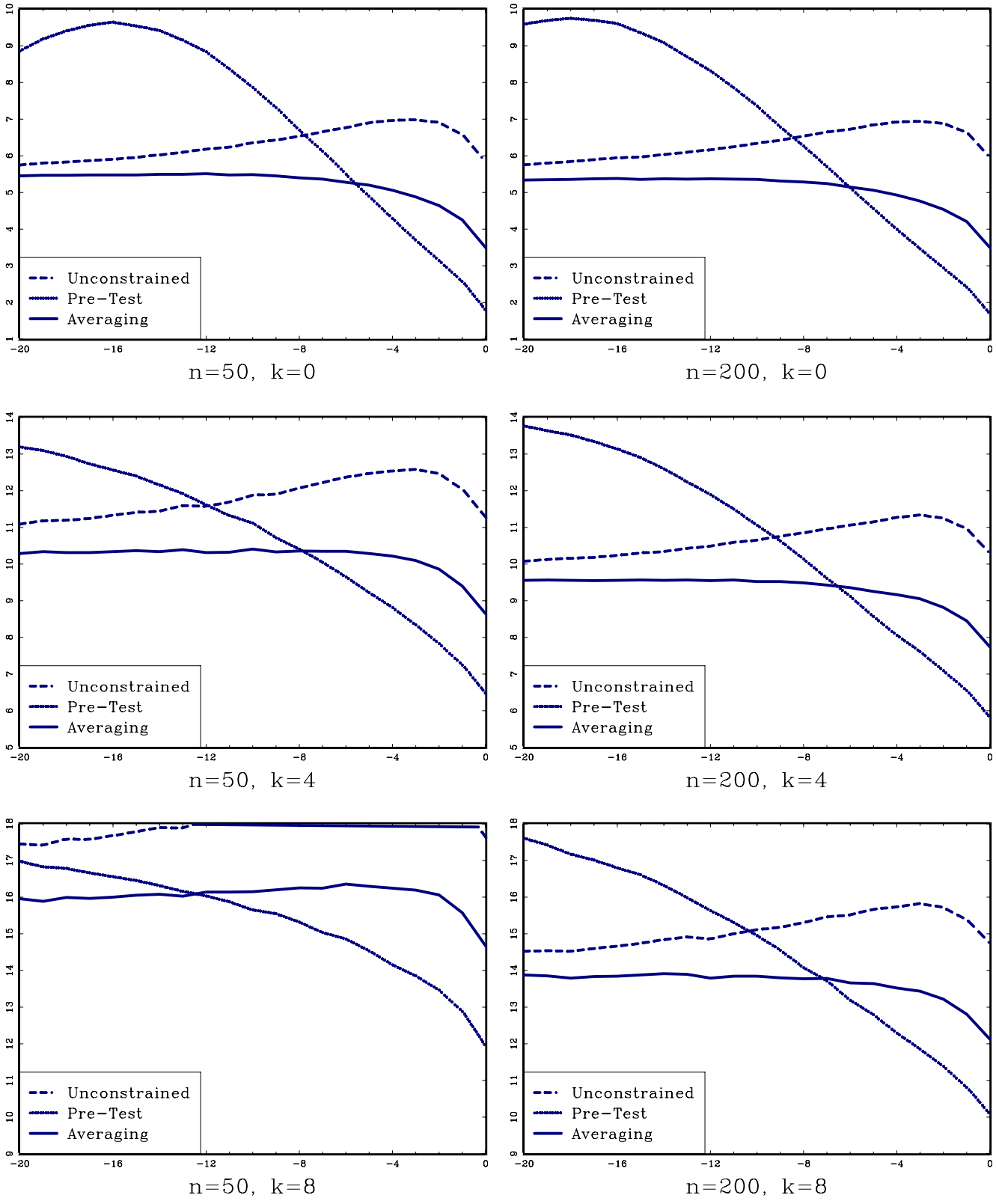

Fig. 3. Finite sample forecast risk.

The Mallows criterion for weight selection as described by Hansen (2007) is

$$M(W) = \sum_{t=1}^{n} \left(y_t - \hat{\mu}_t(W)\right)^2 + 2\hat{\sigma}^2 \left(\sum_{k=0}^{K} (w_{0k}k + w_{1k}(2+k)) + p \right)$$

where

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{t=1}^{n} \left(y_t - \hat{\mu}_t(K)\right)^2$$

is the variance estimator from the unrestricted general model.

A computationally useful alternative formula for $M(W)$ is constructed as follows. Let $\hat{e}(k) = y - \hat{\mu}(k)$ and $\tilde{e}(k) = y - \tilde{\mu}(k)$

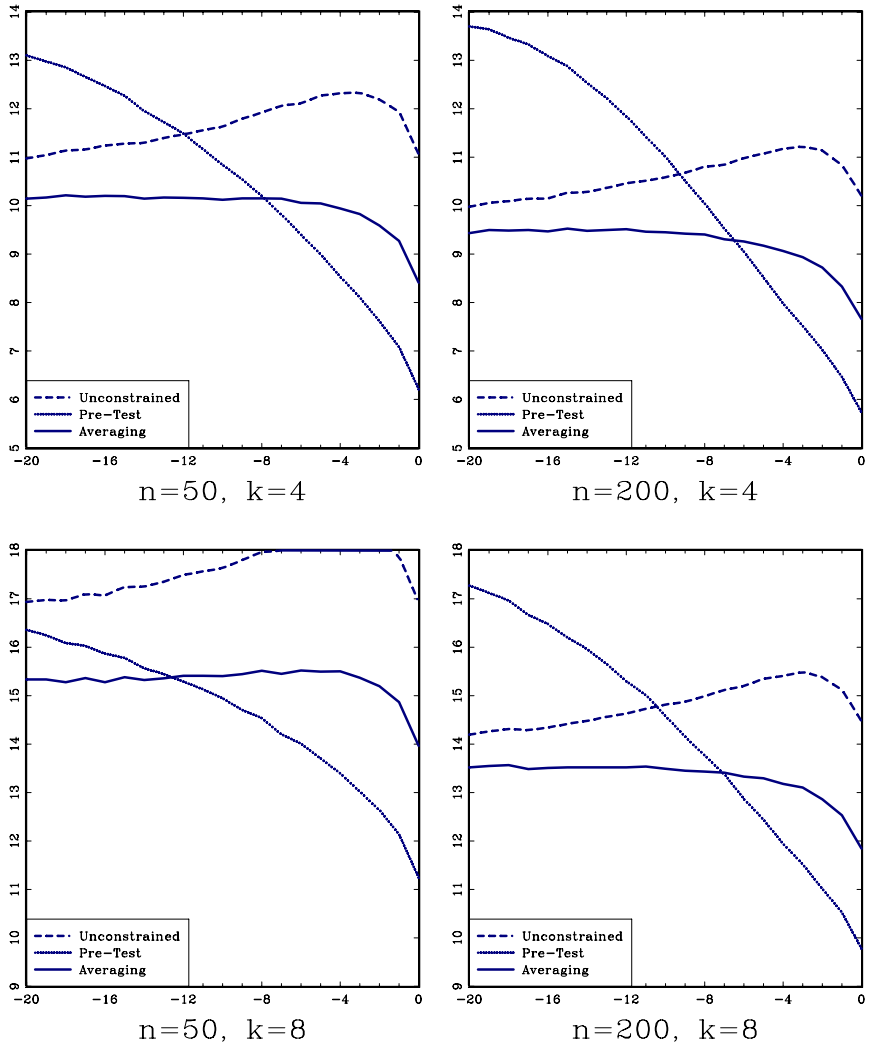

Fig. 4. Finite sample forecast risk, $\Theta = 0.6$.

be the $n \times 1$ vectors of residuals from the individual models, and construct the $n \times (2K+2)$ matrix

$$\hat{e} = \left[\tilde{e}(0), \ldots, \tilde{e}(K), \hat{e}(0), \ldots, \hat{e}(K)\right]$$

and the $(2K+2) \times 1$ vector

$$R = \begin{pmatrix} p \\ \vdots \\ p+K \\ p+2 \\ \vdots \\ p+K+2 \end{pmatrix}.$$

The vector R contains the adjusted Mallows penalties for each model. The criterion can then be written as

$$M(W) = W'\hat{e}'\hat{e}W + 2\hat{\sigma}^2 R'W.$$

The Mallows selected weight vector $\hat{W}$ minimizes $M(W)$ over the set of W which satisfy the constraints (non-negativity and summing to one). This is a linear-quadratic programming problem with inequality constraints, and generally has no closed-form solution. Numerical solutions are readily obtain using linear programming methods. Corner solutions are typical, so many individual selected weights will be zero.

The Mallows estimator is $\hat{\mu}_t = \hat{\mu}(\hat{W})$, the weighted average using these selected weights.

We investigate the finite sample performance of this general Mallows averaging estimator in a Monte Carlo simulation experiment. The same setting was used as in the previous section, and we contrast three estimators. The first estimator is Mallows selection, where the class of models is AR(1) through AR(K) (the estimates $\hat{\mu}_t(1)$ through $\hat{\mu}_t(K)$). The second estimator is Mallows averaging of this set of models (as in Hansen (2007)). The third estimator is the general Mallows averaging estimator as described above. This comparison allows us to disentangle the benefits of selection versus averaging, and the benefits of averaging over the autoregressive order k as well as over the unit root restriction. For this investigation, we consider autoregressions of order $K = 4$, 8, and 12. (Note that in when $K = 12$ the general estimator is averaging over 26 individuals models!)

The asymptotic forecast risk of the methods are presented in Fig. 5 for $\theta = 0.6$. The results are qualitatively similar across K and n. In all cases, the general averaging estimator has the lowest forecast risk, and the selection estimator has the highest forecast risk. We can see that there is a clear improvement by averaging over the autoregressive models (by comparing the selection estimator with the partial averaging estimator) and also a clear improvement by averaging the unrestricted models with those imposing the unit root restriction.

This experiment was repeated for other values of θ. The results are qualitatively similar for other values of θ and therefore omitted.

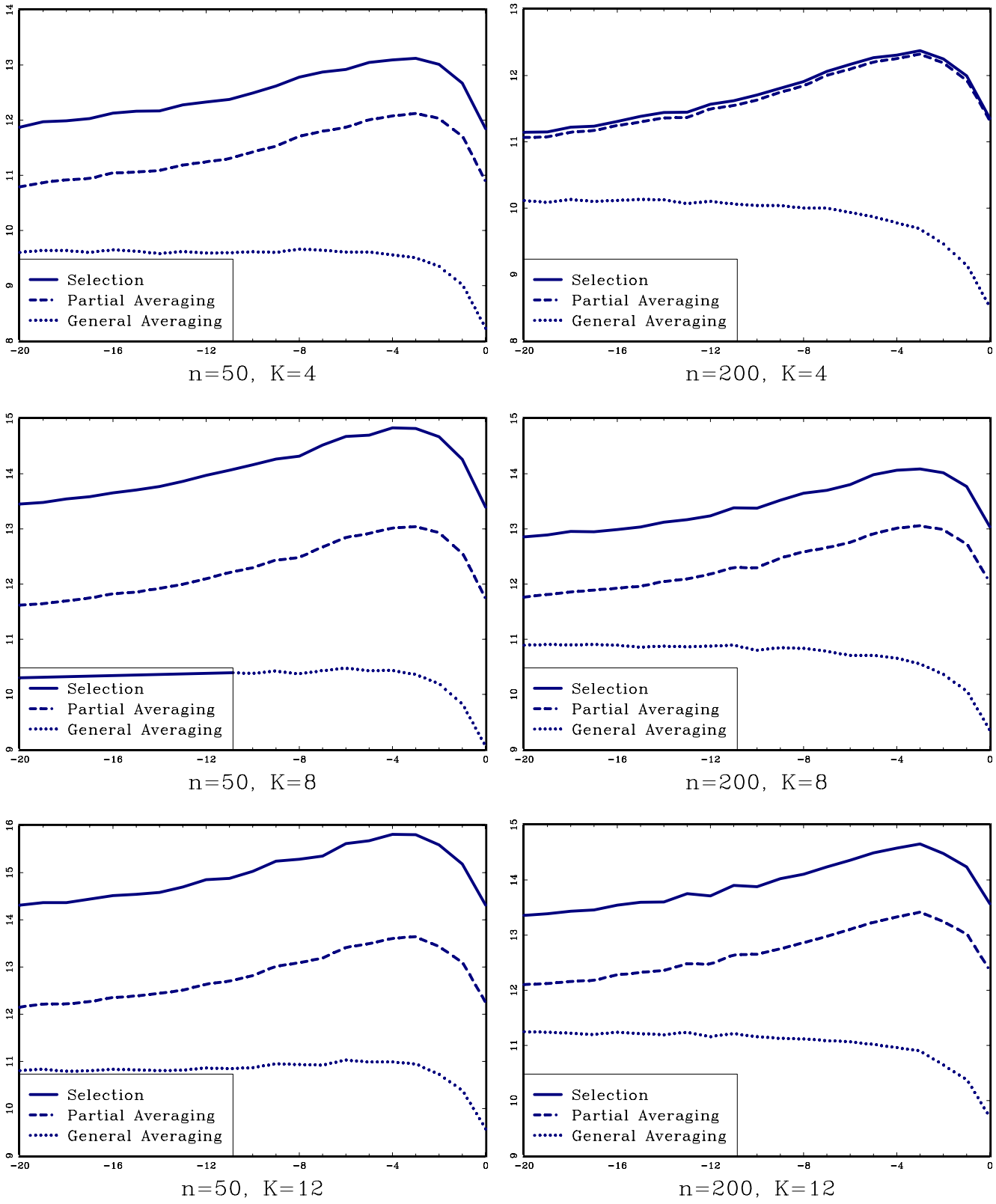

Fig. 5. Selection and averaging over k.

10. Conclusion

This paper examined the question of selection and combination of autoregressive model when the goal is minimizing risk. Using local-to-unity asymptotic methods, we found that two measures of risk of a variety of estimators are functions only of the local-to-unity parameter, facilitating direct comparisons. We examined unconstrained and constrained least-squares estimation, optimal combination, Dickey–Fuller pre-testing, Mallows selection, and Mallows averaging. The numerical comparisons demonstrate the stunning result that the Dickey–Fuller pre-test estimator has particularly high risk, while the Mallows averaging estimator has

uniformly low risk. We conclude that Mallows averaging is a potentially important forecasting method.

The paper has confined attention to one-step-ahead forecasting. It would be useful to use similar methods to study long-horizon forecasting. This presents some special technical challenges and is well beyond the scope of this paper.

Furthermore, the analysis is restricted to univariate autoregressions. A natural extension would be to vector autoregressions, where the question is the number of cointegrating relationships. Based on the analysis in this paper, we expect model averaging methods will have lower risk than estimation based on cointegration pre-testing. This deserves further study.

It is also possible that further improvements can be made by considering alternative estimators to least squares. Stock (1996) documents that using the efficient unit root tests of Elliott et al. (1996) can reduce the forecast risk of the pre-test estimator. Canjels and Watson (1997) develop improved methods for estimation of the trend parameters in models with roots local to unity. These methods may be useful in constructing improved forecasts.

Finally, as documented by Stock and Watson (2005), simple rules for forecast combination (assigning each model equal weight) often achieve lower forecast risk than data-dependent combination methods. This finding suggests that improvements over the Mallows averaging method may be feasible, and calls for further research into improved combination selection.

Acknowledgements

The author's research was supported by the National Science Foundation. The author thanks the guest editors and two referees for the helpful comments.

Appendix

The following results will be useful in subsequent calculations.

Lemma 1.

$$E(W_c(r)^2) = \frac{\exp(2cr) - 1}{2c}, \tag{25}$$

$$E\left(\int_0^1 W_c^2\right) = \frac{1}{2c}\left(\frac{\exp(2c) - 1}{2c} - 1\right), \tag{26}$$

$$E\left(W_c(1)W(1)\right) = \frac{\exp(c) - 1}{c}. \tag{27}$$

Proof. Using (8) and the fact that $\mathrm{d}W(s)$ is an orthogonal process,

$$\begin{aligned} &E(W_c(r)^2) \\ &= E\int_0^r \int_0^r \exp\left(c\left(r-s\right)\right)\exp\left(c\left(r-u\right)\right)\mathrm{d}W(s)\mathrm{d}W(u) \\ &= \int_0^r \exp\left(2c\left(r-s\right)\right)\mathrm{d}s \\ &= \frac{\exp(2cr) - 1}{2c} \end{aligned}$$

which is (25). Eq. (26) follows by integration. To show (27),

$$\begin{aligned} E\left(W_c(1)W(1)\right) &= E\int_0^1 \int_0^1 \exp\left(c\left(1-s\right)\right)\mathrm{d}W(s)\mathrm{d}W(u) \\ &= \int_0^1 \exp(c(1-s))\mathrm{d}s \\ &= \frac{\exp(c) - 1}{c}. \end{aligned}$$

Proof of Theorem 1. First, as shown in Lemma 1 of Hansen (1995), since e_t is a MDS,

$$\frac{1}{\sigma\sqrt{n}}\sum_{t=1}^{[nr]} e_t \xrightarrow{\mathrm{d}} W(r)$$

and

$$\frac{a}{\sigma\sqrt{n}} S_{[nr]} \xrightarrow{\mathrm{d}} W_c(r).$$

Defining the weight matrices $D_{0n} = \mathrm{diag}\{1, n, \ldots, n^{p-1}\}$ and $D_{1n} = \mathrm{diag}\left\{n^p, n^{1/2}\sigma/a\right\}$, we have

$$D_{0n}^{-1}\delta_{[nr]} \xrightarrow{\mathrm{d}} \delta(r),$$

$$D_{1n}^{-1}x_{[nr]} \xrightarrow{\mathrm{d}} X_c(r).$$

Define the orthogonalized series

$$x_t^* = x_t - \delta_t'\left(\sum_{j=1}^n \delta_j\delta_j'\right)^{-1}\sum_{j=1}^n \delta_j x_j$$

$$z_t^* = z_t - \delta_t'\left(\sum_{j=1}^n \delta_j\delta_j'\right)^{-1}\sum_{j=1}^n \delta_j z_j$$

$$S_{t-1}^* = S_{t-1} - \delta_t'\left(\sum_{j=1}^n \delta_j\delta_j'\right)^{-1}\sum_{j=1}^n \delta_j S_{j-1}$$

and observe that

$$\frac{a}{\sigma\sqrt{n}} S_{[nr]}^* \xrightarrow{\mathrm{d}} W_c^*(r),$$

$$D_{1n}^{-1}x_{[nr]}^* \xrightarrow{\mathrm{d}} X_c^*(r).$$

Since the regressions (5) and (6) include δ_t, the fitted means $\hat{\mu}_t$ and $\tilde{\mu}_t$ are unchanged if we replace x_t and z_t with x_t^* and z_t^*, which we now assume for the remainder of the Appendix.

We now examine the constrained estimator. The regression (6) has an effective error of $can^{-1}S_{t-1} + e_t$. Let

$$\tilde{\theta}_0^* = \tilde{\theta}_0 - \frac{ca}{n}\left(\sum_{t=1}^n \delta_t\delta_t'\right)^{-1}\left(\sum_{t=1}^n \delta_t S_{t-1}\right)$$

which satisfies

$$\begin{aligned} &\frac{n^{1/2}}{\sigma} D_{0n}\left(\tilde{\theta}_0^* - \theta_0\right) \\ &= \frac{1}{\sigma} D_{0n}\left(\frac{1}{n}\sum_{t=1}^n \delta_t\delta_t'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{t=1}^n \delta_t e_t\right) + o_p(1) \\ &\xrightarrow{\mathrm{d}} \left(\int_0^1 \delta\delta'\right)^{-1}\left(\int_0^1 \delta\mathrm{d}W\right). \end{aligned} \tag{28}$$

Also

$$\begin{aligned} \frac{n^{1/2}}{\sigma}\left(\tilde{\alpha} - \alpha\right) &= \left(\frac{1}{n}\sum_{t=1}^n z_t^* z_t^{*\prime}\right)^{-1}\left(\frac{1}{\sigma\sqrt{n}}\sum_{t=1}^n z_t^* e_t + o_p(1)\right) \\ &\xrightarrow{\mathrm{d}} Z \sim N\left(0, Q^{-1}\right) \end{aligned} \tag{29}$$

where $Q = E\left(z_t^* z_t^{*\prime}\right)$.

We can write

$$\begin{aligned} \tilde{\mu}_t - \mu_t &= -can^{-1}S_{t-1} + \left(\tilde{\theta}_0 - \theta_0\right)'\delta_t + \left(\tilde{\alpha} - \alpha\right)' z_t^* \\ &= -can^{-1}S_{t-1}^* + \left(\tilde{\theta}_0^* - \theta_0\right)'\delta_t + \left(\tilde{\alpha} - \alpha\right)' z_t^* \end{aligned} \tag{30}$$

so

$$\begin{aligned}
&\frac{1}{\sigma^2}\sum_{t=1}^{n}\left(\tilde{\mu}_t-\mu_t\right)^2\\
&=\frac{c^2a^2}{\sigma^2n^2}\sum_{t=1}^{n}S_{t-1}^{*2}+\frac{1}{\sigma^2}\left(\tilde{\theta}_0^*-\theta_0\right)'\sum_{t=1}^{n}\delta_t\delta_t'\left(\tilde{\theta}_0^*-\theta_0\right)\\
&\quad+\frac{1}{\sigma^2}\left(\tilde{\alpha}-\alpha\right)'\sum_{t=1}^{n}z_t^*z_t^{*\prime}\left(\tilde{\alpha}-\alpha\right)+o_p(1)\\
&\xrightarrow{\mathrm{d}}c^2\int_0^1W_c^{*2}+\left(\int_0^1\mathrm{d}W\delta'\right)\left(\int_0^1\delta\delta'\right)^{-1}\\
&\quad\times\left(\int_0^1\delta\mathrm{d}W\right)+Z'QZ\\
&=F_{0c}+\chi_p^2+\chi_k^2
\end{aligned}\tag{31}$$

where

$$\chi_p^2=\left(\int_0^1\mathrm{d}W\delta'\right)\left(\int_0^1\delta\delta'\right)^{-1}\left(\int_0^1\delta\mathrm{d}W\right)$$

and

$$\chi_k^2=Z'QZ$$

are chi-square with degrees of freedom p and k, respectively. Taking expectations of (31) we obtain (10).

When $p=0$ then there is no $\delta(r)$. Thus using (26),

$$m_0(c,0)=E\left(c^2\int_0^1W_c^2\right)=-\frac{c}{2}+\frac{\exp(2c)-1}{4}$$

which is (12). When $p=1$, $\delta(r)=1$. Thus

$$\begin{aligned}
m_0(c,1)&=E\left(c^2\int_0^1W_c^{*2}\right)+1\\
&=E\left(c^2\int_0^1W_c^2\right)-E\left(c\int_0^1W_c\right)^2+1.
\end{aligned}$$

Eq. (7) implies

$$c\int_0^1W_c=W_c(1)-W(1)\tag{32}$$

and thus using (25), (26) and (27), we find

$$\begin{aligned}
m_0(c,1)&=E\left(c^2\int_0^1W_c^2\right)-EW_c(1)^2\\
&\quad-EW(1)^2+2E\left(W_c(1)W(1)\right)+1\\
&=-\frac{c}{2}+\frac{\exp(2c)-1}{4}-\left(\frac{\exp(2c)-1}{2c}\right)\\
&\quad+2\left(\frac{\exp(c)-1}{c}\right),
\end{aligned}$$

which is (13).

We next consider the unconstrained estimator. Note that

$$\hat{\mu}_t-\mu_t=\delta_t'\left(\hat{\theta}_0-\theta_0\right)+x_t^{*\prime}\left(\hat{\theta}_1-\theta_1\right)+z_t^{*\prime}\left(\hat{\alpha}-\alpha\right).$$

We calculate that

$$\frac{n^{1/2}}{\sigma}D_{0n}\left(\hat{\theta}_0-\theta_0\right)\xrightarrow{\mathrm{d}}\left(\int_0^1\delta\delta'\right)^{-1}\int_0^1\delta\mathrm{d}W,\tag{33}$$

$$\frac{n^{1/2}}{\sigma}D_{1n}\left(\hat{\theta}_1-\theta_1\right)\xrightarrow{\mathrm{d}}\left(\int_0^1X_c^*X_c^{*\prime}\right)^{-1}\int_0^1X_c^*\mathrm{d}W,\tag{34}$$

and

$$\frac{n^{1/2}}{\sigma}\left(\hat{\alpha}-\alpha\right)\xrightarrow{\mathrm{d}}Z\tag{35}$$

as in (29).

We find

$$\begin{aligned}
&\frac{1}{\sigma^2}\sum_{t=1}^{n}\left(\hat{\mu}_t-\mu_t\right)^2\\
&=\frac{1}{\sigma^2}\sum_{t=1}^{n}\left(\delta_t'\left(\hat{\theta}_0-\theta_0\right)+x_t^{*\prime}\left(\hat{\theta}_1-\theta_1\right)+z_t^{*\prime}\left(\hat{\alpha}-\alpha\right)\right)^2\\
&=\frac{1}{\sigma^2}\left(\hat{\theta}_1-\theta_1\right)'\sum_{t=1}^{n}x_t^*x_t^{*\prime}\left(\hat{\theta}_1-\theta_1\right)\\
&\quad+\frac{1}{\sigma^2}\left(\hat{\theta}_0-\theta_0\right)'\sum_{t=1}^{n}\delta_t\delta_t'\left(\hat{\theta}_0-\theta_0\right)\\
&\quad+\frac{1}{\sigma^2}\left(\hat{\alpha}-\alpha\right)'\sum_{t=1}^{n}z_t^*z_t^{*\prime}\left(\hat{\alpha}-\alpha\right)\\
&\quad+2\frac{1}{\sigma^2}\left(\hat{\theta}-\theta\right)'\sum_{t=1}^{n}x_t^*z_t^{*\prime}\left(\hat{\alpha}-\alpha\right)\\
&\xrightarrow{\mathrm{d}}\left(\int_0^1\mathrm{d}WX_c^{*\prime}\right)\left(\int_0^1X_c^*X_c^{*\prime}\right)^{-1}\left(\int_0^1X_c^*\mathrm{d}W\right)\\
&\quad+\left(\int_0^1\mathrm{d}W\delta'\right)\left(\int_0^1\delta\delta'\right)^{-1}\left(\int_0^1\delta\mathrm{d}W\right)+Z'QZ\\
&=F_{1c}+\chi_p^2+\chi_k^2.
\end{aligned}\tag{36}$$

Taking expectations of (36) yields (14).

We now examine the averaging estimator. Let $\hat{\theta}_0(w)=w\hat{\theta}_0+(1-w)\tilde{\theta}_0^*$ and $\hat{\alpha}(w)=w\hat{\alpha}+(1-w)\tilde{\alpha}$. We can see that for any w

$$\frac{n^{1/2}}{\sigma}D_{0n}\left(\hat{\theta}_0(w)-\hat{\theta}_0\right)\xrightarrow{\mathrm{d}}\left(\int_0^1\delta\delta'\right)^{-1}\int_0^1\delta\mathrm{d}W$$

and

$$\frac{n^{1/2}}{\sigma}\left(\hat{\alpha}(w)-\alpha\right)\xrightarrow{\mathrm{d}}Z.$$

Noting that

$$\begin{aligned}
\hat{\mu}_t(w)-\mu_t&=wx_t^{*\prime}\left(\hat{\theta}_1-\theta_1\right)-(1-w)\,can^{-1}S_{t-1}^*\\
&\quad+\left(\hat{\theta}_0(w)-\theta_0\right)'\delta_t+\left(\hat{\alpha}(w)-\alpha\right)'z_t^*,
\end{aligned}\tag{37}$$

we see

$$\begin{aligned}
&\frac{1}{\sigma^2}\sum_{t=1}^{n}\left(\hat{\mu}_t(w)-\mu_t\right)^2=w^2\frac{1}{\sigma^2}\left(\hat{\theta}_1-\theta_1\right)'\sum_{t=1}^{n}x_t^*x_t^{*\prime}\left(\hat{\theta}_1-\theta_1\right)\\
&\quad+(1-w)^2\frac{c^2a^2}{\sigma^2n}\sum_{t=1}^{n}S_{t-1}^{*2}+\frac{1}{\sigma^2}\left(\hat{\theta}_0(w)-\theta_0\right)'\\
&\quad\times\sum_{t=1}^{n}\delta_t\delta_t'\left(\hat{\theta}_0(w)-\theta_0\right)-2w(1-w)\frac{ca}{n\sigma^2}\\
&\quad\times\left(\hat{\theta}_1-\theta_1\right)'\sum_{t=1}^{n}x_t^*S_{t-1}^*+\frac{1}{\sigma^2}\left(\hat{\alpha}(w)-\alpha\right)'\\
&\quad\times\sum_{t=1}^{n}z_t^*z_t^{*\prime}\left(\hat{\alpha}(w)-\alpha\right)+o_p(1)
\end{aligned}$$

$$\begin{aligned}
&\xrightarrow{d} w^2\left(\int_0^1 dW X_c^{*\prime}\right)\left(\int_0^1 X_c^* X_c^{*\prime}\right)^{-1}\\
&\quad\times\left(\int_0^1 X_c^* dW\right)+(1-w)^2 c^2\int_0^1 W_c^{*2}\\
&\quad-2w(1-w)c\left(\int_0^1 dW X_c^{*\prime}\right)\left(\int_0^1 X_c^* X_c^{*\prime}\right)^{-1}\\
&\quad\times\int_0^1 X_c^* W_c^*+\chi_p^2+\chi_k^2\\
&=w^2 F_{1c}+(1-w^2)F_{0c}-2w(1-w)c\\
&\quad\times\int_0^1 dW W_c^*+\chi_p^2+\chi_k^2. \qquad (38)
\end{aligned}$$

Taking expectations establishes (16).

To evaluate $m_{01}(c,p)$, note that

$$\begin{aligned}
m_{01}(c,p) &= -Ec\int_0^1 dW W_c\\
&\quad+E\left(c\int_0^1 dW\delta'\left(\int_0^1\delta\delta'\right)^{-1}\int_0^1\delta W_c\right)+p\\
&= E\left(c\int_0^1 dW\delta'\left(\int_0^1\delta\delta'\right)^{-1}\int_0^1\delta W_c\right)+p \qquad (39)
\end{aligned}$$

since $E\int_0^1 dW W_c=0$ by the definition of the stochastic integral. It follows that when $p=0$, $m_{01}(c,0)=0$. When $p=1$, using (32) and (27),

$$\begin{aligned}
m_{01}(c,1) &= E\left(cW(1)\int_0^1 W_c\right)+1\\
&= -E\left(W(1)^2\right)+E\left(W(1)W_c(1)\right)+1\\
&= \frac{\exp(c)-1}{c},
\end{aligned}$$

which is (16).

The optimal w^* and mean-squared error are found by minimizing $m_w(c,p,k)$ with respect to w. □

Proof of Theorem 2. First, take the unconstrained estimator. Observe that

$$\hat\mu_{n+1}-\mu_{n+1}=\delta_{n+1}'(\hat\theta_0-\theta_0)+x_{n+1}^{*\prime}(\hat\theta_1-\theta_1)+z_{n+1}^{*\prime}(\hat\alpha-\alpha).$$

Using (33) and (34), note that

$$\begin{aligned}
&\frac{n^{1/2}}{\sigma}(\delta_{n+1}'(\hat\theta_0-\theta_0)+x_{n+1}^{*\prime}(\hat\theta_1-\theta_1))\\
&\quad\xrightarrow{d}\delta(1)'\left(\int_0^1\delta\delta'\right)^{-1}\int_0^1\delta dW+X_c^*(1)'\\
&\quad\times\left(\int_0^1 X_c^* X_c^{*\prime}\right)^{-1}\int_0^1 X_c^* dW=T_{1c} \qquad (40)
\end{aligned}$$

and thus

$$\frac{n}{\sigma^2}E(\delta_{n+1}'(\hat\theta_0-\theta_0)+x_{n+1}^{*\prime}(\hat\theta_1-\theta_1))^2\to ET_{1c}^2.$$

Furthermore using (35)

$$\frac{n}{\sigma^2}((\hat\alpha-\alpha)(\hat\alpha-\alpha)')\xrightarrow{d}ZZ',$$

so

$$\begin{aligned}
\frac{n}{\sigma^2}E((\tilde\alpha-\alpha)'z_{n+1}^*)^2 &= \frac{n}{\sigma^2}\mathrm{tr}\,E(z_{n+1}^* z_{n+1}^{*\prime}(\tilde\alpha-\alpha)(\tilde\alpha-\alpha)')\\
&\to \mathrm{tr}\,E(z_{n+1}^* z_{n+1}^{*\prime}ZZ')\\
&= E(Z'QZ)=k. \qquad (41)
\end{aligned}$$

Together,

$$\begin{aligned}
\frac{n}{\sigma^2}E(\hat\mu_{n+1}-\mu_{n+1})^2 &= \frac{n}{\sigma^2}E(\delta_{n+1}'(\hat\theta_0-\theta_0)+x_{n+1}^{*\prime}(\hat\theta_1-\theta_1))^2\\
&\quad+\frac{n}{\sigma^2}E((\hat\alpha-\alpha)z_{n+1}^* z_{n+1}^{*\prime}(\hat\alpha-\alpha))+o(1)\\
&\to ET_{1c}^2+k
\end{aligned}$$

which is (20).

Second, take the constrained estimator. We have

$$\tilde\mu_{n+1}-\mu_{n+1}=-can^{-1}S_n^*+(\tilde\theta_0^*-\theta_0)'\delta_{n+1}+(\tilde\alpha-\alpha)'z_{n+1}^*.$$

Using (28),

$$\begin{aligned}
&\frac{n^{1/2}}{\sigma}(-can^{-1}S_n^*+(\tilde\theta_0^*-\theta_0)'\delta_{n+1})\\
&\quad\xrightarrow{d}-cW_c^*(1)+\delta(1)'\left(\int_0^1\delta\delta'\right)^{-1}\int_0^1\delta dW=T_{0c} \qquad (42)
\end{aligned}$$

and therefore

$$\frac{n}{\sigma^2}E(-can^{-1}S_n^*+(\tilde\theta_0^*-\theta_0)'\delta_{n+1})^2\to ET_{0c}^2.$$

As in (41),

$$\frac{n}{\sigma^2}E((\hat\alpha-\alpha)'z_{n+1}^*)^2\to k.$$

Then

$$\begin{aligned}
\frac{n}{\sigma^2}E(\tilde\mu_{n+1}-\mu_{n+1})^2 &= \frac{n}{\sigma^2}E(-can^{-1}S_n^*+(\tilde\theta_0^*-\theta_0)'\delta_{n+1})^2\\
&\quad+\frac{n}{\sigma^2}E((\hat\alpha-\alpha)'z_{n+1}^*)^2+o(1)\\
&\to ET_{0c}^2+k,
\end{aligned}$$

which is (18). When $p=0$ then $T_{0c}=-cW_c(1)$ so $f_0(c,0)=c^2EW_c(1)^2=c(\exp(2c)-1)/2$ by (25). When $p=1$ then

$$T_{0c}=-cW_c(1)+c\int_0^1 W_c+W(1)=(1-c)W_c(1)$$

and therefore $f_0(c,1)=(1-c)^2EW_c(1)^2=(1-c)^2(\exp(2c)-1)/2c$.

Third, take the averaging estimator. Since

$$\begin{aligned}
\hat\mu_{n+1}(w)-\mu_{n+1} &= w(\delta_{n+1}'(\hat\theta_0-\theta_0)+x_{n+1}^{*\prime}(\hat\theta_1-\theta_1))+(1-w)\\
&\quad\times(-can^{-1}S_n^*+(\tilde\theta_0^*-\theta_0)'\delta_{n+1})+(\hat\alpha(w)-\alpha)'z_{n+1}^*,
\end{aligned}$$

then

$$\begin{aligned}
\frac{n}{\sigma^2}E(\hat\mu_t(w)-\mu_t)^2 &= w^2\frac{n}{\sigma^2}E(\delta_{n+1}'(\hat\theta_0-\theta_0)+x_{n+1}^{*\prime}(\hat\theta_1-\theta_1))^2\\
&\quad+(1-w)^2\frac{n}{\sigma^2}E(-can^{-1}S_n^*+(\tilde\theta_0^*-\theta_0)'\delta_{n+1})^2\\
&\quad+2w(1-w)\frac{n}{\sigma^2}E(\delta_{n+1}'(\hat\theta_0-\theta_0)\\
&\quad+x_{n+1}^{*\prime}(\hat\theta_1-\theta_1))(-can^{-1}S_n^*+(\tilde\theta_0^*-\theta_0)'\delta_{n+1})\\
&\quad+\frac{n}{\sigma^2}E((\hat\alpha(w)-\alpha)z_{n+1}^* z_{n+1}^{*\prime}(\hat\alpha(w)-\alpha))\\
&\quad+o_p(1)\xrightarrow{d}w^2ET_{1c}^2+(1-w)^2ET_{0c}^2\\
&\quad+2w(1-w)E(T_{0c}T_{1c})+k
\end{aligned}$$

which is (22).

The optimal weight and risk are found by minimizing $f_w(c,p,k)$ with respect to w. □

Proof of Theorem 3. From (31) and (36) we have

$$\frac{1}{\sigma^2}\sum_{t=1}^{n}(\tilde{\mu}_t-\mu_t)^2 \xrightarrow{\mathrm{d}} F_{0c}+\chi_p^2+\chi_k^2$$

and

$$\frac{1}{\sigma^2}\sum_{t=1}^{n}(\hat{\mu}_t-\mu_t)^2 \xrightarrow{\mathrm{d}} F_{1c}+\chi_p^2+\chi_k^2.$$

By standard calculations we know that $DF \xrightarrow{\mathrm{d}} DF_c$. Recalling that $\hat{\mu}_{\mathrm{df}}=\hat{\mu}_t 1(DF_n\le r)+\tilde{\mu}_t 1(DF_n>r)$, we then have

$$\begin{aligned}\frac{1}{\sigma^2}\sum_{t=1}^{n}(\hat{\mu}_t^{\mathrm{df}}-\mu_t)^2 &= \frac{1}{\sigma^2}\sum_{t=1}^{n}(\hat{\mu}_t-\mu_t)^2 1(DF_n\le r)\\ &\quad+\frac{1}{\sigma^2}\sum_{t=1}^{n}(\tilde{\mu}_t-\mu_t)^2 1(DF_n>r)\\ &\xrightarrow{\mathrm{d}} F_{1c}1(DF_c\le r)+F_{0c}1(DF_c>r)+\chi_p^2+\chi_k^2\\ &= F_{1c}1(DF_c\le r)+F_{0c}1(DF_c>r)+\chi_p^2+\chi_k^2.\end{aligned}$$

Taking expectations yields the expression for $m_{\mathrm{df}}(c,p,k)$.

Similarly, using (42) and (40),

$$\begin{aligned}\frac{n}{\sigma^2}E(\hat{\mu}_t^{\mathrm{df}}-\mu_{n+1})^2 &= \frac{n}{\sigma^2}E[(\hat{\mu}_{n+1}-\mu_{n+1})^2 1(DF_n\le r)]\\ &\quad+\frac{n}{\sigma^2}E[(\tilde{\mu}_{n+1}-\mu_{n+1})^2 1(DF_n>r)]\\ &= \frac{n}{\sigma^2}E[(\delta_{n+1}'(\hat{\theta}_0-\theta_0)+x_{n+1}'(\hat{\theta}-\theta))^2\\ &\quad\times 1(DF_n\le r)]+\frac{n}{\sigma^2}E[(-can^{-1}S_n^*\\ &\quad+(\tilde{\theta}_0^*-\theta_0)'\delta_{n+1})^2 1(DF_n>r)]\\ &\quad+\frac{n}{\sigma^2}E((\hat{\alpha}-\alpha)z_{n+1}^* z_{n+1}^{*\prime}(\hat{\alpha}-\alpha))+o(1)\\ &\to E\left[T_{1c}^2 1(DF_c\le r)\right]+E[T_{0c}^2 1(DF_c>r)]+k,\end{aligned}$$

which is $f_{\mathrm{df}}(c,p,k)$. □

Proof of Theorem 4. First take $M_0(c)$. Since $y_t-\tilde{\mu}_t=e_t-(\tilde{\mu}_t-\mu_t)$ then

$$\sum_{t=1}^{n}(y_t-\tilde{\mu}_t)^2=\sum_{t=1}^{n}e_t^2+\sum_{t=1}^{n}(\tilde{\mu}_t-\mu_t)^2-2\sum_{t=1}^{n}e_t(\tilde{\mu}_t-\mu_t)$$

and thus

$$\begin{aligned}\frac{M_0(c)-n\sigma^2}{\sigma^2} &= \frac{1}{\sigma^2}\sum_{t=1}^{n}(e_t^2-\sigma^2)+\frac{1}{\sigma^2}\sum_{t=1}^{n}(\tilde{\mu}_t-\mu_t)^2\\ &\quad+\frac{2\hat{\sigma}^2}{\sigma^2}(m_{01}(c,p)+k)-\frac{2}{\sigma^2}\sum_{t=1}^{n}e_t(\tilde{\mu}_t-\mu_t).\end{aligned}\tag{43}$$

The first three terms have expectations tending to $m_0(c,p,k)+2(m_{01}(c,p)+k)$. The fourth term is -2 times

$$\begin{aligned}\frac{1}{\sigma^2}\sum_{t=1}^{n}e_t(\tilde{\mu}_t-\mu_t) &= -\frac{ca}{\sigma^2 n}\sum_{t=1}^{n}e_t S_{t-1}^*+\frac{1}{\sigma^2}\sum_{t=1}^{n}e_t\delta_t'(\tilde{\theta}_0^*-\theta_0)\\ &\quad+\frac{1}{\sigma^2}\sum_{t=1}^{n}e_t z_t^{*\prime}(\tilde{\alpha}-\alpha)\\ &\xrightarrow{\mathrm{d}} -c\int_0^1 \mathrm{d}WW_c^*+\int_0^1 \mathrm{d}W\delta'\left(\int_0^1\delta\delta'\right)^{-1}\int_0^1\delta\mathrm{d}W+Z'QZ\end{aligned}\tag{44}$$

(using (28)), which has expectation $m_{01}(c)+k$. Adding these components, it follows that (43) has expectation tending to $m_0(c,p,k)$ as claimed.

Next consider $M_1(c)$. We have

$$\begin{aligned}\frac{M_1(c)-n\sigma^2}{\sigma^2} &= \frac{1}{\sigma^2}\sum_{t=1}^{n}(e_t^2-\sigma^2)+\frac{1}{\sigma^2}\sum_{t=1}^{n}(\hat{\mu}_t-\mu_t)^2\\ &\quad+\frac{2\hat{\sigma}^2}{\sigma^2}(m_1(c,p)+k)-\frac{2}{\sigma^2}\sum_{t=1}^{n}e_t(\hat{\mu}_t-\mu_t).\end{aligned}\tag{45}$$

The first three terms have expectation tending to $m_1(c,p,k)+2(m_1(c,p)+k)$. The third is -2 times

$$\begin{aligned}&\frac{1}{\sigma^2}\sum_{t=1}^{n}e_t(\hat{\mu}_t-\mu_t)\\ &= \frac{1}{\sigma^2}\sum_{t=1}^{n}e_t x_t^{*\prime}(\hat{\theta}_1-\theta_1)+\frac{1}{\sigma^2}\sum_{t=1}^{n}e_t z_t^{*\prime}(\hat{\alpha}-\alpha)\\ &\xrightarrow{\mathrm{d}} \int_0^1 \mathrm{d}WX_c^*(r)'\left(\int_0^1 X_c^*X_c^{*\prime}\right)^{-1}\int_0^1 X_c^*\mathrm{d}W\\ &\quad+\int_0^1 \mathrm{d}W\delta'\left(\int_0^1\delta\delta'\right)^{-1}\int_0^1\delta\mathrm{d}W+Z'QZ\end{aligned}\tag{46}$$

which has expectation $m_1(c,p)+k$. Adding these two components, we see that (45) has expectation tending to $m_1(c,p,k)$ as claimed. □

Proof of Theorem 5. The argument is the same as for Theorem 3, except that we use the fact that $F_n \xrightarrow{\mathrm{d}} F_c$. □

Proof of Theorem 6. Similar to the argument in the proof of Theorem 4,

$$\begin{aligned}\frac{M_w(c)-n\sigma^2}{\sigma^2} &= \frac{1}{\sigma^2}\sum_{t=1}^{n}(e_t^2-\sigma^2)+\frac{1}{\sigma^2}\sum_{t=1}^{n}(\hat{\mu}_t(w)-\mu_t)^2\\ &\quad+\frac{2\hat{\sigma}^2}{\sigma^2}(w(m_1(c,p)+k)+(1-w)(m_{01}(c,p)+k))\\ &\quad-w\frac{2}{\sigma^2}\sum_{t=1}^{n}e_t(\hat{\mu}_t-\mu_t)-(1-w)\frac{2}{\sigma^2}\sum_{t=1}^{n}e_t(\tilde{\mu}_t-\mu_t).\end{aligned}$$

The first three terms have expectation converging to

$$m_w(c,p,k)+2(w(m_1(c,p)+k)+(1-w)(m_{01}(c,p)+k)).$$

The fourth and fifth terms converge to a random variable with expectation

$$-2(w(m_1(c,p)+k)+(1-w)(m_{01}(c,p)+k))$$

by (44) and (46). Summing, the entire expression converges to a random variable with expectation $m_w(c,p,k)$, as claimed. □

Proof of Theorem 7. Let $\hat{e}_t=y_t-\hat{\mu}_t$ and $\tilde{e}_t=y_t-\tilde{\mu}_t$. Observe that

$$\begin{aligned}\sum_{t=1}^{n}(y_t-\hat{\mu}_t(w))^2 &= \sum_{t=1}^{n}(w\hat{e}_t+(1-w)\tilde{e}_t)^2\\ &= w^2\sum_{t=1}^{n}\hat{e}_t^2+(1-w)^2\sum_{t=1}^{n}\tilde{e}_t^2+2w(1-w)\sum_{t=1}^{n}\hat{e}_t\tilde{e}_t\\ &= w^2\sum_{t=1}^{n}\hat{e}_t^2+(1-w)^2\sum_{t=1}^{n}\tilde{e}_t^2+2w(1-w)\sum_{t=1}^{n}\hat{e}_t^2\\ &= n\hat{\sigma}^2+(1-w)^2 n(\tilde{\sigma}^2-\hat{\sigma}^2).\end{aligned}$$

Thus

$$\frac{M_w}{\hat{\sigma}^2} = n + (1-w)^2 F_n + 2(2w+k).$$

The first-order condition for minimization is $0 = -2(1-\hat{w})F_n + 4$, whose solution is $\hat{w} = 1 - 2/F_n$. If this value is negative, then the constrained minimizer is $\hat{w} = 0$. □

Proof of Theorem 8. Since $F_n \xrightarrow{d} F_c$ it follows directly that $\hat{w} \xrightarrow{d} \pi_c$. Evaluating Eq. (38) at $w = \pi_c$ and then taking expectations we obtain the expression from $m_a(c, p, k)$. The argument for $f_a(c, p, k)$ is similar. □

References

Akaike, H., 1970. Statistical predictor identification. Annals of the Institute of Statistical Mathematics 22, 203–419.

Akaike, H., 1973. Maximum likelihood identification of Gaussian autoregressive moving average models. Biometrika 60, 255–265.

Aznar, Antonio, Salvador, Manuel, 2002. Selecting the rank of the cointegration space and the form of the intercept using an information criterion. Econometric Theory 18, 926–947.

Bates, J.M., Granger, C.M.W., 1969. The combination of forecasts. Operations Research Quarterly 20, 451–468.

Bhansali, R.J., 1996. Asymptotically efficient autoregressive model selection for multistep prediction. Annals of the Institute of Statistical Mathematics 48, 577–602.

Canjels, Eugene, Watson, Mark W., 1997. Estimating deterministic trends in the presence of serially correlated errors. Review of Economics and Statistics 79, 184–200.

Chan, N.H., Wei, Ching-Zong, 1987. Asymptotic inference for nearly nonstationary AR(1) processes. Annals of Statistics 15, 1050–1063.

Chao, John C., Phillips, Peter C.B., 1999. Model selection in partially nonstationary vector autoregressive processes with reduced rank structure. Journal of Econometrics 91, 227–271.

Clemen, R.T., 1989. Combining forecasts: a review and annotated bibliography. International Journal of Forecasting 5, 559–581.

Clements, Michael P., Hendry, David F., 2001. Forecasting with difference-stationary and trend-stationary models. Econometrics Journal 4, S1–S19.

Dickey, David A., Fuller, Wayne A., 1979. Distribution of the estimators for autoregressive time series with a unit root. Journal of the American Statistical Association 74, 427–431.

Dickey, David A., Fuller, Wayne A., 1981. Likelihood ratio statistics for autoregressive time series with a unit root. Econometrica 49, 1057–1072.

Diebold, Francis X., Kilian, Lutz, 2000. Unit-root tests are useful for selecting forecasting models. Journal of Business and Economic Statistics 18, 265–273.

Diebold, Francis X., Lopez, J.A., 1996. Forecast evaluation and combination. In: Maddala, Rao (Eds.), Handbook of Statistics. Elsevier.

Elliott, Graham, 2006. Unit root pre-testing and forecasting. Working Paper. UCSD.

Elliott, Graham, Rothenberg, Thomas J., Stock, James H., 1996. Efficient tests of an autoregressive unit root. Econometrica 64, 813–836.

Engle, Robert F., Granger, Clive W.J., 1987. Co-integration and error correction: representation, estimation and testing. Econometrica 55, 251–276.

Franses, Philip Hans, Kleibergen, Frank, 1996. Unit roots in the Nelson–Plosser data: do they matter for forecasting? International Journal of Forecasting 12, 283–288.

Gonzalo, Jesus, Pitarakis, Jean-Yves, 1998. Specification via model selection in vector error correction models. Economics Letters 60, 321–328.

Granger, Clive W.J., 1989. Combining forecasts—twenty years later. Journal of Forecasting 8, 167–173.

Granger, Clive W.J., Ramanathan, R., 1984. Improved methods of combining forecasts. Journal of Forecasting 3, 197–204.

Hansen, Bruce E., 1995. Rethinking the univariate approach to unit root tests: how to use covariates to increase power. Econometric Theory 11, 1148–1171.

Hansen, Bruce E., 2007. Least squares model averaging. Econometrica 75, 1175–1189.

Hansen, Bruce E., 2008. Least squares forecast averaging. Journal of Econometrics 146, 342–350.

Hansen, Bruce E., 2009. Averaging estimators for regressions with a possible structural break. Econometric Theory 35, 1498–1514.

Hansen, Bruce E., Racine, Jeffrey S., 2007. Jackknife model averaging. Working Paper. University of Wisconsin.

Hendry, D.F., Clements, M.P., 2002. Pooling of forecasts. Econometrics Journal 5, 1–26.

Hjort, Nils Lid, Claeskens, Gerda, 2003. Frequentist model average estimators. Journal of the American Statistical Association 98, 879–899.

Ing, Ching-Kang, 2003. Multistep prediction in autoregressive processes. Econometric Theory 19, 254–279.

Ing, Ching-Kang, 2004. Selecting optimal multistep predictors for autoregressive processes of unknown order. Annals of Statistics 32, 693–722.

Ing, Ching-Kang, Wei, Ching-Zong, 2003. On same-realization prediction in an infinite-order autoregressive process. Journal of Multivariate Analysis 85, 130–155.

Ing, Ching-Kang, Wei, Ching-Zong, 2005. Order selection for same-realization predictions in autoregressive processes. Annals of Statistics 33, 2423–2474.

Inoue, Atsuhsi, Kilian, Lutz, 2006. On the selection of forecasting models. Journal of Econometrics 130, 272–306.

Kapetanios, George, 2004. The asymptotic distribution of the cointegration rank estimator under the Akaike information criterion. Econometric Theory 20, 735–743.

Kemp, Gordon C.R., 1999. The behavior of forecast errors from a nearly integrated AR(1) model as both sample size and forecast horizon become large. Econometric Theory 15, 238–256.

Kim, Tae-Hwan, Leybourne, Stephen J., Newbold, Paul, 2004. Asymptotic mean-squared forecast error when an autoregression with linear trend is fitted to data generated by an I(0) or I(1) process. Journal of Time Series Analysis 25, 583–602.

Lee, Sangyeol, Karagrigoriou, Alex, 2001. An asymptotically optimal selection of the order of a linear process. Sankhya Series A 63, 93–106.

Mallows, C.L., 1973. Some comments on C_p. Technometrics 15, 661–675.

Pesaran, M. Hashem, Timmermann, Allan, 2007. Selection of estimation window in the presence of breaks. Journal of Econometrics 137, 134–161.

Phillips, Peter C.B., 1988a. Regression theory for near-integrated time series. Econometrica 56, 1021–1043.

Phillips, Peter C.B., 1988b. Towards a unified asymptotic theory for autoregression. Biometrika 74, 535–547.

Rissanen, J., 1986. Stochastic complexity and modeling. Annals of Statistics 14, 1080–1100.

Schwarz, G., 1978. Estimating the dimension of a model. Annals of Statistics 6, 461–464.

Shibata, Ritaei, 1980. Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. Annals of Statistics 8, 147–164.

Stock, James H., 1996. VAR, error correction and pretest forecasts at long horizons. Oxford Bulletin of Economics and Statistics 58, 685–701.

Stock, J.H., Watson, M.W., 2006. Forecasting with many predictors. In: Elliott, Granger, Timmermann, (Eds.), Handbook of Economic Forecasting. Elsevier, (Chapter 10).

Stock, J.H., Watson, M.W., 1999. A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series. In: Engle, , White, (Eds.), Cointegration, Causality and Forecasting: A Festschrift for Clive W.J. Granger. Oxford University Press.

Stock, J.H., Watson, M.W., 2004. Combination forecasts of output growth in a seven-country data set. Journal of Forecasting 23, 405–430.

Stock, J.H., Watson, M.W., 2005. An empirical comparison of methods for forecasting using many predictors. Working Paper. NBER.

Timmermann, Allan, 2006. Forecast combinations. In: Elliott, Granger, Timmermann, (Eds.), Handbook of Economic Forecasting. Elsevier (Chapter 4).

Wei, Ching-Zong, 1992. On predictive least squares principles. Annals of Statistics 20, 1–42.