

Averaging Estimators for Autoregressions with a Near Unit Root

Bruce E. Hansen*
University of Wisconsin†

www.ssc.wisc.edu/~bhansen

May 2007

Abstract

This paper uses local-to-unity theory to evaluate the asymptotic mean-squared error and forecast loss 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 MSE and forecast loss depend only on the local-to-unity parameter, facilitating simple graphical comparisons. Our results strongly caution against the use of pretesting. Strong evidence supports averaging based on Mallows weights. This is a new approach to forecast combination.

*Research supported by the National Science Foundation.

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

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 mean-squared error and forecast loss. Our view is that if the purpose of autoregressive estimation is for forecasting, then model selection methods should be designed to minimize forecast loss. 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 the asymptotic mean-squared error and forecast loss 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 selection estimator, and finally the Mallows averaging estimator. In the local-to-unity framework, the normalized asymptotic loss of all of these estimators depends exclusively on the local-to-unity parameter, facilitating graphical comparisons. The conclusions are clear. On one side, we find that the classic Dickey-Fuller pre-test estimator has very poor MSE and forecast loss. On the other side, we find that our new Mallows averaging estimator has the best performance. 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 (1987, 1988).

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 pretesting. 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, Leybourne and Newbold (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 pretest estimators.

Autoregressive models with unit roots are a special case of cointegrated vector autoregressions

(Engle and Granger, 1987). Johansen (1987, 1991, 1995) introduced the likelihood ratio tests for cointegration and argued for their use as a method to determine the cointegration rank. 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 (2006). The idea of using a local-to-zero parameterization to study the asymptotic distribution of pretest 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.

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 loss. Section 5 covers Dickey-Fuller pre-testing. Section 6 presents Mallows section. 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:

$$y_t = \beta_0 + \beta_1 t + \cdots + \beta_p t^p + S_t \quad (1)$$

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

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

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

Differencing (1) and substituting (2) implies

$$\Delta y_t = \delta_t' \theta_0 + x_t' \theta_1 + z_t' \alpha + e_t \quad (3)$$

where

$$\begin{aligned} \delta_t &= \begin{pmatrix} 1 \\ t \\ \vdots \\ t^{p-1} \end{pmatrix}, & x_t &= \begin{pmatrix} t^p \\ y_{t-1} \end{pmatrix}, & z_t &= \begin{pmatrix} \Delta y_{t-1} \\ \vdots \\ \Delta y_{t-k} \end{pmatrix}, \\ \theta_1 &= \begin{pmatrix} -\alpha_0 \beta_p \\ \alpha_0 \end{pmatrix} & \alpha &= \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_k \end{pmatrix}, \end{aligned}$$

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

$$\mu_t = \delta_t' \theta_0 + x_t' \theta_1 + z_t' \alpha. \quad (4)$$

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

$$\Delta y_t = \delta_t' \hat{\theta}_0 + x_t' \hat{\theta}_1 + z_t' \hat{\alpha} + \hat{e}_t. \quad (5)$$

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

$$\Delta y_t = \delta_t' \tilde{\theta}_0 + z_t' \tilde{\alpha} + \tilde{e}_t. \quad (6)$$

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 Mean-Squared Error

To evaluate the quality of our estimators, we use two measures of loss. 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 MSE 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 $n \rightarrow \infty$. Let $W(r)$ denote a standard Brownian motion and define the diffusion process

$$dW_c(r) = cW_c(r) + dW(r) \tag{7}$$

which satisfies

$$W_c(r) = \int_0^r \exp(c(r-s)) dW(s). \tag{8}$$

Also define the trend functions

$$\begin{aligned} \delta(r) &= \begin{pmatrix} 1 \\ r \\ \vdots \\ r^{p-1} \end{pmatrix}, \\ X_c(r) &= \begin{pmatrix} r^p \\ W_c(r) \end{pmatrix}, \end{aligned} \tag{9}$$

and the detrended processes

$$\begin{aligned} W_c^*(r) &= W_c(r) - \int_0^1 W_c \delta' \left(\int_0^1 \delta \delta' \right)^{-1} \delta(r) \\ X_c^*(r) &= X_c(r) - \int_0^1 X_c \delta' \left(\int_0^1 \delta \delta' \right)^{-1} \delta(r). \end{aligned}$$

Theorem 1 *The asymptotic MSE of the constrained estimator is*

$$m_0(c, p, k) \equiv \lim_{n \rightarrow \infty} \frac{1}{\sigma^2} \sum_{t=1}^n E(\tilde{\mu}_t - \mu_t)^2 = m_0(c, p) + p + k \quad (10)$$

where

$$\begin{aligned} m_0(c, p) &= EF_{0c}, \\ F_{0c} &= c^2 \int_0^1 W_c^{*2}. \end{aligned} \quad (11)$$

For $p = 0$ we can calculate

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

and for $p = 1$,

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

The asymptotic MSE of the unconstrained estimator is

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

where

$$\begin{aligned} m_1(c, p) &= EF_{1c}, \\ F_{1c} &= \left(\int_0^1 dW X_c^{*'} \right) \left(\int_0^1 X_c^* X_c^{*'} \right)^{-1} \left(\int_0^1 X_c^* dW \right). \end{aligned} \quad (15)$$

A closed-form expression for $m_1(c, p)$ is not available, but for all p , $\lim_{c \rightarrow -\infty} m_1(c, p) = 2$.

The asymptotic MSE of the averaging estimator is

$$\begin{aligned} m_w(c, p, k) &\equiv \lim_{n \rightarrow \infty} \frac{1}{\sigma^2} \sum_{t=1}^n E(\hat{\mu}_t(w) - \mu_t)^2 \\ &= m_w(c, p) + p + k \end{aligned} \quad (16)$$

where

$$\begin{aligned} m_w(c, p) &= w^2 m_1(c, p) + (1 - w)^2 m_0(c, p) + 2w(1 - w) m_{01}(c, p), \\ m_{01}(c, p) &= -E \left(c \int_0^1 W_c^* dW \right). \end{aligned}$$

When $p = 0$ then

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

and for $p = 1$

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

In general, if $c \leq 0$ then $m_{01}(c, p) \leq 0$. Also

$$\lim_{c \rightarrow -\infty} m_{01}(c, p) = -p. \quad (18)$$

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 mean-squared error is

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

The MSE of all estimators are the sum of $p + k$ plus the additional component $m_0(c, p)$, $m_1(c, p)$, or $m_w(c, p)$. $p + k$ is the normalized variance from the estimation of the coefficients θ_0 and α , which are common across the three models and estimators. For the constrained estimator, $m_0(c, p)$ reflects 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 is thus non-standard. 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 MSE of the optimal averaging estimator is strictly less than both the unrestricted and restricted estimators.

The functions $m_0(c, 1, 0)$, $m_1(c, 1, 0)$ and $m_{w_m}(c, 1, 0)$ are displayed¹ in Figure 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 m_0

¹Figures 1 through 5 are computed on a grid of 201 evenly-spaced points from -20 to 0 . Functions without analytic expressions were calculated by simulation. The asymptotic distributions were approximated by finite-sample counterparts with 1000 observations. 200,000 simulation replications were used.

is approximately linear in c , monotonically increasing as c moves away from zero. The curve m_1 is also monotonic, but with the opposite slope. m_1 obtains its maximal value of 7.3 at $c = 0$, and asymptotically approaches 3 as $c \rightarrow -\infty$. The lines m_0 and m_1 intersect at $c = -8.5$, meaning that for $c > -8.5$, the restricted (unit root) estimator has lower MSE than the unconstrained estimator, while for $c < -8.5$ the unconstrained estimator has lower MSE. For all c , the MSE of the optimal averaging estimator is substantially below the MSE of the other two estimators.

4 Forecast Loss

In this section we consider an alternative measure of loss, the one-step-ahead forecast loss, which is a more direct measure of forecasting ability than in-sample MSE.

Theorem 2 *The asymptotic forecast loss of the constrained estimator is*

$$f_0(c, p, k) \equiv \lim_{n \rightarrow \infty} \frac{n}{\sigma^2} E(\tilde{\mu}_{n+1} - \mu_{n+1})^2 = f_0(c, p) + k \quad (19)$$

where

$$\begin{aligned} f_0(c, p) &= ET_{0c}^2, \\ T_{0c} &= -cW_c^*(1) + \delta(1)' \left(\int_0^1 \delta\delta' \right)^{-1} \int_0^1 \delta dW. \end{aligned} \quad (20)$$

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 loss of the unconstrained estimator is

$$f_1(c, p, k) \equiv \lim_{n \rightarrow \infty} \frac{n}{\sigma^2} E(\hat{\mu}_{n+1} - \mu_{n+1})^2 = f_1(c, p) + k \quad (21)$$

where

$$\begin{aligned} f_1(c, p) &= E(T_{1c}^2) \\ T_{1c} &= \delta(1)' \left(\int_0^1 \delta\delta' \right)^{-1} \int_0^1 \delta dW + X_c^*(1)' \left(\int_0^1 X_c^* X_c^{*'} \right)^{-1} \int_0^1 X_c^* dW. \end{aligned} \quad (22)$$

The asymptotic forecast loss of the averaging estimator is

$$f_w(c, p, k) = f_w(c, p) w^2 f_1(c, p) + (1 - w)^2 f_0(c, p) + 2w(1 - w) f_{01}(c, p) + k \quad (23)$$

where

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

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 mean-squared error 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 functions $f_0(c, 1, 0)$, $f_1(c, 1, 0)$, and $f_{w_f}(c, 1, 0)$ are displayed in Figure 2. The features are qualitatively similar those displayed in Figure 1. One difference is that f_1 is not monotonic in c . The curves f_0 and f_1 intersect at $c = -10.5$, so for a greater range of values of c the restricted estimator has lower forecast loss relative to the unrestricted estimator.

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

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

We now present the MSE and forecast loss for this estimate.

Theorem 3

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

and

$$f_{df}(c, p, k) = \lim_{n \rightarrow \infty} \frac{n}{\sigma^2} E \left(\hat{\mu}_{n+1}^{df} - \mu_{n+1} \right)^2 = E(T_{1c}^2 1(DF_c \leq r)) + E(T_{0c}^2 1(DF_c > r)) + k$$

where F_{0c} , F_{1c} , T_{0c} , and T_{1c} are defined in (11), (15), (20) and (22), respectively,

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

and

$$\begin{aligned} W_c^\tau(r) &= W_c(r) - \int_0^1 W_c \tau' \left(\int_0^1 \tau \tau' \right)^{-1} \tau(r) \\ \tau(r) &= \begin{pmatrix} 1 \\ \vdots \\ r^p \end{pmatrix}. \end{aligned}$$

The function $m_{df}(c, 1, 0)$ is displayed in Figure 3 (the dashed line) and $f_{df}(c, 1, 0)$ is displayed in Figure 4, and can be contrasted with those for the unconstrained estimator (the solid line). While the MSE and forecast loss of the pre-test estimator is quite low for small values of $-c$, it is quite large for moderate to large values of $-c$. (This is similar to the behavior of pre-test estimates in stationary models.) Comparing the pre-test estimator with the unrestricted estimator we see neither uniform dominates the other. The forecast loss of the pre-test estimator is more sensitive to c , and has a meaningfully higher maximum value. We believe that this comparison demonstrates that the DF pre-test estimator is a poor choice relative to unrestricted estimation.

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 Figures 6-9, even for his DF-GLS pretest estimator. Diebold and Kilian (2000) largely miss the difficulty by focusing exclusively on very small and very large values of $|c|$ – both cases where pretesting works well. These papers also focus on the long horizon forecasting case, where they argue that pretesting 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. We recommend the Mallows criterion. Let $\hat{\sigma}^2 = n^{-1} \sum_{t=1}^n (y_t - \hat{\mu}_t)^2$ and $\tilde{\sigma}^2 = n^{-1} \sum_{t=1}^n (y_t - \tilde{\mu}_t)^2$ be the estimates of σ^2 from the unrestricted and restricted estimators. The Mallows criterion for the unit root model and the unrestricted model are

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

and

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

respectively, 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. Classic Mallows selection picks the model with the smallest criterion, which is equivalent to selecting the unrestricted estimator when $F_n \geq 2((2 + p + k) - (p + k)) = 4$ where

$$F_n = n \left(\frac{\tilde{\sigma}^2 - \hat{\sigma}^2}{\hat{\sigma}^2} \right) \quad (24)$$

is the classic Wald statistic for the joint exclusion of y_{t-1} and the time trend from (3). The effective critical value is twice the difference in the number of fitted parameters.

It turns out that since the parameter estimates have nonstandard distributions (due to the nearly integrated regressor), the penalty terms are not quite right. Instead, the optimal Mallows criteria are

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

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

These modification are necessary to obtained unbiased estimates of the in-sample mean-squared error.

Theorem 4

$$\begin{aligned} \frac{EM_0(c)}{\sigma^2} - n &\rightarrow m_0(c, p, k) \\ \frac{EM_1(c)}{\sigma^2} - n &\rightarrow m_1(c, p, k). \end{aligned}$$

Theorem 4 shows that the criteria $M_0(c)$ and $M_1(c)$ are (after normalization) asymptotically unbiased estimates of the mean-squared error. However, they are not feasible criteria as they depend on the unknown c . We suggest using their asymptotes

$$\begin{aligned} M_0^* &= \lim_{c \rightarrow -\infty} M_0(c) = n\tilde{\sigma}^2 + 2\hat{\sigma}^2 k \\ M_1^* &= \lim_{c \rightarrow -\infty} M_1(c) = n\tilde{\sigma}^2 + 2\hat{\sigma}^2(2 + p + k) = M_1 \end{aligned}$$

(Recall from Theorem 1 that $m_{01}(c, p) \rightarrow -p$ and $m_1(c, p) \rightarrow 2$ as $c \rightarrow -\infty$.) This is a conservative choice because it imposes a relatively small penalty on the unrestricted model, and thus favors the unrestricted model. Our modified Mallows criterion has a non-classical penalty for the unit root model. This is analogous to the finding of Chao and Phillips (1999) in their study of Bayesian model selection in reduced rank VARs. A notable difference is that our relative penalty depends on the order of the fitted trend function.

Modified Mallows selection picks $\hat{\mu}_t$ if $M_1^* < M_0^*$, which is equivalent to the event $F_n \geq 2(2+p)$ where F_n was defined in (24). We can write the Mallows selected estimator as

$$\hat{\mu}_t^m = \hat{\mu}_t 1(F_n \geq 2(2+p)) + \tilde{\mu}_t 1(F_n < 2(2+p)).$$

The effective critical value is twice $(2+p)$, or 6 when $p = 1$.

We now characterize the MSE and forecast loss of Modified Mallows selection.

Theorem 5

$$\begin{aligned} m_m(c, p, k) &= \lim_{n \rightarrow \infty} \frac{1}{\sigma^2} \sum_{t=1}^n E(\hat{\mu}_t^m - \mu_t)^2 \\ &= E(F_{1c} 1(F_c \geq 2(2+p))) + E(F_{0c} 1(F_c < 2(2+p))) + p + k \end{aligned}$$

and

$$\begin{aligned} f_m(c, p, k) &= \lim_{n \rightarrow \infty} \frac{n}{\sigma^2} E(\hat{\mu}_t^m - \mu_{n+1})^2 \\ &= E(T_{1c}^2 1(F_c \geq 2(2+p))) + E(T_{0c}^2 1(F_c < 2(2+p))) + k \end{aligned}$$

where F_{0c} , F_{1c} , T_{0c} , and T_{1c} are defined in (11), (15), (20) and (22), and

$$F_c = \left(\int_0^1 dW_c X_c^{*'} \right) \left(\int_0^1 X_c^* X_c^{*'} \right)^{-1} \left(\int_0^1 X_c^* dW_c \right). \quad (25)$$

The function $m_m(c, 1, 0)$ is displayed in Figure 3 (the closely spaced dots) and $f_m(c, 1, 0)$ is displayed in Figure 4. Relative to the unconstrained estimator, the Mallows selection estimator has lower MSE and forecast loss for small c , larger for moderate c , and similar values for large c . The behavior is qualitatively similar to the Dickey-Fuller pre-test estimator but with much reduced variation. The maximum MSE and forecast loss of the Mallows estimator is much smaller than the DF estimator. The Mallows estimator has substantially smaller maximum MSE than the unconstrained estimator, but the maximum forecast loss of the Mallows estimator is similar to the unrestricted estimator. We conclude that the Mallows estimator is preferred to the Dickey-Fuller estimator, but it is not necessarily preferred to the unconstrained estimator.

7 Mallows Averaging

For any w let

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

be the variance estimate using the averaging estimator $\hat{\mu}_t(w)$. The classic Mallows criterion for the averaging estimator (Hansen, 2007) is

$$\begin{aligned} M_w &= n\hat{\sigma}^2(w) + 2\hat{\sigma}^2(w(2+p+k) + (1-w)(p+k)) \\ &= n\hat{\sigma}^2(w) + 2\hat{\sigma}^2(2w+p+k). \end{aligned}$$

However, as discussed in the previous section the penalty terms are incorrect. The optimal Mallows criterion is instead

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

This criterion is asymptotically unbiased for the mean-squared error.

Theorem 6

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

We suggest replacing $M_w(c)$ with its asymptote

$$M_w^* = \lim_{c \rightarrow \infty} M_0(c) = n\hat{\sigma}^2(w) + 2\hat{\sigma}^2((2+p)w+k).$$

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*

$$\hat{w} = \begin{cases} 1 - \frac{(2+p)}{F_n} & \text{if } F_n > (2+p) \\ 0 & \text{otherwise} \end{cases}$$

where F_n is the classic Wald statistic (24).

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

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

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

$$\hat{w} \xrightarrow{d} \pi_c = \begin{cases} 1 - \frac{(2 + p)}{F_c} & \text{if } F_c > (2 + p) \\ 0 & \text{otherwise,} \end{cases}$$

where F_c is defined in (25). The asymptotic mean-squared error of the Mallows averaging estimator is

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

and the asymptotic forecast loss is

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

where F_{0c} , F_{1c} , T_{0c} , and T_{1c} are defined in (11), (15), (20) and (22), respectively.

The functions $m_a(c, 1, 0)$ and $f_a(c, 1, 0)$ are displayed in Figures 3 and 4. The performance of the Mallows averaging estimator is stunning relative to the other estimators. In both figures, the averaging estimator uniformly dominates the Mallows selection estimator and the unrestricted estimator, and for most values of c the improvement is quite significant. The averaging estimator does not uniformly dominate the DF estimator, however, as the latter has somewhat smaller MSE for the smallest values of c . Overall, the averaging estimator has the best MSE and forecast loss performance, and is therefore the best choice among the feasible estimators considered.

8 Finite Sample MSE and Forecast Loss

The analysis of the previous sections has been asymptotic. We have found that the asymptotic MSE and forecast loss are only a function of the local-to-unity parameter c , 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 loss, 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 iid $N(0, 1)$. 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, $\alpha_1 = \dots = \alpha_k = 0$. 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 $[-.4, 0]$ for $n = 50$ and a range of $[-.1, 0]$ for $n = 200$.

For each parameter configuration, we calculate the forecast loss $nE(\hat{\mu}_{n+1} - \mu_{n+1})^2$ for three estimators: the unrestricted least-squares estimator $\hat{\mu}_{n+1}$, the Dickey-Fuller pre-test estimator $\hat{\mu}_{n+1}^{df}$, and the Mallows averaging estimator $\hat{\mu}_{n+1}^a$. The loss was calculated by Monte Carlo simulation, taking the average of $n(\hat{\mu}_{n+1} - \mu_{n+1})^2$ across 200,000 simulation draws.

The results are presented in Figure 5. There are six panels, one for each (n, k) pair. In each panel, the forecast loss is plotted as a function of c . These panels are finite sample analogs of the asymptotic loss as reported in Figure 4. What is striking is that all of the panels in Figure 5 are quite similar to Figure 4. The scaled finite sample forecast loss is nearly identical to the asymptotic loss. 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 loss, and is noticeably dominated by the pre-test and averaging estimators for all values of c .

Our second experiment adds serial correlation. We do this by setting the autoregressive parameters as $\alpha_j = -(-\theta)^j$ for $j = 1, \dots, k$, for $\theta = 0.6$ (the results are not sensitive to this choice). We then set $\alpha_0 = (1 - a_1 - \dots - a_k)c/n$ 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 Figure 6 and are very similar to Figure 5. As predicted by the asymptotic theory, the forecast loss 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 $k \in \{0, 1, \dots, K\}$. For each $k = 0, \dots, K$, 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: $w_{1k} \geq 0$, $w_{0k} \geq 0$, and $\sum_{k=0}^K (w_{0k} + w_{1k}) = 1$. The general averaging estimator of μ_t is

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

where

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

The Mallows criterion for weight selection, as described by Hansen (2007) and amended as suggested in the previous section, is

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

where

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

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 $\tilde{e}(k) = y - \tilde{\mu}(k)$ be the $n \times 1$ vectors of residuals from the individual models, and construct the $n \times (2K + 2)$ matrix

$$\hat{e} = [\tilde{e}(0), \dots, \tilde{e}(K), \hat{e}(0), \dots, \hat{e}(K)]$$

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

$$R = \begin{pmatrix} 0 \\ \vdots \\ K \\ 2 + p \\ \vdots \\ 2 + p + K \end{pmatrix}.$$

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

written as

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

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 results are presented in Figure 7 for $\theta = 0.6$. The results are qualitatively similar across K and n . In all cases, the general averaging estimator has the lowest forecast loss, and the selection estimator has the highest forecast loss. 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.

10 Conclusion

This paper examined the question of selection and combination of autoregressive model when the goal is minimizing squared forecast error. Using local-to-unity asymptotic methods, we found that asymptotic MSE and forecast loss of a variety of estimators are functions only of the local-to-unity parameter, facilitating direct comparisons. We examine 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 poor performance, while the Mallows averaging estimator has the smallest MSE and forecast loss. 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.

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 produce lower forecast MSE than forecasts constructed from models selected by 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, Rothenberg and Stock (1996) can reduce the forecast loss 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 each weight) often achieve lower forecast loss 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.

11 Appendix

The following results will be useful in subsequent calculations.

Lemma 1

$$E(W_c(r)^2) = \frac{\exp(2cr) - 1}{2c}, \quad (26)$$

$$E\left(\int_0^1 W_c^2\right) = \frac{1}{2c} \left(\frac{\exp(2c) - 1}{2c} - 1 \right), \quad (27)$$

$$E(W_c(1)W(1)) = \frac{\exp(c) - 1}{c}. \quad (28)$$

Proof: Using (8) and the fact that $dW(s)$ is an orthogonal process,

$$\begin{aligned} E(W_c(r)^2) &= E \int_0^r \int_0^r \exp(c(r-s)) \exp(c(r-u)) dW(s) dW(u) \\ &= \int_0^r \exp(2c(r-s)) ds \\ &= \frac{\exp(2cr) - 1}{2c} \end{aligned}$$

which is (26). Equation (27) follows by integration. To show (28),

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

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

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

and

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

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

$$\begin{aligned} D_{0n}^{-1} \delta_{[nr]} &\xrightarrow{d} \delta(r), \\ D_{1n}^{-1} x_{[nr]} &\xrightarrow{d} X_c(r). \end{aligned}$$

Define the orthogonalized series

$$\begin{aligned} x_t^* &= x_t - \delta_t' \left(\sum_{j=1}^n \delta_j \delta_j' \right)^{-1} \sum_{j=1}^n \delta_j x_j \\ z_t^* &= z_t - \delta_t' \left(\sum_{j=1}^n \delta_j \delta_j' \right)^{-1} \sum_{j=1}^n \delta_j z_j \\ S_{t-1}^* &= S_{t-1} - \delta_t' \left(\sum_{j=1}^n \delta_j \delta_j' \right)^{-1} \sum_{j=1}^n \delta_j S_{j-1} \end{aligned}$$

and observe that

$$\begin{aligned} \frac{a}{\sigma\sqrt{n}} S_{[nr]}^* &\xrightarrow{d} W_c^*(r), \\ D_{1n}^{-1} x_{[nr]}^* &\xrightarrow{d} X_c^*(r). \end{aligned}$$

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

$$\begin{aligned} \frac{n^{1/2}}{\sigma} D_{0n} (\tilde{\theta}_0^* - \theta_0) &= \frac{1}{\sigma} D_{0n} \left(\frac{1}{n} \sum_{t=1}^n \delta_t \delta_t' \right)^{-1} \left(\frac{1}{\sqrt{n}} \sum_{t=1}^n \delta_t e_t \right) + o_p(1) \\ &\xrightarrow{d} \left(\int_0^1 \delta \delta' \right)^{-1} \left(\int_0^1 \delta dW \right). \end{aligned} \tag{29}$$

Also

$$\begin{aligned} \frac{n^{1/2}}{\sigma} (\tilde{\alpha} - \alpha) &= \left(\frac{1}{n} \sum_{t=1}^n z_t^* z_t^{*'} \right)^{-1} \left(\frac{1}{\sigma\sqrt{n}} \sum_{t=1}^n z_t^* e_t + o_p(1) \right) \\ &\xrightarrow{d} Z \sim N(0, Q^{-1}) \end{aligned} \tag{30}$$

where $Q = E(z_t^* z_t^{*'})$.

We can write

$$\begin{aligned} \tilde{\mu}_t - \mu_t &= -can^{-1}S_{t-1} + \left(\tilde{\theta}_0 - \theta_0 \right)' \delta_t + (\tilde{\alpha} - \alpha)' z_t^* \\ &= -can^{-1}S_{t-1}^* + \left(\tilde{\theta}_0^* - \theta_0 \right)' \delta_t + (\tilde{\alpha} - \alpha)' z_t^*. \end{aligned} \tag{31}$$

so

$$\begin{aligned}
\frac{1}{\sigma^2} \sum_{t=1}^n (\tilde{\mu}_t - \mu_t)^2 &= \frac{c^2 a^1}{\sigma^2 n^2} \sum_{t=1}^n S_{t-1}^{*2} + \frac{1}{\sigma^2} (\tilde{\theta}_0^* - \theta_0)' \sum_{t=1}^n \delta_t \delta_t' (\tilde{\theta}_0^* - \theta_0) \\
&\quad + \frac{1}{\sigma^2} (\tilde{\alpha} - \alpha)' \sum_{t=1}^n z_t^* z_t^{*'} (\tilde{\alpha} - \alpha) + o_p(1) \\
&\xrightarrow{d} c^2 \int_0^1 W_c^{*2} + \left(\int_0^1 dW \delta' \right) \left(\int_0^1 \delta \delta' \right)^{-1} \left(\int_0^1 \delta dW \right) + Z' Q Z \\
&= F_{0c} + \chi_{p+k}^2.
\end{aligned} \tag{32}$$

where

$$\chi_{p+k}^2 = \left(\int_0^1 dW \delta' \right) \left(\int_0^1 \delta \delta' \right)^{-1} \left(\int_0^1 \delta dW \right) + Z' Q Z$$

is chi-square with degrees of freedom $p + k$. Taking expectations of (32) we obtain (10).

When $p = 0$ then there is no $\delta(r)$. Thus using (27),

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

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

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

Equation (7) implies

$$c \int_0^1 W_c = W_c(1) - W(1) \tag{33}$$

and thus using (26), (27), and (28), we find

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

which is (13).

We next consider the unconstrained estimator. Note that

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

We calculate that

$$\frac{n^{1/2}}{\sigma} D_{0n} (\hat{\theta}_0 - \theta_0) \xrightarrow{d} \left(\int_0^1 \delta \delta' \right)^{-1} \int_0^1 \delta dW, \tag{34}$$

$$\frac{n^{1/2}}{\sigma} D_{1n} (\hat{\theta}_1 - \theta_1) \xrightarrow{d} \left(\int_0^1 X_c^* X_c^{*'} \right)^{-1} \int_0^1 X_c^* dW, \quad (35)$$

and

$$\frac{n^{1/2}}{\sigma} (\hat{\alpha} - \alpha) \xrightarrow{d} Z \quad (36)$$

as in (30).

We find

$$\begin{aligned} \frac{1}{\sigma^2} \sum_{t=1}^n (\hat{\mu}_t - \mu_t)^2 &= \frac{1}{\sigma^2} \sum_{t=1}^n \left(\delta_t' (\hat{\theta}_0 - \theta_0) + x_t^{*'} (\hat{\theta}_1 - \theta_1) + z_t^{*'} (\hat{\alpha} - \alpha) \right)^2 \\ &= \frac{1}{\sigma^2} (\hat{\theta}_1 - \theta_1)' \sum_{t=1}^n x_t^* x_t^{*'} (\hat{\theta}_1 - \theta_1) \\ &\quad + \frac{1}{\sigma^2} (\hat{\theta}_0 - \theta_0)' \sum_{t=1}^n \delta_t \delta_t' (\hat{\theta}_0 - \theta_0) \\ &\quad + \frac{1}{\sigma^2} (\hat{\alpha} - \alpha)' \sum_{t=1}^n z_t^* z_t^{*'} (\hat{\alpha} - \alpha) \\ &\quad + 2 \frac{1}{\sigma^2} (\hat{\theta}_1 - \theta_1)' \sum_{t=1}^n x_t^* z_t^{*'} (\hat{\alpha} - \alpha) \\ &\xrightarrow{d} \left(\int_0^1 dW X_c^{*'} \right) \left(\int_0^1 X_c^* X_c^{*'} \right)^{-1} \left(\int_0^1 X_c^* dW \right) \\ &\quad + \left(\int_0^1 dW \delta' \right) \left(\int_0^1 \delta \delta' \right)^{-1} \left(\int_0^1 \delta dW \right) + Z' Q Z \\ &= F_{1c} + \chi_{p+k}^2. \end{aligned} \quad (37)$$

Taking expectations of (37) 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

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

and

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

Noting that

$$\hat{\mu}_t(w) - \mu_t = w x_t^{*'} (\hat{\theta}_1 - \theta_1) - (1-w) c a n^{-1} S_{t-1}^* + (\hat{\theta}_0(w) - \theta_0)' \delta_t + (\hat{\alpha}(w) - \alpha)' z_t^*, \quad (38)$$

we see

$$\begin{aligned}
\frac{1}{\sigma^2} \sum_{t=1}^n (\hat{\mu}_t(w) - \mu_t)^2 &= w^2 \frac{1}{\sigma^2} (\hat{\theta}_1 - \theta_1)' \sum_{t=1}^n x_t^* x_t^{*'} (\hat{\theta}_1 - \theta_1) \\
&\quad + (1-w)^2 \frac{c^2 a^2}{\sigma^2 n} \sum_{t=1}^n S_{t-1}^{*2} \\
&\quad + \frac{1}{\sigma^2} (\hat{\theta}_0(w) - \theta_0)' \sum_{t=1}^n \delta_t \delta_t' (\hat{\theta}_0(w) - \theta_0) \\
&\quad - 2w(1-w) \frac{ca}{n\sigma^2} (\hat{\theta}_1 - \theta_1)' \sum_{t=1}^n x_t^* S_{t-1}^* \\
&\quad + \frac{1}{\sigma^2} (\hat{\alpha}(w) - \alpha)' \sum_{t=1}^n z_t^* z_t^{*'} (\hat{\alpha}(w) - \alpha) \\
&\quad + o_p(1) \\
&\xrightarrow{d} w^2 \left(\int_0^1 dW X_c^{*'} \right) \left(\int_0^1 X_c^* X_c^{*'} \right)^{-1} \left(\int_0^1 X_c^* dW \right) \\
&\quad + (1-w)^2 c^2 \int_0^1 W_c^{*2} \\
&\quad - 2w(1-w)c \left(\int_0^1 dW X_c^{*'} \right) \left(\int_0^1 X_c^* X_c^{*'} \right)^{-1} \int_0^1 X_c^* W_c^* + \chi_{p+k}^2 \\
&= w^2 F_{1c} + (1-w^2) F_{0c} - 2w(1-w)c \int_0^1 dW W_c^* + \chi_{p+k}^2. \tag{39}
\end{aligned}$$

Taking expectations establishes (16).

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

$$\begin{aligned}
m_{01}(c, p) &= -Ec \int_0^1 dW W_c + E \left(c \int_0^1 dW \delta' \left(\int_0^1 \delta \delta' \right)^{-1} \int_0^1 \delta W_c \right) \\
&= E \left(c \int_0^1 dW \delta' \left(\int_0^1 \delta \delta' \right)^{-1} \int_0^1 \delta W_c \right) \tag{40}
\end{aligned}$$

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 (33) and (28),

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

which is (16).

To see that $m_{01}(c, p) \leq 0$ when $c \leq 0$, using (40), expanding the integrals, noting that $E(W_c(r)dW(s)) = \exp(c(r-s))$ for $s \leq r$ and zero otherwise, we find

$$\begin{aligned}
m_{01}(c, p) &= c \operatorname{tr} \left[\left(\int_0^1 \delta \delta' \right)^{-1} E \left(\int_0^1 \delta W_c \right) \left(\int_0^1 dW \delta' \right) \right] \\
&= c \operatorname{tr} \left[\left(\int_0^1 \delta \delta' \right)^{-1} E \left(\int_0^1 \int_0^1 \delta(s) \delta(r)' W_c(r) dW(s) ds dr \right) \right] \\
&= c \operatorname{tr} \left[\left(\int_0^1 \delta \delta' \right)^{-1} \left(\int_0^1 \int_0^r \delta(s) \delta(r)' \exp(c(r-s)) ds dr \right) \right] \tag{41}
\end{aligned}$$

which has the same sign as c since the function $\delta(s)$ and the exponential are non-negative. The function $-c \exp(ca)$ is a kernel function on $a \in [0, \infty)$ when $c < 0$. As $c \rightarrow -\infty$, it degenerates to the dirac function with unit mass at $a = 0$. Thus the limit as $c \rightarrow -\infty$ of (41) is $-\operatorname{tr} \left[\left(\int_0^1 \delta \delta' \right)^{-1} \left(\int_0^1 \delta \delta' \right) \right] = -p$, which is (18).

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

$$\hat{\mu}_{n+1} - \mu_{n+1} = \delta'_{n+1} (\hat{\theta}_0 - \theta_0) + x_{n+1}^{*'} (\hat{\theta}_1 - \theta_1) + z_{n+1}^{*'} (\hat{\alpha} - \alpha).$$

Using (34) and (35), note that

$$\begin{aligned}
&\frac{n^{1/2}}{\sigma} \left(\delta'_{n+1} (\hat{\theta}_0 - \theta_0) + x_{n+1}^{*'} (\hat{\theta}_1 - \theta_1) \right) \\
&\xrightarrow{d} \delta(1)' \left(\int_0^1 \delta \delta' \right)^{-1} \int_0^1 \delta dW + X_c^*(1)' \left(\int_0^1 X_c^* X_c^{*'} \right)^{-1} \int_0^1 X_c^* dW = T_{1c} \tag{42}
\end{aligned}$$

and thus

$$\frac{n}{\sigma^2} E \left(\delta'_{n+1} (\hat{\theta}_0 - \theta_0) + x_{n+1}^{*'} (\hat{\theta}_1 - \theta_1) \right)^2 \rightarrow ET_{1c}^2.$$

Furthermore using (36)

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

so

$$\begin{aligned}
\frac{n}{\sigma^2} E ((\tilde{\alpha} - \alpha)' z_{n+1}^*)^2 &= \frac{n}{\sigma^2} \operatorname{tr} E (z_{n+1}^* z_{n+1}^{*'} (\tilde{\alpha} - \alpha) (\tilde{\alpha} - \alpha)') \\
&\rightarrow \operatorname{tr} E (z_{n+1}^* z_{n+1}^{*'} ZZ') \\
&= E (Z' QZ) = k. \tag{43}
\end{aligned}$$

Together,

$$\begin{aligned}\frac{n}{\sigma^2} E (\hat{\mu}_{n+1} - \mu_{n+1})^2 &= \frac{n}{\sigma^2} E \left(\delta'_{n+1} (\hat{\theta}_0 - \theta_0) + x_{n+1}^{*'} (\hat{\theta}_1 - \theta_1) \right)^2 \\ &\quad + \frac{n}{\sigma^2} E \left((\hat{\alpha} - \alpha) z_{n+1}^* z_{n+1}^{*'} (\hat{\alpha} - \alpha) \right) + o(1) \\ &\rightarrow ET_{1c}^2 + k\end{aligned}$$

which is (21)

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 (29),

$$\frac{n^{1/2}}{\sigma} \left(-can^{-1} S_n^* + (\tilde{\theta}_0^* - \theta_0)' \delta_{n+1} \right) \xrightarrow{d} -cW_c^*(1) + \delta(1)' \left(\int_0^1 \delta \delta' \right)^{-1} \int_0^1 \delta dW = T_{0c} \quad (44)$$

and therefore

$$\frac{n}{\sigma^2} E \left(-can^{-1} S_n^* + (\tilde{\theta}_0^* - \theta_0)' \delta_{n+1} \right)^2 \rightarrow ET_{0c}^2.$$

As in (43),

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

Then

$$\begin{aligned}\frac{n}{\sigma^2} E (\tilde{\mu}_{n+1} - \mu_{n+1})^2 &= \frac{n}{\sigma^2} E \left(-can^{-1} S_n^* + (\tilde{\theta}_0^* - \theta_0)' \delta_{n+1} \right)^2 \\ &\quad + \frac{n}{\sigma^2} E \left((\hat{\alpha} - \alpha)' z_{n+1}^* \right)^2 + o(1) \\ &\rightarrow ET_{0c}^2 + k.\end{aligned}$$

which is (19). When $p = 0$ then $T_{0c} = -cW_c(1)$ so $f_0(c, 0) = c^2 EW_c(1)^2 = c(\exp(2c) - 1)/2$ by (26). When $p = 1$ then

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

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

Third, take the averaging estimator. Since

$$\begin{aligned}\hat{\mu}_{n+1}(w) - \mu_{n+1} &= w \left(\delta'_{n+1} (\hat{\theta}_0 - \theta_0) + x_{n+1}^{*'} (\hat{\theta}_1 - \theta_1) \right) \\ &\quad + (1-w) \left(-can^{-1} S_n^* + (\tilde{\theta}_0^* - \theta_0)' \delta_{n+1} \right) + (\hat{\alpha}(w) - \alpha)' z_{n+1}^*,\end{aligned}$$

then

$$\begin{aligned}
\frac{n}{\sigma^2} E(\hat{\mu}_t(w) - \mu_t)^2 &= w^2 \frac{n}{\sigma^2} E \left(\delta'_{n+1} (\hat{\theta}_0 - \theta_0) + x'_{n+1} (\hat{\theta}_1 - \theta_1) \right)^2 \\
&\quad + (1-w)^2 \frac{n}{\sigma^2} E \left(-can^{-1} S_n^* + (\tilde{\theta}_0^* - \theta_0)' \delta_{n+1} \right)^2 \\
&\quad + 2w(1-w) \frac{n}{\sigma^2} E \left(\delta'_{n+1} (\hat{\theta}_0 - \theta_0) + x'_{n+1} (\hat{\theta}_1 - \theta_1) \right) \left(-can^{-1} S_n^* + (\tilde{\theta}_0^* - \theta_0)' \delta_{n+1} \right) \\
&\quad + \frac{n}{\sigma^2} E \left((\hat{\alpha}(w) - \alpha) z_{n+1}^* z'_{n+1} (\hat{\alpha}(w) - \alpha) \right) \\
&\quad + o_p(1) \\
&\xrightarrow{d} w^2 ET_{1c}^2 + (1-w)^2 ET_{0c}^2 + 2w(1-w) E(T_{0c}T_{1c}) + k
\end{aligned}$$

which is (23).

The optimal weight and mean-squared error are found by minimizing $f_w(c, p, k)$ with respect to w . ■

Proof of Theorem 3: From (32) and (37) we have

$$\frac{1}{\sigma^2} \sum_{t=1}^n (\tilde{\mu}_t - \mu_t)^2 \xrightarrow{d} F_{0c} + \chi_{p+k}^2.$$

and

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

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

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

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

Similarly, using (44) and (42),

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

which is $f_{df}(c, p, k)$. ■

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

$$\begin{aligned}
\frac{M_0(c) - n\sigma^2}{\sigma^2} &= \frac{1}{\sigma^2} \sum_{t=1}^n (e_t^2 - \sigma^2) + \frac{1}{\sigma^2} \sum_{t=1}^n (\tilde{\mu}_t - \mu_t)^2 \\
&\quad + \frac{2\hat{\sigma}^2}{\sigma^2} (m_{01}(c, p) + p + k) - \frac{2}{\sigma^2} \sum_{t=1}^n e_t (\tilde{\mu}_t - \mu_t). \tag{45}
\end{aligned}$$

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

$$\begin{aligned}
\frac{1}{\sigma^2} \sum_{t=1}^n e_t (\tilde{\mu}_t - \mu_t) &= -\frac{ca}{\sigma^2 n} \sum_{t=1}^n e_t S_{t-1}^* + \frac{1}{\sigma^2} \sum_{t=1}^n e_t \delta'_t \left(\tilde{\theta}_0^* - \theta_0 \right) + \frac{1}{\sigma^2} \sum_{t=1}^n e_t z_t^{*'} (\tilde{\alpha} - \alpha) \\
&\xrightarrow{d} -c \int_0^1 dW W_c^* + \int_0^1 dW \delta' \left(\int_0^1 \delta \delta' \right)^{-1} \int_0^1 \delta dW + Z' Q Z \tag{46}
\end{aligned}$$

(using (29)), which has expectation $m_{01}(c) + p + k$. Adding these components, it follows that (45) has expectation tending to $m_0(c, p, k)$ as claimed.

Next consider $M_1(c)$. We have

$$\frac{M_1(c) - n\sigma^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{t=1}^n (e_t^2 - \sigma^2) + \frac{1}{\sigma^2} \sum_{t=1}^n (\hat{\mu}_t - \mu_t)^2 + \frac{2\hat{\sigma}^2}{\sigma^2} (m_1(c, p) + p + k) - \frac{2}{\sigma^2} \sum_{t=1}^n e_t (\hat{\mu}_t - \mu_t). \tag{47}$$

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

$$\begin{aligned}
\frac{1}{\sigma^2} \sum_{t=1}^n e_t (\hat{\mu}_t - \mu_t) &= \frac{1}{\sigma^2} \sum_{t=1}^n e_t x_t^{*'} (\hat{\theta}_1 - \theta_1) + \frac{1}{\sigma^2} \sum_{t=1}^n e_t z_t^{*'} (\hat{\alpha} - \alpha) \\
&\xrightarrow{d} \int_0^1 dW X_c^*(r)' \left(\int_0^1 X_c^* X_c^{*'} \right)^{-1} \int_0^1 X_c^* dW \\
&\quad + \int_0^1 dW \delta' \left(\int_0^1 \delta \delta' \right)^{-1} \int_0^1 \delta dW + Z' Q Z
\end{aligned} \tag{48}$$

which has expectation $m_1(c, p) + p + k$. Adding these two components, we see that (47) 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 $F_n \xrightarrow{d} F_c$. ■

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

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

The first three terms have expectation converging to

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

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

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

by (46) and (48). 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

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

The first-order condition for minimization is $0 = -2(1-\hat{w})F_n + 2(2+p)$, whose solution is $\hat{w} = 1 - (2+p)/F_n$. 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 equation (39) 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

- [1] Akaike, H. (1970): "Statistical predictor identification," *Annals of the Institute of Statistical Mathematics*, 22, 203-419.
- [2] Akaike, H. (1973): "Maximum likelihood identification of Gaussian autoregressive moving average models," *Biometrika*, 60, 255-265.
- [3] Bates, J.M. and C.M.W. Granger (1969): "The combination of forecasts," *Operations Research Quarterly*, 20, 451-468.
- [4] Bhansali, R. J. (1996): "Asymptotically efficient autoregressive model selection for multistep prediction," *Annals of the Institute of Statistical Mathematics*, 48, 577-602.
- [5] Canjels, Eugene and Mark W. Watson (1997): "Estimating deterministic trends in the presence of serially correlated errors," *Review of Economics and Statistics*, 79, 184-200.
- [6] Chan, N.H. and Ching-Zong Wei (1987): "Asymptotic inference for nearly nonstationary AR(1) processes," *Annals of Statistics*, 15, 1050-1063.
- [7] Chao, John C. and Peter C.B. Phillips (1999): "Model selection in partially nonstationary vector autoregressive processes with reduced rank structure," *Journal of Econometrics*. 91, 227-271.
- [8] Clemen, R.T. (1989): "Combining forecasts: A review and annotated bibliography," *International Journal of Forecasting*, 5, 559-581.
- [9] Clements, Michael P. and David F. Hendry (2001): "Forecasting with difference-stationary and trend-stationary models," *Econometrics Journal*, 4, S1-S19.
- [10] Dickey, David A. and Wayne A. Fuller (1979): "Distribution of the estimators for autoregressive time series with a unit root," *Journal of the American Statistical Association*, 74, 427-431.
- [11] Dickey, David A. and Wayne A. Fuller (1981): "Likelihood ratio statistics for autoregressive time series with a unit root," *Econometrica*, 49, 1057-1072.
- [12] Diebold, Francis X. and Lutz Kilian (2000): "Unit-root tests are useful for selecting forecasting models," *Journal of Business and Economic Statistics*, 18, 265-273.
- [13] Diebold, Francis X. and J. A. Lopez (1996): "Forecast evaluation and combination," in Maddala and Rao, eds., *Handbook of Statistics*, Elsevier.
- [14] Elliott, Graham (2006): "Unit root pre-testing and forecasting," working paper, UCSD.
- [15] Elliott, Graham, Thomas J. Rothenberg, and James H. Stock (1996): "Efficient tests of an autoregressive unit root," *Econometrica*, 64, 813-836.

- [16] Franses, Philip Hans, and Frank Kleibergen (1996): "Unit roots in the Nelson-Plosser data: Do they matter for forecasting?" *International Journal of Forecasting*, 12, 283-288.
- [17] Engle, Robert F. and Clive W.J. Granger (1987): "Co-integration and error correction: Representation, estimation and testing," *Econometrica*, 55, 251-276.
- [18] Granger, Clive W.J. (1989): "Combining forecasts – Twenty years later," *Journal of Forecasting*, 8, 167-173.
- [19] Granger, Clive W.J. and R. Ramanathan (1984): "Improved methods of combining forecasts," *Journal of Forecasting*, 3, 197-204.
- [20] Hansen, Bruce E. (1995): "Rethinking the univariate approach to unit root tests: How to use covariates to increase power," *Econometric Theory*, 11, 1148-1171.
- [21] Hansen, Bruce E. (2007): "Least Squares Model Averaging," *Econometrica*, forthcoming.
- [22] Hansen, Bruce E. (2006): "Least Squares Forecast Averaging," working paper, University of Wisconsin.
- [23] Hendry, D.F. and M. P. Clements (2002): "Pooling of forecasts," *Econometrics Journal*, 5, 1-26.
- [24] Hjort, Nils Lid and Gerda Claeskens (2003): "Frequentist Model Average Estimators," *Journal of the American Statistical Association*, 98, 879-899.
- [25] Ing, Ching-Kang (2003): "Multistep prediction in autoregressive processes," *Econometric Theory*, 19, 254-279.
- [26] Ing, Ching-Kang (2004): "Selecting optimal multistep predictors for autoregressive processes of unknown order," *Annals of Statistics*, 32, 693-722.
- [27] Ing, Ching-Kang, and Ching-Zong Wei (2003): "On same-realization prediction in an infinite-order autoregressive process," *Journal of Multivariate Analysis*, 85, 130-155.
- [28] Ing, Ching-Kang, and Ching-Zong Wei (2005): "Order selection for same-realization predictions in autoregressive processes," *Annals of Statistics*, 33, 2423-2474.
- [29] Inoue, Atsuhi, and Lutz Kilian (2006): "On the selection of forecasting models," *Journal of Econometrics*, 130, 272-306.
- [30] Johansen, Soren (1988): "Statistical analysis of cointegrating vectors," *Journal of Economic Dynamics and Control*, 12, 231-254.
- [31] Johansen, Soren (1991): "Estimation and hypothesis testing of cointegration vectors in the presence of linear trend," *Econometrica*, 59, 1551-1580.

- [32] Johansen, Soren (1995): *Likelihood-Based Inference in Cointegrated Vector Auto-Regressive Models*, Oxford University Press.
- [33] 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.
- [34] Kim, Tae-Hwan, Stephen J. Leybourne, and Paul Newbold (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.
- [35] Lee, Sangyeol and Alex Karagrigoriou (2001): "An asymptotically optimal selection of the order of a linear process," *Sankhya*, 63, Series A, 93-106.
- [36] Mallows, C.L. (1973): "Some comments on C_p ," *Technometrics*, 15, 661-675.
- [37] Phillips, Peter C.B. (1988): "Towards a unified asymptotic theory for autoregression," *Biometrika*, 74, 535-547.
- [38] Rissanen, J. (1986): "Stochastic complexity and modeling," *Annals of Statistics*, 14, 1080-1100.
- [39] Phillips, Peter C.B. (1988): "Regression theory for near-integrated time series," *Econometrica*, 56, 1021-1043.
- [40] Schwarz, G. (1978): "Estimating the dimension of a model," *Annals of Statistics*, 6, 461-464.
- [41] 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.
- [42] Stock, James H. (1996): "VAR, error correction and pretest forecasts at long horizons," *Oxford Bulletin of Economics and Statistics*, 58, 685-701.
- [43] Stock, J.H. and M. W. Watson (1999): "A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series," in Engle and White, eds., *Cointegration, Causality and Forecasting: A Festschrift for Clive W.J. Granger*, Oxford University Press.
- [44] Stock, J.H. and M. W. Watson (2004): "Combination forecasts of output growth in a seven-country data set," *Journal of Forecasting*, forthcoming.
- [45] Stock, J.H. and M. W. Watson (2005): "An empirical comparison of methods for forecasting using many predictors," working paper, NBER.
- [46] Stock, J.H. and M. W. Watson (2006): "Forecasting with many predictors," in Elliott, Granger and Timmermann, eds., *Handbook of Economic Forecasting*, Chapter 10, Elsevier.

- [47] Timmermann, Allan (2006): “Forecast Combinations,” in Elliott, Granger and Timmermann, eds., *Handbook of Economic Forecasting*, Chapter 4, Elsevier.
- [48] Wei, Ching-Zong Wei (1992): “On predictive least squares principles,” *Annals of Statistics*, 20, 1-42.

Figure 1: Asymptotic Mean-Squared Error

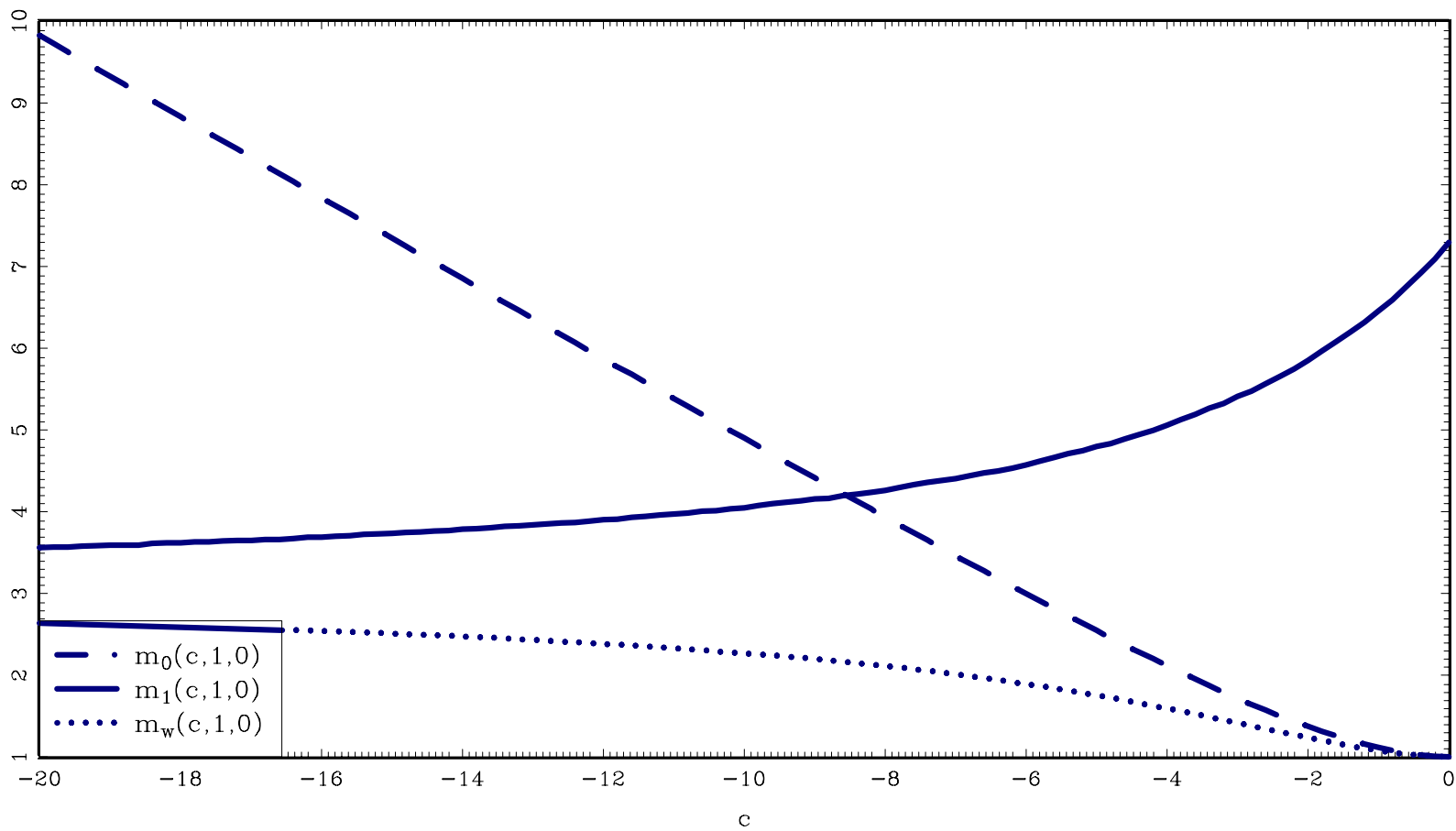


Figure 2: Asymptotic Forecast Loss

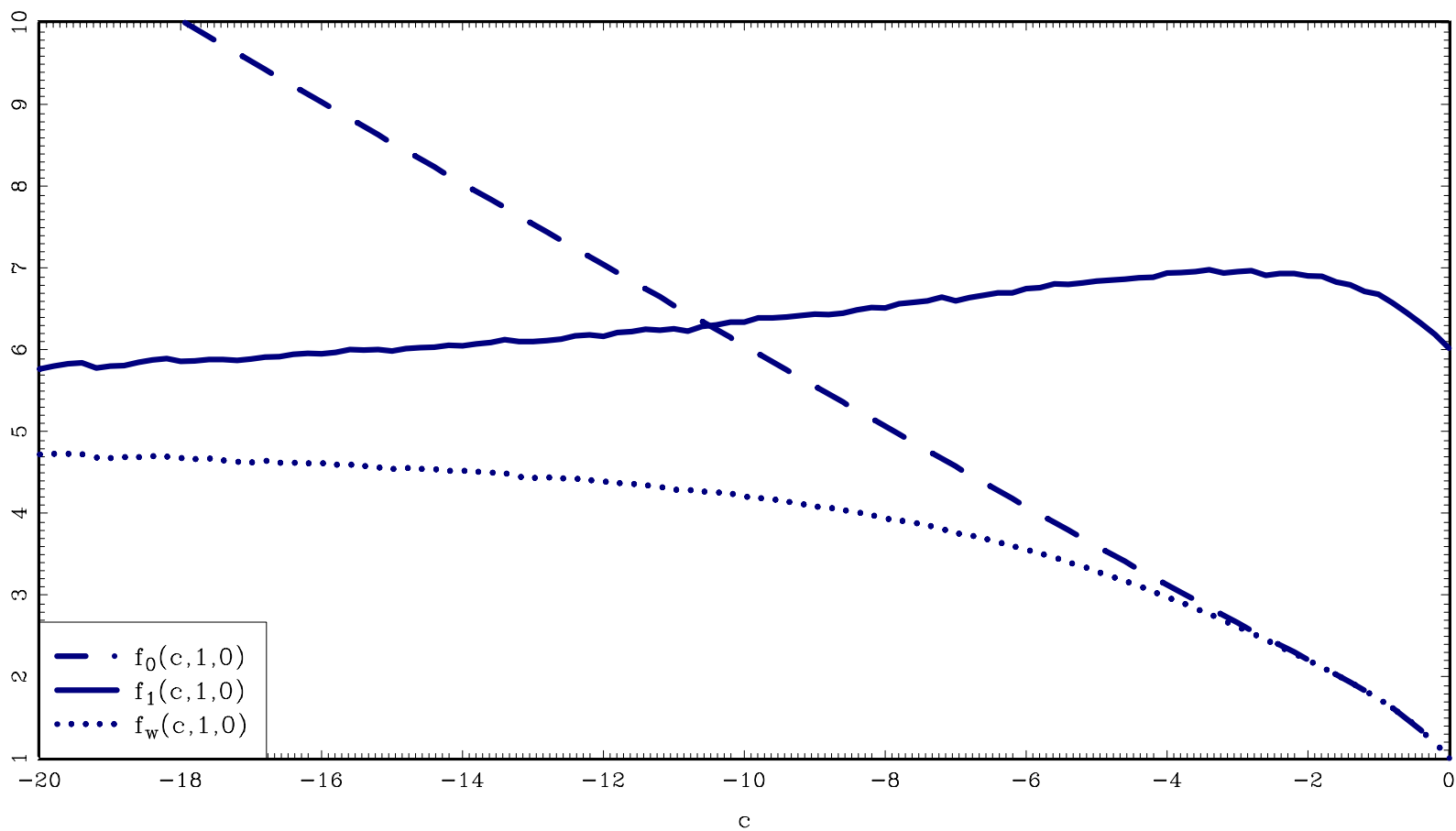
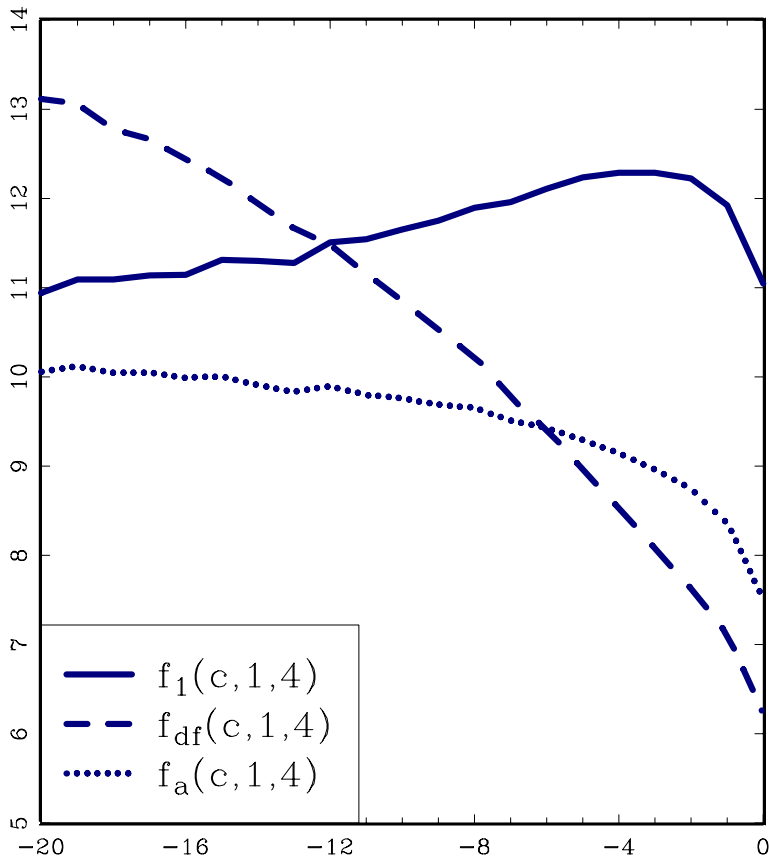
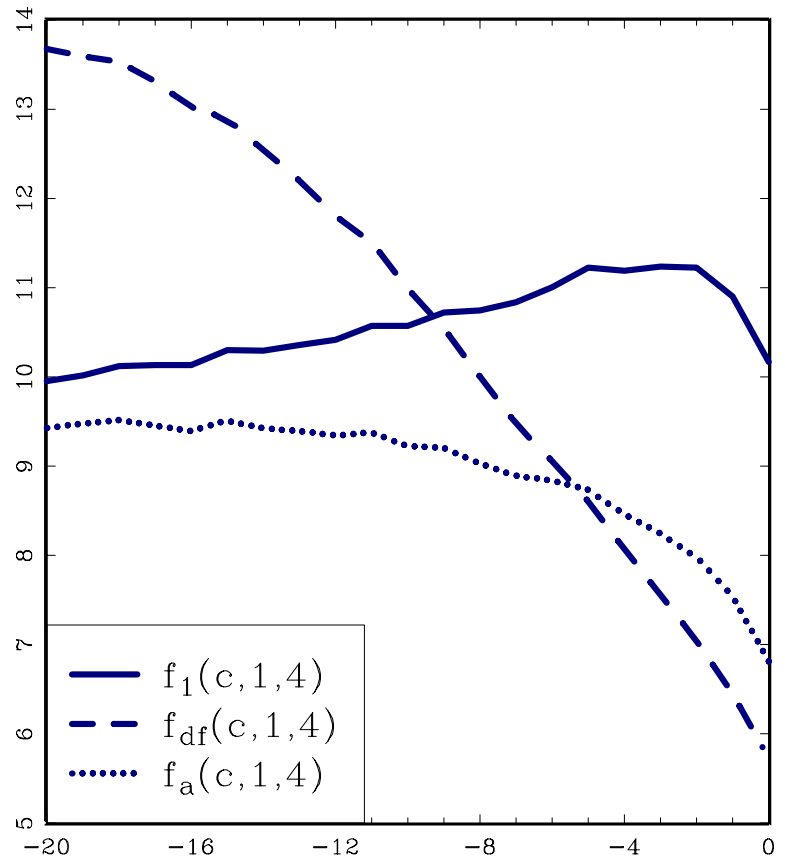


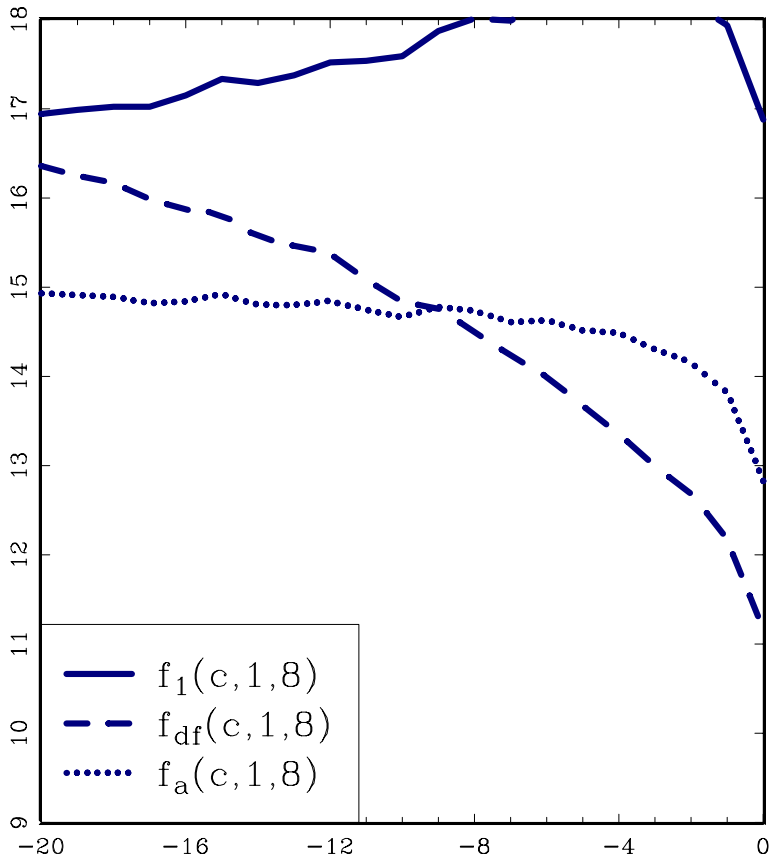
Figure 6: Finite Sample Forecast Loss, $\Theta=0.6$



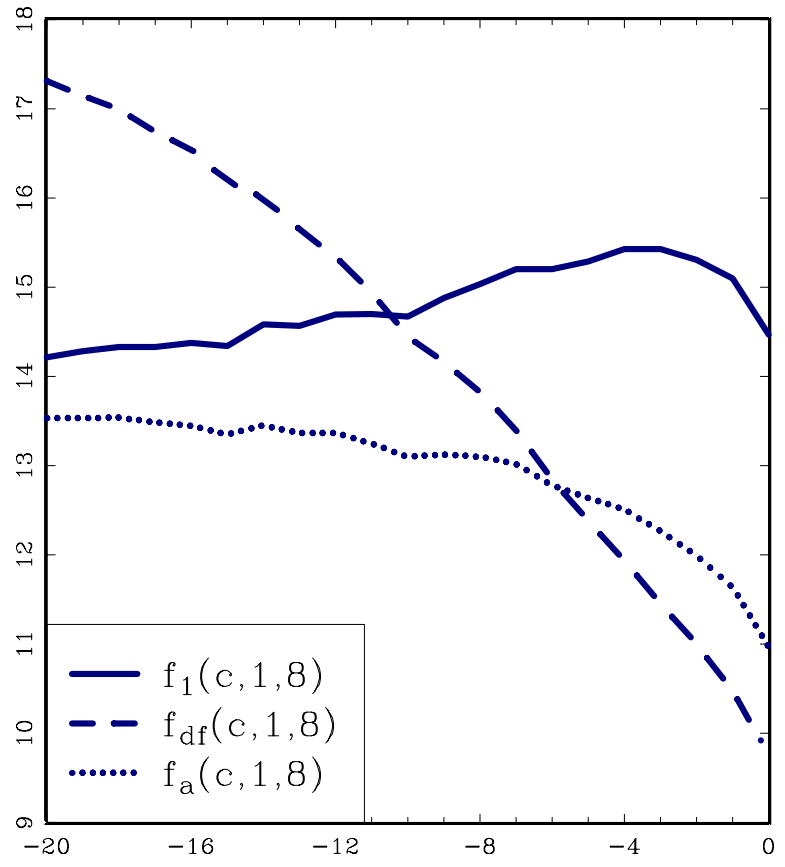
$n=50, k=4$



$n=200, k=4$



$n=50, k=8$



$n=200, k=8$

Figure 3: Asymptotic MSE of Feasible Estimators

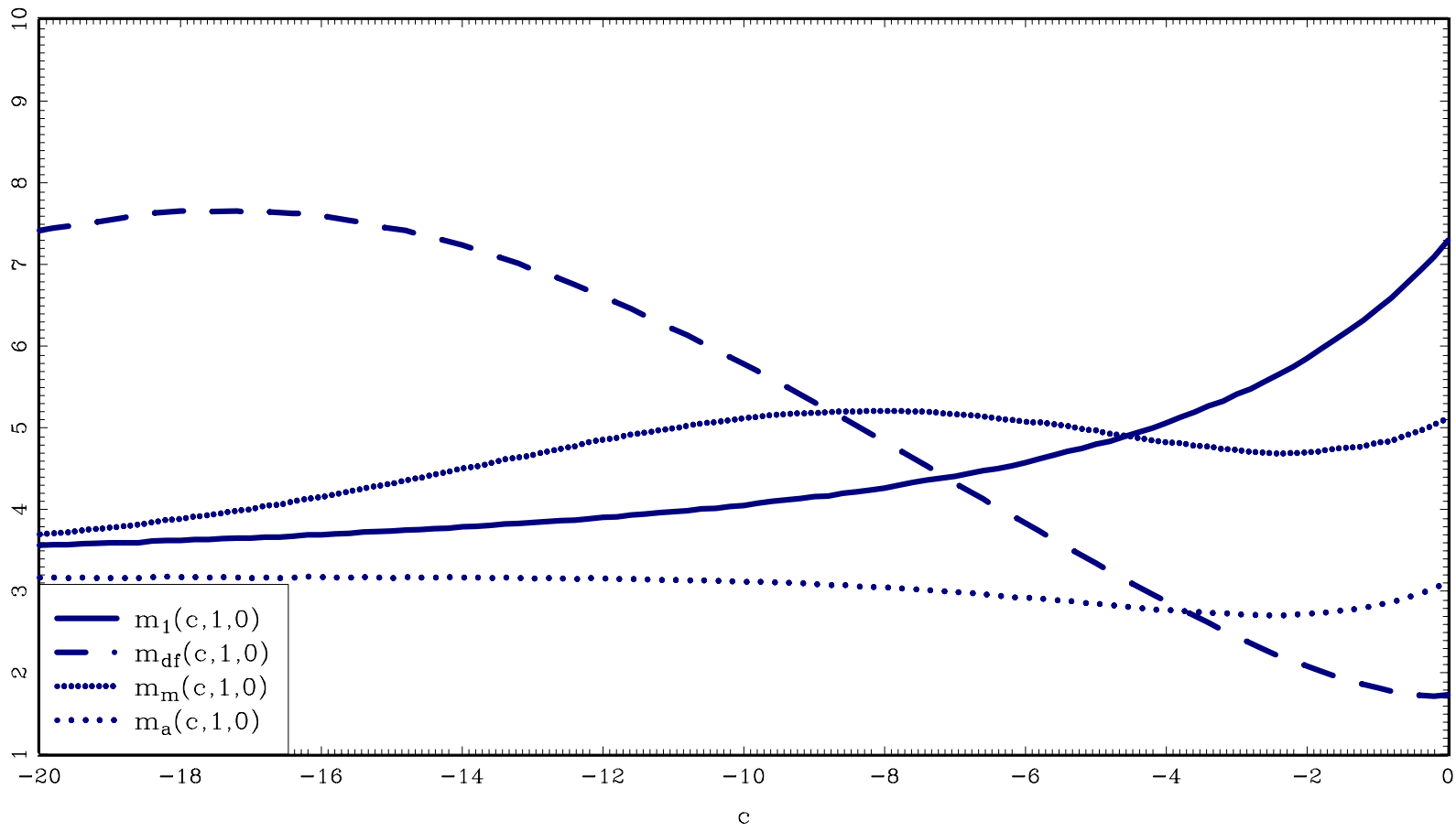


Figure 4: Asymptotic Forecast Loss of Feasible Estimators

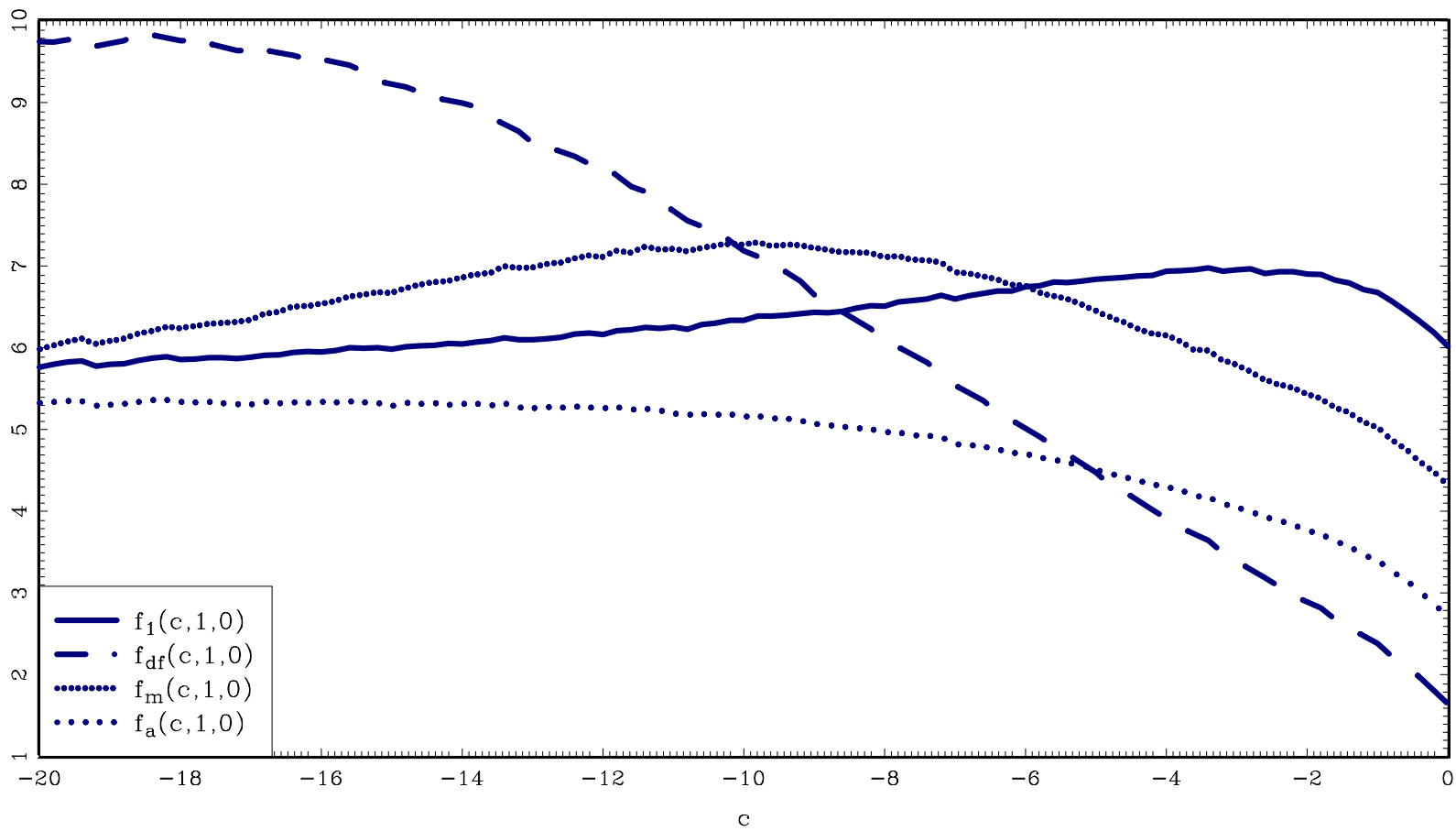


Figure 5: Finite Sample Forecast Loss

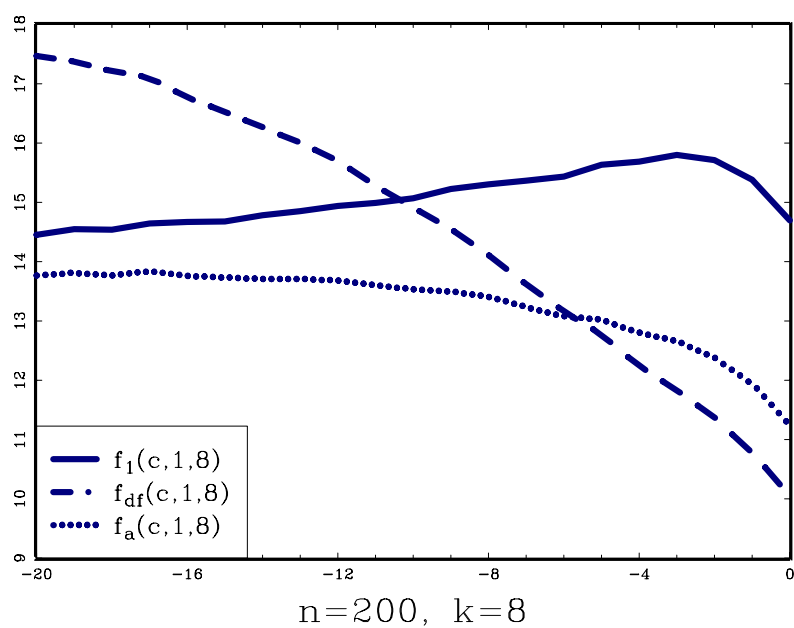
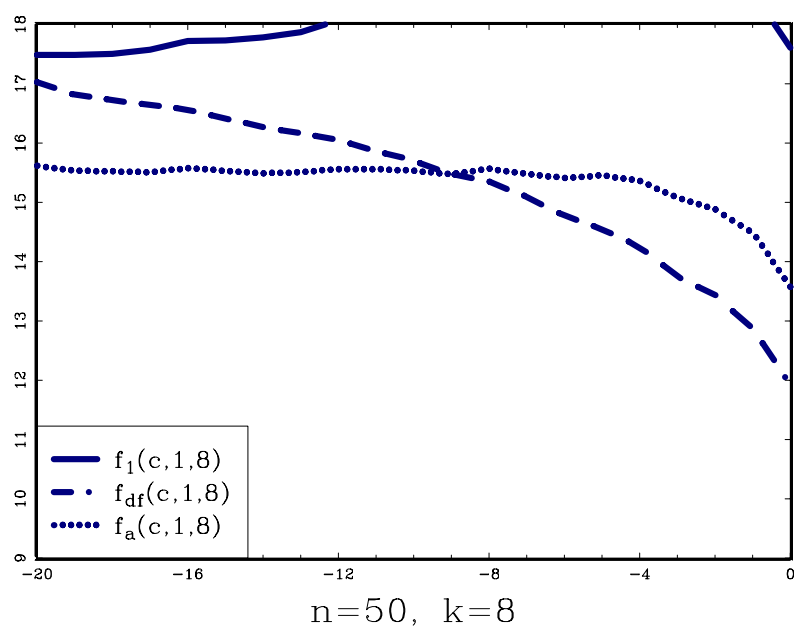
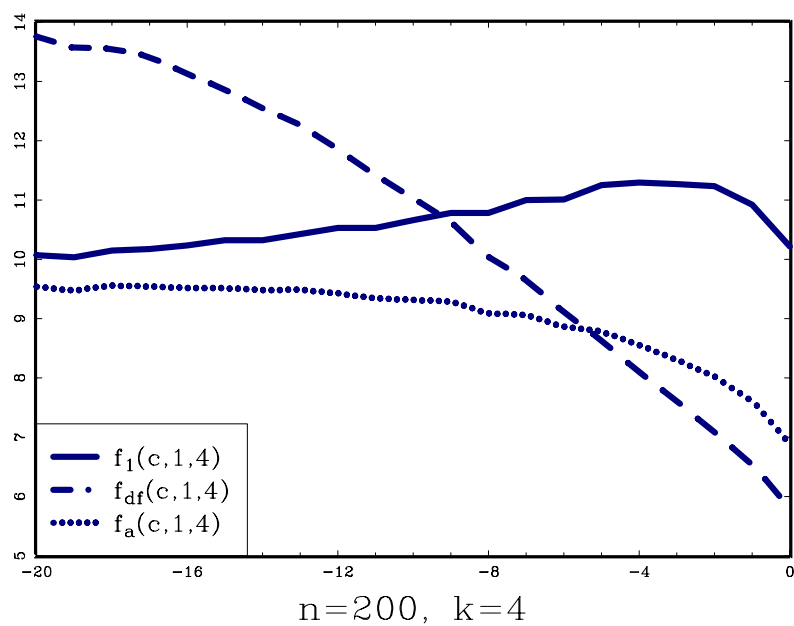
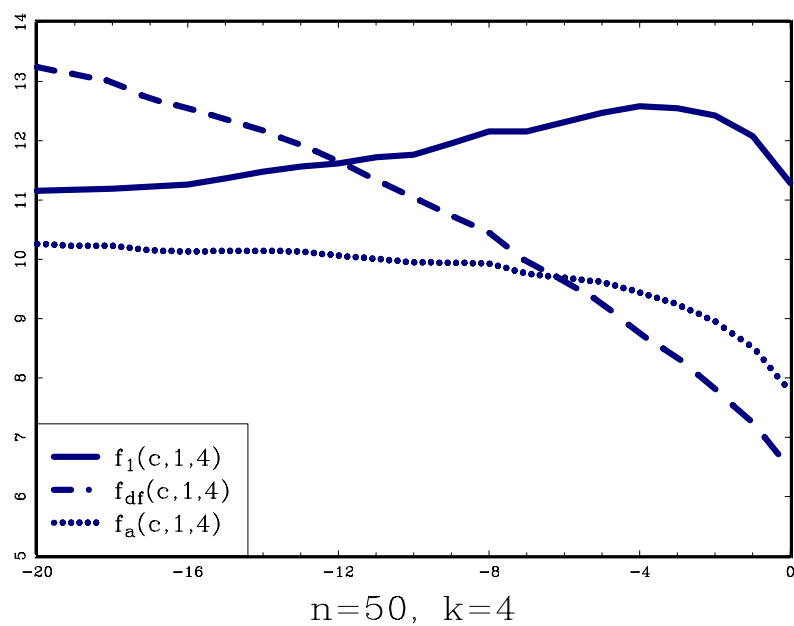
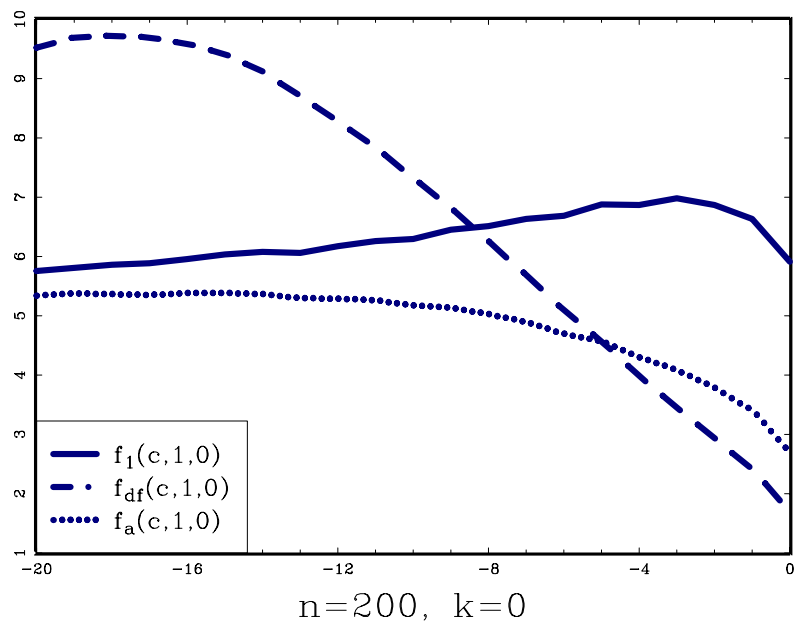
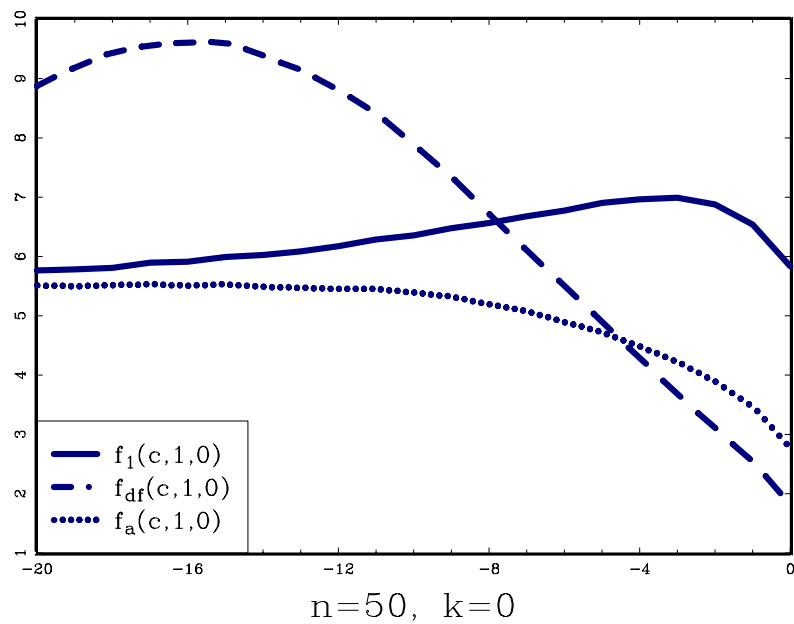


Figure 7: Selection and Averaging over k , $\Theta=0.6$

