Predictive Ability With Cointegrated Variables[*]

Valentina Corradi[1], Norman R. Swanson[2] and Claudia Olivetti[3]

[1] University of Exeter, Exeter EX4 4PU, UK
[2] Texas A&M University, College Station, TX 77843
[3] University of Pennsylvania, Philadelphia, PA 19104

January 2001

Abstract

In this paper we outline conditions under which the Diebold and Mariano (DM: 1995) test for predictive ability can be extended to the case of two forecasting models, each of which may include cointegrating relations, when allowing for parameter estimation error. We show that in the cases where either the loss function is quadratic or the length of the prediction period, P, grows at a slower rate than the length of the regression period, R, the standard DM test can be used. On the other hand, in the case of a generic loss function, if $\frac{P}{R} \to \pi$ as $T \to \infty$, $0 < \pi < \infty$, then the asymptotic normality result of West (1996) no longer holds. We also extend the "data snooping" technique of White (2000) for comparing the predictive ability of multiple forecasting models to the case of cointegrated variables. In a series of Monte Carlo experiments, we examine the impact of both short run and cointegrating vector parameter estimation error on DM, data snooping, and related tests. Our results suggest that size is reasonable for R and P greater than 50, and power improves with P, as expected. Furthermore, the additional cost, in terms of size distortion, due to the estimation of the cointegrating relations is not substantial. We illustrate the use of the tests in a nonnested cointegration framework by forming prediction models for industrial production which include two interest rate variables, prices, and either M1, M2, or M3.

JEL classification: C22, C51.
Keywords: almost sure convergence, forecasting, cointegration, parameter estimation error.

----

1 Introduction

The evaluation of economic forecasts is an area of research which has been of interest to applied economists for many decades. However, it is only in recent years that attention has been given in economics to the study of formal tests of out of sample fit (see e.g. Meese and Rogoff (1983) and the references contained therein). From an applied perspective, much of the recent literature focuses on nonlinear modelling (e.g. Granger (1995), Granger and Swanson (1996), and Kuan and Liu (1996)), loss functions (e.g. Granger (1969), Christoffersen and Diebold (1996,1997) and Weiss (1996)), model evaluation and selection (e.g. Swanson and White (1995) and Sullivan, Timmerman, and White (1999)), for example. In a recent series of papers, Clements and Hendry (1996,1999a,b,2001) address the issue of forecast comparison in the presence of integrated and cointegrated variables, and highlight the fact that deteministic shifts (e.g. shifts in intercept and trend components) often account for systematic forecast failure. From a theoretical perspective, two areas in which recent progress has been made are the construction of forecast model comparison tests (which may account for parameter estimation error) (e.g. Diebold and Mariano (1995), West (1996), McCracken (1998,2000), Clark and McCracken (2000), Clark (1999), Chao, Corradi and Swanson (2000) and Rossi (2000)), and test size adjustment when multiple models are jointly evaluated (e.g. White (2000)).

From a Bayesian perspective, the posterior odds approach has been extensively used for out of sample evaluation of competing economic models. For example, Palm and Zellner (1992), Min and Zellner (1993) and Zellner and Min (1999) employ the posterior odds approach to analyze the issue of optimal forecast combinations. Phillips (1995), proposes a new forecast encompassing test called the forecast posterior information criterion (PICF) to evaluate the relative out of sample performance of two alternative models. As pointed out by Palm (1995), the PICF is a special case of posterior odds, characterized by flat priors on the coefficients and by a priori independence among the regressors. More recently, posterior odds have been used to evaluate the out of sample performance of alternative dynamic stochastic (general equilibrium) models (see e.g. De Jong, Ingram and Whiteman (2000), Schorfheide (2000)). Finally Bayesian methods for assessing the finite sample behavior of out of sample forecasts have been proposed by Sims and Zha (1998, 1999) for the unconditional case and by Waggoner and Zha (1999) for the conditional case. For a general overview of Bayesian forecasting, the reader is referred to Geweke (2000).

In this paper we begin by addressing two theoretical issues concerning forecasting in cointegrated systems. First, we outline the conditions under which the Diebold and Mariano (DM: 1995) test for comparative predictive ability can be extended to the case of two forecasting models, each of which may include cointegrating relations, when allowing for parameter estimation error. This is done by examining the effect of parameter estimation error in a cointegrating framework, using the methods proposed by West (1996). We show that the standard DM test can be used in the cases where either the loss function, f, is quadratic, or the length of the prediction period, P, grows at a slower rate than the length of the regression period, R. On the other hand, in the case of a generic loss function, if $\frac{P}{R} \to \pi$ as $T \to \infty$, $0 < \pi < \infty$, the asymptotic normality result of West (1996) no longer holds. The effect of parameter estimation error in the presence of (nearly) integrated, but not cointegrated variables, has been recently addressed by Rossi (2000). Second, we show that when $\frac{P}{R^{2/3} \log \log R} \to 0$, parameter estimation error approaches zero not only in probability, but also almost surely. This allows us to propose an extension (to the case of cointegrated models) of the "data snooping" technique of White (2000) for comparing the predictive ability of multiple forecasting models with a given "benchmark" model. In particular, we show that the Politis and Romano (1994) stationary bootstrap can be applied to cointegrated variables, thus validating not only White's technique, but also suggesting an avenue for applying a wide variety of bootstrapping techniques to problems involving estimated cointegrating vectors.

We then conduct a series of Monte Carlo experiments in order to examine the impact of both short run and cointegrating vector parameter estimation error on DM, data snooping, and related tests. Our results suggest that size is reasonable for R and P greater than 50, and power improves with P, as expected. Furthermore, the additional cost, in terms of size distorsion, due to the estimation of the cointegrating relations is not substantial, although the tests do generally perform more poorly when parameters are not known. Additional findings in a number of related areas are discussed in West and McCracken (1998), McCracken (1998,2000) and White (2000). In an empirical illustration, we construct nonnested forecasting models of industrial production using either M1, M2, or M3, as well as the t-bill rate, the commercial-paper rate, and prices. These models are then compared using DM and data snooping tests. We present evidence supportive of the hypothesis that the broader M2 and M3 money measures dominate M1 when forecasts are compared using a mean square error criterion for forecast horizons of 1- and 3-months. On the other hand, there is little to choose between the alternative money measures when 1-year ahead

forecasts are constructed.

By examining predictive ability within the context of cointegrated variables, we believe that we contribute in a number of ways to the discussion of the use of cointegration models for forecasting. One dimension of this contribution is that we outline the cases for which the DM test can be straightforwardly applied in the context of cointegrated variables. Contributions are also attempted in several other related areas. For example, we establish the validity of using the stationary bootstrap in analyses of cointegrated variables, and propose a simple test for the null hypothesis of equal predictive ability when f is non quadratic. The rest of the paper is organized as follow. In Section 2, we examine DM type tests for comparing two or multiple forecasting models. Section 3 contains our Monte Carlo findings. In Section 4, an empirical illustration is provided. Some concluding remarks are given in Section 5. All proofs are collected in the Appendix. Hereafter, $\Rightarrow$ denotes weak convergence of probability measures and W denotes a standard Brownian motion process.

2 Predictive Ability and Parameter Variability

Let $\widehat{v}_{0,t+h}$ and $\widehat{v}_{k,t+h}$, $k = 1, \dots, l$, denote h step ahead estimated forecasting errors from models 0 (the benchmark model) and k, where $v_{0,t+h}$ and $v_{k,t+h}$, are the errors in the absence of parameter uncertainty, respectively. We begin by considering the relative predictive accuracy of two forecasting models, and later compare the predictive accuracy of several forecasting models by assessing whether any model, k, $k = 1, \dots, l$ has better predictive accuracy than the benchmark model.

2.1 Comparing Two Forecasting Models

Consider the DM test statistic:

$$\widehat{d}_P = P^{-1/2} \frac{\sum_{t=R-h+1}^{T-1} (f(\widehat{v}_{0,t+h}) - f(\widehat{v}_{k,t+h}))}{\widehat{\sigma}_P}, \tag{1}$$

where R denotes the length of estimation period, P is the length of the prediction period, f is some generic loss function, $h \geq 1$ is the forecast horizon, $\widehat{v}_{0,t+h}$ and $\widehat{v}_{k,t+h}$ are constructed using consistent estimators, and $\widehat{\sigma}_P^2$ is defined as

$$\frac{1}{P} \sum_{t=R-h+1}^{T-1} (f(\widehat{v}_{0,t+h}) - f(\widehat{v}_{k,t+h}))^2 + \frac{2}{P} \sum_{j=1}^{l_P} w_j \sum_{t=R-h+1+j}^{T-1} (f(\widehat{v}_{0,t+h}) - f(\widehat{v}_{k,t+h}))(f(\widehat{v}_{0,t+h-j}) - f(\widehat{v}_{k,t+h-j})) \tag{2}$$

3

where $w_j = 1 - \frac{j}{l_P+1}$, $l_P = o(P^{1/4})$. As we shall assume strict stationarity (see below), the null of equal predictive ability is[1]:

$$H_0 : E(f(v_{0,t+h}) - f(v_{k,t+h})) = 0,$$

while the alternative is:

$$H_A : E(f(v_{0,t+h}) - f(v_{k,t+h})) \neq 0.$$

In the sequel, we shall use the convenient triangular error correction representation of Phillips (1991). This allows us to disregard parameters describing short-run dynamics and focus on cointegrating vector estimators. Consider the following pseudo DGPs[2]:

$$Y_t = \alpha_{0,0} + \beta_{0,0} X_{0,t} + u_{0,t},$$

$$X_{0,t} = X_{0,0} + \sum_{j=1}^{t} w_{0,j},$$

$$\ldots$$

$$Y_t = \alpha_{l,0} + \beta_{l,0} X_{l,t} + u_{l,t},$$

$$X_{l,t} = \frac{\beta_{0,0}}{\beta_{l,0}} X_{0,0} + \frac{\beta_{0,0}}{\beta_{l,0}} \sum_{j=1}^{t} w_{0,j} - \frac{1}{\beta_{l,0}}(u_{l,t} - u_{0,t}) - \frac{1}{\beta_{l,0}}(\alpha_{l,0} - \alpha_{0,0}), \tag{3}$$

so that $X_{k,t} = \frac{\beta_{0,0}}{\beta_{k,0}} X_{0,0} + \frac{\beta_{0,0}}{\beta_{k,0}} \sum_{j=1}^{t} w_{0,j} - \frac{1}{\beta_{k,0}}(u_{k,t} - u_{0,t}) - \frac{1}{\beta_{k,0}}(\alpha_{k,0} - \alpha_{0,0})$, for $k = 0, ..., l$. Here, we assume that $X_{k,t}, k = 0, ..., l$ are scalars (see below for a discussion of vector valued $X_{k,t}$). The error correction representations implied by the DGPs above are:

$$\Delta Y_{t+1} = \alpha_{k,0} - (Y_t - \beta_{k,0} X_{k,t}) + v_{k,t+1}, \ k = 0, ..., l, \tag{4}$$

where $v_{k,t} = \beta_{0,0} w_{0,t} + \Delta u_{0,t} + u_{k,t-1}, k = 0, ..., l$. Consider constructing $\widehat{d}_P$ for $h = 1$. Define $\overline{X}_{k,t} = \frac{1}{t} \sum_{j=1}^{t} X_{k,j}$, $\overline{Y}_t = \frac{1}{t} \sum_{j=1}^{t} Y_j$, $t = R, R+1, ...T$, $\widehat{\beta}_{k,t} = \sum_{j=1}^{t}(X_{k,j} - \overline{X}_{k,t})(Y_j - \overline{Y}_t)(\sum_{j=1}^{t}(X_{k,j} - \overline{X}_{k,t})^2)^{-1}$, and $\widehat{\alpha}_{k,t} = \overline{Y}_t - \widehat{\beta}_{k,t} \overline{X}_{k,t}$. We begin by estimating two forecasting models using $t =$

---

[1]We could have alternatively tested the null that model 0 is not outperformed by model k. In this case the null should be stated with the $\leq$ sign and the alternative with the $>$ sign. In this case, the statistic diverges to minus infinity whenever the null holds with the $<$ sign.

[2]We use the term pseudo DGPs as we do not require (dynamic) correct specification of the conditional mean. The estimation of short-run dynamics is discussed below. Henceforth, we are implicitly assuming that $h = 1$. The case where $h > 1$ is also discussed below.

R observations, and constructing a 1-step ahead forecast error for each of the models. Then, observations up to $t = R+1$ are used to re-estimate the models and another 1-step ahead forecast error is constructed for each model. This procedure is continued until the entire original sample of data is exhausted. Thus, for two models (say models 0 and k) we obtain two sequences of forecast errors of length P, namely $\widehat{v}_{0,t+1} = \Delta Y_{t+1} + (Y_t - \widehat{\beta}_{0,t} X_{0,t}) - \widehat{\alpha}_{0,t}$ and $\widehat{v}_{k,t+1} = \Delta Y_{t+1} + (Y_t - \widehat{\beta}_{k,t} X_{k,t}) - \widehat{\alpha}_{k,t}$, for $t = R, \ldots, T-1$. In the sequel we shall assume the following:

Assumption 1 (A1): $(u_{k,t}, w_{0,t})$, $k = 0, \ldots, l$, are zero mean, strictly stationary strong mixing processes, with mixing coefficient α_m of size $-\frac{8(8+\delta)}{\delta}$, with $\delta > 0$. Additionally, $E(|u_{k,t}|^{8+\delta}) \leq C_1 < \infty$, $E(|w_{0,t}|^{8+\delta}) \leq C_2 < \infty$, $E(u_{k,t}^2) = \sigma_{k,u}^2 > 0$, and $E(w_{0,t}^2) = \sigma_w^2 > 0$ for $0 < C_i < \infty$, $i = 1, 2$.

Assumption 2 (A2): $T = R + P$, as $T \to \infty$, $P, R \to \infty$, and $\frac{P}{R} \to \pi$, with $0 \leq \pi < \infty$.

Assumption 3 (A3): Define $\theta = (\alpha, \beta)'$, where $\theta \in \Theta$, Θ compact, and assume that $f : R^2 \times \Theta \to R^+$ is twice continuously differentiable in the interior of Θ. Define $f_{k,t}(\theta) = f(\Delta Y_t + (Y_t - X_{k,t}\beta) - \alpha)$, and assume that $E(f(v_{k,t})^s) \leq C_3 < \infty, k = 0, \ldots, l$, and $\sup_{\theta \in \Theta} E(|f''_{k,t}(\theta)|^s) < \infty$, with $s > 4$.[3]

Assumption 4 (A4): Let $F_t = \sigma(u_{k,s}, w_{0,s}, k = 0, 1, \ldots l, 0 \leq s \leq t)$, so that $v_{k,t}$ is F_t−measurable for $k = 0, 1, \ldots, l$. Also, $E(E(|v_{k,t}|^{21}|F_{t-1})) \leq C_4 < \infty$, $\forall k$.

Assumption 1 gives moment and memory conditions which suffice for both strong and weak invariance principles to hold. Assumption 3 imposes additional moment conditions on the second derivative of f. While the condition on the moments of the second derivative is innocuous, the differentiability assumption is not. In the sequel we require the following two lemmas.

Lemma 2.1: Assume that A1 and A2 hold, then:

(i) $\sup_{t \geq R} R|\widehat{\beta}_{k,t} - \beta_{k,0}| = O_p(1)$, $\sup_{t \geq R} \sqrt{R}|\widehat{\alpha}_{k,t} - \alpha_{k,0}| = O_p(1)$, for $k = 0, 1, \ldots, l$,[4]

(ii) $R^{1-\gamma}|\widehat{\beta}_{k,R} - \beta_{k,0}| = o_{a.s.}(1)$, $\sqrt{R^{1-\gamma}}|\widehat{\alpha}_{k,R} - \alpha_{k,0}| = o_{a.s.}(1)$, for $k = 0, 1, \ldots, l$, and $\forall \gamma > 0$.

Lemma 2.2: Assume that A1-A4 hold, then:

$$P^{-1/2} \sum_{t=R}^{T-1} f(\widehat{v}_{k,t+1}) = P^{-1/2} \sum_{t=R}^{T-1} f(v_{k,t+1}) - P^{-1/2} \sum_{t=R}^{T-1} f'(v_{k,t+1}) X_{k,t}(\widehat{\beta}_{k,t} - \beta_{k,0})$$

$$- P^{-1/2} \sum_{t=R}^{T-1} f'(v_{k,t+1})(\widehat{\alpha}_{k,t} - \alpha_{k,0}) + o_p(1), \tag{5}$$

where $\widehat{v}_{k,t+1} = \Delta Y_{t+1} + (Y_t - \widehat{\beta}_{k,t} X_{k,t}) - \widehat{\alpha}_{k,t}$, for any $k = 0, ..., l$. It turns out that whenever either f is quadratic or $\pi = 0$, the second and third term on the right hand side of (5) vanish in probability,

---

[3]Note that f' and f'' denote derivatives with respect to the argument, and not derivatives with respect to θ.

[4]Hereafter, with the notation $\sup_{t \geq R}$ we mean $\sup_{R \leq t \leq T-1}$.

so that parameter estimation error vanishes asymptotically, as shown in the proof of the following proposition.

Proposition 2.3: Assume that A1-A4 hold, and either (i) f is a quadratic function (i.e. $f(\widehat{v}_{k,t+1}) = \widehat{v}_{k,t+1}^2$), or (ii) $\pi = 0$. Then as $T \to \infty$, under H_0: $\widehat{d}_P \overset{d}{\to} N(0,1)$, and under H_A, $\Pr\left[|P^{-\gamma}\widehat{d}_P| > \varepsilon\right] \to 1$, for $\varepsilon > 0$ and $\gamma < 1/2$.

In addition, it can easily be shown that:

Proposition 2.4: Assume that A1-A4 hold and that f is not quadratic. For $0 < \pi < \infty$,

(i) $P^{-1/2} \sum_{t=R}^{T-1} f(\widetilde{v}_{k,t+1}) = P^{-1/2} \sum_{t=R}^{T-1} f(v_{k,t+1}) + P^{-1/2} \sum_{t=R}^{T-1} E(f'_{v_k}) X_{k,t}(\widetilde{\beta}_{k,t} - \beta_{k,0}) + o_p(1)$ and

(ii) $P^{-1/2} \sum_{t=R}^{T-1} E(f'_{v_k}) X_{k,t}(\widetilde{\beta}_{k,t} - \beta_{k,0}) \Rightarrow E(f'_{v_k}) \left(\frac{1+\pi}{\pi}\right)^{1/2} \int_{\frac{1}{1+\pi}}^{1} B_{u_k}(s) \left(\frac{\int_0^s B_{u_k}(r)dB_{w.u_k}(r)}{\int_0^s B_{u_k}^2(r)dr}\right) ds$,

where $\widetilde{\beta}_{k,t}$ is an efficient estimator of β (e.g. see Phillips (1991) pp. 419-420)[5], $\widetilde{v}_{k,t}$ is the corresponding regression residual, $E(f_{v_k})$ is the expectation of $f(v_{k,t})$, and $E(f'_{v_k})$ is the expectation of $f'(v_{k,t})$. Also, $B_{w.u_k}$ is a Brownian motion which is independent of B_{u_k}.

It follows that, $P^{-1/2} \sum_{t=R}^{T-1} E(f'_{v_k}) X_{k,t}(\widetilde{\beta}_{k,t} - \beta_{k,0})$ is asymptotically mixed normal, and is normal when conditioned on B_{u_k}. Also, $P^{-1/2} \sum_{t=R}^{T-1} f(v_{k,t+1})$ converges in distribution to a zero mean normal. This ensures that $P^{-1/2} \sum_{t=R}^{T-1} f(\widetilde{v}_{k,t+1})$ is $O_p(1)$. Nevertheless, an asymptotic mixed normality result for $P^{-1/2} \sum_{t=R}^{T-1} (f(\widetilde{v}_{k,t+1}) - E(f(v_{k,t+1})))$ cannot be provided. The reason for this is that $P^{-1/2} \sum_{t=R}^{T-1} (f(v_{k,t+1}) - E(f(v_{k,t+1})))$ converges to a normal which is not independent of B_{u_k}, so that we cannot obtain an explicit expression for the asymptotic covariance of the two right hand side terms. Thus, $P^{-1/2} \sum_{t=R}^{T-1} (f(\widetilde{v}_{k,t+1}) - E(f(v_{k,t+1})))$ does not necessarily converge in distribution to a mixed normal, and so in general we cannot get a mixed normality result for the DM statistic. Based on his version of Proposition 2.4 (i), West (1996) obtains an asymptotic normality result for the parameter estimation error component. Although the above discussion suggests such a result is not generally valid in our framework, the results of West (1988,1996) can be used to obtain an asymptotic normality result when $X_{k,t}$ contains a trend component, so that it becomes possible to obtain a limiting normal distribution for $\widehat{d}_P$, with an asymptotic covariance matrix which reflects parameter estimation error. Consider the following version of (3)[6]:

$$Y_t = \beta_{0,0} X_{0,t} + u_{0,t},$$

---

[5]Note that $\widetilde{\beta}_{k,t}$ is used here rather than $\widehat{\beta}_{k,t}$ in order to ensure that Phillips (1991) mixed normality result holds for the cointegrating vector estimator.

[6]Although we add a deterministic trend to (3) in the following discussion, we drop the intercept. The result extends immediately to the case where an intercept is included if the deviations from the sample means are used.

$$X_{0,t} = \sum_{j=1}^{t} w_{0,j} + \delta_0 t,$$

$$\ldots$$

$$Y_t = \beta_{l,0} X_{1,t} + u_{l,t},$$

$$X_{l,t} = \frac{\beta_{0,0}}{\beta_{l,0}} \sum_{j=1}^{t} w_{0,j} - \frac{1}{\beta_{l,0}} (u_{l,t} - u_{0,t}) + \frac{\beta_{0,0}\delta_0}{\beta_{l,0}} t, \tag{6}$$

and let $\widehat{\overline{\beta}}_{k,t} = \frac{\sum_{j=1}^{t} X_{k,j} Y_j}{\sum_{j=1}^{t} X_{k,j}^2}$. Following standard arguments we see that,

$$\widehat{\overline{\beta}}_{k,t} - \beta_0 = \frac{\sum_{j=1}^{t} X_{k,j} u_{kj}}{\sum_{j=1}^{t} X_{k,j}^2} = \frac{\sum_{j=1}^{t} j u_{k,j}}{\sum_{j=1}^{t} \delta_k j^2} + o_p(1), \tag{7}$$

where the $o_p(1)$ term holds uniformly in t, and $\delta_k = \frac{\beta_{0,0}}{\beta_{k,0}} \delta_0$. By similar arguments to those used in the proof of Proposition 2.4 we have that

$$P^{-1/2} \sum_{t=R}^{T-1} (f(\widehat{v}_{k,t+1}) - E(f_k))$$

$$= P^{-1/2} \sum_{t=R}^{T-1} (f(v_{k,t+1}) - E(f_k)) + P^{-1/2} E(f'(v_{k,t+1})) \sum_{t=R}^{T-1} X_{k,t} (\widehat{\overline{\beta}}_{k,t} - \beta_0) + o_p(1)$$

$$= P^{-1/2} \sum_{t=R}^{T-1} (f(v_{k,t+1}) - E(f_k)) + \mu_k P^{-1/2} \sum_{t=R}^{T-1} t \left(\frac{\sum_{j=1}^{t} j u_{kj}}{\sum_{j=1}^{t} j^2} \right) + o_p(1), \tag{8}$$

where $\mu_k = E(f'(v_{k,t+1}))$. Also, note that $\frac{1}{t^3} \sum_{j=1}^{t} j^2 \to \frac{1}{3}$. Thus, we can rewrite the second RHS part of (8) as

$$3\mu_k (P^{-1/2} \sum_{t=R}^{T-1} \frac{1}{t^2} u_{k,1} + P^{-1/2} \sum_{t=R}^{T-1} \frac{2}{t^2} u_{k,2} + \ldots + P^{-1/2} \sum_{t=R}^{T-1} \frac{R}{t^2} u_{k,R}$$

$$+ P^{-1/2} \sum_{t=R+1}^{T-1} \frac{R+1}{t^2} u_{k,R+1} + P^{-1/2} \sum_{t=R+2}^{T-1} \frac{R+2}{t^2} u_{k,R+2} + \ldots + P^{-1/2} \frac{1}{R+P-1} u_{k,R+P-1})$$

As each term in this sum is asymptotically normal, we have that the second RHS term in (8) is asymptotically normal with a covariance matrix equal to

$$V_2 = 9\mu_k^2 \lim_{T \to \infty} \left(P^{-1} \sum_{j=1}^{T-1} a_j^2 \gamma_0 + 2 \sum_{i=1}^{T-1} a_i a_j \gamma_j \right),$$

7

where $\gamma_j = E(u_{k,i}u_{k,i+j})$, $a_j = \sum_{t=R}^{T-1} \frac{j}{t^2}$, for $j = 1, \ldots, R$, and $a_j = \sum_{t=j}^{T-1} \frac{j}{t^2}$, for $j = R+1, \ldots, R+P-1$. Now let $\overline{f}_{k,t} = f(v_{k,t}) - E(f(v_{k,t}))$. Then,

$$\mu_k P^{-1} \sum_{t=R}^{T-1} \overline{f}_{k,t+1} X_{k,t}(\widehat{\widehat{\beta}}_{k,t} - \beta_0) = 3\mu_k P^{-1} \sum_{t=R}^{T-1} \overline{f}_{k,t+1} \frac{1}{t} \sum_{j=1}^{t} \frac{j}{t} u_{k,j} + o_p(1).$$

Now,

$$P^{-1} \sum_{t=R}^{T-1} \overline{f}_{k,t+1} \frac{1}{t} \sum_{j=1}^{t} \frac{j}{t} u_{k,j} = P^{-1}\overline{f}_{k,R+1}\frac{1}{R}\left(\frac{1}{R}u_{k,1} + \ldots \frac{R-1}{R}u_{k,R-1} + u_{k,R}\right) + \ldots$$

$$+ P^{-1}\overline{f}_{k,R+P}\frac{1}{R+P}\left(\frac{1}{R+P}u_{k,1} + \ldots \frac{R+P-1}{R+P}u_{k,R+P-1} + u_{k,R+P}\right)$$

Thus,

$$\lim_{T\to\infty} E(\mu_k P^{-1} \sum_{t=R}^{T-1} f(v_{k,t+1})X_{k,t}(\widehat{\widehat{\beta}}_{k,t} - \beta_0)) = C,$$

with

$$C = \lim_{T\to\infty} \left(P^{-1} \sum_{t=R}^{T-1} \sum_{i=0}^{t-1} \frac{t-i}{t^2}\gamma_i^f\right),$$

where $\gamma_i^f = E(\overline{f}_{k,t+i+1}u_{k,t})$. Finally let $V_1 = 2\sum_{j=0}^{\infty} \gamma_j^{ff}$, where $\gamma_j^{ff} = E(\overline{f}_t\overline{f}_{t+j})$. Thus,

$$P^{-1/2} \sum_{t=R}^{T-1}(f(\widehat{v}_{k,t+1}) - E(f_k)) \xrightarrow{d} N(V_1 + V_2 + 2C).$$

Although we do not extensively examine the case where f is non quadratic and $\pi > 0$, the following approach may be worthy of further investigation. Consider the following statistic. Let $\xi_i, i = 1, \ldots, P$ be an iid $N(0,1)$ random variable, and construct:

$$S_P = \frac{1}{\sqrt{P}} \sum_{i=1}^{P} \xi_i + P^\gamma(P^{-1} \sum_{t=R}^{T-1}(f(\widetilde{v}_{0,t+1}) - f(\widetilde{v}_{k,t+1}))), \ \gamma \in (0, 1/2).$$

Do not reject H_0 if $|S_P| \leq c_\alpha$, and reject if $|S_P| > c_\alpha$, where c_α is the $(1-\alpha)^{th}$ percentile of the standard normal distribution. This provides a rule with type I error approaching α and type II error approaching zero, as $P \to \infty$. The reason for this is that $\frac{1}{\sqrt{P}}\sum_{i=1}^{P}\xi_i$ is a standard normal random variable for any P, while $P^{-1}\sum_{t=R}^{T-1}(f(\widetilde{v}_{k,t+1}) - E(f(v_k))) = O_p(P^{-1/2})$. Thus, $P^{-1}\sum_{t=R}^{T-1}(f(\widetilde{v}_{0,t+1}) - f(\widetilde{v}_{k,t+1}))$ vanishes in probability at rate $P^{\gamma-1/2}$ under the null, and diverges at rate P^γ under the alternative. Finite sample properties of S_P are discussed in Section 3.

Thus far, we have not addressed a number of issues which arise in practical applications of the DM test. First, we have assumed that $h = 1$. The extension of our analysis to the multiple horizon case follows directly. Assume that (4) is true, and estimate a model of the form $\Delta Y_{t+2} = \alpha_{k,0} - (Y_t - \beta_{k,0}X_{k,t}) + v_{k,t+2}$.[7] It follows that $v_{k,t+2} = \beta_{0,0}w_{0,t+2} + \Delta u_{0,t+1} + u_{k,t}$. Thus, $\widehat{v}_{k,t+2} = v_{k,t+2} + (\widehat{\beta}_{k,t} - \beta_{k,0})X_{k,t} + (\widehat{\alpha}_{k,t} - \alpha_{k,0})$, and so Lemma 2.2 and Proposition 2.3 (and Proposition 2.5 below) follow directly. In addition, extensions to cases where $h > 2$ follow directly. Second, consider estimating models of the form given by (4), but with addition short-run dynamics terms. In particular, consider models of the form:

$$\Delta Y_{t+1} = \alpha_{k,0} - (Y_t - \beta_{k,0}X_{k,t}) + \sum_{i=0}^{m_1} \delta_{k,i}\Delta X_{k,t-i} + \sum_{i=0}^{m_2} \gamma_{k,i}\Delta Y_{t-i} + v_{k,t+1}, \ k = 0,...,l.$$

Let $\widehat{v}_{k,t+1} = v_{k,t+1} + (\widehat{\beta}_{k,t} - \beta_{k,0})X_{k,t} + (\widehat{\alpha}_{k,t} - \alpha_{k,0}) + \sum_{i=0}^{m_1}(\widehat{\delta}_{k,i,t} - \delta_{k,i})\Delta X_{k,t-i} + \sum_{i=0}^{m_1}(\widehat{\delta}_{k,i,t} - \delta_{k,i})\Delta Y_{t-i}$, where $\widehat{\beta}_{k,t}$ is estimated as above and all of the other coefficients are then estimated by least squares. From the definition of $\widehat{v}_{k,t+1}$, it is clear that the parameter estimation error terms that arise from the short run dynamics can be treated as in West (1996), so that Lemma 2.2 and Proposition 2.3 (and Lemma 2.5) follow directly. Third, all of the above results extend immediately to the case of non scalar X, as the proofs to our propositions can be applied component by component.

2.2   Comparing Multiple Forecasting Models

In this subsection, we address the problem of comparing multiple forecasting models against a given benchmark model. It is well known that sequential use of statistical tests (such as the DM test) results in sequential test bias. Given this fact, we draw on recent results from the stimulating paper by White (2000) to provide a testing framework which is suitable in our context.

Consider the following statistic:

$$\max_{k=1,...,l} \widehat{V}_{P,k} = \max_{k=1,...l} P^{-1/2} \sum_{t=R}^{T-1} \left((f(\widehat{v}_{0,t+1}) - E(f(v_{0,t}))) - (f(\widehat{v}_{k,t+1} - E(f(v_{k,t})))) \right).$$

From Proposition 2.3 we know that for f quadratic and/or $\pi = 0$, $\widehat{V}_{P,k}$ is asymptotically normal. As k is finite, it follows that $\max_{k=1,...,l} \widehat{V}_{P,k} \Rightarrow V_{Max}$, where $V_{Max} = \max_{k=1,...,l}\{Z_k\}$, with Z_k a one-dimensional Gaussian process for any given k. Also, $(Z_1, \ldots, Z_l)'$ is an l-dimensional Gaussian

---

[7]Estimating models of this type in order to construct multi-step ahead forecasts is common.

process, with a covariance structure which depends on the covariance between the forecast errors from the different models. However, recall that the maximum of a Gaussian process is not Gaussian in general, so that standard critical values cannot be used to conduct inference on $\max_{k=1,\dots,l} \widehat{V}_{P,k}$. As pointed out by White (2000), one possibility in this case is to first estimate the covariance structure and then draw 1 realization from an l-dimensional normal with covariance equal to the estimated covariance structure. From this realization, pick the maximum value over $k = 1, \dots, l$. Repeat this a large number of times, form an empirical distribution using the maximum values over $k = 1, \dots, l$, and obtain critical values in the usual way. A drawback to this approach is that we need to rely on an estimator of the covariance structure based on the available sample of observations, which in many cases may be small relative to the number of models being compared. Furthermore whenever the forecasting errors are not martingale difference sequences (as in our context), heteroskedasticity and autocorrelation consistent covariance matrices should be estimated, and thus a lag truncation parameter must be chosen. Another approach which avoids these problems involves using the stationary bootstrap of Politis and Romano (1994). In general, bootstrap procedures have been shown to perform well in a variety of finite sample contexts. For example, see Diebold and Chen (1996), where bootstrap procedures for testing structural stability are assessed, and are found to perform better than procedures based on asymptotic approximations. White (2000) uses this approach in the context of short-memory observations (software implementing White's technique is available from *NeuralNet R&D Associates*, patent pending). We here show that White's (2000) approach also holds for the case of cointegrated variables.

To begin, consider the stationary bootstrap of Politis and Romano (1994). From a geometric distribution (i.e. $\Pr(L_i = m) = (1 - q)^{m-1} q$, m a positive integer) with parameter $q \in (0, 1)$ (see below), randomly draw block sizes, say $L_1, L_2 \dots$. Then draw a sequence of identically and independently distributed random variables, $I_1, I_2, \dots$, from a discrete uniform distribution on $\{R+1, \dots, T\}$. Recall that the $\widehat{v}_{k,t}$ are the forecasting errors from models $k = 0, \dots, l$. Pick a first realization of L, say L_1, and then pick a first realization of I, say I_1. The first L_1 observations of a pseudo series (say $\widehat{v}^*_{k,t}$, $t = R + 1, \dots, T$) of length P, are defined to be $\widehat{v}^*_{k,1}, \dots \widehat{v}^*_{k,L_1}$, $k = 0, \dots, l$, and are set equal to $\widehat{v}_{k,I_1}, \dots \widehat{v}_{k,I_1+L_1-1}$. Then L_2 observations, beginning from I_2, are concatenated onto the pseudo time series, $\widehat{v}^*_{k,t}$. This process is continued across all $k = 0, \dots, l$ using the same L and I values to construct pseudo time series for each model. Finally, compute $f^*(\widehat{v}^*_{0,t}) - f^*(\widehat{v}^*_{k,t})$,

$k = 1, \ldots, l$, and construct the resampled statistic,

$$P^{-1/2} \sum_{t=R}^{T-1} \left(f^*(\widehat{v}_{0,t+1}^*) - f^*(\widehat{v}_{k,t+1}^*) \right) = P^{-1/2} \sum_{t=R}^{T-1} f^*(\widehat{\theta}_{k,t}^*), \quad k = 1, \ldots, l.$$

Collecting these statistics into a vector, define,

$$P^{-1/2} \sum_{t=R}^{T-1} f^*(\widehat{\theta}_t^*) = P^{-1/2} \sum_{t=R}^{T-1} (f^*(\widehat{\theta}_{1,t}^*), \ldots, f^*(\widehat{\theta}_{l,t}^*))',$$

where $\widehat{\theta}_t^* = (\widehat{\theta}_{0,t}^*, \ldots, \widehat{\theta}_{l,t}^*)'$, $\widehat{\theta}_{k,t}^* = (\widehat{\alpha}_{k,t}^*, \widehat{\beta}_{k,t}^*)'$, $\widehat{\alpha}_{k,t}^* = \widehat{\alpha}_{\tau(t)}$, and $\widehat{\beta}_{k,t}^* = \widehat{\beta}_{\tau(t)}$, with $\tau(t)$ a discrete uniform random variable on $[R+1, T]$. Also, define,

$$\widehat{G}_{P,k} = P^{-1/2} \sum_{t=R}^{T-1} (f(\widehat{v}_{0,t+1}) - f(\widehat{v}_{k,t+1})) = P^{-1/2} \sum_{t=R}^{T-1} f(\widehat{\theta}_{k,t}), \quad k = 1, \ldots, l. \qquad (9)$$

Then, let $\widehat{G}_{\max} = \max_{k=1,\ldots,l} \widehat{G}_{P,k}$. Finally, define,

$$P^{-1/2} \sum_{t=R}^{T-1} f(\widehat{\theta}_t) = P^{-1/2} \sum_{t=R}^{T-1} (f(\widehat{\theta}_{1,t}), \ldots, f(\widehat{\theta}_{l,t}))',$$

and let $E(f) = (E(f(v_{0,t}) - f(v_{1,t})), \ldots E(f(v_{0,t}) - f(v_{l,t})))'$. Again exploiting strict stationarity, consider the following hypotheses:

$$H_0 : \max_{k=1,\ldots,l} E(f(v_{0,t}) - f(v_{k,t})) \leq 0$$

$$H_A : \max_{k=1,\ldots,l} E(f(v_{0,t}) - f(v_{k,t})) > 0.$$

Thus, the best competing model does not outperform the benchmark model under H_0, while it has better predictive accuracy under H_A. It follows that if H_0 is not rejected, the benchmark model is preferred, while it should be discarded otherwise. Our objective is to show that in the cointegration case, the limiting distribution of $P^{-1/2} \sum_{t=R}^{T-1} (f^*(\widehat{\theta}_t^*) - f(\widehat{\theta}_t))$ approaches the limiting distribution of $P^{-1/2} \sum_{t=R}^{T-1} (f(\widehat{\theta}_t) - E(f))$, which is a k−dimensional normal under both H_0 and H_A, when conditioning on the data. This ensures that it is valid to construct $P^{-1/2} \sum_{t=R}^{T-1} (f^*(\widehat{\theta}_t^*) - f(\widehat{\theta}_t))$, maximize this expression over k, repeat this a large number of times, and then compute the appropriate percentiles. Then, choose H_0 for $\widehat{G}_{\max}$ values below the 95% percentile, say, and reject otherwise. Under H_0, $\widehat{G}_{\max}$ either diverges to minus infinity, or has the same limiting distribution as $\max_k P^{-1/2} \sum_{t=R}^{T-1} (f^*(\widehat{\theta}_t^*) - f(\widehat{\theta}_t))$ (i.e. when $E(f(v_0) - f(v_k)) = 0, \forall k$). On the other hand, $\widehat{G}_{\max}$

diverges to infinity under H_A. Thus, the procedure outlined above results in tests with asymptotic size equal to or smaller than the nominal size, and with unitary asymptotic power.

In order to show the desired result, we need to address the issue of estimated parameters. In particular, terms involving $(\widehat{\theta}^*_{k,t} - \theta_{k,0})$ must be appropriately treated. White's proof relies on the fact that $(\widehat{\theta}_{k,t} - \theta_{k,0})$ satisfies the law of the iterated logarithm (LIL), when properly scaled, so that $(\widehat{\theta}^*_{k,t} - \theta_{k,0})$ also satisfies the LIL. As our variables are I(1), this fact is not generally applicable. Nevertheless, we are able to show that if P does not grow too fast relative to R, parameter estimation error vanishes not only in probability, but also almost surely. Our results are gathered in the following lemma and theorem:

Lemma 2.5: Assume that A1-A4 hold. If $P = o(T^{2/3}/\log\log T)$, then for any $k = 0, \ldots, l$,

$$P^{-1/2} \sum_{t=R}^{T-1} f(\widehat{v}_{k,t+1}) = P^{-1/2} \sum_{t=R}^{T-1} f(v_{k,t+1}) + o_{a.s.}(1).$$

Lemma 2.5 is an almost sure version of Lemma 2.2. The cost of strengthening the weak result to a strong result is the requirement that P grow at a slower rate than previously assumed.

Theorem 2.6: Assume that A1-A4 hold. Also, assume that $P = o(T^{2/3}/\log\log T)$. If $q \to 0$ and $Pq \to \infty$ as $P \to \infty$ then[8]:

$$\rho\left(L\left(P^{-1/2}\sum_{t=R}^{T-1}(f^*(\widehat{\theta}^*_t) - f(\widehat{\theta}_t))|data\right), L\left(P^{-1/2}\sum_{t=R}^{T-1}(f(\widehat{\theta}_t) - E(f))\right)\right) \overset{pr}{\to} 0,$$

where ρ is a metric metricizing convergence in distribution, and L denotes the distributional law. As a straightforward consequence of the continuous mapping theorem, the result above generalizes to the max statistic. Thus, we can proceed as outlined above.

3   Monte Carlo Evidence

McCracken (1998,2000), Clark and McCracken (2000), West (1996) and West and McCracken (1998) provide various theoretical results and Monte Carlo evidence underscoring the importance of short run parameter estimation error in the context of predictive ability tests. In this section, we examine the impact of both short run and cointegrating vector parameter estimation error (see e.g. Gonzalo (1995) and Gonzalo and Lee (1998)) on DM and data snooping type tests constructed

---

[8]Recall that q is the parameter characterizing the geometric distribution from which the block size is drawn.

using a quadratic loss function (mean squared forecast error: MSE), and on S_P type tests using MSE and mean absolute forecast error (MAD) loss functions.

Following the notation of Section 2.2, consider the following three pseudo DGPs[9]: (1) $Y_t = \beta_{0,0}X_{0,t} + u_{0,t}$, $X_{0,t} = \sum_{j=1}^{t} w_{0,j}$;(2) $Y_t = \beta_{1,0}X_{1,t} + u_{1,t}$, $X_{1,t} = \frac{\beta_{0,0}}{\beta_{1,0}}\sum_{j=1}^{t} w_{0,j} - \frac{1}{\beta_{1,0}}(u_{1,t} - u_{0,t})$; and (3) $Y_t = \beta_{2,0}X_{2,t} + u_{2,t}$, $X_{2,t} = \frac{\beta_{0,0}}{\beta_{2,0}}\sum_{j=1}^{t} w_{0,j} - \frac{1}{\beta_{2,0}}(u_{2,t} - u_{0,t})$. Direct substitution yields the corresponding error correction models: (1) $\Delta Y_{t+1} = \delta_{0,0}(Y_t - \beta_{0,0}X_{0,t}) + v_{0,t+1}$; (2) $\Delta Y_{t+1} = \delta_{1,0}(Y_t - \beta_{1,0}X_{1,t}) + v_{1,t+1}$,and (3) $\Delta Y_{t+1} = \delta_{2,0}(Y_t - \beta_{2,0}X_{2,t}) + v_{2,t+1}$,where $v_{0,t} = u_{0,t} + \beta_{0,0}w_{0,t}$, $v_{1,t} = u_{1,t-1} + \beta_{0,0}w_{0,t} + \Delta u_{0,t}$, and $v_{2,t} = u_{2,t-1} + \beta_{0,0}w_{0,t} + \Delta u_{0,t}$,with $\delta_{0,0} = \delta_{1,0} = \delta_{2,0} = -1$. Henceforth, set $\beta_{0,0} = \beta_{1,0} = \beta_{2,0} = 1$. For convenience, assume that $E(u_{0,t}, u_{1,s}) = E(u_{0,t}, u_{2,s}) = E(u_{2,t}, u_{3,s}) = 0$, $\forall\, t, s$[10]. Denoting variances and covariances using standard notation it follows that,

$$\sigma_{v0}^2 = \sigma_{u0}^2 + \sigma_{w0}^2 + 2\sigma_{u_0,w_0},$$

$$\sigma_{v1}^2 = \sigma_{u1}^2 + \sigma_{w0}^2 + 2\sigma_{u0}^2 + 2\sigma_{u_0,w_0} - 2\sigma_{u_0,u_{0-1}},$$

$$\sigma_{v2}^2 = \sigma_{u2}^2 + \sigma_{w0}^2 + 2\sigma_{u0}^2 + 2\sigma_{u_0,w_0} - 2\sigma_{u_0,u_{0-1}},$$

where $\sigma_{u_0,u_{0-1}} = cov(u_{0,t}, u_{0,t-1})$. Thus, the null and alternative hypotheses for the DM and S_P tests can be written as:

$H_0 : \sigma_{v1}^2 - \sigma_{v0}^2 = \sigma_{u_0}^2 + \sigma_{u_1}^2 - 2\sigma_{u_0,u_{0-1}} = 0,$

$H_A : \sigma_{v1}^2 - \sigma_{v0}^2 = \sigma_{u_0}^2 + \sigma_{u_1}^2 - 2\sigma_{u_0,u_{0-1}} \neq 0.$

The hypotheses for the data snooping test, similarly, can be written as:

$H_0 : \sigma_{v1}^2 - \sigma_{v0}^2 = \sigma_{u_0}^2 + \sigma_{u_1}^2 - 2\sigma_{u_0,u_{0-1}} = 0$ and $\sigma_{v2}^2 - \sigma_{v0}^2 = \sigma_{u_0}^2 + \sigma_{u2}^2 - 2\sigma_{u_0,u_{0-1}} \geq 0,$

$H_A : \sigma_{v1}^2 - \sigma_{v0}^2 = \sigma_{u_0}^2 + \sigma_{u_1}^2 - 2\sigma_{u_0,u_{0-1}} < 0$ and $\sigma_{v2}^2 - \sigma_{v0}^2 = \sigma_{u_0}^2 + \sigma_{u2}^2 - 2\sigma_{u_0,u_{0-1}} \leq 0.$[11]

Our experiments are based on the following parameterizations, with $w_0 \sim iidN(0,4)$, and all varibles generated as Gaussian:

Empirical Size (DM and S_P Tests): $\sigma_{u_0}^2 = 2$, $\sigma_{u_1}^2 = 1$, $\sigma_{u_0,u_{0-1}} = 1.5$, and $u_{0,t} = 0.75u_{0,t-1} + \xi_t$, where $\xi_t \sim iidN(0, 7/8)$, as $\sigma_{u_0}^2 = \sigma_\xi^2/(1 - \rho^2)$, $\rho = 0.75$.

Empirical Power (DM and S_P Tests): As above with $\sigma_{u_0,u_{0-1}} = 0$ and hence $\rho = 0$ in both cases.

---

[9]The first two DGPs are used for DM and S_P experiments, while all three are used for data snooping experiments.

[10]Note that that even though these cross correlations are set equal to zero, we allow for cross correlation among the $X_{i,t}$, $i = 0, 1, 2$, while retaining the property that the three models are nonnested.

[11]Note that H_0 for the data snooping case could be written using $\geq$ instead of equality signs. In addition, note that a particular alternative is given as H_A, although in general H_A is simply the negation of H_0.

Empirical Size (Data Snooping Tests): As empirical size experiment above, but additionally set $\sigma_{u_2}^2 = 1.0$.

Empirical Power (Data Snooping Tests): As empirical power experiment above, but additionally set $\sigma_{u_1}^2 = 0.1$ and $\sigma_{u_2}^2 = 5.0$.

Sequences of P ex ante 1-step ahead forecasts of Y_t are constructed. All forecasts are formed using models estimated with R observations for the first forecast, $R+1$ observations for the second forecast, etc., for $R = \{25, 50, 100, 150, 200, 250\}$ and $P = \{25, 50, 100, 150, 200, 250\}$. The following three cases are considered: *Case a* - No parameter estimation error: β and δ known; *Case b* - Cointegrating parameter estimation error: β estimated, δ known; and *Case c* - Cointegrating and short run parameter estimation error: β and δ estimated. For Cases b and c, forecasting models are estimated using a two step procedure with cointegrating vectors estimated first using OLS. Additionally, all experiments are based on 5000 Monte Carlo simulations, and values of $\gamma = \{0.1, 0.2, 0.3, 0.4, \text{and } 0.45\}$ are used for the S_P test. Additional experiments were also run with different parameterizations, and results are available upon request from the authors (see also Corradi, Swanson, and Olivetti (1999)). Given the DGPs above, $u_{i,t}, v_{i,t}, w_{0,t}, i = 0, 1, 2$ are geometrically mixing with all finite moments, so **A1** and **A4** are satisfied. **A3** is trivially satisfied for a quadratic loss function. In practice we do not know π, although we can set $\widehat{\pi} = \frac{P}{R}$, so that **A2** is not immediately verifiable in finite sample. For the case of the DM test with quadratic loss function, and for the S_p (Tables 1,3,4) test, any finite value of π is valid. On the other hand, for the data snooping test (Table 2) we require $P = o(R^{2/3})$ and so $\pi = 0$. In this sense, findings for the case where $P/R < 1$ are more reliable than those for the case where $P/R \geq 1$.

The findings from our experiments are summarized in Tables 1-4. All reported rejection frequencies are based on 5% nominal size tests, and power rejection frequencies are size adjusted (except for the results based on the data snooping test in Table 2). Results based on 10% nominal size tests are available upon request. As expected, DM and data snooping tests perform more poorly when parameters are not known, with the data snooping tests being severely affected when short run parameters are estimated. (For a complete Monte Carlo analysis of data snooping tests in which a broad set of DGPs are specified, the reader is referred to the working paper version of White (2000).) However, note that little is given up when only the cointegrating vector parameter is unknown (Case b). In addition, empirical size generally improves with R for fixed P (e.g. see Table 1). Comparison of empirical power entries in Table 1 suggests that power increases with P, as

14

expected, and that estimating cointegrating vectors in addition to short run parameters does affect power, which is around 0.71 for Case c, when $P = R = 250$. Note also that parameter estimation error plays a much more important role when using the data snooping test than when using the simpler DM test. For the S_P tests (see Tables 3 and 4), we report results for only one γ value for MSE loss (i.e. $\gamma = 0.1$) and one value for MAD loss (i.e. $\gamma = 0.4$). Other results are available upon request, and as expected conform to the convention that lower γ results in improved size, while higher γ results in improved power. The results which we report essentially choose the best γ values for the two different loss functions, suggesting that for MSE loss, values of γ even lower than 0.1 should result in even better size, possibly at little expense in power, given the rather high finite sample power figures even when $\gamma = 0.1$ (e.g. power is 0.80 when all parameters are estimated for $P = R = 250$). However, for the nondifferential loss function (MAD), improvement in size comes at the expense of even further reduction of the already relatively low power (note that power is 0.41 when all parameters are estimated for $P = R = 250$). We leave further analysis of the tradeoffs between the various tests to future research.

4   Empirical Illustration

In this section, we examine the extent to which different money measures are useful for predicting industrial production. This question is answered by examining models of the variety used previously to assess the extent to which fluctuations in money stock anticipate fluctuations in real income (see for example, Christiano and Ljungqvist (1988), Friedman and Kuttner (1993), Stock and Watson (1989), Hafer and Jansen (1991), Thoma (1994), Hafer and Kutan (1997), and the references contained therein). One feature which links these papers is the assumption that either M1 or M2 (but usually not both) are adequate measures of money. Given that outside money (i.e. the monetary base) accounts only for around 10% of the broader M2 aggregate, and that the behavior of M2 since around 1985 has been rather erratic (partly because of well documented recent shifts in the public's demand for money balances), it should be of interest to assess whether income growth is more closely tied to M1 or M2 growth. This question can be answered by comparing alternative nonnested models using DM tests.

In the following example, we consider VAR models similar in spirit to those examined by Swanson (1998) and Amato and Swanson(2000). In particular, monthly observations from 1959:1

- 1997:12 on seasonally adjusted and logged divisia M1 ($m1_t$), M2 ($m2_t$), M3 ($m3_t$), the log of seasonally adjusted industrial production (y_t), the log of the wholesale price index (p_t), the secondary market rate on 90-day U.S. Treasury bills (R_t), and the three-month financial commercial paper (C_t) are used to specify single equation forecasting models of industrial production. The commercial paper rate is taken from the Federal Reserve Board of Governors (see e.g. http://bos.business.uab.edu/browse), while the other variables (including the divisia monetary aggregates) are available from the St. Louis Federal Reserve Bank website.[12] Based on augmented Dickey-Fuller tests, we found that all of our series are best characterized as being I(1), although the deterministic components vary depending on the variable examined. A summary of these results is available upon request from the authors.

Our empirical approach can be summarized as follows. We construct sequences of 180 (corresponding to the period 1983:1-1997:12) ex-ante h-step ahead (h=1,3,12) forecasts for industrial production growth based on models of the form:

$$\Delta y_t = \alpha_0 + B(L)\Delta x_{t-1} + \sum_{i=1}^{r} \delta_i z_{i,t-1} + u_t,$$

where x_t is alternately $x1_t$, $x2_t$, or $x3_t$ with $x1_t = (y_t, m1_t, p_t, R_t, C_t)'$, $x2_t = (y_t, m2_t, p_t, R_t, C_t)'$, and $x3_t = (y_t, m3_t, p_t, R_t, C_t)'$, u_t is an error term, and $z_{i,t-1} = \widehat{\beta}'_i x_{t-1}$, i=1,...,r, is a vector of error-correction terms estimated before the primary forecasting model is estimated (see above discussion) using least squares, r is the estimated rank of the cointegrating space ($0 \leq r < 5$), and $B(L)$ is a matrix polynomial in the lag operator L. Note that the number of cointegrating vectors was fixed to be 2 in all cases, and the variables used in the construction of the error correction terms are (i) y_t, mi_t, p_t, R_t, $i = 1, 2, 3$ and (ii) R_t and C_t, corresponding to the empirical findings of Swanson (1998). The lag order of our models is chosen using either the AIC or the BIC. It is perhaps worth stressing that all models are re-estimated at each point in time before each new forecast is constructed, resulting in new parameter estimates, estimates of r, and estimates of the lags in the system. Also, the sample is increased by one observation before each new forecast is constructed. For example, when we compare models using the entire ex-ante forecast period, the first 1-step ahead forecast is based on a model estimated using data from 1959:1-1982:12, while the last 1-step ahead forecast is based on the sample 1959:1-1997:11. Similarly, for h=3, the first

---

[12]We also fitted models using log of seasonally adjusted nominal M1, M2, and M3. Our findings based on these alternative money measures were qualitatively the same (complete results are available upon request from the authors).

sample used runs from 1959:1-1982:10. Alternatively, when we are comparing models using only a three year ex-ante period, the first 1-step ahead forecast is based on a model estimated using data from 1959:1-1994:12, while the last 1-step ahead forecast is based on a forecasting model estimated using data from 1959:1-1997:11. The resulting ex-ante forecast errors are then compared on a model by model basis using the DM test (i.e. M1 versus M2 models as well as M1 versus M3 models). Finally, all three models are compared using the data snooping test.[13]

Our empirical findings are summarized in Table 5, with additional results available in Corradi, Swanson and Olivetti (1999). The table contains MSE values for different forecast horizons and lag selection criteria - all based on the entire 15 year ex-ante period. Our interpretaion of the findings summarized in the table are as follows:

· $\Delta m2_t$ and $\Delta m3_t$ appear to be more useful than $\Delta m1_t$ for predicting Δy_t, although there is little to choose between the broader money measures. For example, note that based on the AIC lag selection criterion, forecasts made using $\Delta m2_t$ and $\Delta m3_t$ dominate those made using $\Delta m1_t$ at a 5% significance level (based on the DM test) for h=1 and h=3. The same is true when considering $h = 1$ in the case where lags are selected using the BIC (see lower panel in Table 3). In addition, the data snooping test was also performed with $\Delta m2_t$ as the benchmark model, and the null hypothesis failed to reject at a 5% level, suggesting that neither the $\Delta m1_t$ or the $\Delta m3_t$ forecasts MSE-dominated the $\Delta m2_t$ forecasts at the 1 and 3-month horizons.

· The exception to the above finding appears to be $h = 12$. For this longer forecast horizon, models based on the AIC have lower point MSEs when $\Delta m1_t$ is used, although not significantly so (based on DM tests). On the other hand, models based on the BIC have significantly lower point MSEs based on a comparison of forecasts made using $\Delta m2_t$, but not based on $\Delta m3_t$. Thus, the evidence is rather mixed, with broader money measures being preferred for shorter forecasting horizons, and narrower measures MSE-dominating at the longer 1-year ahead horizon.

· As expected, the BIC usually selects models with lower MSEs relative to analogous models

---

[13]It should perhaps be noted that in an interesting paper, Christoffersen and Diebold (1998) discuss cointegration and long-horizon forecasting. They show that in certain contexts, nothing is lost at long horizons by ignoring cointegration. Their findings highlight an important deficiency among standard forecast accuracy measures, namely that they fail to adequately account for the usefulness of maintaining cointegrating relations among economic variables. Our approach differs from theirs in a number of ways. For example, we focus on how best to select lags when different forecast criteria are used and whether or not forecast horizon plays a role in the selection of parsimonious versus less parsimonious models.

constructed based on the AIC.

5   Conclusions

We have used the framework introduced by Diebold and Mariano (1995), West (1996), and White (2000) to examine tests of predictive ability in cointegrated systems with parameter estimation error. Our conclusions can be summarized as follows: First, when the loss function of the forecaster is quadratic and/or the length of the forecast period, P, grows at a slower rate than the length of the regression period, then the standard Diebold-Mariano test can be used. On the other hand, in the case of a generic loss function, the asymptotic normality result of West (1996) no longer holds. Second, when P grows at an appropriate rate, we show that parameter estimation error vanishes almost surely. This allows us to extend the "data snooping" technique of White (2000) for comparing the predictive ability of multiple forecasting models to the case of cointegrated variables. Third, via a series of Monte Carlo experiments, we show how cointegrating vector and short run parameter estimation error affect the finite sample performance of DM and data snooping tests. Fourth, we find that M2 serves as a better predictor of income growth than does M1.

This work is a beginning. Many questions still remain for subsequent research. On the theoretical side it is of interest to extend our procedures to DGPs evolving in continuous time. In addition, non differentiable loss functions pose a series of interesting problems within the context of constructing Diebold-Mariano type test statistics which yet remain to be answered. A few interesting empirical projects include assessing the dependence of forecast model selection on the ex-ante sample period used and determining the usefulness of the types of tests discussed above for constructing ex-ante based causality tests, for example.

6 Appendix

Hereafter C denotes a positive constant, i.e. $0 < C < \infty$.

Proof of Lemma 2.1: Recall that $\overline{X}_{k,t} = \frac{1}{t}\sum_{j=1}^{t} X_{k,j}$, $\widehat{\beta}_{k,t} - \beta_{k,0} = \frac{\sum_{j=1}^{t}(X_{k,j} - \overline{X}_{k,t})u_{k,j}}{\sum_{j=1}^{t}(X_{k,j} - \overline{X}_{k,t})^2}$, and $\widehat{\alpha}_{k,t} - \alpha_{k,0} = \frac{1}{t}\sum_{j=1}^{t} u_{k,j} - (\widehat{\beta}_{k,t} - \beta_{k,0})\frac{1}{t}\sum_{j=1}^{t} X_{k,j}$. Also recall that $X_{k,t} = \frac{\beta_{0,0}}{\beta_{k,0}}\sum_{j=1}^{t} w_{0,j} + \frac{1}{\beta_{k,0}}(u_{0,t} - u_{k,t}) + \frac{1}{\beta_{k,0}}(\alpha_{0,0} - \alpha_{k,0})$. As the proofs of Lemmas 2.1 and 2.2 do not depend on k, the subscript is dropped in this and the next proof.

(i) Given A1 and A2,

$$\sup_{t \geq R} R|\widehat{\beta}_t - \beta_0| \leq \sup_{t \geq R} \left| \frac{R \frac{1}{T}\sum_{j=1}^{t}(X_j - \overline{X}_t)u_j}{T \frac{1}{T^2}\sum_{j=1}^{R}(X_j - \overline{X}_R)^2} \right| \overset{d}{\to} \sup_{r \in (\frac{1}{1+\pi}, 1)} \left| \frac{1}{1+\pi} \frac{\int_0^r B_w^\mu(s)dB_u(s) + rC}{\int_0^{(1+\pi)^{-1}} B_w^{\mu,2}(s)d(s)} \right|, \tag{10}$$

where C captures the long run covariance between w and u, B_u and B_w are Brownian motions with variances σ_u^2 and σ_v^2, respectively, B_w^μ is a demeaned Brownian motion (i.e. $B_w^\mu(s) = B_w(s) - \int_0^1 B_w(s)ds$), and henceforth $\sup_{t \geq R}$ is used to mean $\sup_{R \leq t \leq T-1}$. Also,

$$\sup_{t \geq R} \sqrt{R}|\widehat{\alpha}_{k,t} - \alpha_{k,0}| \leq \sup_{t \geq R} |\frac{1}{\sqrt{R}}\sum_{j=1}^{t} u_{k,j}| + \sup_{t \geq R} R|\widehat{\beta}_t - \beta_0| \sup_{t \geq R} \frac{1}{R^{3/2}}\sum_{j=1}^{t} |X_{k,j}|. \tag{11}$$

The second term on the RHS of (11) is $O_p(1)$ because of (10). In addition, the first term is $O_p(1)$, given the functional central limit theorem (FCLT) for strong mixing processes.

(ii) From part (i), we have that $\sup_{t \geq R} R^{1-\gamma}|\widehat{\beta}_t - \beta_0| \overset{pr}{\to} 0$, $\sup_{t \geq R} \sqrt{R^{1-\gamma}}|\widehat{\alpha}_t - \alpha_0| \overset{pr}{\to} 0$, $\forall \gamma > 0$. This implies that $R^{1-\gamma}|\widehat{\beta}_R - \beta_0| \overset{a.s.}{\to} 0$, $\sqrt{R^{1-\gamma}}|\widehat{\alpha}_R - \alpha_0| \overset{a.s.}{\to} 0$ (see Theorem 2.1.2 of Lukacs (1969)).

Proof of Lemma 2.2: Recall that $\theta = (\alpha, \beta)'$ $f_t(\theta) = f(\Delta Y_{t+1} + (Y_t - \beta X_t) - \alpha)$, and $f_t(\widehat{\theta}_t) = f(\Delta Y_{t+1} + (Y_t - \widehat{\beta}_t X_t) - \widehat{\alpha}_t) = f(\widehat{v}_{k,t+1})$. Taking a second order mean expansion of $f(\widehat{\theta}_t)$ around θ_0 yields,

$$P^{-1/2}\sum_{t=R}^{T-1} f_t(\widehat{\theta}_t) = P^{-1/2}\sum_{t=R}^{T-1} f_t(\theta_0) - P^{-3/2}\sum_{t=R}^{T-1} f_t'(\theta_0)X_t P(\widehat{\beta}_t - \beta_0)$$

$$-P^{-1/2}\sum_{t=R}^{T-1} f_t'(\theta_0)(\widehat{\alpha}_{k,t} - \alpha_{k,0}) - P^{-1/2}\sum_{t=R}^{T-1} \frac{\partial^2 f_t}{\partial \beta^2}|_{\overline{\theta}_t}(\widehat{\beta}_t - \beta_0)^2$$

$$-P^{-1/2}\sum_{t=R}^{T-1} \frac{\partial^2 f_t}{\partial \alpha^2}|_{\overline{\theta}_t}(\widehat{\alpha}_t - \alpha_0)^2 + P^{-1/2}\sum_{t=R}^{T-1} \frac{\partial^2 f_t}{\partial \alpha \partial \beta}|_{\overline{\theta}_t}(\widehat{\alpha}_t - \alpha_0)(\widehat{\beta}_t - \beta_0),$$

where $\overline{\theta}_t = (\overline{\alpha}_t, \overline{\beta}_t)'$, and $\overline{\alpha}_t \in (\widehat{\alpha}_t, \alpha_0)$ and $\overline{\beta}_t \in (\widehat{\beta}_t, \beta_0)$. From Lemma 2.1(i), note that $R\sup_{t \geq R} |\widehat{\beta}_t -$

19

$\beta_0| = O_p(1)$. Now for $\overline{\beta}_t \in (\widehat{\beta}_t, \beta_0)$,

$$\left| P^{-1/2} \sum_{t=R}^{T-1} \frac{\partial^2 f_t}{\partial \beta^2} |_{\overline{\theta}_t} (\widehat{\beta}_t - \beta_0)^2 \right| = \left| P^{-1/2} \sum_{t=R}^{T-1} f_t''(\overline{\theta}_t) X_t^2 (\widehat{\beta}_t - \beta_0)^2 \right|$$

$$\leq \sup_{t \geq R} R^2 (\widehat{\beta}_t - \beta_0)^2 \sup_{t \geq R} \left| P^{-1/2} f_t''(\overline{\theta}_t) \right| \sup_{t \geq R} \frac{T^2}{R^2} \frac{1}{T^2} \sum_{t=R}^{T-1} X_t^2 = O_p(1) o_{a.s.}(1)(1+\pi)^2 O_p(1) = o_p(1).$$

In fact,

$$\Pr\left(\sup_{t \geq R} \left| P^{-1/2} f_t''(\overline{\theta}_t) \right| > \eta \right) \leq \sum_{t=R}^{T-1} \Pr\left(\left| P^{-1/2} f_t''(\overline{\theta}_t) \right| > \eta \right) \leq \frac{1}{\eta^s P^{s/2}} \sup_{\theta \in \Theta} E|f_t''(\theta)|^s,$$

and given A3, $\sup_{t \geq R} \left| P^{-1/2} f_t''(\overline{\theta}_t) \right| = o_{a.s.}(1)$ (by the Borel Cantelli lemma). Similarly,

$$\left| P^{-1/2} \sum_{t=R}^{T} \frac{\partial^2 f_{t+1}}{\partial \alpha^2} |_{\overline{\theta}_t} (\widehat{\alpha}_t - \alpha_0)^2 \right| \leq \sup_{t \geq R} P(\widehat{\alpha}_t - \alpha_0)^2 \sup_{t \geq R} \left| P^{-1/2} f_t''(\overline{\theta}_t) \right| = o_p(1)$$

and

$$P^{-1/2} \sum_{t=R}^{T-1} \frac{\partial^2 f_t}{\partial \alpha \partial \beta} |_{\overline{\theta}_t} |\widehat{\alpha}_t - \alpha_0| |\widehat{\beta}_t - \beta_0| \leq \sup_{t \geq R} R |\widehat{\beta}_t - \beta_0| \sup_{t \geq R} \sqrt{R} |\widehat{\alpha}_t - \alpha_0|$$

$$\times \sup_{t \geq R} \left| P^{-1/2} f_t''(\overline{\theta}_t) \right| \sup_{t \geq R} \frac{T^{3/2}}{R^{3/2}} \frac{1}{T^{3/2}} \sum_{t=R}^{T-1} |X_t| = o_p(1).$$

Proof of Proposition 2.3: Suppose that f is quadratic and recall that $\widehat{v}_{k,t+1} = v_{k,t+1} - (\widehat{\beta}_{k,t} - \beta_{k,0}) X_{k,t} - (\widehat{\alpha}_{k,t} - \alpha_{k,0})$. It follows that,

$$P^{-1/2} \sum_{t=R}^{T-1} \widehat{v}_{k,t+1}^2 = P^{-1/2} \sum_{t=R}^{T-1} v_{k,t+1}^2 + P^{-1/2} \sum_{t=R}^{T-1} (\widehat{\beta}_{k,t} - \beta_{k,0})^2 X_{k,t}^2 + P^{-1/2} \sum_{t=R}^{T-1} (\widehat{\alpha}_{k,t} - \alpha_{k,0})^2$$

$$+ 2P^{-1/2} \sum_{t=R}^{T-1} (\widehat{\beta}_{k,t} - \beta_{k,0})(\widehat{\alpha}_{k,t} - \alpha_{k,0}) X_{k,t}$$

$$- 2P^{-1/2} \sum_{t=R}^{T-1} (\widehat{\beta}_{k,t} - \beta_{k,0}) v_{k,t+1} X_{k,t} - 2P^{-1/2} \sum_{t=R}^{T-1} (\widehat{\alpha}_{k,t} - \alpha_{k,0}) v_{k,t+1}. \qquad (12)$$

We want to show that

$$P^{-1/2} \sum_{t=R}^{T-1} \widehat{v}_{k,t+1}^2 = P^{-1/2} \sum_{t=R}^{T-1} v_{k,t+1}^2 + o_p(1), \forall k. \qquad (13)$$

20

Given A1-A2, Lemma 2.1 and 2.2 ensure that the second, third and fourth terms on the RHS of (12) are $o_p(1)$. Thus, it remains to show that the fifth and sixth terms on the RHS of (12) are $o_p(1)$. For notational simplicity, and when there is no ambiguity, we omit the subscript k. Along the lines of Hansen (1992), we approximate the strong mixing process, v_t, by the sum of a martingale difference sequence (mds) plus a remainder term. Let $F_i = \sigma(u_1, \ldots u_i, w_{01}, \ldots w_{0i})$, $E_i(\cdot) = E(\cdot|F_i)$ and,

$$\varepsilon_i = \sum_{j=0}^{\infty} \left(E_t v_{t+j} - E_{t-1} v_{t+j} \right), z_t = \sum_{j=1}^{\infty} E_t v_{t+j}, \tag{14}$$

so that $v_t = \varepsilon_t + (z_{t-1} - z_t)$. Thus,

$$P^{-1/2} \sum_{t=R}^{T-1} (\widehat{\beta}_t - \beta_0) v_{t+1} X_t = P^{-1/2} \sum_{t=R}^{T-1} (\widehat{\beta}_t - \beta_0) X_t \varepsilon_{t+1} + P^{-1/2} \sum_{t=R}^{T-1} (\widehat{\beta}_t - \beta_0) X_t (z_t - z_{t+1}). \tag{15}$$

We begin by showing that the first term on the RHS of (15) is $o_p(1)$. As ε_t is an F_t-adapted martingale difference sequence, $E(X_t \varepsilon_{t+1}) = E(X_t E_t \varepsilon_{t+1}) = 0$, and for $s < t$, $E(X_t \varepsilon_{t+1} X_s \varepsilon_{s+1}) = E(X_t X_s \varepsilon_{s+1} E_t \varepsilon_{t+1}) = 0$. Given A4, and for some constants, Δ and Δ', application of the Chebyshev and Hölder inequalities yields that,

$$\Pr\left[\left| P^{-1/2} \sum_{t=R}^{T-1} (\widehat{\beta}_t - \beta_0) X_t \varepsilon_{t+1} \right| > \eta \right] \leq \frac{1}{\eta^2} Var\left(P^{-1/2} \sum_{t=R}^{T-1} (\widehat{\beta}_t - \beta_0) X_t \varepsilon_{t+1} \right)$$

$$\leq \frac{1}{\eta^2} \left(P^{-1} \sum_{t=R}^{T-1} E(X_t^2 (\widehat{\beta}_t - \beta_0)^2 E_t \varepsilon_{t+1}^2) \right)$$

$$\leq \frac{1}{\eta^2} \frac{1}{P} \sum_{t=R}^{T-1} \left(E\left(|X_t|^{2.1} |\widehat{\beta}_t - \beta_0|^{2.1} \right) \right)^{20/21} (E((E_t \varepsilon_{t+1}^2)^{21}))^{1/21}$$

$$\leq \frac{\Delta}{\eta^2} \frac{1}{P} \sum_{t=R}^{T-1} (E(|X_t|^7)^{2/7} (E(|\widehat{\beta}_t - \beta_0|^3))^{2/3} \leq \frac{\Delta'}{\eta^2} \frac{1}{T^{1/4}}.$$

(We show in Lemma A below that $\max_{R \leq t \leq T} E(|X_t|^7) = O(T^{7/2})$ and $\max_{R \leq t \leq T} E(|\widehat{\beta}_t - \beta_0|^3) = O(T^{-15/8})$.) By a similar argument it follows that $P^{-1/2} \sum_{t=R}^{T-1} (\widehat{\alpha}_{k,t} - \alpha_{k,0}) \varepsilon_{t+1} = o_p(1)$. We now need to show that the second term on the RHS of 15 is $o_p(1)$. Now,

$$P^{-1/2} \sum_{t=R}^{T-1} (\widehat{\beta}_t - \beta_0) X_t (z_t - z_{t+1}) = P^{-1/2} \sum_{t=R}^{T-1} (X_t - X_{t-1}) z_t (\widehat{\beta}_t - \beta_0)$$

$$+P^{-1/2}\sum_{t=R}^{T-1}X_{t-1}z_t\left((\widehat{\beta}_t-\beta_0)-(\widehat{\beta}_{t-1}-\beta_0)\right)-P^{-1/2}X_{T-1}z_T(\widehat{\beta}_{T-1}-\beta_0). \tag{16}$$

Thus, it suffices to show that all the terms on the RHS of (16) are $o_p(1)$. Consider first,

$$P^{-1/2}|\sum_{t=R}^{T-1}(X_t-X_{t-1})z_t(\widehat{\beta}_t-\beta_0)|=P^{-1/2}|\sum_{t=R}^{T-1}w_tz_t(\widehat{\beta}_t-\beta_0)|\leq\sup_{t\geq R}|\widehat{\beta}_t-\beta_0|P^{-1/2}\sum_{t=R}^{T}|w_tz_t|,$$

where $w_t=\frac{\beta_{0,0}}{\beta_{k,0}}w_{0,t}+\frac{1}{\beta_{k,0}}(\Delta u_{0,t}-\Delta u_{k,t})$. By Lemma 2.1(i), it suffices to show that $P^{-3/2}\sum_{t=R}^{T}|w_tz_t|=o_p(1)$. Given A1, the strong law of large numbers ensures that $P^{-1}\sum_{t=R}^{T}|w_t|=O_{a.s.}(1)$, and thus,

$$P^{-3/2}\sum_{t=R}^{T}|w_tz_t|\leq P^{-1/2}\sup_{t\geq R}|z_t|P^{-1}\sum_{t=R}^{T}|w_t|\leq P^{-1/2}\sup_{t\geq R}|z_t|O_{a.s.}(1).$$

Recalling that $z_t=\sum_{k=1}^{\infty}E_tv_{t+k}$, the McLeish (1975) mixing inequalities imply that

$$\Pr\left(\sup_{t\geq R}\frac{|z_t|}{P^{1/2}}>\eta\right)\leq\sum_{t=R}^{T}\Pr\left(\frac{|z_t|}{P^{1/2}}>\eta\right)\leq\frac{1}{\eta^{\psi}P^{\psi/2-1}}\left(C\sum_{k=1}^{\infty}\alpha_k^{1/\psi-1/q}\right)^{\psi}\|v_t\|_q,$$

so that $P^{-1/2}\sup_{t\geq R}|z_t|=o_{a.s.}(1)$, as $\psi>4$, and from application of the first Borel Cantelli Lemma. Now, we show that the second term on the RHS of (16) is $o_p(1)$. After some simple manipulation, note that,

$$\left((\widehat{\beta}_t-\beta_0)-(\widehat{\beta}_{t-1}-\beta_0)\right)=\frac{X_tu_t}{\sum_{j=1}^{t}X_j^2}-\frac{X_t^2\sum_{j=1}^{t-1}X_ju_j}{\sum_{j=1}^{t}X_j^2\sum_{j=1}^{t-1}X_j^2}. \tag{17}$$

In addition, the second term on the RHS of (16) is majorized by

$$\left(\sup_{t\geq R}\left|\frac{X_tu_t}{\sum_{j=1}^{R}X_j^2}\right|+\sup_{t\geq R}\left|\frac{X_t^2\sum_{j=1}^{t-1}X_ju_j}{(\sum_{j=1}^{R}X_j^2)^2}\right|\right)\times\frac{1}{P^{1/2}}\sum_{t=R}^{T}|X_{t-1}z_t|.$$

Now, $\frac{1}{T^2}\sum_{j=1}^{R}X_j^2\overset{d}{\to}\frac{1}{(1+\pi)^2}\int_0^{\frac{1}{1+\pi}}B_u(s)ds$, so that,

$$\sup_{t\geq R}\left|\frac{X_tu_t}{T^2}\right|\leq\frac{1}{T^{5/4}}\sup_{t\geq R}\left|\frac{X_t}{T^{1/2}}\right|\sup_{t\geq R}\left|\frac{u_t}{T^{1/4}}\right|=O(T^{-5/4})O_p(1)o_p(1),$$

given that $\Pr\left[\sup_{t\geq R}\frac{|u_t|}{T^{1/4}}>\eta\right]\leq\sum_{t=R+1}^{P}\Pr\left[\frac{|u_t|}{T^{1/4}}>\eta\right]\leq\frac{P}{T^{\frac{\psi}{4}}}\|u_t\|_{\psi}$, with $\psi>4$. In addition,

$$\sup_{t\geq R}\left|\frac{X_t^2\sum_{j=1}^{t-1}X_ju_j}{T^4}\right|\leq\sup_{t\geq R}\left|\frac{X_t^2}{T}\right|\sup_{t\geq R}\left|\frac{X_t}{T^{1/2}}\right|\sup_{t\geq R}\frac{\sum_{j=1}^{t-1}|u_j|}{T}\times\frac{1}{T^{3/2}}=O_p(T^{-3/2}).$$

Finally,

$$\frac{1}{T^{5/4}P^{1/2}}\sum_{t=R}^{T}|X_{t-1}z_t|\leq\sup_{t\geq R}\left|\frac{X_t}{T^{1/2}}\right|\frac{1}{T^{3/4}P^{1/2}}\sum_{t=R}^{T-1}|z_t|=O_p(1)o_p(1),$$

as z_t is a $\psi/2$-mixingale. Thus, the second term on the RHS of (16) is $o_p(1)$ by the law of large numbers for mixingales. Now, consider the third term on the RHS of (16), which can be written as,

$$|(\widehat{\beta}_{T-1} - \beta_0)\frac{1}{P^{1/2}}X_{T-1}z_T| \leq R|(\widehat{\beta}_{T-1} - \beta_0)|\frac{T}{R}\left|\frac{X_t}{T}\right|\left|\frac{z_T}{P^{1/2}}\right| = O_p(1)O(1)o_p(1) = o_p(1).$$

Now,

$$P^{-1/2}\sum_{t=R}^{T-1}(\widehat{\alpha}_t - \alpha_0)(z_t - z_{t+1}) = P^{-1/2}\sum_{t=R}^{T-1}z_t\left((\widehat{\alpha}_t - \alpha_0) - (\widehat{\alpha}_{t-1} - \alpha_0)\right) - P^{-1/2}z_T(\widehat{\alpha}_{T-1} - \alpha_0). \tag{18}$$

It follows immediately that the lead term on the RHS of (18) is $o_p(1)$, and the first term can be written as

$$P^{-1/2}\sum_{t=R}^{T-1}z_t\left(\frac{1}{t}u_t - \frac{1}{t(t-1)}\sum_{j=1}^{t-1}u_j + \left((\widehat{\beta}_t - \beta_0) - (\widehat{\beta}_{t-1} - \beta_0)\right)\left(\frac{1}{t}X_t - \frac{1}{t(t-1)}\sum_{j=1}^{t-1}X_j)\right),$$

which is $o_p(1)$. We now move to the case of a generic loss function, f, which satisfies A3 and A4 (and $\pi = 0$). From Lemma 2.2 it suffices to show that $P^{-3/2}\sum_{t=R}^{T-1}f'_t(\theta_0)X_tP(\widehat{\beta}_t - \beta_0) = o_p(1)$ and $P^{-1/2}\sum_{t=R}^{T-1}f'_t(\theta_0)(\widehat{\alpha}_{k,t} - \alpha_{k,0}) = o_p(1)$ for $\frac{P}{R} \to \pi = 0$. In fact,

$$P^{-1/2}\left|\sum_{t=R}^{T-1}f'_t(\theta_0)X_tP(\widehat{\beta}_t - \beta_0)\right| \leq \sup_{t\geq R}\sqrt{P}\sqrt{R}|\widehat{\beta}_t - \beta_0|\sup_{t\geq R}|\frac{X_t}{\sqrt{R}}|P^{-1}\sum_{t=R}^{T-1}|f'_t(\theta_0)| = o_p(1) \text{ and}$$

$$P^{-1/2}\left|\sum_{t=R}^{T-1}f'_t(\theta_0)(\widehat{\alpha}_{k,t} - \alpha_{k,0})\right| \leq \sup_{t\geq R}\sqrt{P}|\widehat{\alpha}_t - \alpha_0|P^{-1}\sum_{t=R}^{T-1}|f'_t(\theta_0)| = o_p(1),$$

for $P/R \to 0$. Given A1-A4, $\widehat{\sigma}_P - \widetilde{\sigma}_P = o_p(1)$, where

$$\widetilde{\sigma}_P^2 = \frac{1}{P}\sum_{t=R-h+1}^{T-1}(f(v_{0,t+h})-f(v_{k,t+h}))^2+\frac{2}{P}\sum_{j=1}^{l_P}w_j\sum_{t=R-h+1+j}^{T-1}(f(v_{0,t+h})-f(v_{k,t+h}))(f(v_{0,t+h-j})-f(v_{k,t+h-j})),$$

for $l_p = o(P^{1/4})$. Now, $\widetilde{\sigma}_P - \sigma_P = o_p(1)$ by the same argument as in Newey and West (1997). Finally, $\forall k = 1, 2, \ldots l$, models k and 0 are nonnested, so that the desired result follows directly as an application of the central limit theorem.

Lemma A: Under the assumptions of Proposition 2.3, (i) $\max_{R\leq t\leq T} E(|X_t|^7) = O(T^{7/2})$ and (ii) $\max_{R\leq t\leq T} E(|\widehat{\beta}_t - \beta_0|^3) = O(T^{-15/8})$.

Proof of Lemma A: Hereafter, C denotes a generic constant, which differs across cases considered.

(i) Recall that $X_t = \sum_{j=1}^{t} w_j$, where $w_t = \frac{\beta_{0,0}}{\beta_{k,0}}w_{0,t} + \frac{1}{\beta_{k,0}}(\Delta u_{0,t} - \Delta u_{k,t})$. By McLeish's (1975) inequality (see e.g. Davidson (1994, Lemma 16.8)),

$$\max_{R\leq t\leq T} E(|X_t|^7) \leq E(\max_{R\leq t\leq T}|X_t|^7) \leq \left(\frac{7}{6}\right)^7\left(\sum_{k=0}^{\infty}\alpha_k\right)^6\sum_{k=0}^{\infty}\alpha_k^6 E(|Y_{t,k}|^7),$$

where α_k denotes the mixing coefficients (as in Assumption **A1**) and

$$Y_{t,k} = \sum_{j=1}^{t} \left(E_{j-k} w_j - E_{j-k-1} w_j \right).$$

By Burkholder's inequality (Burkholder, 1973, p.23) and Love's c_r inequality (see e.g. Davidson (1994, p.140)),

$$
\begin{aligned}
E(|Y_{t,k}|^7) &\leq CE | \sum_{j=1}^{t} \left(E_{j-k} w_j - E_{j-k-1} w_j \right)^2 |^{7/2} \\
&\leq Ct^{5/2} \sum_{j=1}^{t} E(|E_{j-k} w_j - E_{j-k-1} w_j|)^7.
\end{aligned}
$$

Now, by Minkowski's inequality and McLeish's mixing inequalities (McLeish (1975, Lemma 3.5)),

$$\| E_{j-k} w_j - E_{j-k-1} w_j \|_7 \leq C \alpha_k^{\frac{\delta}{7(7+\delta)}} \| w_j \|_{7+\delta},$$

where $\|u\|_p = (E(|u|^p))^{1/p}$, so that from **A1** we see that,

$$E(|Y_{t,k}|^7) \leq Ct^{7/2} \alpha_k^{\frac{7\delta}{7(7+\delta)}} \quad \text{and} \quad \max_{R \leq t \leq T} E(|X_t|^7) \leq CT^{7/2} \sum_{k=0}^{\infty} \alpha_k^{\frac{42\delta}{7(7+\delta)}}.$$

The statement in (i) follows, as $\sum_{k=0}^{\infty} \alpha_k^{\frac{7\delta}{7(7+\delta)}} < \infty$ (given **A1**).

(ii) Note that

$$\max_{R \leq t \leq T} E(|\widehat{\beta}_t - \beta_0|^3) \leq \left(E \left(\frac{1}{(T^{-2} \sum_{j=1}^{R} (X_j - \overline{X}_R)^2)^{12}} \right) \right)^{1/4} \left(E \left(T^{-8} \left(\sum_{j=1}^{t} (X_j - \overline{X}_t) u_j \right)^4 \right) \right)^{3/4}.$$

For any given T, $R \, T^{-2} \sum_{j=1}^{R} (X_j - \overline{X}_R)^2$ is strictly positive with probability one, as it is a continuous random variable. Thus, $E \left(\frac{1}{(T^{-2} \sum_{j=1}^{R} (X_j - \overline{X}_R)^2)^{13}} \right) < \infty$, and so by Theorem 3.4.1 in Amemya (1985),

$$\lim_{T \to \infty} E \left(\frac{1}{(T^{-2} \sum_{j=1}^{R} (X_j - \overline{X}_R)^2)^{12}} \right) = E \left(\lim_{T \to \infty} \left(\frac{1}{(T^{-2} \sum_{j=1}^{R} (X_j - \overline{X}_R)^2)^{12}} \right) \right) < \infty.$$

In fact, as shown below in the proof of Proposition 2.5, $\lim_{T \to \infty} T^{-2} \sum_{j=1}^{R} (X_j - \overline{X}_R)^2 > 0$ almost surely. Thus, it suffices to show that

$$E \left(T^{-8} \left(\sum_{j=1}^{T} X_j u_j \right)^4 \right) = O(T^{-5/2}), \tag{19}$$

as $E\left(T^{-8}\left(\sum_{j=1}^{T}\overline{X}_T u_j\right)^4\right)$ and the cross terms are of the same order of magnitude as (19). Hereafter, $\simeq$ means "of the same order of magnitude". Now,

$$E\left(\left(\sum_{j=1}^{T}X_j u_j\right)^4\right) \simeq \sum_{j=1}^{T}E\left(X_j^4 u_j^4\right) + \sum_{j=1}^{T}\sum_{s>j}E(X_j^2 u_j^2 X_s^2 u_s^2)$$

$$+\sum_{j=1}^{T}\sum_{s>j}E(X_j^3 u_j^3 X_s u_s) + \sum_{j=1}^{T}\sum_{s>j}\sum_{i>j}\sum_{l>i}E(X_j u_j X_s u_s X_i u_i X_l u_l). \tag{20}$$

We thus need to show that all of the terms on the RHS of (20) are $O(T^{11/2})$. Given the Hölder inequality it follows that the second term is majorized by

$$\sum_{j=1}^{T}\sum_{s>j}\left(E(X_j^4 u_j^4)\right)^{1/2}\left(E(X_s^4 u_s^4)\right)^{1/2} \leq T^2 \max_{t\leq T}\left(E(X_j^4 u_j^4)\right)$$

$$\leq T^2 (E(u_j^8))^{1/2}\max_{t\leq T}\left(E(X_j^8)\right)^{1/2} \leq CT^4,$$

as $\max_{t\leq T}\left(E(X_j^8)\right) = O(T^4)$ by the same argument used in part (i). By the same argument the first and third terms are $O(T^3)$ and $O(T^4)$, respectively. For the last term, it suffices to show that

$$\sum_{j=1}^{T}\sum_{s>j}\sum_{i>j}\sum_{l>i}E\left(\frac{X_j}{T^{1/2}}u_j\frac{X_s}{T^{1/2}}u_s\frac{X_i}{T^{1/2}}u_i X_l u_l E\left(X_l u_l|F_i\right)\right) = O(T^4), \tag{21}$$

where $F_i = \sigma(w_j, u_j, j = 1, \dots i)$. By Hölder's inequality the left hand side of (21) is majorized by

$$\sum_{j=1}^{T}\sum_{s>j}\sum_{i>j}\sum_{l>i}\left(E\left(\left|\frac{X_j}{T^{1/2}}u_j\frac{X_s}{T^{1/2}}u_s\frac{X_i}{T^{1/2}}\right|^{4/3}\right)\right)^{3/4}\left(E\left(E\left(X_l u_l|F_i\right)\right)^4\right)^{1/4}$$

$$\leq T^3 \max_{k\leq T}\left(E\left|\frac{X_k^4 u_k^4}{T^2}\right|\right)^{3/4}\max_{3\leq i<T}\sum_{l>i}\left(E\left(E\left(X_l u_l|F_i\right)\right)^4\right)^{1/4}$$

$$\leq CT^3 \max_{3\leq i<T}\sum_{l>i}\left(E\left(E\left(X_l u_l|F_i\right)\right)^4\right)^{1/4}, \tag{22}$$

as $\max_{k\leq T}\left(E\left|\frac{X_k^4 u_k^4}{T^2}\right|\right)^{3/4}$ is bounded, by the same argument used in the proof of part (i). Note that $X_l u_l = \sum_{j=1}^{l}w_j u_l$,

$$\max_{3\leq i<T}\sum_{l>i}\left(E\left(E\left(X_l u_l|F_i\right)\right)^4\right)^{1/4} = \max_{3\leq i<T}\sum_{l>i}\left(E\left(\sum_{j=1}^{l}E\left(w_j u_l|F_i\right)\right)^4\right)^{1/4}$$

25

and

$$E\left(\sum_{j=1}^{l} E\left(w_j u_l | F_i\right)\right)^4 \simeq \sum_{j=1}^{l} E\left(E\left(w_j u_l | F_i\right)\right)^4 + \sum_{j=1}^{l}\sum_{s>j}^{l} E\left(\left(E\left(w_j u_l | F_i\right)\right)^2\left(E\left(w_s u_l | F_i\right)\right)^2\right)$$

$$+ \sum_{j=1}^{l}\sum_{s>j}^{l} E\left(\left(E\left(w_j u_l | F_i\right)\right)^3 E\left(w_s u_l | F_i\right)\right)$$

$$+ \sum_{j=1}^{l}\sum_{s>j}^{l}\sum_{h>s}^{l}\sum_{m>h}^{l} E\left(E\left(w_j u_l | F_i\right) E\left(w_s u_l | F_i\right) E\left(w_h u_l | F_i\right) E\left(w_m u_l | F_i\right)\right). \tag{23}$$

We shall show that $\max_{3\leq i<T}\sum_{l>i}^{T}$ of the last term in (23) is less than or equal to CT. As the first, second and third terms in (23) are at most of the same order of magnitude as the last one, it also follows that $\max_{3\leq i<T}\sum_{l>i}^{T}$ of these three terms is less than or equal to CT. This ensures that the right hand side of (22) is majorized by CT^4, thus ensuring that the statement in (21), and hence the desired outcome, follows. Now, by Hölder's inequality, the last term above is majorized by

$$\sum_{j=1}^{l}\sum_{s>j}^{l}\sum_{h>s}^{l}\sum_{m>h}^{l}\left(E(|E\left(w_j u_l | F_i\right)|^2|E\left(w_s u_l | F_i\right)|^2)\right)^{1/2}\left(E(|E\left(w_h u_l | F_i\right)|^2|E\left(w_m u_l | F_i\right)|^2)\right)^{1/2} \leq$$

$$\sum_{j=1}^{l}\left(E(|E\left(w_j u_l | F_i\right)|^4)\right)^{1/4}\sum_{s>j}^{l}\left(E(|E\left(w_s u_l | F_i\right)|^4)\right)^{1/4}\sum_{h>s}^{l}\left(E(|E\left(w_h u_l | F_i\right)|^4)\right)^{1/4}\sum_{m>h}^{l}\left(E(|E\left(w_m u_l | F_i\right)|^4)\right)^{1/4}. \tag{24}$$

We shall show that $\lim_{T\to\infty}\max_{3\leq i<T}\frac{1}{T}\sum_{l>i}^{T}$ of all of the summations in (24) are bounded. Recall that $l > i$,

$$\sum_{j=1}^{l}\left(E(|E\left(w_j u_l | F_i\right)|^4)\right)^{1/4} = \sum_{j=1}^{i}\left(E(|w_j E\left(u_l | F_i\right)|^4)\right)^{1/4} + \sum_{j=i+1}^{l}\left(E(|E\left(w_j u_l | F_i\right)|^4)\right)^{1/4}$$

$$\sum_{j=1}^{i}\left(E(|w_j E\left(u_l | F_i\right)|^4)\right)^{1/4} \leq \sum_{j=1}^{i}\left(E(w_j^8)\right)^{1/8} E\left(\left(E\left(u_l | F_i\right)\right)^8\right)^{1/8} \leq C\sum_{j=0}^{i}||w_j||_8 \alpha_{l-i}^{\frac{\delta}{8(8+\delta)}}||u_j||_{8+\delta},$$

and, given the moment and size conditions in **A1**,

$$\lim_{T\to\infty}\max_{3<i<T}\frac{1}{T}\sum_{l>i}^{T}\sum_{j=0}^{i}||w_j||_8 \alpha_{l-i}^{\frac{\delta}{8(8+\delta)}}||u_j||_{8+\delta}$$

$$\leq \sup_{j}(||w_j||_8 ||u_j||_{8+\delta})\lim_{T\to\infty}\max_{3<i<T}\frac{1}{T}\sum_{l>i}^{T}\frac{i}{(l-i)^{1+\eta}} \leq C \text{ as } \eta > 0.$$

26

Also, by Minkowski's inequality,

$$\sum_{j=i+1}^{l} (E(|E(w_j u_l|F_i)|^4))^{1/4} \leq \sum_{j=i+1}^{l} (E(E((w_j u_l - E(w_j u_l))|F_i))^4)^{1/4} + \sum_{j=i+1}^{l} |E(w_j u_l)|$$

and

$$\sum_{j=i+1}^{l} (E(E((w_j u_l - E(w_j u_l))|F_i))^4)^{1/4} \leq C \|w_j u_l\|_4 \sum_{k=1}^{l-i} \alpha_k^{\frac{\delta}{8(8+\delta)}} \leq C. \tag{25}$$

By Corollary 6.16 in White (1984),

$$\sum_{j=i+1}^{l} |E(w_j u_l)| \leq Var(w_j) \|u_l\|_{2(2+\delta)} \sum_{j=i+1}^{l} \alpha_{l-j}^{\frac{\delta}{2(2+\delta)}} \leq C. \tag{26}$$

Thus, $\max_{3 \leq i < T} \sum_{l>i}^{T}$ of the right hand sides of (25) and (26) are $\leq CT$.

Proof of Proposition 2.4: (i) Note that

$$P^{-1/2} \sum_{t=R}^{T} f(\tilde{v}_{k,t+1}) = P^{-1/2} \sum_{t=R}^{T} f(v_{k,t+1}) + P^{-1} \sum_{t=R}^{T} E(f'_{v_k}) P^{-1/2} X_{k,t} P(\tilde{\beta}_{k,t} - \beta_{k,0}) + o_p(1),$$

as $\left| P^{-1/2} \sum_{t=R}^{T} (f'(v_{k,t+1}) - E(f'_{v_k})) X_{k,t}(\tilde{\beta}_{k,t} - \beta_{k,0}) \right| = o_p(1)$, under A1-A4, by the same argument used in the proof of Proposition 2.3.

(ii) By Theorem 3.1 in Phillips (1991) and the continuous mapping theorem,

$$\frac{T^{1/2}}{P^{1/2}} \frac{1}{T} \sum_{t=R}^{T-1} \frac{X_{k,t}}{T^{1/2}} T(\tilde{\beta}_{k,t} - \beta_{k,0}) \Rightarrow \sqrt{\frac{1+\pi}{\pi}} \int_{\frac{1}{1+\pi}}^{1} B_{u_k}(s) \left(\frac{\int_0^s B_{u_k}(r) B_{w.u_k}(r)}{\int_0^s B_{u_k}^2(r) dr} \right) ds.$$

Proof of Lemma 2.5: As above, the subscript k is dropped for notational simplicity. Note that,

$$\begin{aligned} P^{-1/2} \sum_{t=R}^{T-1} f_t(\hat{\theta}_t) =\ & P^{-1/2} \sum_{t=R}^{T-1} f_t(\theta_0) - P^{-1/2} \sum_{t=R}^{T-1} (f'_t(\theta_0) - E(f'_t(\theta_0))) X_t(\hat{\beta}_t - \beta_0) \\ & - P^{-1/2} \sum_{t=R}^{T-1} E(f'_t(\theta_0)) X_t(\hat{\beta}_t - \beta_0) - P^{-1/2} \sum_{t=R}^{T-1} (f'_t(\theta_0) - E(f'_t(\theta_0)))(\hat{\alpha}_t - \alpha_0) \\ & - P^{-1/2} \sum_{t=R}^{T-1} E(f'_t(\theta_0))(\hat{\alpha}_t - \alpha_0) + P^{-1/2} \sum_{t=R}^{T-1} \frac{\partial f_t^2}{\partial \beta^2} |_{\bar{\theta}_t} (\hat{\beta}_t - \beta_0)^2 \\ & + P^{-1/2} \sum_{t=R}^{T-1} \frac{\partial f_t^2}{\partial \alpha^2} |_{\bar{\theta}_t} (\hat{\alpha}_t - \alpha_0)^2 + P^{-1/2} \sum_{t=R}^{T-1} \frac{\partial f_t^2}{\partial \alpha \partial \beta} |_{\bar{\theta}_t} (\hat{\alpha}_t - \alpha_0)(\hat{\beta}_t - \beta_0). \end{aligned} \tag{27}$$

It suffices to show that, with the exception of the first one, all of the terms on the RHS of (27) are $o_{a.s.}(1)$. Note that the second term can be written as

$$\left| P^{-1/2} \sum_{t=R}^{T-1} (f'_t(\theta_0) - E(f'_t(\theta_0))X_t(\widehat{\beta}_t - \beta_0) \right|$$

$$\leq \sup_{t \geq R} \frac{|X_t|}{\sqrt{T \log \log T}} P^{1/2} \sqrt{T \log \log T} \sup_{t \geq R} \left| \widehat{\beta}_t - \beta_0 \right| \frac{1}{P} \sum_{t=R}^{T-1} |f'_t(\theta_0) - E(f'_t(\theta_0))| .$$

Given **A1** and of the strong law of large numbers, $\frac{1}{P} \sum_{t=R}^{T-1} |f'_t(\theta_0) - E(f'_t(\theta_0))| = O_{a.s.}(1)$, and by the strong invariance principle for mixing processes (Eberlain (1986, Theorem 1)), $\sup_{t \geq R} \frac{|X_t|}{\sqrt{T \log \log T}} = O_{a.s.}(1)$. It remains to show that

$$P^{1/2} \sqrt{T \log \log T} \sup_{t \geq R} \left| \widehat{\beta}_t - \beta_0 \right| = o_{a.s.}(1).$$

Now,

$$P^{1/2} \sqrt{T \log \log T} \sup_{t \geq R} \left| (\widehat{\beta}_t - \widehat{\beta}_R) + (\widehat{\beta}_R - \beta_0) \right|$$

$$\leq P^{1/2} \sqrt{T \log \log T} \sup_{t \geq R} |\widehat{\beta}_t - \widehat{\beta}_R| + P^{1/2} \sqrt{T \log \log T} \sup_{t \geq R} |\widehat{\beta}_R - \widehat{\beta}_0|. \qquad (28)$$

The second term term on the RHS of (28) is $o_{a.s.}(1)$ for $P = o(T^\gamma)$, $\gamma < 1$, because of Lemma 2.1(ii). The first term on the RHS of (28) can be written as,

$$P^{1/2} \sqrt{2T \log \log T} \sup_{t \geq R} |\widehat{\beta}_t - \widehat{\beta}_R|$$

$$= P^{1/2} \sqrt{2T \log \log T} \sup_{t \geq R} \left| \frac{\sum_{j=R+1}^t (X_j - \overline{X}_t) u_j}{\sum_{j=1}^t (X_j - \overline{X}_t)^2} - \frac{\sum_{j=R+1}^t (X_j - \overline{X}_t)^2 \sum_{j=1}^R (X_j - \overline{X}_R) u_j}{\sum_{j=1}^R (X_j - \overline{X}_R)^2 \sum_{j=1}^t (X_j - \overline{X}_t)^2} \right|.$$

We first need to show that $\lim_{T \to \infty} \frac{1}{T^2} \sum_{j=1}^R (X_j - \overline{X}_R)^2$ is almost surely positive. Write,

$$\frac{1}{T^2} \sum_{t=1}^R (X_t - \overline{X}_R)^2 = \frac{1}{T} \sum_{t=1}^R \frac{t}{T} \frac{(X_t - \overline{X}_R)^2}{t} = \frac{1}{T} \sum_{t=1}^R \frac{t}{T} \left(\frac{\sigma_u W_t^\mu}{\sqrt{t}} + \left(\frac{(X_t - \overline{X}_R)}{\sqrt{t}} - \frac{\sigma_u W_t^\mu}{\sqrt{t}} \right) \right)^2$$

$$= \frac{1}{T} \sum_{t=1}^R \frac{t}{T} \left(\frac{\sigma_u W_t^\mu}{\sqrt{t}} + O_{a.s.}(t^{-\vartheta}) \right)^2$$

$$= \frac{1}{T} \sum_{t=1}^R \frac{t}{T} \left(\frac{\sigma_u W_t^\mu}{\sqrt{t}} \right)^2 + \frac{1}{T} \sum_{t=1}^R O_{a.s.}(t^{-2\vartheta}) + \frac{1}{T} \sum_{t=1}^R \frac{t}{T} \frac{\sigma_u W_t^\mu}{\sqrt{t}} O_{a.s.}(t^{-\vartheta}), \qquad (29)$$

with $\vartheta > 0$. Also, the $O_{a.s.}$ terms above hold uniformly in t. In fact, A1 implies the assumptions of Theorem 2 in Eberlain (1986), according to which $\frac{(X_t - \overline{X}_R)}{\sqrt{t}} - \frac{\sigma_u W_t^\mu}{\sqrt{t}} = O_{a.s.}(t^{-\vartheta})$, for $\vartheta > 0$, and for

28

$W_t^\mu = W_t - R^{-1} \sum_{t=1}^{R} W_t$, with W_t a standard Brownian motion. Since, $\frac{W_t^\mu}{\sqrt{t}}$ is a continuous random variable, it follows that the first term on the RHS of (29) is strictly positive with probability one, given that the weighted average of an almost surely strictly positive term is almost surely strictly positive. As the second and third terms on the RHS of the last equality in (29) are almost surely zero, it follows that $\Pr(\lim_{T\to\infty} \frac{1}{T^2} \sum_{t=1}^{R}(X_t - \overline{X}_R)^2 > 0) = 1$. Also,

$$\frac{P^{1/2}\sqrt{2\log\log T}}{T^{3/2}} \left| \sum_{j=R+1}^{t} (X_j - \overline{X}_t)u_j \right| \leq \sup_{t\geq R} \frac{|X_t - \overline{X}_t|}{\sqrt{2T\log\log T}} \frac{2P^{3/2}\log\log T}{T} \frac{1}{P} \sum_{j=R+1}^{T-1} |u_j|$$

$$= O_{a.s.}(1)o_{a.s.}(1) = o_{a.s}(1),$$

as $P = o(T^{2/3}/\log\log T)$.

By a similar argument, $\sup_{t\geq R} P^{1/2}\sqrt{T\log\log T} \left| \frac{\sum_{j=R+1}^{t}(X_j-\overline{X}_t)^2 \sum_{j=1}^{R}(X_j-\overline{X}_R)u_j}{\sum_{j=1}^{R}(X_j-\overline{X}_R)^2 \sum_{j=1}^{t}(X_j-\overline{X}_t)^2} \right| = o_{a.s.}(1)$. Thus, the last term on the RHS of (28) is $o_{a.s.}(1)$. The third term on the RHS of (27) is $o_{a.s.}(1)$ by the same argument used above.

It remains to show that the fourth, fifth, sixth, seventh and eighth, terms on the RHS of (27) are $o_{a.s.}(1)$. Now, for the fourth and the fifth terms in (27), it suffices to show that

$$\sup_{t\geq R} P^{1/2}|\widehat{\alpha}_t - \alpha_0| = o_{a.s.}(1). \tag{30}$$

Note that

$$\sup_{t\geq R} P^{1/2}|\widehat{\alpha}_t - \alpha_0| \leq P^{1/2}|\widehat{\alpha}_R - \alpha_0| + \sup_{t\geq R} P^{1/2}|\widehat{\alpha}_t - \widehat{\alpha}_R|.$$

The first term on the RHS above is $o_{a.s.}(1)$ by Lemma 2.1(ii). In addition, note that

$$(\widehat{\alpha}_t - \widehat{\alpha}_R) = \frac{1}{t} \sum_{j=R+1}^{t} u_j + \frac{t-R}{tR} \sum_{j=1}^{R} u_j + (\widehat{\beta}_t - \widehat{\beta}_R)\frac{1}{t} \sum_{j=R+1}^{t} X_j - (\widehat{\beta}_t - \widehat{\beta}_R)\frac{t-R}{tR} \sum_{j=1}^{R} X_j,$$

so that

$$\sup_{t\geq R} P^{1/2}|\widehat{\alpha}_t - \alpha_R| \leq \sup_{t\geq R} \frac{P^{3/2}}{R} \frac{1}{P} \sum_{j=R}^{T-1} |u_j| + \frac{P^{3/2}}{R} \frac{1}{R} \sum_{j=1}^{R} |u_j|$$

$$+ P^{1/2}\sqrt{2T\log\log T} \sup_{t\geq R}|\widehat{\beta}_t - \widehat{\beta}_R| \sup_{t\geq R} |\frac{X_t}{\sqrt{2T\log\log T}}|$$

$$+ P^{1/2}\sqrt{2T\log\log T} \sup_{t\geq R}|\widehat{\beta}_t - \widehat{\beta}_R|\frac{P}{R} \sup_{1\leq t\leq R} |\frac{X_t}{\sqrt{2T\log\log T}}|. \tag{31}$$

For $P = o(T^{2/3}/\log\log T)$, the first two terms on the RHS of (31) are $o_{a.s.}(1)$, by the strong law of large numbers for $\alpha-$mixing sequences. The last two terms are $o_{a.s.}(1)$ because $P^{1/2}\sqrt{2T\log\log T} \sup_{t\geq R}|\widehat{\beta}_t -$

29

$\widehat{\beta}_R| = o_{a.s.}(1)$, and because of the strong invariance principle for mixing processes. This shows that the fourth and the fifth terms on the RHS of (27) are $o_{a.s.}(1)$. Now, the sixth term on the RHS of (27) is,

$$\left| P^{-1/2} \sum_{t=R}^{T-1} \frac{\partial^2 f_t}{\partial \beta^2} |_{\overline{\theta}_t} (\widehat{\beta}_t - \beta_0)^2 \right| = \left| P^{-1/2} \sum_{t=R}^{T-1} f_t''(\overline{\theta}_t) X_t^2 (\widehat{\beta}_t - \beta_0)^2 \right|$$

$$\leq \sup_{t \geq R} TP(2 \log \log T)(\widehat{\beta}_t - \beta_0)^2 \sup_{t \geq R} \left| P^{-1/2} f_t''(\overline{\theta}_t) \right| \frac{1}{TP(2 \log \log T)} \sum_{t=R}^{T-1} X_t^2$$

$$= o_{a.s.}(1) o_{a.s.}(1) O_{a.s.}(1),$$

which follows because $\frac{1}{TP(2 \log \log T)} \sum_{t=R}^{T-1} X_t^2 \leq \sup_{t \geq R} X_t^2 / (2T \log \log T) = O_{a.s.}(1)$. The middle term of the inequality is $o_{a.s.}(1)$ because of A3, and the first term is $o_{a.s.}(1)$ for $P = o(T^{2/3}/\log \log T)$, by the same argument used above. Finally, the seventh and eighth term are $o_{a.s.}(1)$ by the same argument used above.

Proof of Theorem 2.6: Hereafter, let S be the probability law of the sample, and let Q be the probability law of the pseudo-time series $f_t^*(\theta_0)$. We have that

$$P^{-1/2} \sum_{t=R}^{T-1} \left(f_t^*(\widehat{\theta}_t^*) - f_t(\widehat{\theta}_t) \right) = P^{-1/2} \sum_{t=R}^{T-1} (f_t^*(\theta_0) - f_t(\theta_0))$$

$$- P^{-1/2} \sum_{t=R}^{T-1} \left(f_t(\widehat{\theta}_t) - f_t(\theta_0) \right) + P^{-1/2} \sum_{t=R}^{T-1} \left(f_t^*(\widehat{\theta}_t^*) - f_t^*(\theta_0) \right) \qquad (32)$$

The desired result follows from Theorem A3 and Theorem 2.3 in White (2000), once we have shown that the second term on the RHS of (32) is $o_{a.s.-S}(1)$, the third term vanishes in probability$-Q$, $a.s - S$, and the first has the same limiting distribution as $P^{-1/2} \sum_{t=R}^{T-1} (f_t^*(\widehat{\theta}_t) - E(f))$. By Lemma 2.5, the second term on the RHS of (32) is $o_{a.s.-S}(1)$. Application of the mean value theorem to the third term yields,

$$P^{-1/2} \sum_{t=R}^{T-1} \left(f_t^*(\widehat{\theta}_t^*) - f_t^*(\theta_0) \right) = -P^{-1/2} \sum_{t=R}^{T-1} f_t^{*\prime}(\theta_0) X_t^* (\widehat{\beta}_t^* - \beta_0) - P^{-1/2} \sum_{t=R}^{T-1} f_t^{*\prime}(\theta_0)(\widehat{\alpha}_t^* - \alpha_0)$$

$$+ P^{-1/2} \sum_{t=R}^{T-1} \frac{\partial^2 f_t^*}{\partial \beta^2} |_{\overline{\theta}_t} (\widehat{\beta}_t^* - \beta_0)^2 + P^{-1/2} \sum_{t=R}^{T-1} \frac{\partial^2 f_t^*}{\partial \alpha^2} |_{\overline{\theta}_t} (\widehat{\alpha}_t^* - \alpha_0)^2$$

$$+ P^{-1/2} \sum_{t=R}^{T-1} \frac{\partial^2 f_t^*}{\partial \beta \partial \alpha} |_{\overline{\theta}_t} (\widehat{\alpha}_t^* - \alpha_0)(\widehat{\beta}_t^* - \beta_0). \qquad (33)$$

By construction,

$\widehat{v}^*_{k,t} = \Delta Y_{\tau(t)+1} + (Y_{\tau(t)} - \widehat{\beta}_{k,\tau(t)} X_{k,\tau(t)}) - \widehat{\alpha}_{k,\tau(t)} = v^*_{k,t} - (\widehat{\beta}_{k,\tau(t)} - \beta_{k,0}) X_{k,\tau(t)} - (\widehat{\alpha}_{k,\tau(t)} - \alpha_0)$, for any $k = 0, 1, \ldots, l$. From the proof of Lemma 2.5, recall that for all k, $P^{1/2}\sqrt{2T \log\log T}\sup_{t \geq R}(\widehat{\beta}_{k,t} - \beta_{k,0})$ and $P^{1/2}\sup_{t \geq R}(\widehat{\alpha}_{k,t} - \alpha_{k,0})$ approach zero $a.s. - S$. As $\widehat{\beta}^*_{k,t} = \widehat{\beta}_{k,\tau(t)}$, $\widehat{\alpha}^*_{k,t} = \widehat{\alpha}_{k,\tau(t)}$, where $\tau(t)$ (hereafter τ for notational simplicity) is a discrete uniform random variable over $R + 1, \ldots, T$, it follows that $P^{1/2}\sqrt{2T \log\log T}\sup_{t \geq R}(\widehat{\beta}^*_t - \beta_0)$ and $P^{1/2}\sup_{t \geq R}(\widehat{\alpha}^*_t - \alpha_0)$ approache zero $a.s. - Q$ for any sample realization, as $P \to \infty$. Also, $\sup_{\tau \in [R+1, \ldots T]} \frac{|X_\tau|}{\sqrt{2T\log\log T}} = O_{a.s.-Q}(1)$, almost surely$-S$. Thus, $\widehat{v}^*_{k,t} = v^*_{k,t} + o_{a.s.-Q}(1)$, for any k, $a.s. - S$, as $P \to \infty$. Now $v^*_{k,t} = v_{k,\tau}$. As v_t is a continuous function of u_t and $w_{0,t}$, it satisfies A1, and so is strictly stationary and strong mixing. Thus, by Lemma A1 in White (2000), $v^*_{k,t}$ is a strictly stationary strong mixing process, with mixing coefficients decaying at a geometric rate. In addition, $v^*_{k,t}$ satisfies A1. Given A3-A4 and the condition on the rate of growth of P stated in the theorem, it follows by the same argument used in the proof of Lemma 2.5 that as $P \to \infty$, the RHS of (33) approaches 0 in probability-Q, almost surely$-S$. Finally from Proposition 2.3, we know that $P^{-1/2}\sum_{t=R}^{T-1}(f_t(\widehat{\theta}_t) - E(f))$ is asymptotically normal. It then follows from Theorem A2 in White (2000) that the first term on the RHS of (32) has the same limiting distribution as $P^{-1/2}\sum_{t=R}^{T-1}(f_t(\widehat{\theta}_t) - E(f))$.

7 References

Amato, J. and N.R. Swanson, (2000), The Real Time Predictive Content of Money for Output, *Journal of Monetary Economics,* forthcoming.

Amemyia, T., (1985), *Advanced Econometrics,* Harvard University Press, Cambridge.

Burkholder, D.L., (1973), Distribution Function Inequalities for Martingales, *Annals of Probability,* 1, 19-42.

Chao, J.C., V. Corradi and N.R. Swanson (2000), An Out of Sample Test for Granger Causality, *Macroeconomic Dynamics*, forthcoming.

Christiano, L.J. and L. Ljungqvist, (1988), Money Does Granger-Cause Output in the Bivariate Money-Output Relation, *Journal of Monetary Economics*, 22, 217-235.

Christoffersen, P. and F.X. Diebold, (1996), Further Results on Forecasting and Model Selection Under Asymmetric Loss, *Journal of Applied Econometrics*, 11, 651-572.

Christoffersen, P. and F.X. Diebold, (1997), Optimal Prediction Under Asymmetric Loss, *Econometric Theory*, 13, 808-817.

Christoffersen, P. and F.X. Diebold, (1998), Cointegration and Long-Horizon Forecasting, *Journal of Business and Economic Statistics*, 16, 450-458.

Clark, T.E. (1999), Finite Sample Properties of Tests for Equal Forecasts Accuracy, *Journal of Forecasting,* 18, 489-504.

Clark, T.E. and M.W. McCracken, (2000), Tests for Equal Forecast Accuracy and Encompassing for Nested Models, *Journal of Econometrics,* forthcoming.

Clements, M.P. and D.F. Hendry, (1996), Intercept Correction and Structural Change, *Journal of Applied Econometrics,* 11, 127-146.

Clements, M.P. and D.F. Hendry, (1999a), *Forecasting Non-stationary Economic Time Series*: *The Zeuthen Lectures on Economic Forecasting,* MIT Press: Cambridge.

Clements, M.P. and D.F. Hendry, (1999b), On Winning Forecasting Competitions in Economics, *Spanish Economic Review*, 1, 123-160.

Clements, M.P. and D.F. Hendry, (2001), Forecasting with Difference and Trend Stationary Models, *The Econometrics Journal,* forthcoming.

Corradi, V., N.R. Swanson and C. Olivetti, (1999), Predictive Ability with Cointegrated Variables, Working Paper, Texas A&M University.

Davidson, J., (1994), *Stochastic Limit Theory,* Oxford University Press, Oxford.

De Jong, D., B.F. Ingram and C.H. Whiteman, (2000), A Bayesian Approach to Dynamic Macroeconomics, *Journal of Econometrics,* 98, 203-223.

Diebold, F.X. and C. Chen, (1996), Testing Structural Stability with Endogenous Breakpoint: A Size Comparison of Analytic and Bootstrap Procedures, *Journal of Econometrics*, 70, 221-241.

Diebold, F.X. and R.S. Mariano, (1995), Comparing Predictive Accuracy, *Journal of Business and Economic Statistics*, 13, 253-263.

Eberlain, E., (1986), On Strong Invariance Principles under Dependence Assumptions, *Annals of Probability*, 14, 260-270.

Friedman, B.M. and K. Kuttner, (1993), Another Look at The Evidence on the Money-Income Causality, *Journal of Econometrics*, 57, 189-203.

Gonzalo, J., (1995), Five Alternative Methods of Estimating Long-Run Equilibrium Relationships, *Journal of Econometrics*, 60, 203-233.

Gonzalo, J. and T.-H. Lee, (1998), Pitfalls in Testing for Long Run Relationships, *Journal of Econometrics*, 86, 129-154.

Geweke, J., (2000), Bayesian Econometrics and Forecasting, *Journal of Econometrics, 100, 11-15.*

Granger, C.W.J., (1969), Prediction with Generalized Cost of Error Functions, *Operational Research Quarterly*, 20, 451-468.

Granger, C.W.J., (1995), Modelling Nonlinear Relationships Between Extended-Memory Variables, *Econometrica*, 63, 265-279.

Granger, C.W.J. and N.R. Swanson, (1996), Further Developments in the Study of Cointegrated Variables, *Oxford Bulletin of Economics and Statistics*, 58, 537-553.

Hafer, R.W. and D.W. Jansen, (1991), The Demand for Money in the United States: Evidence from Cointegration Tests, *Journal of Money, Credit, and Banking*, 23, 155-168.

Hafer, R.W. and A.M. Kutan, (1997), More Evidence on the Money-Output Relationship, *Economic Inquiry*, 35, 48-58.

Hansen, B., (1992), Convergence to Stochastic Integrals for Dependent Heterogeneous Processes, *Econometric Theory*, 8, 489-500.

Johansen, S., (1988), Statistical Analysis of Cointegration Vectors, *Journal of Economic Dynamics and Control*, 12, 231-254.

Johansen, S., (1991), Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian

Vector Autoregressive Models, *Econometrica*, 59, 1551-1580.

Kuan, C.-M. and T. Liu, (1995), Forecasting Exchange Rates Using Feedforward and Recurrent Neural Networks, *Journal of Applied Econometrics*, 10, 347-364.

Lukacs, E., (1969), *Stochastic Convergence,* Academic Press, New York.

McCracken, M.W., (1998), Asymptotics for Out of Sample Tests of Causality, Working Paper, Louisiana State University.

McCracken, M.W., (2000), Robust Out of Sample Inference, *Journal of Econometrics,* 99, 195-223.

McLeish, D.L., (1975), A Maximal Inequality and Dependent Strong Laws, *Annals of Probability*, 3, 829-839.

Meese, R.A. and K. Rogoff, (1983), Empirical Exchange Rate Models of the Seventies: Do They Fit Out of Sample, *Journal of International Economics*, 14, 3-24.

Min, C.K. and A. Zellner, (1993), Bayesian and non-Bayesian Methods for Combining Models and Forecasts with Applications to Forecasting International Growth Rates, *Journal of Econometrics,* 56, 89-118.

Newey W.K. and K.D. West, (1987), A Simple, Positive Semi-Definite Heteroskedasticity and Autocorrelation Consistent Covariance Matrix, *Econometrica,* 55, 703-708.

Palm, F.C. and A. Zellner, (1992), To Combine or Not to Combine: Issues of Combining Forecasts, *Journal of Forecasting,* 11, 687-701.

Palm, F.C., (1995), Bayesian Model Selection and Prediction with Empirical Applications: Comments, *Journal of Econometrics,* 69, 333-336.

Phillips, P.C.B., (1991), Spectral Regression for Cointegrated Time Series, in W.A. Barnett, J. Powell, and G. Tauchen, eds. *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, 413-435, New York, Cambridge University Press.

Phillips, P.C.B., (1995), Bayesian Model Selection and Prediction with Empirical Applications, *Journal of Econometrics,* 69, 289-332.

Politis, D. and J. Romano, (1994), The Stationary Bootstrap, *Journal of the American Statistical Association*, 89, 1303-1313.

Rossi, B., (2000), Testing Out of Sample Predictive Ability with High Persistence. An Application to Models of Nominal Exchange Rate Determination. Manuscript, Princeton University.

Schorfheide, F., (2000), Loss Function Based Evaluation of DSGE, *Journal of Applied Econometrics,* forthcoming.

Sims, C.A. and T. Zha, (1998), Bayesian Methods for Dynamic Multivariate Models, *International Economic Review,* 39, 949-968.

Sims, C.A. and T. Zha, (1999), Error Bands for Impulse Responses, *Econometrica,* 67, 1113-1156.

Stock, J.H. and M.M. Watson, (1989), Interpreting the Evidence on Money−Income Causality, *Journal of Econometrics*, 40, 161-181.

Sullivan, R., A. Timmermann and H. White, (1999), Dangers of Data-Driven Inference: The Case of Calendar Effects in Stock Returns, *Journal of Finance,* 54, 1647-1691.

Swanson, N.R., (1998), Money and Output Viewed Through a Rolling Window, *Journal of Monetary Economics* 41, 455-474.

Swanson, N.R. and H. White, (1995), A Model Selection Approach to Assessing the Information in the Term Structure Using Linear Models and Artificial Neural Networks, *Journal of Business and Economic and Statistics*, 13, 265-275.

Thoma, M.A., (1994), Subsample Instability and Asymmetries in Money-Income Causality, *Journal of Econometrics*, 64, 279-306.

Waggoner, D.F. and T. Zha, (1999), Conditional Forecasts in Dynamic Multivariate Models, *Review of Economics and Statistics,* vol.LXXXI, 639-651.

Weiss, A.A., (1996), Estimating Time Series Models Using the Relevant Cost Function, *Journal of Applied Econometrics*, 11, 539-560.

West, K., (1988), Asymptotic Normality, When Regressors Have a Unit Root, *Econometrica*, 56, 1397-1417.

West, K., (1996), Asymptotic Inference About Predictive Ability, *Econometrica*, 64, 1067-1084.

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

White, H., (2000), A Reality Check for Data Snooping, *Econometrica*, 68, 1097-1126.

White, H., (1984), *Asymptotic Theory for Econometricians,* Academic Press, San Diego.

Zellner, A. and C.K. Min, (1999), Forecasting Turning Points in Countries' Output Growth Rates: A Response to Milton Friedman, *Journal of Econometrics,* 88, 203-206.

Table 1: Monte Carlo Results - DM Test: MSE Loss Function[*]

P	R					
	25	50	100	150	200	250

Empirical Size - Case a

P	25	50	100	150	200	250
25	0.114	0.086	0.106	0.102	0.076	0.076
50	0.088	0.092	0.062	0.076	0.082	0.040
100	0.086	0.076	0.078	0.066	0.048	0.044
150	0.072	0.078	0.062	0.068	0.058	0.054
200	0.064	0.066	0.060	0.060	0.048	0.050
250	0.060	0.064	0.040	0.052	0.062	0.054

Empirical Power - Case a

P	25	50	100	150	200	250
25	0.174	0.204	0.220	0.174	0.238	0.264
50	0.334	0.276	0.344	0.316	0.308	0.424
100	0.562	0.546	0.522	0.538	0.636	0.628
150	0.726	0.672	0.724	0.722	0.764	0.832
200	0.852	0.860	0.868	0.880	0.884	0.902
250	0.930	0.930	0.940	0.926	0.942	0.946

Empirical Size - Case b

P	25	50	100	150	200	250
25	0.112	0.082	0.100	0.098	0.078	0.078
50	0.092	0.094	0.066	0.078	0.088	0.044
100	0.084	0.072	0.072	0.062	0.056	0.044
150	0.074	0.060	0.062	0.062	0.056	0.050
200	0.060	0.068	0.066	0.062	0.044	0.050
250	0.068	0.066	0.046	0.046	0.056	0.054

Empirical Power - Case b

P	25	50	100	150	200	250
25	0.158	0.192	0.234	0.178	0.254	0.262
50	0.278	0.288	0.400	0.350	0.306	0.448
100	0.556	0.588	0.534	0.552	0.616	0.586
150	0.722	0.724	0.764	0.734	0.748	0.840
200	0.862	0.864	0.832	0.868	0.886	0.914
250	0.924	0.926	0.940	0.934	0.946	0.950

Empirical Size - Case c

P	25	50	100	150	200	250
25	0.054	0.078	0.066	0.066	0.064	0.084
50	0.040	0.054	0.046	0.044	0.064	0.052
100	0.028	0.046	0.026	0.060	0.038	0.050
150	0.026	0.028	0.028	0.044	0.034	0.052
200	0.016	0.034	0.024	0.034	0.046	0.044
250	0.020	0.026	0.026	0.032	0.056	0.076

Empirical Power - Case c

P	25	50	100	150	200	250
25	0.174	0.164	0.202	0.200	0.182	0.156
50	0.252	0.240	0.306	0.258	0.234	0.246
100	0.474	0.378	0.426	0.378	0.480	0.358
150	0.642	0.522	0.568	0.606	0.596	0.572
200	0.726	0.714	0.732	0.690	0.708	0.666
250	0.870	0.850	0.866	0.816	0.784	0.708

[*]Notes: Reported rejection frequencies are based on 5% nominal size DM tests. See Section 3 for further details.

Table 2: Monte Carlo Results - Data Snooping Test: MSE Loss Function*

P	R					
	25	50	100	150	200	250

			.			
Empirical Size - Case a						
25	0.146	0.120	0.152	0.170	0.142	0.130
50	0.106	0.104	0.090	0.124	0.108	0.098
100	0.070	0.076	0.064	0.106	0.074	0.096
150	0.080	0.064	0.076	0.068	0.088	0.072
200	0.062	0.068	0.068	0.082	0.082	0.062
250	0.054	0.052	0.064	0.074	0.066	0.068
Empirical Power - Case a						
25	0.922	0.944	0.924	0.942	0.920	0.958
50	0.986	0.994	0.990	0.990	0.992	0.996
100	1.000	1.000	1.000	1.000	1.000	1.000
150	1.000	1.000	1.000	1.000	1.000	1.000
200	1.000	1.000	1.000	1.000	1.000	1.000
250	1.000	1.000	1.000	1.000	1.000	1.000
			.			
Empirical Size - Case b						
25	0.140	0.136	0.150	0.168	0.142	0.134
50	0.092	0.102	0.110	0.120	0.110	0.116
100	0.076	0.090	0.068	0.100	0.066	0.098
150	0.072	0.066	0.074	0.072	0.090	0.072
200	0.060	0.068	0.064	0.080	0.084	0.066
250	0.062	0.058	0.064	0.076	0.068	0.068
Empirical Power - Case b						
25	0.896	0.934	0.912	0.936	0.902	0.954
50	0.980	0.990	0.990	0.990	0.988	0.994
100	1.000	1.000	1.000	1.000	1.000	1.000
150	1.000	1.000	1.000	1.000	1.000	1.000
200	1.000	1.000	1.000	1.000	1.000	1.000
250	1.000	1.000	1.000	1.000	1.000	1.000
			.			
Empirical Size - Case c						
25	0.080	0.070	0.120	0.090	0.130	0.140
50	0.030	0.030	0.100	0.080	0.100	0.110
100	0.010	0.030	0.040	0.040	0.060	0.020
150	0.020	0.040	0.030	0.020	0.050	0.030
200	0.010	0.020	0.000	0.020	0.060	0.000
250	0.000	0.010	0.000	0.020	0.030	0.000
Empirical Power - Case c						
25	0.110	0.140	0.120	0.200	0.170	0.180
50	0.020	0.060	0.070	0.140	0.160	0.170
100	0.010	0.040	0.080	0.170	0.160	0.080
150	0.010	0.050	0.110	0.110	0.110	0.170
200	0.040	0.030	0.090	0.080	0.090	0.160
250	0.040	0.060	0.100	0.130	0.150	0.170

*Notes: See notes to Table 1. Results for the data snooping test with 1 benchmark model and 2 alternative models are reported. Note that the poor results reported in the third panel of this table are due to very inaccurate estimation of γ, which affects both size and power results. Further analysis of this feature of our experimental setup is left to future research.

Table 3: Monte Carlo Results - $S_P(\gamma = 0.1)$ Test: MSE Loss Function [*]

P	R					
	25	50	100	150	200	250

.

Empirical Size - Case a

25	0.622	0.580	0.618	0.620	0.610	0.610
50	0.538	0.556	0.510	0.558	0.572	0.516
100	0.466	0.468	0.448	0.448	0.468	0.428
150	0.408	0.430	0.364	0.388	0.370	0.324
200	0.354	0.366	0.324	0.352	0.328	0.326
250	0.328	0.322	0.298	0.294	0.322	0.298

Empirical Power - Case a

25	0.290	0.292	0.282	0.260	0.352	0.262
50	0.380	0.380	0.482	0.362	0.388	0.440
100	0.564	0.596	0.550	0.508	0.588	0.606
150	0.700	0.678	0.742	0.734	0.706	0.790
200	0.772	0.764	0.826	0.802	0.852	0.826
250	0.856	0.858	0.884	0.896	0.896	0.850

.

Empirical Size - Case b

25	0.590	0.592	0.612	0.630	0.626	0.634
50	0.524	0.548	0.492	0.544	0.572	0.552
100	0.460	0.448	0.448	0.456	0.464	0.424
150	0.394	0.412	0.380	0.404	0.368	0.330
200	0.366	0.346	0.340	0.350	0.336	0.326
250	0.318	0.326	0.304	0.288	0.320	0.286

Empirical Power - Case b

25	0.288	0.262	0.286	0.270	0.354	0.280
50	0.342	0.388	0.470	0.348	0.378	0.440
100	0.624	0.598	0.598	0.510	0.576	0.590
150	0.700	0.712	0.758	0.734	0.666	0.786
200	0.808	0.790	0.836	0.800	0.852	0.840
250	0.860	0.866	0.872	0.868	0.888	0.852

.

Empirical Size - Case c

25	0.358	0.322	0.234	0.194	0.216	0.202
50	0.276	0.200	0.168	0.164	0.152	0.132
100	0.146	0.128	0.130	0.120	0.126	0.100
150	0.138	0.146	0.106	0.088	0.092	0.084
200	0.130	0.110	0.092	0.084	0.084	0.090
250	0.106	0.100	0.088	0.074	0.076	0.074

Empirical Power - Case c

25	0.280	0.346	0.470	0.462	0.518	0.480
50	0.372	0.458	0.576	0.500	0.578	0.572
100	0.580	0.618	0.622	0.640	0.648	0.682
150	0.664	0.690	0.710	0.718	0.708	0.738
200	0.732	0.732	0.748	0.762	0.764	0.768
250	0.762	0.738	0.800	0.780	0.806	0.796

[*]Notes: See notes to Table 1. Results for the S_P test with $\gamma = 0.1$ and a mean square forecast error loss function are reported.

Table 4: Monte Carlo Results - $S_P(\gamma = 0.4)$ Test: MAD Loss Function [*]

P	R					
	25	50	100	150	200	250

.

Empirical Size - Case a

25	0.240	0.178	0.202	0.220	0.208	0.206
50	0.178	0.180	0.180	0.206	0.212	0.158
100	0.170	0.144	0.162	0.188	0.178	0.152
150	0.162	0.170	0.138	0.150	0.164	0.152
200	0.160	0.144	0.146	0.166	0.156	0.150
250	0.158	0.160	0.154	0.156	0.184	0.156

Empirical Power - Case a

25	0.124	0.166	0.160	0.174	0.190	0.202
50	0.194	0.186	0.242	0.216	0.186	0.226
100	0.356	0.306	0.374	0.320	0.332	0.326
150	0.408	0.438	0.428	0.460	0.430	0.436
200	0.516	0.494	0.510	0.512	0.494	0.498
250	0.602	0.574	0.562	0.612	0.538	0.562

.

Empirical Size - Case b

25	0.222	0.178	0.194	0.218	0.204	0.190
50	0.178	0.186	0.152	0.198	0.202	0.170
100	0.148	0.140	0.154	0.180	0.170	0.154
150	0.168	0.174	0.136	0.152	0.168	0.142
200	0.152	0.148	0.146	0.164	0.156	0.158
250	0.148	0.154	0.156	0.154	0.178	0.158

Empirical Power - Case b

25	0.124	0.160	0.170	0.176	0.202	0.174
50	0.174	0.188	0.248	0.234	0.198	0.232
100	0.348	0.328	0.384	0.334	0.326	0.324
150	0.424	0.438	0.460	0.448	0.420	0.444
200	0.534	0.516	0.494	0.512	0.494	0.484
250	0.616	0.568	0.558	0.618	0.534	0.574

.

Empirical Size - Case c

25	0.126	0.098	0.064	0.076	0.078	0.072
50	0.096	0.098	0.066	0.068	0.078	0.058
100	0.096	0.088	0.080	0.076	0.080	0.066
150	0.086	0.082	0.070	0.082	0.072	0.070
200	0.080	0.076	0.080	0.074	0.064	0.072
250	0.086	0.082	0.068	0.066	0.060	0.064

Empirical Power - Case c

25	0.114	0.184	0.190	0.216	0.226	0.202
50	0.190	0.232	0.248	0.274	0.252	0.220
100	0.264	0.266	0.304	0.306	0.256	0.254
150	0.328	0.342	0.344	0.292	0.308	0.324
200	0.374	0.388	0.376	0.364	0.370	0.346
250	0.400	0.400	0.426	0.416	0.442	0.412

[*]Notes: See notes to Table 1. Results for the S_P test with $\gamma = 0.4$ and a mean absolute deviation forecast error loss function are reported.

Table 5: Summary of Ex-Ante Forecasting Results[*]

Lag Selec	Horizon	MSE		
	(in months)	M1	M2	M3
AIC	h=1	59.41	51.74^a	51.51^b
AIC	h=3	62.44	52.32^a	53.19^b
AIC	h=12	49.40	52.80	52.17
BIC	h=1	54.36	44.65^a	45.59^a
BIC	h=3	53.43	48.41	49.16
BIC	h=12	44.40^a	52.08	46.90

[*] Notes: Selected ex-ante forecasting summary statistics are presented for rolling ex-ante industrial production forecasts based on $M1$ models (estimated using $(y_t, m1_t, p_t, R_t, C_t)'$), $M2$ models (estimated using $(y_t, m2_t, p_t, R_t, C_t)'$), and $M3$ models (estimated using $(y_t, m3_t, p_t, R_t, C_t)'$). All models include an intercept term, are in differenced form, and are re-estimated before each new 1-, 3-, or 12-step ahead forecast (h=1,3,12) is constructed. Lags are selected using either the AIC or the BIC (see column 1 of table). Mean squared forecast errors (MSEs) are reported in columns 3-5 (actual forecast errors are multiplied by 1200 *before* MSFEs are formed). Entries with the superscipt "a" denote significantly smallers MSEs based on a pairwise comparison of IP forecasts constructed using models which include either $m1_t$ or $m2_t$ to form DM tests at 5% significance levels. Entries superscripted with "b" likewise denote smaller MSEs based on a comparison of IP forecasts constructed using models including either $m1_t$ or $m3_t$. For example, in the first row, an ^{a} beside the MSE entry of 51.74 signifies that the model with $m2_t$ DM-dominates the corresponding model with $m1_t$ (which has an MSE of 59.41). Results based on the application of data snooping tests are reported in Section 4. The ex-ante period is 1983:1-1997:12 (15 years - 180 monthly observations).