Chapter 3

FORECAST EVALUATION

KENNETH D. WEST

University of Wisconsin

Contents

Handbook of Economic Forecasting, Volume 1
Edited by Graham Elliott, Clive W.J. Granger and Allan Timmermann

DOI: 10.1016/S1574-0706(05)01003-7

Abstract

This chapter summarizes recent literature on asymptotic inference about forecasts. Both analytical and simulation based methods are discussed. The emphasis is on techniques applicable when the number of competing models is small. Techniques applicable when a large number of models is compared to a benchmark are also briefly discussed.

Keywords

forecast, prediction, out of sample, prediction error, forecast error, parameter estimation error, asymptotic irrelevance, hypothesis test, inference

JEL classification: C220, C320, C520, C530

1. Introduction

This chapter reviews asymptotic methods for inference about moments of functions of predictions and prediction errors. The methods may rely on conventional asymptotics or they may be bootstrap based. The relevant class of applications are ones in which the investigator uses a long time series of predictions and prediction errors as a model evaluation tool. Typically the evaluation is done retrospectively rather than in real time. A classic example is Meese and Rogoff's (1983) evaluation of exchange rate models.

In most applications, the investigator aims to compare two or more models. Measures of relative model quality might include ratios or differences of mean, mean-squared or mean-absolute prediction errors; correlation between one model's prediction and another model's realization (also known as forecast encompassing); or comparisons of utility or profit-based measures of predictive ability. In other applications, the investigator focuses on a single model, in which case measures of model quality might include correlation between prediction and realization, lack of serial correlation in one step ahead prediction errors, ability to predict direction of change, or bias in predictions.

Predictive ability has long played a role in evaluation of econometric models. An early example of a study that retrospectively set aside a large number of observations for predictive evaluation is Wilson (1934, pp. 307–308). Wilson, who studied monthly price data spanning more than a century, used estimates from the first half of his data to forecast the next twenty years. He then evaluated his model by computing the correlation between prediction and realization.[1] Growth in data and computing power has led to widespread use of similar predictive evaluation techniques, as is indicated by the applications cited below.

To prevent misunderstanding, it may help to stress that the techniques discussed here are probably of little relevance to studies that set aside one or two or a handful of observations for out of sample evaluation. The reader is referred to textbook expositions about confidence intervals around a prediction, or to proposals for simulation methods such as Fair (1980). As well, the paper does not cover density forecasts. Inference about such forecasts is covered in the Handbook Chapter 5 by Corradi and Swanson (2006). Finally, the paper takes for granted that one wishes to perform out of sample analysis. My purpose is to describe techniques that can be used by researchers who have decided, for reasons not discussed in this chapter, to use a non-trivial portion of their samples for prediction. See recent work by Chen (2004), Clark and McCracken (2005b) and Inoue and Kilian (2004a, 2004b) for different takes on the possible power advantages of using out of sample tests.

Much of the paper uses tests for equal mean squared prediction error (MSPE) for illustration. MSPE is not only simple, but it is also arguably the most commonly used measure of predictive ability. The focus on MSPE, however, is done purely for expositional reasons. This paper is intended to be useful for practitioners interested in a

[1] Which, incidentally and regrettably, turned out to be negative.

wide range of functions of predictions and prediction errors that have appeared in the literature. Consequently, results that are quite general are presented. Because the target audience is practitioners, I do not give technical details. Instead, I give examples, summarize findings and present guidelines.

Section 2 illustrates the evolution of the relevant methodology. Sections 3–8 discuss inference when the number of models under evaluation is small. "Small" is not precisely defined, but in sample sizes typically available in economics suggests a number in the single digits. Section 3 discusses inference in the unusual, but conceptually simple, case in which none of the models under consideration rely on estimated regression parameters to make predictions. Sections 4 and 5 relax this assumption, but for reasons described in those sections assume that the models under consideration are nonnested. Section 4 describes when reliance on estimated regression parameters is irrelevant asymptotically, so that Section 3 procedures may still be applied. Section 5 describes how to account for reliance on estimated regression parameters. Sections 6 and 7 consider nested models. Section 6 focuses on MSPE, Section 7 other loss functions. Section 8 summarizes the results of previous sections. Section 9 briefly discusses inference when the number of models being evaluated is large, possibly larger than the sample size. Section 10 concludes.

2. A brief history

I begin the discussion with a brief history of methodology for inference, focusing on mean squared prediction errors (MSPE).

Let e_{1t} and e_{2t} denote one step ahead prediction errors from two competing models. Let their corresponding second moments be

$$\sigma_1^2 \equiv \mathrm{E}e_{1t}^2 \quad \text{and} \quad \sigma_2^2 \equiv \mathrm{E}e_{2t}^2.$$

(For reasons explained below, the assumption of stationarity – the absence of a t subscript on σ_1^2 and σ_2^2 – is not always innocuous. For the moment, I maintain it for consistency with the literature about to be reviewed.) One wishes to test the null

$$\mathrm{H}_0\colon \quad \sigma_1^2 - \sigma_2^2 = 0,$$

or perhaps construct a confidence interval around the point estimate of $\sigma_1^2 - \sigma_2^2$.

Observe that $\mathrm{E}(e_{1t} - e_{2t})(e_{1t} + e_{2t}) = \sigma_1^2 - \sigma_2^2$. Thus $\sigma_1^2 - \sigma_2^2 = 0$ if and only if the covariance or correlation between $e_{1t} - e_{2t}$ and $e_{1t} + e_{2t}$ is zero. Let us suppose initially that (e_{1t}, e_{2t}) is i.i.d. Granger and Newbold (1977) used this observation to suggest testing H_0: $\sigma_1^2 - \sigma_2^2 = 0$ by testing for zero correlation between $e_{1t} - e_{2t}$ and $e_{1t} + e_{2t}$. This procedure was earlier proposed by Morgan (1939) in the context of testing for equality between variances of two normal random variables. Granger and Newbold (1977) assumed that the forecast errors had zero mean, but Morgan (1939) indicates that this assumption is not essential. The Granger and Newbold test was extended to multistep, serially correlated and possibly non-normal prediction errors by Meese and Rogoff (1988) and Mizrach (1995).

Ashley, Granger and Schmalensee (1980) proposed a test of equal MSPE in the context of nested models. For nested models, equal MSPE is theoretically equivalent to a test of Granger non-causality. Ashley, Granger and Schmalensee (1980) proposed executing a standard F-test, but with out of sample prediction errors used to compute restricted and unrestricted error variances. Ashley, Granger and Schmalensee (1980) recommended that tests be one-sided, testing whether the unrestricted model has smaller MSPE than the restricted (nested) model: it is not clear what it means if the restricted model has a significantly smaller MSPE than the unrestricted model.

The literature on predictive inference that is a focus of this chapter draws on now standard central limit theory introduced into econometrics research by Hansen (1982) – what I will call "standard results" in the rest of the discussion. Perhaps the first explicit use of standard results in predictive inference is Christiano (1989). Let $f_t = e_{1t}^2 - e_{2t}^2$. Christiano observed that we are interested in the mean of f_t, call it $\mathrm{E}f_t \equiv \sigma_1^2 - \sigma_2^2$.[2] And there are standard results on inference about means – indeed, if f_t is i.i.d. with finite variance, introductory econometrics texts describe how to conduct inference about $\mathrm{E}f_t$ given a sample of $\{f_t\}$. A random variable like $e_{1t}^2 - e_{2t}^2$ may be non-normal and serially correlated. But results in Hansen (1982) apply to non-i.i.d. time series data. (Details below.)

One of Hansen's (1982) conditions is stationarity. Christiano acknowledged that standard results might not apply to his empirical application because of a possible failure of stationarity. Specifically, Christiano compared predictions of models estimated over samples of increasing size: the first of his 96 predictions relied on models estimated on quarterly data running from 1960 to 1969, the last from 1960 to 1988. Because of increasing precision of estimates of the models, forecast error variances might decline over time. (This is one sense in which the assumption of stationarity was described as "not obviously innocuous" above.)

West, Edison and Cho (1993) and West and Cho (1995) independently used standard results to compute test statistics. The objects of interest were MSPEs and a certain utility based measure of predictive ability. Diebold and Mariano (1995) proposed using the same standard results, also independently, but in a general context that allows one to be interested in the mean of a general loss or utility function. As detailed below, these papers explained either in context or as a general principle how to allow for multistep, non-normal, and conditionally heteroskedastic prediction errors.

The papers cited in the preceding two paragraphs all proceed without proof. None directly address the possible complications from parameter estimation noted by Christiano (1989). A possible approach to allowing for these complications in special cases is in Hoffman and Pagan (1989) and Ghysels and Hall (1990). These papers showed how

[2] Actually, Christiano looked at root mean squared prediction errors, testing whether $\sigma_1 - \sigma_2 = 0$. For clarity and consistency with the rest of my discussion, I cast his analysis in terms of MSPE.

standard results from Hansen (1982) can be extended to account for parameter estimation in out of sample tests of instrument residual orthogonality when a fixed parameter estimate is used to construct the test. [Christiano (1989), and most of the forecasting literature, by contrast updates parameter estimate as forecasts progress through the sample.] A general analysis was first presented in West (1996), who showed how standard results can be extended when a sequence of parameter estimates is used, and for the mean of a general loss or utility function.

Further explication of developments in inference about predictive ability requires me to start writing out some results. I therefore call a halt to the historical summary. The next section begins the discussion of analytical results related to the papers cited here.

3. A small number of nonnested models, Part I

Analytical results are clearest in the unusual (in economics) case in which predictions do not rely on estimated regression parameters, an assumption maintained in this section but relaxed in future sections.

Notation is as follows. The object of interest is $\mathrm{E}f_t$, an $(m \times 1)$ vector of moments of predictions or prediction errors. Examples include MSPE, mean prediction error, mean absolute prediction error, covariance between one model's prediction and another model's prediction error, mean utility or profit, and means of loss functions that weight positive and negative errors asymmetrically as in Elliott and Timmermann (2003). If one is comparing models, then the elements of $\mathrm{E}f_t$ are expected differences in performance. For MSPE comparisons, and using the notation of the previous section, for example, $\mathrm{E}f_t = \mathrm{E}e_{1t}^2 - \mathrm{E}e_{2t}^2$. As stressed by Diebold and Mariano (1995), this framework also accommodates general loss functions or measures of performance. Let $\mathrm{E}g_{it}$ be the measure of performance of model i – perhaps MSPE, perhaps mean absolute error, perhaps expected utility. Then when there are two models, $m = 1$ and $\mathrm{E}f_t = \mathrm{E}g_{1t} - \mathrm{E}g_{2t}$.

We have a sample of predictions of size P. Let $\bar{f}^* \equiv P^{-1} \sum_t f_t$ denote the $m \times 1$ sample mean of f_t. (The reason for the "*" superscript will become apparent below.) If we are comparing two models with performance of model i measured by $\mathrm{E}g_{it}$, then of course $\bar{f}^* \equiv P^{-1} \sum_t (g_{1t} - g_{2t}) \equiv \bar{g}_1 - \bar{g}_2 =$ the difference in performance of the two models, over the sample. For simplicity and clarity, assume covariance stationarity – neither the first nor second moments of f_t depend on t. At present (predictions do not depend on estimated regression parameters), this assumption is innocuous. It allows simplification of formulas. The results below can be extended to allow moment drift as long as time series averages converge to suitable constants. See Giacomini and White (2003). Then under well-understood and seemingly weak conditions, a central limit theorem holds:

$$\sqrt{P}\bigl(\bar{f}^* - \mathrm{E}f_t\bigr) \sim_A \mathrm{N}\bigl(0, V^*\bigr), \quad V^* \equiv \sum_{j=-\infty}^{\infty} \mathrm{E}(f_t - \mathrm{E}f_t)(f_{t-j} - \mathrm{E}f_t)'. \tag{3.1}$$

See, for example, White (1984) for the "well-understood" phrase of the sentence prior to (3.1); see below for the "seemingly weak" phrase. Equation (3.1) is the "standard result" referenced above. The $m \times m$ positive semidefinite matrix V^* is sometimes called the long run variance of f_t. If f_t is serially uncorrelated (perhaps i.i.d.), then $V^* = \mathrm{E}(f_t - \mathrm{E}f_t)(f_t - \mathrm{E}f_t)'$. If, further, $m = 1$ so that f_t is a scalar, $V^* = \mathrm{E}(f_t - \mathrm{E}f_t)^2$.

Suppose that V^* is positive definite. Let $\hat{V}^*$ be a consistent estimator of V^*. Typically $\hat{V}^*$ will be constructed with a heteroskedasticity and autocorrelation consistent covariance matrix estimator. Then one can test the null

$$\mathrm{H}_0\colon \quad \mathrm{E}f_t = 0 \tag{3.2}$$

with a Wald test:

$$\bar{f}^{*\prime}\hat{V}^{*-1}\bar{f}^* \sim_A \chi^2(m). \tag{3.3}$$

If $m = 1$ so that f_t is a scalar, one can test the null with a t-test:

$$\begin{aligned} &\bar{f}^*/\big[\hat{V}^*/P\big]^{1/2} \sim_A \mathrm{N}(0,1), \\ &\hat{V}^* \to_p V^* \equiv \sum_{j=-\infty}^{\infty} \mathrm{E}(f_t - \mathrm{E}f_t)(f_{t-j} - \mathrm{E}f_t). \end{aligned} \tag{3.4}$$

Confidence intervals can be constructed in obvious fashion from $[\hat{V}^*/P]^{1/2}$.

As noted above, the example of the previous section maps into this notation with $m = 1$, $f_t = e_{1t}^2 - e_{2t}^2$, $\mathrm{E}f_t = \sigma_1^2 - \sigma_2^2$, and the null of equal predictive ability is that $\mathrm{E}f_t = 0$, i.e., $\sigma_1^2 = \sigma_2^2$. Testing for equality of MSPE in a set of $m + 1$ models for $m > 1$ is straightforward, as described in the next section. To give an illustration or two of other possible definitions of f_t, sticking for simplicity with $m = 1$: If one is interested in whether a forecast is unbiased, then $f_t = e_{1t}$ and $\mathrm{E}f_t = 0$ is the hypothesis that the model 1 forecast error is unbiased. If one is interested in mean absolute error, $f_t = |e_{1t}| - |e_{2t}|$, and $\mathrm{E}f_t = 0$ is the hypothesis of equal mean absolute prediction error. Additional examples are presented in a subsequent section below.

For concreteness, let me return to MSPE, with $m = 1$, $f_t = e_{1t}^2 - e_{2t}^2$, $\bar{f}^* \equiv P^{-1}\sum_t(e_{1t}^2 - e_{2t}^2)$. Suppose first that (e_{1t}, e_{2t}) is i.i.d. Then so, too, is $e_{1t}^2 - e_{2t}^2$, and $V^* = \mathrm{E}(f_t - \mathrm{E}f_t)^2 = \mathrm{variance}(e_{1t}^2 - e_{2t}^2)$. In such a case, as the number of forecast errors $P \to \infty$ one can estimate V^* consistently with $\hat{V}^* = P^{-1}\sum_t(f_t - \bar{f}^*)^2$. Suppose next that (e_{1t}, e_{2t}) is a vector of τ step ahead forecast errors whose (2×1) vector of Wold innovations is i.i.d. Then (e_{1t}, e_{2t}) and $e_{1t}^2 - e_{2t}^2$ follow MA$(\tau - 1)$ processes, and $V^* = \sum_{j=-\tau+1}^{\tau-1} \mathrm{E}(f_t - \mathrm{E}f_t)(f_{t-j} - \mathrm{E}f_t)$. One possible estimator of V^* is the sample analogue. Let $\hat{\Gamma}_j = P^{-1}\sum_{t>|j|}(f_t - \bar{f}^*)(f_{t-|j|} - \bar{f}^*)$ be an estimate of $\mathrm{E}(f_t - \mathrm{E}f_t)(f_{t-j} - \mathrm{E}f_t)$, and set $\hat{V}^* = \sum_{j=-\tau+1}^{\tau-1}\hat{\Gamma}_j$. It is well known, however, that this estimator may not be positive definite if $\tau > 0$. Hence one may wish to use an estimator that is both consistent and positive semidefinite by construction [Newey and West (1987, 1994), Andrews (1991), Andrews and Monahan (1994), den Haan and

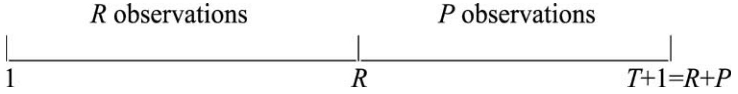
Levin (2000)]. Finally, under some circumstances, one will wish to use a heteroskedasticity and autocorrelation consistent estimator of V^* even when (e_{1t}, e_{2t}) is a one step forecast error. This will be the case if the second moments follow a GARCH or related process, in which case there will be serial correlation in $f_t = e_{1t}^2 - e_{2t}^2$ even if there is no serial correlation in (e_{1t}, e_{2t}).

But such results are well known, for f_t a scalar or vector, and for f_t relevant for MSPE or other moments of predictions and prediction errors. The "seemingly weak" conditions referenced above Equation (3.1) allow for quite general forms of dependence and heterogeneity in forecasts and forecast errors. I use the word "seemingly" because of some ancillary assumptions that are not satisfied in some relevant applications. First, the number of models m must be "small" relative to the number of predictions P. In an extreme case in which $m > P$, conventional estimators will yield $\hat{V}^*$ that is not of full rank. As well, and more informally, one suspects that conventional asymptotics will yield a poor approximation if m is large relative to P. Section 9 briefly discusses alternative approaches likely to be useful in such contexts.

Second, and more generally, V^* must be full rank. When the number of models $m = 2$, and MSPE is the object of interest, this rules out $e_{1t}^2 = e_{2t}^2$ with probability 1 (obviously). It also rules out pairs of models in which $\sqrt{P}(\hat{\sigma}_1^2 - \hat{\sigma}_2^2) \rightarrow_p 0$. This latter condition is violated in applications in which one or both models make predictions based on estimated regression parameters, and the models are nested. This is discussed in Sections 6 and 7 below.

4. A small number of nonnested models, Part II

In the vast majority of economic applications, one or more of the models under consideration rely on estimated regression parameters when making predictions. To spell out the implications for inference, it is necessary to define some additional notation. For simplicity, assume that one step ahead prediction errors are the object of interest. Let the total sample size be $T + 1$. The last P observations of this sample are used for forecast evaluation. The first R observations are used to construct an initial set of regression estimates that are then used for the first prediction. We have $R + P = T + 1$. Schematically:



(4.1)

Division of the available data into R and P is taken as given.

In the forecasting literature, three distinct schemes figure prominently in how one generates the sequence of regression estimates necessary to make predictions. Asymptotic results differ slightly for the three, so it is necessary to distinguish between them. Let β denote the vector of regression parameters whose estimates are used to make predictions. In the *recursive* scheme, the size of the sample used to estimate β grows as one makes predictions for successive observations. One first estimates β with data from 1 to R and uses the estimate to predict observation $R+1$ (recall that I am assuming one step ahead predictions, for simplicity); one then estimates β with data from 1 to $R+1$, with the new estimate used to predict observation $R+2; \ldots$; finally, one estimates β with data from 1 to T, with the final estimate used to predict observation $T+1$. In the *rolling* scheme, the sequence of β's is always generated from a sample of size R. The first estimate of β is obtained with a sample running from 1 to R, the next with a sample running from 2 to $R+1, \ldots$, the final with a sample running from $T-R+1$ to T. In the *fixed* scheme, one estimates β just once, using data from 1 to R. In all three schemes, the number of predictions is P and the size of the smallest regression sample is R. Examples of applications using each of these schemes include Faust, Rogers and Wright (2004) (recursive), Cheung, Chinn and Pascual (2003) (rolling) and Ashley, Granger and Schmalensee (1980) (fixed). The fixed scheme is relatively attractive when it is computationally difficult to update parameter estimates. The rolling scheme is relatively attractive when one wishes to guard against moment or parameter drift that is difficult to model explicitly.

It may help to illustrate with a simple example. Suppose one model under consideration is a univariate zero mean AR(1): $y_t = \beta^* y_{t-1} + e_{1t}$. Suppose further that the estimator is ordinary least squares. Then the sequence of P estimates of β^* are generated as follows for $t = R, \ldots, T$:

$$\begin{aligned}
&\text{recursive:} \quad \hat{\beta}_t = \left[\sum_{s=1}^{t}\left(y_{s-1}^2\right)\right]^{-1}\left[\sum_{s=1}^{t} y_{s-1}y_s\right];\\
&\text{rolling:} \quad \hat{\beta}_t = \left[\sum_{s=t-R+1}^{t}\left(y_{s-1}^2\right)\right]^{-1}\left[\sum_{s=t-R+1}^{t} y_{s-1}y_s\right];\\
&\text{fixed:} \quad \hat{\beta}_t = \left[\sum_{s=1}^{R}\left(y_{s-1}^2\right)\right]^{-1}\left[\sum_{s=1}^{R} y_{s-1}y_s\right].
\end{aligned} \tag{4.2}$$

In each case, the one step ahead prediction error is $\hat{e}_{t+1} \equiv y_{t+1} - y_t\hat{\beta}_t$. Observe that for the fixed scheme $\hat{\beta}_t = \hat{\beta}_R$ for all t, while $\hat{\beta}_t$ changes with t for the rolling and recursive schemes.

I will illustrate with a simple MSPE example comparing two linear models. I then introduce notation necessary to define other moments of interest, sticking with linear models for expositional convenience. An important asymptotic result is then stated. The next section outlines a general framework that covers all the simple examples in this section, and allows for nonlinear models and estimators.

So suppose there are two least squares models, say $y_t = X_{1t}'\beta_1^* + e_{1t}$ and $y_t = X_{2t}'\beta_2^* + e_{2t}$. (Note the dating convention: X_{1t} and X_{2t} can be used to predict y_t, for example $X_{1t} = y_{t-1}$ if model 1 is an AR(1).) The population MSPEs are $\sigma_1^2 \equiv \mathrm{E}e_{1t}^2$ and $\sigma_2^2 \equiv \mathrm{E}e_{2t}^2$. (Absence of a subscript t on the MSPEs is for simplicity and without substance.) Define the sample one step ahead forecast errors and sample MSPEs as

$$\hat{e}_{1t+1} \equiv y_{t+1} - X'_{1t+1}\hat{\beta}_{1t}, \quad \hat{e}_{2t+1} \equiv y_{t+1} - X'_{2t+1}\hat{\beta}_{2t},$$
$$\hat{\sigma}_1^2 = P^{-1}\sum_{t=R}^{T}\hat{e}_{1t+1}^2, \quad \hat{\sigma}_2^2 = P^{-1}\sum_{t=R}^{T}\hat{e}_{2t+1}^2. \tag{4.3}$$

With MSPE the object of interest, one examines the difference between the sample MSPEs $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$. Let

$$\hat{f}_t \equiv \hat{e}_{1t}^2 - \hat{e}_{2t}^2, \quad \bar{f} \equiv P^{-1}\sum_{t=R}^{T}\hat{f}_{t+1} \equiv \hat{\sigma}_1^2 - \hat{\sigma}_2^2. \tag{4.4}$$

Observe that $\bar{f}$ defined in (4.4) differs from $\bar{f}^*$ defined above (3.1) in that $\bar{f}$ relies on $\hat{e}$'s, whereas $\bar{f}^*$ relies on e's.

The null hypothesis is $\sigma_1^2 - \sigma_2^2 = 0$. One way to test the null would be to substitute $\hat{e}_{1t}$ and $\hat{e}_{2t}$ for e_{1t} and e_{2t} in the formulas presented in the previous section. If $(e_{1t}, e_{2t})'$ is i.i.d., for example, one would set $\hat{V}^* = P^{-1}\sum_{t=R}^{T}(\hat{f}_{t+1} - \bar{f})^2$, compute the t-statistic

$$\bar{f}/\left[\hat{V}^*/P\right]^{1/2} \tag{4.5}$$

and use standard normal critical values. [I use the "*" in $\hat{V}^*$ for both $P^{-1}\sum_{t=R}^{T}(\hat{f}_{t+1} - \bar{f})^2$ (this section) and for $P^{-1}\sum_{t=R}^{T}(f_{t+1} - \bar{f}^*)^2$ (previous section) because under the asymptotic approximations described below, both are consistent for the long run variance of f_{t+1}.]

Use of (4.5) is not obviously an advisable approach. Clearly, $\hat{e}_{1t}^2 - \hat{e}_{2t}^2$ is polluted by error in estimation of β_1 and β_2. It is not obvious that sample averages of $\hat{e}_{1t}^2 - \hat{e}_{2t}^2$ (i.e., $\bar{f}$) have the same asymptotic distribution as those of $e_{1t}^2 - e_{2t}^2$ (i.e., $\bar{f}^*$). Under suitable conditions (see below), a key factor determining whether the asymptotic distributions are equivalent is whether or not the two models are nested. If they are nested, the distributions are not equivalent. Use of (4.5) with normal critical values is not advised. This is discussed in a subsequent section.

If the models are not nested, West (1996) showed that when conducting inference about MSPE, parameter estimation error is *asymptotically irrelevant*. I put the phrase in italics because I will have frequent recourse to it in the sequel: "asymptotic irrelevance" means that one conduct inference by applying standard results to the mean of the loss function of interest, treating parameter estimation error as irrelevant.

To explain this result, as well as to illustrate when asymptotic irrelevance does not apply, requires some – actually, considerable – notation. I will phase in some of this notation in this section, with most of the algebra deferred to the next section. Let β^* denote the $k \times 1$ population value of the parameter vector used to make predictions. Suppose for expositional simplicity that the model(s) used to make predictions are linear,

$$y_t = X_t'\beta^* + e_t \tag{4.6a}$$

if there is a single model,

$$y_t = X_{1t}'\beta_1^* + e_{1t}, \quad y_t = X_{2t}'\beta_2^* + e_{2t}, \quad \beta^* \equiv \left(\beta_1^{*\prime}, \beta_2^{*\prime}\right)', \tag{4.6b}$$

if there are two competing models. Let $f_t(\beta^*)$ be the random variable whose expectation is of interest. Then leading scalar $(m = 1)$ examples of $f_t(\beta^*)$ include:

$$f_t(\beta^*) = e_{1t}^2 - e_{2t}^2 = \left(y_t - X_{1t}'\beta_1^*\right)^2 - \left(y_t - X_{2t}'\beta_2^*\right)^2 \tag{4.7a}$$

($\mathrm{E}f_t = 0$ means equal MSPE);

$$f_t(\beta^*) = e_t = y_t - X_t'\beta^* \tag{4.7b}$$

($\mathrm{E}f_t = 0$ means zero mean prediction error);

$$f_t(\beta^*) = e_{1t}X_{2t}'\beta_2^* = \left(y_t - X_{1t}'\beta_1^*\right)X_{2t}'\beta_2^* \tag{4.7c}$$

[$\mathrm{E}f_t = 0$ means zero correlation between one model's prediction error and another model's prediction, an implication of forecast encompassing proposed by Chong and Hendry (1986)];

$$f_t(\beta^*) = e_{1t}(e_{1t} - e_{2t}) = \left(y_t - X_{1t}'\beta_1^*\right)\left[\left(y_t - X_{1t}'\beta_1^*\right) - \left(y_t - X_{2t}'\beta_2^*\right)\right] \tag{4.7d}$$

[$\mathrm{E}f_t = 0$ is an implication of forecast encompassing proposed by Harvey, Leybourne and Newbold (1998)];

$$f_t(\beta^*) = e_{t+1}e_t = \left(y_{t+1} - X_{t+1}'\beta^*\right)\left(y_t - X_t'\beta^*\right) \tag{4.7e}$$

($\mathrm{E}f_t = 0$ means zero first order serial correlation);

$$f_t(\beta^*) = e_tX_t'\beta^* = \left(y_t - X_t'\beta^*\right)X_t'\beta^* \tag{4.7f}$$

($\mathrm{E}f_t = 0$ means the prediction and prediction error are uncorrelated);

$$f_t(\beta^*) = |e_{1t}| - |e_{2t}| = \left|y_t - X_{1t}'\beta_1^*\right| - \left|y_t - X_{2t}'\beta_2^*\right| \tag{4.7g}$$

($\mathrm{E}f_t = 0$ means equal mean absolute error).

More generally, $f_t(\beta^*)$ can be per period utility or profit, or differences across models of per period utility or profit, as in Leitch and Tanner (1991) or West, Edison and Cho (1993).

Let $\hat{f}_{t+1} \equiv f_{t+1}(\hat{\beta}_t)$ denote the sample counterpart of $f_{t+1}(\beta^*)$, with $\bar{f} \equiv P^{-1}\sum_{t=R}^{T}\hat{f}_{t+1}$ the sample mean evaluated at the series of estimates of β^*. Let $\bar{f}^* = P^{-1}\sum_{t=R}^{T}f_{t+1}(\beta^*)$ denote the sample mean evaluated at β^*. Let F denote the $(1 \times k)$ derivative of the expectation of f_t, evaluated at β^*:

$$F = \frac{\partial \mathrm{E}f_t(\beta^*)}{\partial \beta}. \tag{4.8}$$

For example, $F = -\mathrm{E}X_t'$ for mean prediction error (4.7b).

Then under mild conditions,

$$\begin{aligned}\sqrt{P}\left(\bar{f} - \mathrm{E}f_t\right) &= \sqrt{P}\left(\bar{f}^* - \mathrm{E}f_t\right) + F \times (P/R)^{1/2} \\ &\quad \times \left[\mathrm{O}_p(1) \text{ terms from the sequence of estimates of } \beta^*\right] + \mathrm{o}_p(1).\end{aligned} \tag{4.9}$$

Some specific formulas are in the next section. Result (4.9) holds not only when f_t is a scalar, as I have been assuming, but as well when f_t is a vector. (When f_t is a vector of dimension (say) m, F has dimension $m \times k$.)

Thus, uncertainty about the estimate of $\mathrm{E}f_t$ can be decomposed into uncertainty that would be present even if β^* were known and, possibly, additional uncertainty due to estimation of β^*. The qualifier "possibly" results from at least three sets of circumstances in which error in estimation of β^* is asymptotically irrelevant: (1) $F = 0$; (2) $P/R \to 0$; (3) the variance of the terms due to estimation of β^* is exactly offset by the covariance between these terms and $\sqrt{P}(\bar{f}^* - \mathrm{E}f_t)$. For cases (1) and (2), the middle term in (4.9) is identically zero ($F = 0$) or vanishes asymptotically ($P/R \to 0$), implying that $\sqrt{P}(\bar{f} - \mathrm{E}f_t) - \sqrt{P}(\bar{f}^* - \mathrm{E}f_t) \to_p 0$; for case (3) the asymptotic variances of $\sqrt{P}(\bar{f} - \mathrm{E}f_t)$ and $\sqrt{P}(\bar{f}^* - \mathrm{E}f_t)$ happen to be the same. In any of the three sets of circumstances, *inference can proceed as described in the previous section*. This is important because it simplifies matters if one can abstract from uncertainty about β^* when conducting inference.

To illustrate each of the three circumstances:

1. For MSPE in our linear example $F = (-2\mathrm{E}X_{1t}'e_{1t}, 2\mathrm{E}X_{2t}'e_{2t})'$. So $F = 0_{1\times k}$ if the predictors are uncorrelated with the prediction error.[3] Similarly, $F = 0$ for mean absolute prediction error (4.7g) ($\mathrm{E}[|e_{1t}| - |e_{2t}|]$) when the prediction errors have a median of zero, conditional on the predictors. (To prevent confusion, it is to be emphasized that MSPE and mean absolute error are unusual in that asymptotic irrelevance applies even when P/R is not small. In this sense, my focus on MSPE is a bit misleading.)

Let me illustrate the implications with an example in which f_t is a vector rather than a scalar. Suppose that we wish to test equality of MSPEs from $m+1$ competing models, under the assumption that the forecast error vector $(e_{1t}, \ldots, e_{m+1,t})'$ is i.i.d. Define the $m \times 1$ vectors

$$f_t \equiv \left(e_{1t}^2 - e_{2t}^2, \ldots, e_{1t}^2 - e_{m+1,t}^2\right)', \quad \hat{f}_t = \left(\hat{e}_{1t}^2 - \hat{e}_{2t}^2, \ldots, \hat{e}_{1t}^2 - \hat{e}_{m+1,t}^2\right)',$$
$$\bar{f} = P^{-1}\sum_{t=R}^{T} \hat{f}_{t+1}. \tag{4.10}$$

The null is that $\mathrm{E}f_t = 0_{m\times 1}$. (Of course, it is arbitrary that the null is defined as discrepancies from model 1's squared prediction errors; test statistics are identical regardless of the model used to define f_t.) Then under the null

$$\bar{f}'\hat{V}^{*-1}\bar{f} \sim_A \chi^2(m), \quad \hat{V}^* \to_p V^* \equiv \sum_{j=-\infty}^{\infty} \mathrm{E}(f_t - \mathrm{E}f_t)(f_{t-j} - \mathrm{E}f_t)', \tag{4.11}$$

[3] Of course, one would be unlikely to forecast with a model that *a priori* is expected to violate this condition, though prediction is sometimes done with realized right hand side endogenous variables [e.g., Meese and Rogoff (1983)]. But prediction exercise do sometimes find that this condition does not hold. That is, out of sample prediction errors display correlation with the predictors (even though in sample residuals often display zero correlation by construction). So even for MSPE, one might want to account for parameter estimation error when conducting inference.

where, as indicated, $\hat{V}^*$ is a consistent estimate of the $m \times m$ long run variance of f_t. If $f_t \equiv (e_{1t}^2 - e_{2t}^2, \ldots, e_{1t}^2 - e_{m+1,t}^2)'$ is serially uncorrelated (sufficient for which is that $(e_{1t}, \ldots, e_{m+1,t})'$ is i.i.d.), then a possible estimator of V is simply

$$\hat{V}^* = P^{-1} \sum_{t=R}^{T} \left(\hat{f}_{t+1} - \bar{f}\right)\left(\hat{f}_{t+1} - \bar{f}\right)'.$$

If the squared forecast errors display persistence (GARCH and all that), a robust estimator of the variance-covariance matrix should be used [Hueng (1999), West and Cho (1995)].

2. One can see in (4.9) that asymptotic irrelevance holds quite generally when $P/R \to 0$. The intuition is that the relatively large sample (big R) used to estimate β produces small uncertainty relative to uncertainty that would be present in the relatively small sample (small P) even if one knew β. The result was noted informally by Chong and Hendry (1986). Simulation evidence in West (1996, 2001), McCracken (2004) and Clark and McCracken (2001) suggests that $P/R < 0.1$ more or less justifies using the asymptotic approximation that assumes asymptotic irrelevance.

3. This fortunate cancellation of variance and covariance terms occurs for certain moments and loss functions, when estimates of parameters needed to make predictions are generated by the recursive scheme (but not by the rolling or fixed schemes), and when forecast errors are conditionally homoskedastic. These loss functions are: mean prediction error; serial correlation of one step ahead prediction errors; zero correlation between one model's forecast error and another model's forecast. This is illustrated in the discussion of Equation (7.2) below.

To repeat: When asymptotic irrelevance applies, one can proceed as described in Section 3. One need not account for dependence of forecasts on estimated parameter vectors. When asymptotic irrelevance does not apply, matters are more complicated. This is discussed in the next sections.

5. A small number of nonnested models, Part III

Asymptotic irrelevance fails in a number of important cases, at least according to the asymptotics of West (1996). Under the rolling and fixed schemes, it fails quite generally. For example, it fails for mean prediction error, correlation between realization and prediction, encompassing, and zero correlation in one step ahead prediction errors [West and McCracken (1998)]. Under the recursive scheme, it similarly fails for such moments when prediction errors are not conditionally homoskedastic. In such cases, asymptotic inference requires accounting for uncertainty about parameters used to make predictions.

The general result is as follows. One is interested in an $(m \times 1)$ vector of moments $\mathrm{E}f_t$, where f_t now depends on observable data through a $(k \times 1)$ unknown parameter vector β^*. If moments of predictions or prediction errors of competing sets of regressions are to be compared, the parameter vectors from the various regressions are stacked to form β^*. It is assumed that $\mathrm{E}f_t$ is differentiable in a neighborhood around β^*. Let $\hat{\beta}_t$ denote an estimate of β^* that relies on data from period t and earlier. Let $\tau \geqslant 1$ be the forecast horizon of interest; $\tau = 1$ has been assumed in the discussion so far. Let the total sample available be of size $T + \tau$. The estimate of $\mathrm{E}f_t$ is constructed as

$$\bar{f} = P^{-1} \sum_{t=R}^{T} f_{t+\tau}(\hat{\beta}_t) \equiv P^{-1} \sum_{t=R}^{T} \hat{f}_{t+\tau}. \tag{5.1}$$

The models are assumed to be parametric. The estimator of the regression parameters satisfies

$$\hat{\beta}_t - \beta^* = B(t)H(t), \tag{5.2}$$

where $B(t)$ is $k \times q$, $H(t)$ is $q \times 1$ with

(a) $B(t) \overset{\text{a.s.}}{\to} B$, B a matrix of rank k;

(b) $H(t) = t^{-1} \sum_{s=1}^{t} h_s(\beta^*)$ (recursive), $H(t) = R^{-1} \sum_{s=t-R+1}^{t} h_s(\beta^*)$ (rolling), $H(t) = R^{-1} \sum_{s=1}^{R} h_s(\beta^*)$ (fixed), for a $(q \times 1)$ orthogonality condition $h_s(\beta^*)$ orthogonality condition that satisfies

(c) $\mathrm{E}h_s(\beta^*) = 0$.

Here, h_t is the score if the estimation method is maximum likelihood, or the GMM orthogonality condition if GMM is the estimator. The matrix $B(t)$ is the inverse of the Hessian (ML) or linear combination of orthogonality conditions (GMM), with large sample counterpart B. In exactly identified models, $q = k$. Allowance for overidentified GMM models is necessary to permit prediction from the reduced form of simultaneous equations models, for example. For the results below, various moment and mixing conditions are required. See West (1996) and Giacomini and White (2003) for details.

It may help to pause to illustrate with linear least squares examples. For the least squares model (4.6a), in which $y_t = X_t' \beta^* + e_t$,

$$h_t = X_t e_t. \tag{5.3a}$$

In (4.6b), in which there are two models $y_t = X_{1t}' \beta_1^* + e_{1t}$, $y_t = X_{2t}' \beta_2^* + e_{2t}$, $\beta^* \equiv (\beta_1^{*\prime}, \beta_2^{*\prime})'$,

$$h_t = (X_{1t}' e_{1t}, X_{2t}' e_{2t})', \tag{5.3b}$$

where $h_t = h_t(\beta^*)$ is suppressed for simplicity. The matrix B is $k \times k$:

$$\begin{aligned} B &= (\mathrm{E}X_{1t}X_{1t}')^{-1} \quad \text{(model (4.6a))}, \\ B &= \operatorname{diag}\left[(\mathrm{E}X_{1t}X_{1t}')^{-1}, (\mathrm{E}X_{2t}X_{2t}')^{-1}\right] \quad \text{(model (4.6b))}. \end{aligned} \tag{5.4}$$

If one is comparing two models with $\mathrm{E}g_{it}$ and $\bar{g}_i$ the expected and sample mean performance measure for model i, $i = 1, 2$, then $\mathrm{E}f_t = \mathrm{E}g_{1t} - \mathrm{E}g_{2t}$ and $\bar{f} = \bar{g}_1 - \bar{g}_2$.

To return to the statement of results, which require conditions such as those in West (1996), and which are noted in the bullet points at the end of this section. Assume a large sample of both predictions and prediction errors,

$$P \to \infty, \quad R \to \infty, \quad \lim_{T\to\infty} \frac{P}{R} = \pi, \quad 0 \leqslant \pi < \infty. \tag{5.5}$$

An expansion of $\bar{f}$ around $\bar{f}^*$ yields

$$\sqrt{P}\big(\bar{f} - \mathrm{E}f_t\big) = \sqrt{P}\big(\bar{f}^* - \mathrm{E}f_t\big) + F(P/R)^{1/2}\big[BR^{1/2}\bar{H}\big] + \mathrm{o}_p(1) \tag{5.6}$$

which may also be written

$$\begin{aligned} &P^{-1/2}\sum_{t=R}^{T}\big[f\big(\hat{\beta}_{t+1}\big) - \mathrm{E}f_t\big] \\ &\quad = P^{-1/2}\sum_{t=R}^{T}\big[f_{t+1}\big(\beta^*\big) - \mathrm{E}f_t\big] + F(P/R)^{1/2}\big[BR^{1/2}\bar{H}\big] + \mathrm{o}_p(1). \end{aligned} \tag{5.6$'$}$$

The first term on the right-hand side of (5.6) and (5.6)$'$ – henceforth (5.6), for short – represents uncertainty that would be present even if predictions relied on the population value of the parameter vector β^*. The limiting distribution of this term was given in (3.1). The second term on the right-hand side of (5.6) results from reliance of predictions on estimates of β^*. To account for the effects of this second term requires yet more notation. Write the long run variance of $(f'_{t+1}, h'_t)'$ as

$$S = \begin{bmatrix} V^* & S_{fh} \\ S'_{fh} & S_{hh} \end{bmatrix}. \tag{5.7}$$

Here, $V^* \equiv \sum_{j=-\infty}^{\infty} \mathrm{E}(f_t - \mathrm{E}f_t)(f_{t-j} - \mathrm{E}f_t)'$ is $m \times m$, $S_{fh} = \sum_{j=-\infty}^{\infty} \mathrm{E}(f_t - \mathrm{E}f_t)h'_{t-j}$ is $m \times k$, and $S_{hh} \equiv \sum_{j=-\infty}^{\infty} \mathrm{E}h_t h'_{t-j}$ is $k \times k$, and f_t and h_t are understood to be evaluated at β^*. The asymptotic ($R \to \infty$) variance–covariance matrix of the estimator of β^* is

$$V_\beta \equiv BS_{hh}B'. \tag{5.8}$$

With π defined in (5.5), define the scalars λ_{fh}, λ_{hh} and $\lambda \equiv (1 + \lambda_{hh} - 2\lambda_{fh})$, as in the following table:

Sampling scheme	λ_{fh}	λ_{hh}	λ
Recursive	$1 - \pi^{-1}\ln(1+\pi)$	$2[1 - \pi^{-1}\ln(1+\pi)]$	1
Rolling, $\pi \leqslant 1$	$\frac{\pi}{2}$	$\pi - \frac{\pi^2}{3}$	$1 - \frac{\pi^2}{3}$
Rolling, $\pi > 1$	$1 - \frac{1}{2\pi}$	$1 - \frac{1}{3\pi}$	$\frac{2}{3\pi}$
Fixed	0	π	$1 + \pi$

(5.9)

Finally, define the $m \times k$ matrix F as in (4.8), $F \equiv \partial \mathrm{E}f_t(\beta^*)/\partial\beta$.

Then $P^{-1/2}\sum_{t=R}^{T}[f(\hat{\beta}_{t+1}) - \mathrm{E}f_t]$ is asymptotically normal with variance-covariance matrix

$$V = V^* + \lambda_{fh}\big(FBS'_{fh} + S_{fh}B'F'\big) + \lambda_{hh}FV_\beta F'. \tag{5.10}$$

V^* is the long run variance of $P^{-1/2}\big[\sum_{t=R}^{T} f_{t+1}(\beta^*) - \mathrm{E}f_t\big]$ and is the same object as V^* defined in (3.1), $\lambda_{hh}FV_\beta F'$ is the long run variance of $F(P/R)^{1/2}[BR^{1/2}\bar{H}]$, and $\lambda_{fh}(FBS'_{fh} + S_{fh}B'F')$ is the covariance between the two.

This completes the statement of the general result. To illustrate the expansion (5.6) and the asymptotic variance (5.10), I will temporarily switch from my example of comparison of MSPEs to one in which one is looking at mean prediction error. The variable f_t is thus redefined to equal the prediction error, $f_t = e_t$, and $\mathrm{E}f_t$ is the moment of interest. I will further use a trivial example, in which the only predictor is the constant term, $y_t = \beta^* + e_t$. Let us assume as well, as in the Hoffman and Pagan (1989) and Ghysels and Hall (1990) analyses of predictive tests of instrument-residual orthogonality, that the fixed scheme is used and predictions are made using a single estimate of β^*. This single estimate is the least squares estimate on the sample running from 1 to R, $\hat{\beta}_R \equiv R^{-1}\sum_{s=1}^{R} y_s$. Now, $\hat{e}_{t+1} = e_{t+1} - (\hat{\beta}_R - \beta^*) = e_{t+1} - R^{-1}\sum_{s=1}^{R} e_s$. So

$$P^{-1/2}\sum_{t=R}^{T}\hat{e}_{t+1} = P^{-1/2}\sum_{t=R}^{T} e_{t+1} - (P/R)^{1/2}\left(R^{-1/2}\sum_{s=1}^{R} e_s\right). \tag{5.11}$$

This is in the form (4.9) or (5.6)$'$, with: $F = -1$, $R^{-1/2}\sum_{s=1}^{R} e_s$ = [$\mathrm{O}_p(1)$ terms due to the sequence of estimates of β^*], $B \equiv 1$, $\bar{H} = (R^{-1}\sum_{s=1}^{R} e_s)$ and the $\mathrm{o}_p(1)$ term identically zero.

If e_t is well behaved, say i.i.d. with finite variance σ^2, the bivariate vector $(P^{-1/2}\sum_{t=R}^{T} e_{t+1}, R^{-1/2}\sum_{s=1}^{R} e_s)'$ is asymptotically normal with variance covariance matrix $\sigma^2 I_2$. It follows that

$$P^{-1/2}\sum_{t=R}^{T} e_{t+1} - (P/R)^{1/2}\left(R^{-1/2}\sum_{s=1}^{R} e_s\right) \sim_A \mathrm{N}\big(0, (1+\pi)\sigma^2\big). \tag{5.12}$$

The variance in the normal distribution is in the form (5.10), with $\lambda_{fh} = 0$, $\lambda_{hh} = \pi$, $V^* = FV_\beta F' = \sigma^2$. Thus, use of $\hat{\beta}_R$ rather than β^* in predictions inflates the asymptotic variance of the estimator of mean prediction error by a factor of $1 + \pi$.

In general, when uncertainty about β^* matters asymptotically, the adjustment to the standard error that would be appropriate if predictions were based on population rather than estimated parameters is increasing in:

- The ratio of number of predictions P to number of observations in smallest regression sample R. Note that in (5.10) as $\pi \to 0$, $\lambda_{fh} \to 0$ and $\lambda_{hh} \to 0$; in the specific example (5.12) we see that if P/R is small, the implied value of π is small and the adjustment to the usual asymptotic variance of σ^2 is small; otherwise the adjustment can be big.

- The variance–covariance matrix of the estimator of the parameters used to make predictions.

Both conditions are intuitive. Simulations in West (1996, 2001), West and McCracken (1998), McCracken (2000), Chao, Corradi and Swanson (2001) and Clark and McCracken (2001, 2003) indicate that with plausible parameterizations for P/R and uncertainty about β^*, failure to adjust the standard error can result in very substantial size distortions. It is possible that $V < V^*$ – that is, accounting for uncertainty about regression parameters may *lower* the asymptotic variance of the estimator.[4] This happens in some leading cases of practical interest when the rolling scheme is used. See the discussion of Equation (7.2) below for an illustration.

A consistent estimator of V results from using the obvious sample analogues. A possibility is to compute λ_{fh} and λ_{hh} from (5.10) setting $\pi = P/R$. (See Table 1 for the implied formulas for λ_{fh}, λ_{hh} and λ.) As well, one can estimate F from the sample average of $\partial f(\hat{\beta}_t)/\partial\beta$, $\hat{F} = P^{-1}\sum_{t=R}^{T}\partial f(\hat{\beta}_t)/\partial\beta$;[5] estimate V_β and B from one of the sequence of estimates of β^*. For example, for mean prediction error, for the fixed scheme, one might set

$$\hat{F} = -P^{-1}\sum_{t=R}^{T} X'_{t+1}, \quad \hat{B} = \left(R^{-1}\sum_{s=1}^{R} X_s X'_s\right)^{-1},$$

Table 1
Sample analogues for λ_{fh}, λ_{hh} and λ

	Recursive	Rolling, $P \leqslant R$	Rolling, $P > R$	Fixed
λ_{fh}	$1 - \frac{R}{P}\ln(1+\frac{P}{R})$	$\frac{1}{2}\frac{P}{R}$	$1 - \frac{1}{2}\frac{R}{P}$	0
λ_{hh}	$2[1 - \frac{R}{P}\ln(1+\frac{P}{R})]$	$\frac{P}{R} - \frac{1}{3}\frac{P^2}{R^2}$	$1 - \frac{1}{3}\frac{R}{P}$	$\frac{P}{R}$
λ	1	$1 - \frac{1}{3}\frac{P^2}{R^2}$	$\frac{2R}{3P}$	$1 + \frac{P}{R}$

Notes:
1. The recursive, rolling and fixed schemes are defined in Section 4 and illustrated for an AR(1) in Equation (4.2).
2. P is the number of predictions, R the size of the smallest regression sample. See Section 4 and Equation (4.1).
3. The parameters λ_{fh}, λ_{hh} and λ are used to adjust the asymptotic variance covariance matrix for uncertainty about regression parameters used to make predictions. See Section 5 and Tables 2 and 3.

[4] Mechanically, such a fall in asymptotic variance indicates that the variance of terms resulting from estimation of β^* is more than offset by a negative covariance between such terms and terms that would be present even if β^* were known.

[5] See McCracken (2000) for an illustration of estimation of F for a non-differentiable function.

$$\hat{V}_\beta \equiv \left(R^{-1}\sum_{s=1}^{R} X_s X_s'\right)^{-1}\left(R^{-1}\sum_{s=1}^{R} X_s X_s'\hat{e}_s^2\right)\left(R^{-1}\sum_{s=1}^{R} X_s X_s'\right)^{-1}.$$

Here, $\hat{e}_s$, $1 \leqslant s \leqslant R$, is the in-sample least squares residual associated with the parameter vector $\hat{\beta}_R$ that is used to make predictions and the formula for $\hat{V}_\beta$ is the usual heteroskedasticity consistent covariance matrix for $\hat{\beta}_R$. (Other estimators are also consistent, for example sample averages running from 1 to T.) Finally, one can combine these with an estimate of the long run variance S constructed using a heteroskedasticity and autocorrelation consistent covariance matrix estimator [Newey and West (1987, 1994), Andrews (1991), Andrews and Monahan (1994), den Haan and Levin (2000)].

Alternatively, one can compute a smaller dimension long run variance as follows. Let us assume for the moment that f_t and hence V are scalar. Define the (2×1) vector $\hat{g}_t$ as

$$\hat{g}_t = \begin{bmatrix} \hat{f}_t \\ \hat{F}\hat{B}\hat{h}_t \end{bmatrix}. \tag{5.13}$$

Let g_t be the population counterpart of $\hat{g}_t$, $g_t \equiv (f_t, FBh_t)'$. Let Ω be the (2×2) long run variance of g_t, $\Omega \equiv \sum_{j=-\infty}^{\infty} \mathrm{E}g_t g_{t-j}'$. Let $\hat{\Omega}$ be an estimate of Ω. Let $\hat{\Omega}_{ij}$ be the (i, j) element of $\hat{\Omega}$. Then one can consistently estimate V with

$$\hat{V} = \hat{\Omega}_{11} + 2\lambda_{fh}\hat{\Omega}_{12} + \lambda_{hh}\hat{\Omega}_{22}. \tag{5.14}$$

The generalization to vector f_t is straightforward. Suppose f_t is say $m \times 1$ for $m \geqslant 1$. Then

$$\hat{g}_t = \begin{bmatrix} f_t \\ FBh_t \end{bmatrix}.$$

is $2m \times 1$, as is $\hat{g}_t$; Ω and $\hat{\Omega}$ are $2m \times 2m$. One divides $\hat{\Omega}$ into four $(m \times m)$ blocks, and computes

$$\hat{V} = \hat{\Omega}(1,1) + \lambda_{fh}\big[\hat{\Omega}(1,2) + \hat{\Omega}(2,1)\big] + \lambda_{hh}\hat{\Omega}(2,2). \tag{5.15}$$

In (5.15), $\hat{\Omega}(1,1)$ is the $m \times m$ block in the upper left hand corner of $\hat{\Omega}$, $\hat{\Omega}(1,2)$ is the $m \times m$ block in the upper right hand corner of $\hat{\Omega}$, and so on.

Alternatively, in some common problems, and if the models are linear, regression based tests can be used. By judicious choice of additional regressors [as suggested for in-sample tests by Pagan and Hall (1983), Davidson and MacKinnon (1984) and Wooldridge (1990)], one can "trick" standard regression packages into computing standard errors that properly reflect uncertainty about β^*. See West and McCracken (1998) and Table 3 below for details, Hueng and Wong (2000), Avramov (2002) and Ferreira (2004) for applications.

Conditions for the expansion (5.6) and the central limit result (5.10) include the following.

- Parametric models and estimators of β are required. Similar results may hold with nonparametric estimators, but, if so, these have yet to be established. Linearity is not required. One might be basing predictions on nonlinear time series models, for example, or restricted reduced forms of simultaneous equations models estimated by GMM.
- At present, results with I(1) data are restricted to linear models [Corradi, Swanson and Olivetti (2001), Rossi (2003)]. Asymptotic irrelevance continues to apply when $F = 0$ or $\pi = 0$. When those conditions fail, however, the normalized estimator of $\mathrm{E} f_t$ typically is no longer asymptotically normal. (By I(1) data, I mean I(1) data entered in levels in the regression model. Of course, if one induces stationarity by taking differences or imposing cointegrating relationships prior to estimating β^*, the theory in the present section is applicable quite generally.)
- Condition (5.5) holds. Section 7 discusses implications of an alternative asymptotic approximation due to Giacomini and White (2003) that holds R fixed.
- For the recursive scheme, condition (5.5) can be generalized to allow $\pi = \infty$, with the same asymptotic approximation. (Recall that π is the limiting value of P/R.) Since $\pi < \infty$ has been assumed in existing theoretical results for rolling and fixed, researchers using those schemes should treat the asymptotic approximation with extra caution if $P \gg R$.
- The expectation of the loss function f must be differentiable in a neighborhood of β^*. This rules out direction of change as a loss function.
- A full rank condition on the long run variance of $(f'_{t+1}, (Bh_t)')'$. A necessary condition is that the long run variance of f_{t+1} is full rank. For MSPE, and i.i.d. forecast errors, this means that the variance of $e^2_{1t} - e^2_{2t}$ is positive (note the absence of a "$\hat{}$" over e^2_{1t} and e^2_{2t}). This condition will fail in applications in which the models are nested, for in that case $e_{1t} \equiv e_{2t}$. Of course, for the sample forecast errors, $\hat{e}_{1t} \neq \hat{e}_{2t}$ (note the "$\hat{}$") because of sampling error in estimation of β^*_1 and β^*_2. So the failure of the rank condition may not be apparent in practice. McCracken's (2004) analysis of nested models shows that under the conditions of the present section apart from the rank condition, $\sqrt{P}(\hat{\sigma}^2_1 - \hat{\sigma}^2_2) \rightarrow_p 0$. The next two sections discuss inference for predictions from such nested models.

6. A small number of models, nested: MSPE

Analysis of nested models per se does not invalidate the results of the previous sections. A rule of thumb is: if the rank of the data becomes degenerate when regression parameters are set at their population values, then a rank condition assumed in the previous sections likely is violated. When only two models are being compared, "degenerate" means identically zero.

Consider, as an example, out of sample tests of Granger causality [e.g., Stock and Watson (1999, 2002)]. In this case, model 2 might be a bivariate VAR, model 1 a univariate AR that is nested in model 2 by imposing suitable zeroes in the model 2 regression vector. If the lag length is 1, for example:

$$\text{Model 1:}\quad y_t = \beta_{10} + \beta_{11} y_{t-1} + e_{1t} \equiv X_{1t}' \beta_1^* + e_{1t}, \quad X_{1t} \equiv (1, y_{t-1})', \quad \beta_1^* \equiv (\beta_{10}, \beta_{11})'; \tag{6.1a}$$

$$\text{Model 2:}\quad y_t = \beta_{20} + \beta_{21} y_{t-1} + \beta_{22} x_{t-1} + e_{2t} \equiv X_{2t}' \beta_2^* + e_{2t}, \quad X_{2t} \equiv (1, y_{t-1}, x_{t-1})', \quad \beta_2^* \equiv (\beta_{20}, \beta_{21}, \beta_{22})'. \tag{6.1b}$$

Under the null of no Granger causality from x to y, $\beta_{22} = 0$ in model 2. Model 1 is then nested in model 2. Under the null, then,

$$\beta_2^{*\prime} = \left(\beta_1^{*\prime}, 0\right), \quad X_{1t}' \beta_1^* = X_{2t}' \beta_2^*,$$

and the disturbances of model 2 and model 1 are identical: $e_{2t}^2 - e_{1t}^2 \equiv 0$, $e_{1t}(e_{1t} - e_{2t}) = 0$ and $|e_{1t}| - |e_{2t}| = 0$ for all t. So the theory of the previous sections does not apply if MSPE, $\text{cov}(e_{1t}, e_{1t} - e_{2t})$ or mean absolute error is the moment of interest. On the other hand, the random variable $e_{1t+1} x_t$ is nondegenerate under the null, so one can use the theory of the previous sections to examine whether $\text{E} e_{1t+1} x_t = 0$. Indeed, Chao, Corradi and Swanson (2001) show that (5.6) and (5.10) apply when testing $\text{E} e_{1t+1} x_t = 0$ with out of sample prediction errors.

The remainder of this section considers the implications of a test that does fail the rank condition of the theory of the previous section – specifically, MSPE in nested models. This is a common occurrence in papers on forecasting asset prices, which often use MSPE to test a random walk null against models that use past data to try to predict changes in asset prices. It is also a common occurrence in macro applications, which, as in example (6.1), compare univariate to multivariate forecasts. In such applications, the asymptotic results described in the previous section will no longer apply. In particular, and under essentially the technical conditions of that section (apart from the rank condition), when $\hat{\sigma}_1^2 - \hat{\sigma}_2^2$ is normalized so that its limiting distribution is non-degenerate, that distribution is non-normal.

Formal characterization of limiting distributions has been accomplished in McCracken (2004) and Clark and McCracken (2001, 2003, 2005a, 2005b). This characterization relies on restrictions not required by the theory discussed in the previous section. These restrictions include:

(6.2a) The objective function used to estimate regression parameters must be the same quadratic as that used to evaluate prediction. That is:

- The estimator must be nonlinear least squares (ordinary least squares of course a special case).
- For multistep predictions, the "direct" rather than "iterated" method must be used.[6]

[6] To illustrate these terms, consider the univariate example of forecasting $y_{t+\tau}$ using y_t, assuming that mathematical expectations and linear projections coincide. The objective function used to evaluate predictions is $\text{E}[y_{t+\tau} - \text{E}(y_{t+\tau} \mid y_t)]^2$. The "direct" method estimates $y_{t+\tau} = y_t \gamma + u_{t+\tau}$ by least squares, uses $y_t \hat{\gamma}_t$

(6.2b) A pair of models is being compared. That is, results have not been extended to multi-model comparisons along the lines of (3.3).

McCracken (2004) shows that under such conditions, $\sqrt{P}(\hat{\sigma}_1^2 - \hat{\sigma}_2^2) \rightarrow_p 0$, and derives the asymptotic distribution of $P(\hat{\sigma}_1^2 - \hat{\sigma}_2^2)$ and certain related quantities. (Note that the normalizing factor is the prediction sample size P rather than the usual $\sqrt{P}$.) He writes test statistics as functionals of Brownian motion. He establishes limiting distributions that are asymptotically free of nuisance parameters under certain additional conditions:

(6.2c) one step ahead predictions and conditionally homoskedastic prediction errors, or

(6.2d) the number of additional regressors in the larger model is exactly 1 [Clark and McCracken (2005a)].

Condition (6.2d) allows use of the results about to be cited, in conditionally heteroskedastic as well as conditionally homoskedastic environments, and for multiple as well as one step ahead forecasts. Under the additional restrictions (6.2c) or (6.2d), McCracken (2004) tabulates the quantiles of $P(\hat{\sigma}_1^2 - \hat{\sigma}_2^2)/\hat{\sigma}_2^2$. These quantiles depend on the number of additional parameters in the larger model and on the limiting ratio of P/R. For conciseness, I will use "(6.2)" to mean

Conditions (6.2a) and (6.2b) hold, as does either or both of conditions (6.2c) and (6.2d). (6.2)

Simulation evidence in Clark and McCracken (2001, 2003, 2005b), McCracken (2004), Clark and West (2005a, 2005b) and Corradi and Swanson (2005) indicates that in MSPE comparisons in nested models the usual statistic (4.5) is non-normal not only in a technical but in an essential practical sense: use of standard critical values usually results in very poorly sized tests, with *far* too few rejections. As well, the usual statistic has very poor power. For both size and power, the usual statistic performs worse the larger the number of irrelevant regressors included in model 2. The evidence relies on one-sided tests, in which the alternative to H$_0$: $\mathrm{E}e_{1t}^2 - \mathrm{E}e_{2t}^2 = 0$ is

$$\mathrm{H_A}: \quad \mathrm{E}e_{1t}^2 - \mathrm{E}e_{2t}^2 > 0. \tag{6.3}$$

Ashley, Granger and Schmalensee (1980) argued that in nested models, the alternative to equal MSPE is that the larger model outpredicts the smaller model: it does not make sense for the population MSPE of the parsimonious model to be smaller than that of the larger model.

to forecast, and computes a sample average of $(y_{t+\tau} - y_t\hat{\gamma}_t)^2$. The "iterated" method estimates $y_{t+1} = y_t\beta + e_{t+1}$, uses $y_t(\hat{\beta}_t)^\tau$ to forecast, and computes a sample average of $[y_{t+\tau} - y_t(\hat{\beta}_t)^\tau]^2$. Of course, if the AR(1) model for y_t is correct, then $\gamma = \beta^\tau$ and $u_{t+\tau} = e_{t+\tau} + \beta e_{t+\tau-1} + \cdots + \beta^{\tau-1}e_{t+1}$. But if the AR(1) model is incorrect, the two forecasts may differ, even in a large sample. See Ing (2003) and Marcellino, Stock and Watson (2004) for theoretical and empirical comparison of direct and iterated methods.

To illustrate the sources of these results, consider the following simple example. The two models are:

$$\text{Model 1: } y_t = e_t; \quad \text{Model 2: } y_t = \beta^* x_t + e_t; \quad \beta^* = 0;$$
$$e_t \text{ a martingale difference sequence with respect to past } y\text{'s and } x\text{'s.} \tag{6.4}$$

In (6.4), all variables are scalars. I use x_t instead of X_{2t} to keep notation relatively uncluttered. For concreteness, one can assume $x_t = y_{t-1}$, but that is not required. I write the disturbance to model 2 as e_t rather than e_{2t} because the null (equal MSPE) implies $\beta^* = 0$ and hence that the disturbance to model 2 is identically equal to e_t. Nonetheless, for clarity and emphasis I use the "2" subscript for the sample forecast error from model 2, $\hat{e}_{2t+1} \equiv y_{t+1} - x_{t+1}\hat{\beta}_t$. In a finite sample, the model 2 sample forecast error differs from the model 1 forecast error, which is simply y_{t+1}. The model 1 and model 2 MSPEs are

$$\hat{\sigma}_1^2 \equiv P^{-1}\sum_{t=R}^{T} y_{t+1}^2, \quad \hat{\sigma}_2^2 \equiv P^{-1}\sum_{t=R}^{T} \hat{e}_{2t+1}^2 \equiv P^{-1}\sum_{t=R}^{T}\left(y_{t+1} - x_{t+1}\hat{\beta}_t\right)^2. \tag{6.5}$$

Since

$$\hat{f}_{t+1} \equiv y_{t+1}^2 - \left(y_{t+1} - x_{t+1}\hat{\beta}_t\right)^2 = 2y_{t+1}x_{t+1}\hat{\beta}_t - \left(x_{t+1}\hat{\beta}_t\right)^2$$

we have

$$\bar{f} \equiv \hat{\sigma}_1^2 - \hat{\sigma}_2^2 = 2\left(P^{-1}\sum_{t=R}^{T} y_{t+1}x_{t+1}\hat{\beta}_t\right) - \left[P^{-1}\sum_{t=R}^{T}\left(x_{t+1}\hat{\beta}_t\right)^2\right]. \tag{6.6}$$

Now,

$$-\left[P^{-1}\sum_{t=R}^{T}\left(x_{t+1}\hat{\beta}_t\right)^2\right] \leqslant 0$$

and under the null ($y_{t+1} = e_{t+1} \sim$ i.i.d.)

$$2\left(P^{-1}\sum_{t=R}^{T} y_{t+1}x_{t+1}\hat{\beta}_t\right) \approx 0.$$

So under the null it will generally be the case that

$$\bar{f} \equiv \hat{\sigma}_1^2 - \hat{\sigma}_2^2 < 0 \tag{6.7}$$

or: the *sample* MSPE from the null model will tend to be *less* than that from the alternative model.

The intuition will be unsurprising to those familiar with forecasting. If the null is true, the alternative model introduces noise into the forecasting process: the alternative model attempts to estimate parameters that are zero in population. In finite samples, use of the noisy estimate of the parameter will *raise* the estimated MSPE of the alternative model relative to the null model. So if the null is true, the model 1 MSPE should be smaller by the amount of estimation noise.

To illustrate concretely, let me use the simulation results in Clark and West (2005b). As stated in (6.3), one tailed tests were used. That is, the null of equal MSPE is rejected at (say) the 10 percent level only if the alternative model predicts better than model 1:

$$\begin{aligned}
&\bar{f}/\big[\hat{V}^*/P\big]^{1/2} = \big(\hat{\sigma}_1^2 - \hat{\sigma}_2^2\big)/\big[\hat{V}^*/P\big]^{1/2} > 1.282,\\
&\quad \hat{V}^* = \text{estimate of long run variance of } \hat{\sigma}_1^2 - \hat{\sigma}_2^2, \text{ say,}\\
&\quad \hat{V}^* = P^{-1}\sum_{t=R}^{T}\big(\hat{f}_{t+1} - \bar{f}\big)^2 = P^{-1}\sum_{t=R}^{T}\big[\hat{f}_{t+1} - \big(\hat{\sigma}_1^2 - \hat{\sigma}_2^2\big)\big]^2 \quad \text{if } e_t \text{ is i.i.d.}
\end{aligned} \tag{6.8}$$

Since (6.8) is motivated by an asymptotic approximation in which $\hat{\sigma}_1^2 - \hat{\sigma}_2^2$ is centered around zero, we see from (6.7) that the test will tend to be undersized (reject too infrequently). Across 48 sets of simulations, with DGPs calibrated to match key characteristics of asset price data, Clark and West (2005b) found that the median size of a nominal 10% test using the standard result (6.8) was less than 1%. The size was better with bigger R and worse with bigger P. (Some alternative procedures (described below) had median sizes of 8–13%.) The power of tests using "standard results" was poor: rejection of about 9%, versus 50–80% for alternatives.[7] Non-normality also applies if one normalizes differences in MSPEs by the unrestricted MSPE to produce an out of sample F-test. See Clark and McCracken (2001, 2003), and McCracken (2004) for analytical and simulation evidence of marked departures from normality.

Clark and West (2005a, 2005b) suggest adjusting the difference in MSPEs to account for the noise introduced by the inclusion of irrelevant regressors in the alternative model. If the null model has a forecast $\hat{y}_{1t+1}$, then (6.6), which assumes $\hat{y}_{1t+1} = 0$, generalizes to

$$\hat{\sigma}_1^2 - \hat{\sigma}_2^2 = -2P^{-1}\sum_{t=R}^{T}\hat{e}_{1t+1}\big(\hat{y}_{1t+1} - \hat{y}_{2t+1}\big) - P^{-1}\sum_{t=R}^{T}\big(\hat{y}_{1t+1} - \hat{y}_{2t+1}\big)^2. \tag{6.9}$$

To yield a statistic better centered around zero, Clark and West (2005a, 2005b) propose adjusting for the negative term $-P^{-1}\sum_{t=R}^{T}(\hat{y}_{1t+1} - \hat{y}_{2t+1})^2$. They call the result *MSPE-adjusted*:

$$\begin{aligned}
&P^{-1}\sum_{t=R}^{T}\hat{e}_{1t+1}^2 - \left[P^{-1}\sum_{t=R}^{T}\hat{e}_{2t+1}^2 - P^{-1}\sum_{t=R}^{T}\big(\hat{y}_{1t+1} - \hat{y}_{2t+1}\big)^2\right]\\
&\quad \equiv \hat{\sigma}_1^2 - \big(\hat{\sigma}_2^2\text{-adj}\big).
\end{aligned} \tag{6.10}$$

[7] Note that (4.5) and the left-hand side of (6.8) are identical, but that Section 4 recommends the use of (4.5) while the present section recommends against use of (6.8). At the risk of beating a dead horse, the reason is that Section 4 assumed that models are non-nested, while the present section assumes that they are nested.

$\hat{\sigma}_2^2$-adj, which is smaller than $\hat{\sigma}_2^2$ by construction, can be thought of as the MSPE from the larger model, adjusted downwards for estimation noise attributable to inclusion of irrelevant parameters.

Viable approaches to testing equal MSPE in nested models include the following (with the first two summarizing the previous paragraphs):

1. Under condition (6.2), use critical values from Clark and McCracken (2001) and McCracken (2004), [e.g., Lettau and Ludvigson (2001)].
2. Under condition (6.2), or when the null model is a martingale difference, adjust the differences in MSPEs as in (6.10), and compute a standard error in the usual way. The implied t-statistic can be obtained by regressing $\hat{e}_{1t+1}^2 - [\hat{e}_{2t+1}^2 - (\hat{y}_{1t+1} - \hat{y}_{2t+1})^2]$ on a constant and computing the t-statistic for a coefficient of zero. Clark and West (2005a, 2005b) argue that standard normal critical values are approximately correct, even though the statistic is non-normal according to asymptotics of Clark and McCracken (2001).

   It remains to be seen whether the approaches just listed in points 1 and 2 perform reasonably well in more general circumstances – for example, when the larger model contains several extra parameters, and there is conditional heteroskedasticity. But even if so other procedures are possible.
3. If $P/R \to 0$, Clark and McCracken (2001) and McCracken (2004) show that asymptotic irrelevance applies. So for small P/R, use standard critical values [e.g., Clements and Galvao (2004)]. Simulations in various papers suggest that it generally does little harm to ignore effects from estimation of regression parameters if $P/R \leqslant 0.1$. Of course, this cutoff is arbitrary. For some data, a larger value is appropriate, for others a smaller value.
4. For MSPE and one step ahead forecasts, use the standard test if it rejects: if the standard test rejects, a properly sized test most likely will as well [e.g., Shintani (2004)].[8]
5. Simulate/bootstrap your own standard errors [e.g., Mark (1995), Sarno, Thornton and Valente (2005)]. Conditions for the validity of the bootstrap are established in Corradi and Swanson (2005).

Alternatively, one can swear off MSPE. This is discussed in the next section.

7. A small number of models, nested, Part II

Leading competitors of MSPE for the most part are encompassing tests of various forms. Theoretical results for the first two statistics listed below require condition (6.2),

[8] The restriction to one step ahead forecasts is for the following reason. For multiple step forecasts, the difference between model 1 and model 2 MSPEs presumably has a negative expectation. And simulations in Clark and McCracken (2003) generally find that use of standard critical values results in too few rejections. But sometimes there are too many rejections. This apparently results because of problems with HAC estimation of the standard error of the MSPE difference (private communication from Todd Clark).

and are asymptotically non-normal under those conditions. The remaining statistics are asymptotically normal, and under conditions that do not require (6.2).

1. Of various variants of encompassing tests, Clark and McCracken (2001) find that power is best using the Harvey, Leybourne and Newbold (1998) version of an encompassing test, normalized by unrestricted variance. So for those who use a non-normal test, Clark and McCracken (2001) recommend the statistic that they call "Enc-new":

$$\text{Enc-new} = \bar{f} = \frac{P^{-1}\sum_{t=R}^{T} \hat{e}_{1t+1}(\hat{e}_{1t+1} - \hat{e}_{2t+1})}{\hat{\sigma}_2^2},$$
$$\hat{\sigma}_2^2 \equiv P^{-1}\sum_{t=R}^{T} \hat{e}_{2t+1}^2. \tag{7.1}$$

2. It is easily seen that MSPE-adjusted (6.10) is algebraically identical to $2P^{-1} \times \sum_{t=R}^{T} \hat{e}_{1t+1}(\hat{e}_{1t+1} - \hat{e}_{2t+1})$. This is the sample moment for the Harvey, Leybourne and Newbold (1998) encompassing test (4.7d). So the conditions described in point (2) at the end of the previous section are applicable.
3. Test whether model 1's prediction error is uncorrelated with model 2's predictors or the subset of model 2's predictors not included in model 1 [Chao, Corradi and Swanson (2001)], $f_t = e_{1t}X'_{2t}$ in our linear example or $f_t = e_{1t}x_{t-1}$ in example (6.1). When both models use estimated parameters for prediction (in contrast to (6.4), in which model 1 does not rely on estimated parameters), the Chao, Corradi and Swanson (2001) procedure requires adjusting the variance–covariance matrix for parameter estimation error, as described in Section 5. Chao, Corradi and Swanson (2001) relies on the less restricted environment described in the section on nonnested models; for example, it can be applied in straightforward fashion to joint testing of multiple models.
4. If $\beta_2^* \neq 0$, apply an encompassing test in the form (4.7c), $0 = \mathrm{E}e_{1t}X'_{2t}\beta_2^*$. Simulation evidence to date indicates that in samples of size typically available, this statistic performs poorly with respect to both size and power [Clark and McCracken (2001), Clark and West (2005a)]. But this statistic also neatly illustrates some results stated in general terms for nonnested models. So to illustrate those results: With computation and technical conditions similar to those in West and McCracken (1998), it may be shown that when $\bar{f} = P^{-1}\sum_{t=R}^{T} \hat{e}_{1t+1}X'_{2t+1}\hat{\beta}_{2t}$, $\beta_2^* \neq 0$, and the models are nested, then

$$\sqrt{P}\bar{f} \sim_A \mathrm{N}(0, V), \qquad V \equiv \lambda V^*, \qquad \lambda \text{ defined in (5.9)},$$
$$V^* \equiv \sum_{j=-\infty}^{\infty} \mathrm{E}e_t e_{t-j}\left(X'_{2t}\beta_2^*\right)\left(X'_{2t-j}\beta_2^*\right). \tag{7.2}$$

Given an estimate of V^*, one multiplies the estimate by λ to obtain an estimate of the asymptotic variance of $\sqrt{P}\bar{f}$. Alternatively, one divides the t-statistic by $\sqrt{\lambda}$.

Observe that $\lambda = 1$ for the recursive scheme: this is an example in which there is the cancellation of variance and covariance terms noted in point 3 at the end of Section 4. For the fixed scheme, $\lambda > 1$, with λ increasing in P/R. So uncertainty about parameter estimates inflates the variance, with the inflation factor increasing in the ratio of the size of the prediction to regression sample. Finally, for the rolling scheme $\lambda < 1$. So use of (6.8) will result in *smaller* standard errors and larger t-statistics than would use of a statistic that ignores the effect of uncertainty about β^*. The magnitude of the adjustment to standard errors and t-statistics is increasing in the ratio of the size of the prediction to regression sample.

5. If $\beta_2^* = 0$, and if the rolling or fixed (but *not* the recursive) scheme is used, apply the encompassing test just discussed, setting $\bar{f} = P^{-1}\sum_{t=R}^{T} e_{1t+1}X_{2t+1}'\hat{\beta}_{2t}$. Note that in contrast to the discussion just completed, there is no "ˆ" over e_{1t+1}: because model 1 is nested in model 2, $\beta_2^* = 0$ means $\beta_1^* = 0$, so $e_{1t+1} = y_{t+1}$ and e_{1t+1} is observable. One can use standard results – asymptotic irrelevance applies. The factor of λ that appears in (7.2) resulted from estimation of β_1^*, and is now absent. So $V = V^*$; if, for example, e_{1t} is i.i.d., one can consistently estimate V with $\hat{V} = P^{-1}\sum_{t=R}^{T}(e_{1t+1}X_{2t+1}'\hat{\beta}_{2t})^2$.[9]
6. If the rolling or fixed regression scheme is used, construct a conditional rather than unconditional test [Giacomini and White (2003)]. This paper makes both methodological and substantive contributions. The methodological contributions are twofold. First, the paper explicitly allows data heterogeneity (e.g., slow drift in moments). This seems to be a characteristic of much economic data. Second, while the paper's conditions are broadly similar to those of the work cited above, its asymptotic approximation holds R fixed while letting $P \to \infty$.

   The substantive contribution is also twofold. First, the objects of interest are moments of $\hat{e}_{1t}$ and $\hat{e}_{2t}$ rather than e_t. (Even in nested models, $\hat{e}_{1t}$ and $\hat{e}_{2t}$ are distinct because of sampling error in estimation of regression parameters used to make forecasts.) Second, and related, the moments of interest are conditional ones, say $\mathrm{E}(\hat{\sigma}_1^2 - \hat{\sigma}_2^2 \mid \text{lagged } y\text{'s and } x\text{'s})$. The Giacomini and White (2003) framework allows general conditional loss functions, and may be used in nonnested as well as nested frameworks.

[9] The reader may wonder whether asymptotic normality violates the rule of thumb enunciated at the beginning of this section, because $f_t = e_{1t}X_{2t}'\beta_2^*$ is identically zero when evaluated at population $\beta_2^* = 0$. At the risk of confusing rather than clarifying, let me briefly note that the rule of thumb still applies, but only with a twist on the conditions given in the previous section. This twist, which is due to Giacomini and White (2003), holds R fixed as the sample size grows. Thus in population the random variable of interest is $f_t = e_{1t}X_{2t}'\hat{\beta}_{2t}$, which for the fixed or rolling schemes is nondegenerate for all t. (Under the recursive scheme, $\hat{\beta}_{2t} \to_p 0$ as $t \to \infty$, which implies that f_t is degenerate for large t.) It is to be emphasized that technical conditions (R fixed vs. $R \to \infty$) are not arbitrary. Reasonable technical conditions should reasonably rationalize finite sample behavior. For tests of equal MSPE discussed in the previous section, a vast range of simulation evidence suggests that the $R \to \infty$ condition generates a reasonably accurate asymptotic approximation (i.e., non-normality is implied by the theory and is found in the simulations). The more modest array of simulation evidence for the R fixed approximation suggests that this approximation might work tolerably for the moment $\mathrm{E}e_{1t}X_{2t}'\beta_{2t}^*$, provided the rolling scheme is used.

8. Summary on small number of models

Let me close with a summary. An expansion and application of the asymptotic analysis of the preceding four sections is given in Tables 2 and 3A–3C. The rows of Table 2 are organized by sources of critical values. The first row is for tests that rely on standard results. As described in Sections 3 and 4, this means that asymptotic normal critical values are used without explicitly taking into account uncertainty about regression parameters used to make forecasts. The second row is for tests that rely on asymptotic normality, but only after adjusting for such uncertainty as described in Section 5 and in some of the final points of this section. The third row is for tests for which it would be ill-advised to use asymptotic normal critical values, as described in preceding sections.

Tables 3A–3C present recommended procedures in settings with a small number of models. They are organized by class of application: Table 3A for a single model, Table 3B for a pair of nonnested models, and Table 3C for a pair of nested models. Within each table, rows are organized by the moment being studied.

Tables 2 and 3A–3C aim to make specific recommendations. While the tables are self-explanatory, some qualifications should be noted. First, the rule of thumb that asymptotic irrelevance applies when $P/R < 0.1$ (point A1 in Table 2, note to Table 3A) is just a rule of thumb. Second, as noted in Section 4, asymptotic irrelevance for MSPE or mean absolute error (point A2 in Table 2, rows 1 and 2 in Table 3B) requires that the prediction error is uncorrelated with the predictors (MSPE) or that the disturbance is symmetric conditional on the predictors (mean absolute error). Otherwise, one will need to account for uncertainty about parameters used to make predictions. Third, some of the results in A3 and A4 (Table 2) and the regression results in Table 3A, rows 1–3, and Table 3B, row 3, have yet to be noted. They are established in West and McCracken (1998). Fourth, the suggestion to run a regression on a constant and compute a HAC t-stat (e.g., Table 3B, row 1) is just one way to operationalize a recommendation to use standard results. This recommendation is given in non-regression form in Equation (4.5). Finally, the tables are driven mainly by asymptotic results. The reader should be advised that simulation evidence to date seems to suggest that in seemingly reasonable sample sizes the asymptotic approximations sometimes work poorly. The approximations generally work poorly for long horizon forecasts [e.g., Clark and McCracken (2003), Clark and West (2005a)], and also sometimes work poorly even for one step ahead forecasts [e.g., rolling scheme, forecast encompassing (Table 3B, line 3, and Table 3C, line 3), West and McCracken (1998), Clark and West (2005a)].

9. Large number of models

Sometimes an investigator will wish to compare a large number of models. There is no precise definition of large. But for samples of size typical in economics research, procedures in this section probably have limited appeal when the number of models is say in the single digits, and have a great deal of appeal when the number of models is

Table 2
Recommended sources of critical values, small number of models

Source of critical values	Conditions for use
A. Use critical values associated with asymptotic normality, abstracting from any dependence of predictions on estimated regression parameters, as illustrated for scalar hypothesis test in (4.5) and a vector test in (4.11).	1. Prediction sample size P is small relative to regression sample size R, say $P/R < 0.1$ (any sampling scheme or moment, nested or nonnested models). 2. MSPE or mean absolute error in nonnested models. 3. Sampling scheme is recursive, moment of interest is mean prediction error or correlation between a given model's prediction error and prediction. 4. Sampling scheme is recursive, one step ahead conditionally homoskedastic prediction errors, moment of interest is either: (a) first order autocorrelation or (b) encompassing in the form (4.7c). 5. MSPE, nested models, equality of MSPE rejects (implying that it will also reject with an even smaller p-value if an asymptotically valid test is used).
B. Use critical values associated with asymptotic normality, but adjust test statistics to account for the effects of uncertainty about regression parameters.	1. Mean prediction error, first order autocorrelation of one step ahead prediction errors, zero correlation between a prediction error and prediction, encompassing in the form (4.7c) (with the exception of point C3), encompassing in the form (4.7d) for nonnested models. 2. Zero correlation between a prediction error and another model's vector of predictors (nested or nonnested) [Chao, Corradi and Swanson (2001)]. 3. A general vector of moments or a loss or utility function that satisfies a suitable rank condition. 4. MSPE, nested models, under condition (6.2), after adjustment as in (6.10).
C. Use non-standard critical values.	1. MSPE or encompassing in the form (4.7d), nested models, under condition (6.2): use critical values from McCracken (2004) or Clark and McCracken (2001). 2. MSPE, encompassing in the form (4.7d) or mean absolute error, nested models, and in contexts not covered by A5, B4 or C1: simulate/bootstrap your own critical values. 3. Recursive scheme, $\beta_1^* = 0$, encompassing in the form (4.7c): simulate/bootstrap your own critical values.

Note: Rows B and C assume that P/R is sufficiently large, say $P/R \geqslant 0.1$, that there may be nonnegligible effects of estimation uncertainty about parameters used to make forecasts. The results in row A, points 2–5, apply whether or not P/R is large.

Table 3A

Recommended procedures, small number of models.

Tests of adequacy of a single model, $y_t = X_t'\beta^* + e_t$

Description	Null hypothesis	Recommended procedure	Asymptotic normal critical values?
1. Mean prediction error (bias)	$\mathrm{E}(y_t - X_t'\beta^*) = 0$, or $\mathrm{E}e_t = 0$	Regress prediction error on a constant, divide HAC t-stat by $\sqrt{\lambda}$.	Y
2. Correlation between prediction error and prediction (efficiency)	$\mathrm{E}(y_t - X_t'\beta^*)X_t'\beta^* = 0$, or $\mathrm{E}e_t X_t'\beta^* = 0$	Regress $\hat{e}_{t+1}$ on $X_{t+1}'\hat{\beta}_t$, divide HAC t-stat by $\sqrt{\lambda}$, or regress y_{t+1} on prediction $X_{t+1}'\hat{\beta}_t$, divide HAC t-stat (for testing coefficient value of 1) by $\sqrt{\lambda}$.	Y
3. First order correlation of one step ahead prediction errors	$\mathrm{E}(y_{t+1} - X_{t+1}'\beta^*)(y_t - X_t'\beta^*) = 0$, or $\mathrm{E}e_{t+1}e_t = 0$.	a. Prediction error conditionally homoskedastic: 1. Recursive scheme: regress $\hat{e}_{t+1}$ on $\hat{e}_t$, use OLS t-stat. 2. Rolling or fixed schemes: regress $\hat{e}_{t+1}$ on $\hat{e}_t$ and X_t, use OLS t-tstat on coefficient on $\hat{e}_t$. b. Prediction error conditionally heteroskedastic: adjust standard errors as described in Section 5 above.	Y

Notes:

1. The quantity λ is computed as described in Table 1. "HAC" refers to a heteroskedasticity and autocorrelation consistent covariance matrix. Throughout, it is assumed that predictions rely on estimated regression parameters and that P/R is large enough, say $P/R \geqslant 0.1$, that there may be nonnegligible effects of such estimation. If P/R is small, say $P/R < 0.1$, any such effects may well be negligible, and one can use standard results as described in Sections 3 and 4.

2. Throughout, the alternative hypothesis is the two-sided one that the indicated expectation is nonzero (e.g., for row 1, $\mathrm{H_A}$: $\mathrm{E}e_t \neq 0$).

Table 3B
Recommended procedures, small number of models.
Tests comparing a pair of nonnested models, $y_t = X'_{1t}\beta^*_1 + e_{1t}$ vs. $y_t = X'_{2t}\beta^*_2 + e_{2t}$, $X'_{1t}\beta^*_1 \neq X'_{2t}\beta^*_2$, $\beta^*_2 \neq 0$

Description	Null hypothesis	Recommended procedure	Asymptotic normal critical values?
1. Mean squared prediction error (MSPE)	$\mathrm{E}(y_t - X'_{1t}\beta^*_1)^2 - \mathrm{E}(y_t - X'_{2t}\beta^*_2)^2 = 0$, or $\mathrm{E}e^2_{1t} - \mathrm{E}e^2_{2t} = 0$	Regress $\hat{e}^2_{1t+1} - \hat{e}^2_{2t+1}$ on a constant, use HAC t-stat.	Y
2. Mean absolute prediction error (MAPE)	$\mathrm{E}\lvert y_t - X'_{1t}\beta^*_1\rvert - \mathrm{E}\lvert y_t - X'_{2t}\beta^*_2\rvert = 0$, or $\mathrm{E}\lvert e_{1t}\rvert - \mathrm{E}\lvert e_{2t}\rvert = 0$	Regress $\lvert\hat{e}_{1t}\rvert - \lvert\hat{e}_{2t}\rvert$ on a constant, use HAC t-stat.	Y
3. Zero correlation between model 1's prediction error and the prediction from model 2 (forecast encompassing)	$\mathrm{E}(y_t - X'_{1t}\beta^*_1)X'_{2t}\beta^*_2 = 0$, or $\mathrm{E}e_{1t}X'_{2t}\beta^*_2 = 0$	a. Recursive scheme, prediction error e_{1t} homoskedastic conditional on both X_{1t} and X_{2t}: regress $\hat{e}_{1t+1}$ on $X'_{2t+1}\hat{\beta}_{2t}$, use OLS t-stat. b. Recursive scheme, prediction error e_{1t} conditionally heteroskedastic, or rolling or fixed scheme: regress $\hat{e}_{1t+1}$ on $X'_{2t+1}\hat{\beta}_{2t}$ and X_{1t}, use HAC t-stat on coefficient on $X'_{2t+1}\hat{\beta}_{2t}$.	Y
4. Zero correlation between model 1's prediction error and the difference between the prediction errors of the two models (another form of forecast encompassing)	$\mathrm{E}(y_t - X'_{1t}\beta^*_1) \times [(y_t - X'_{1t}\beta^*_1) - (y_t - X'_{2t}\beta^*_2)] = 0$, or $\mathrm{E}e_{1t}(e_{1t} - e_{2t}) = 0$	Adjust standard errors as described in Section 5 above and illustrated in West (2001).	Y
5. Zero correlation between model 1's prediction error and the model 2 predictors	$\mathrm{E}(y_t - X'_{1t}\beta^*_1)X_{2t} = 0$, or $\mathrm{E}e_{1t}X_{2t} = 0$	Adjust standard errors as described in Section 5 above and illustrated in Chao, Corradi and Swanson (2001).	Y

See notes to Table 3A.

Table 3C
Recommended procedures, small number of models.
Tests of comparing a pair of nested models, $y_t = X'_{1t}\beta^*_1 + e_{1t}$ vs. $y_t = X'_{2t}\beta^*_2 + e_{2t}$, $X_{1t} \subset X_{2t}$, $X'_{2t} = (X'_{1t}, X'_{22t})'$

Description	Null hypothesis	Recommended procedure	Asympt. normal critical values?
1. Mean squared prediction error (MSPE)	$\mathrm{E}(y_t - X'_{1t}\beta^*_1)^2 - \mathrm{E}(y_t - X'_{2t}\beta^*_2)^2 = 0$, or $\mathrm{E}e^2_{1t} - \mathrm{E}e^2_{2t} = 0$	a. If condition (6.2) applies: either (1) use critical values from McCracken (2004), or (2) compute MSPE-adjusted (6.10). b. Equality of MSPE rejects (implying that it will also reject with an even smaller p-value if an asymptotically valid test is used). c. Simulate/bootstrap your own critical values.	N Y Y N
2. Mean absolute prediction error (MAPE)	$\mathrm{E}\|y_t - X'_{1t}\beta^*_1\| - \mathrm{E}\|y_t - X'_{2t}\beta^*_2\| = 0$, or $\mathrm{E}\|e_{1t}\| - \mathrm{E}\|e_{2t}\| = 0$	Simulate/bootstrap your own critical values.	N
3. Zero correlation between model 1's prediction error and the prediction from model 2 (forecast encompassing)	$\mathrm{E}(y_t - X'_{1t}\beta^*_1)X'_{2t}\beta^*_2 = 0$, or $\mathrm{E}e_{1t}X'_{2t}\beta^*_2 = 0$	a. $\beta^*_1 \neq 0$: regress $\hat{e}_{1t+1}$ on $X'_{2t+1}\hat{\beta}_{2t}$, divide HAC t-stat by $\sqrt{\lambda}$. b. $\beta^*_1 = 0$ ($\Rightarrow \beta^*_2 = 0$): (1) Rolling or fixed scheme: regress $\hat{e}_{1t+1}$ on $X'_{2t+1}\hat{\beta}_{2t}$, use HAC t-stat. (2) $\beta^*_1 = 0$, recursive scheme: simulate/bootstrap your own critical values.	Y Y N
4. Zero correlation between model 1's prediction error and the difference between the prediction errors of the two models (another form of forecast encompassing)	$\mathrm{E}(y_t - X'_{1t}\beta^*_1)$ $\times [(y_t - X'_{1t}\beta^*_1) - (y_t - X'_{2t}\beta^*_2)] = 0$ or $\mathrm{E}e_{1t}(e_{1t} - e_{2t}) = 0$	a. If condition (6.2) applies: either (1) use critical values from Clark and McCracken (2001), or (2) use standard normal critical values. b. Simulate/bootstrap your own critical values.	N Y N
5. Zero correlation between model 1's prediction error and the model 2 predictors	$\mathrm{E}(y_t - X'_{1t}\beta^*_1)X_{22t} = 0$, or $\mathrm{E}e_{1t}X_{22t} = 0$	Adjust standard errors as described in Section 5 above and illustrated in Chao et al. (2001).	Y

1. See note 1 to Table 3A. 2. Under the null, the coefficients on X_{22t} (the regressors included in model 2 but not model 1) are zero. Thus, $X'_{1t}\beta^*_1 = X'_{2t}\beta^*_2$ and $e_{1t} = e_{2t}$. 3. Under the alternative, one or more of the coefficients on X_{22t} are nonzero. In rows 1–4, the implied alternative is one sided: $\mathrm{E}e^2_{1t} - \mathrm{E}e^2_{2t} > 0$, $\mathrm{E}|e_{1t}| - \mathrm{E}|e_{2t}| > 0$, $\mathrm{E}e_{1t}X'_{2t}\beta^*_2 > 0$, $\mathrm{E}e_{1t}(e_{1t} - e_{2t}) > 0$. In row 5, the alternative is two sided, $\mathrm{E}e_{1t}X_{22t} \neq 0$.

into double digits or above. White's (2000) empirical example examined 3654 models using a sample of size 1560. An obvious problem is controlling size, and, independently, computational feasibility.

I divide the discussion into (A) applications in which there is a natural null model, and (B) applications in which there is no natural null.

(A) Sometimes one has a natural null, or benchmark, model, which is to be compared to an array of competitors. The leading example is a martingale difference model for an asset price, to be compared to a long list of methods claimed in the past to help predict returns. Let model 1 be the benchmark model. Other notation is familiar: For model i, $i = 1, \ldots, m + 1$, let $\hat{g}_{it}$ be an observation on a prediction or prediction error whose sample mean will measure performance. For example, for MSPE, one step ahead predictions and linear models, $\hat{g}_{it} = \hat{e}_{it}^2 = (y_t - X_{it}'\hat{\beta}_{i,t-1})^2$. Measure performance so that smaller values are preferred to larger values – a natural normalization for MSPE, and one that can be accomplished for other measures simply by multiplying by -1 if necessary. Let $\hat{f}_{it} = \hat{g}_{1t} - \hat{g}_{i+1,t}$ be the difference in period t between the benchmark model and model $i + 1$.

One wishes to test the null that the benchmark model is expected to perform at least as well as any other model. One aims to test

$$\mathrm{H}_0\colon \quad \max_{i=1,\ldots,m} \mathrm{E}g_{it} \leqslant 0 \tag{9.1}$$

against

$$\mathrm{H}_A\colon \quad \max_{i=1,\ldots,m} \mathrm{E}g_{it} > 0. \tag{9.2}$$

The approach of previous sections would be as follows. Define an $m \times 1$ vector

$$\hat{f}_t = \left(\hat{f}_{1t}, \hat{f}_{2t}, \ldots, \hat{f}_{mt}\right)'; \tag{9.3}$$

compute

$$\begin{aligned}\bar{f} &\equiv P^{-1}\sum \hat{f}_t \equiv \left(\bar{f}_1, \bar{f}_2, \ldots, \bar{f}_m\right)' \\ &\equiv (\bar{g}_1 - \bar{g}_2, \bar{g}_1 - \bar{g}_3, \ldots, \bar{g}_1 - \bar{g}_{m+1})';\end{aligned} \tag{9.4}$$

construct the asymptotic variance covariance matrix of $\bar{f}$. With small m, one could evaluate

$$\bar{\nu} \equiv \max_{i=1,\ldots,m} \sqrt{P}\bar{f}_i \tag{9.5}$$

via the distribution of the maximum of a correlated set of normals. If $P \ll R$, one could likely even do so for nested models and with MSPE as the measure of performance (per note 1 in Table 2A). But that is computationally difficult. And in any event, when m is large, the asymptotic theory relied upon in previous sections is doubtful.

White's (2000) "reality check" is a computationally convenient bootstrap method for construction of p-values for (9.1). It assumes asymptotic irrelevance $P \ll R$ though the actual asymptotic condition requires $P/R \to 0$ at a sufficiently rapid rate [White (2000, p. 1105)]. The basic mechanics are as follows:

(1) Generate prediction errors, using the scheme of choice (recursive, rolling, fixed).

(2) Generate a series of bootstrap samples as follows. For bootstrap repetitions $j = 1, \ldots, N$:

   (a) Generate a new sample by sampling with replacement from the prediction errors. There is no need to generate bootstrap samples of parameters used for prediction because asymptotic irrelevance is assumed to hold. The bootstrap generally needs to account for possible dependency of the data. White (2000) recommends the stationary bootstrap of Politis and Romano (1994).

   (b) Compute the difference in performance between the benchmark model and model $i + 1$, for $i = 1, \ldots, m$. For bootstrap repetition j and model $i + 1$, call the difference $\bar{f}^*_{ij}$.

   (c) For $\bar{f}_i$ defined in (9.4), compute and save $\bar{\nu}^*_j \equiv \max_{i=1,\ldots,m} \sqrt{P}(\bar{f}^*_{ij} - \bar{f}_i)$.

(3) To test whether the benchmark can be beaten, compare $\bar{\nu}$ defined in (9.5) to the quantiles of the $\bar{\nu}^*_j$.

While White (2000) motivates the method for its ability to tractably handle situations where the number of models is large relative to sample size, the method can be used in applications with a small number of models as well [e.g., Hong and Lee (2003)].

White's (2000) results have stimulated the development of similar procedures. Corradi and Swanson (2005) indicate how to account for parameter estimation error, when asymptotic irrelevance does not apply. Corradi, Swanson and Olivetti (2001) present extensions to cointegrated environments. Hansen (2003) proposes studentization, and suggests an alternative formulation that has better power when testing for superior, rather than equal, predictive ability. Romano and Wolf (2003) also argue that test statistics be studentized, to better exploit the benefits of bootstrapping.

(B) Sometimes there is no natural null. McCracken and Sapp (2003) propose that one gauge the "false discovery rate" of Storey (2002). That is, one should control the fraction of rejections that are due to type I error. Hansen, Lunde and Nason (2004) propose constructing a set of models that contain the best forecasting model with prespecified asymptotic probability.

10. Conclusions

This paper has summarized some recent work about inference about forecasts. The emphasis has been on the effects of uncertainty about regression parameters used to make forecasts, when one is comparing a small number of models. Results applicable for a comparison of a large number of models were also discussed. One of the highest priorities for future work is development of asymptotically normal or otherwise nuisance parameter free tests for equal MSPE or mean absolute error in a pair of nested models. At present only special case results are available.

Acknowledgements

I thank participants in the January 2004 preconference, two anonymous referees, Pablo M. Pincheira-Brown, Todd E. Clark, Peter Hansen and Michael W. McCracken for helpful comments. I also thank Pablo M. Pincheira-Brown for research assistance and the National Science Foundation for financial support.

References

Andrews, D.W.K. (1991). “Heteroskedasticity and autocorrelation consistent covariance matrix estimation”. Econometrica 59, 1465–1471.

Andrews, D.W.K., Monahan, J.C. (1994). “An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator”. Econometrica 60, 953–966.

Ashley, R., Granger, C.W.J., Schmalensee, R. (1980). “Advertising and aggregate consumption: An analysis of causality”. Econometrica 48, 1149–1168.

Avramov, D. (2002). “Stock return predictability and model uncertainty”. Journal of Financial Economics 64, 423–458.

Chao, J., Corradi, V., Swanson, N.R. (2001). “Out-of-sample tests for Granger causality”. Macroeconomic Dynamics 5, 598–620.

Chen, S.-S. (2004). “A note on in-sample and out-of-sample tests for Granger causality”. Journal of Forecasting. In press.

Cheung, Y.-W., Chinn, M.D., Pascual, A.G. (2003). “Empirical exchange rate models of the nineties: Are any fit to survive?”. Journal of International Money and Finance. In press.

Chong, Y.Y., Hendry, D.F. (1986). “Econometric evaluation of linear macro-economic models”. Review of Economic Studies 53, 671–690.

Christiano, L.J. (1989). “P^*: Not the inflation forecaster’s Holy Grail”. Federal Reserve Bank of Minneapolis Quarterly Review 13, 3–18.

Clark, T.E., McCracken, M.W. (2001). “Tests of equal forecast accuracy and encompassing for nested models”. Journal of Econometrics 105, 85–110.

Clark, T.E., McCracken, M.W. (2003). “Evaluating long horizon forecasts”. Manuscript, University of Missouri.

Clark, T.E., McCracken, M.W. (2005a). “Evaluating direct multistep forecasts”. Manuscript, Federal Reserve Bank of Kansas City.

Clark, T.E., McCracken, M.W. (2005b). “The power of tests of predictive ability in the presence of structural breaks”. Journal of Econometrics 124, 1–31.

Clark, T.E., West, K.D. (2005a). “Approximately normal tests for equal predictive accuracy in nested models”. Manuscript, University of Wisconsin.

Clark, T.E., West, K.D. (2005b). “Using out-of-sample mean squared prediction errors to test the martingale difference hypothesis”. Journal of Econometrics. In press.

Clements, M.P., Galvao, A.B. (2004). “A comparison of tests of nonlinear cointegration with application to the predictability of US interest rates using the term structure”. International Journal of Forecasting 20, 219–236.

Corradi, V., Swanson, N.R., Olivetti, C. (2001). “Predictive ability with cointegrated variables”. Journal of Econometrics 104, 315–358.

Corradi V., Swanson N.R. (2005). “Nonparametric bootstrap procedures for predictive inference based on recursive estimation schemes”. Manuscript, Rutgers University.

Corradi, V., Swanson, N.R. (2006). “Predictive density evaluation”. In: Elliott, G., Granger, C.W.J., Timmermann, A. (Eds.), Handbook of Economic Forecasting. Elsevier, Amsterdam. Chapter 5 in this volume.

Davidson, R., MacKinnon, J.G. (1984). "Model specification tests based on artificial linear regressions". International Economic Review 25, 485–502.

den Haan, W.J, Levin, A.T. (2000). "Robust covariance matrix estimation with data-dependent VAR prewhitening order". NBER Technical Working Paper No. 255.

Diebold, F.X., Mariano, R.S. (1995). "Comparing predictive accuracy". Journal of Business and Economic Statistics 13, 253–263.

Elliott, G., Timmermann, A. (2003). "Optimal forecast combinations under general loss functions and forecast error distributions". Journal of Econometrics. In press.

Fair, R.C. (1980). "Estimating the predictive accuracy of econometric models". International Economic Review 21, 355–378.

Faust, J., Rogers, J.H., Wright, J.H. (2004). "News and noise in G-7 GDP announcements". Journal of Money, Credit and Banking. In press.

Ferreira, M.A. (2004). "Forecasting the comovements of spot interest rates". Journal of International Money and Finance. In press.

Ghysels, E., Hall, A. (1990). "A test for structural stability of Euler conditions parameters estimated via the generalized method of moments estimator". International Economic Review 31, 355–364.

Giacomini, R. White, H. (2003). "Tests of conditional predictive ability". Manuscript, University of California at San Diego.

Granger, C.W.J., Newbold, P. (1977). Forecasting Economic Time Series. Academic Press, New York.

Hansen, L.P. (1982). "Large sample properties of generalized method of moments estimators". Econometrica 50, 1029–1054.

Hansen, P.R. (2003). "A test for superior predictive ability". Manuscript, Stanford University.

Hansen, P.R., Lunde, A., Nason, J. (2004). "Model confidence sets for forecasting models". Manuscript, Stanford University.

Harvey, D.I., Leybourne, S.J., Newbold, P. (1998). "Tests for forecast encompassing". Journal of Business and Economic Statistics 16, 254–259.

Hoffman, D.L., Pagan, A.R. (1989). "Practitioners corner: Post sample prediction tests for generalized method of moments estimators". Oxford Bulletin of Economics and Statistics 51, 333–343.

Hong, Y., Lee, T.-H. (2003). "Inference on predictability of foreign exchange rates via generalized spectrum and nonlinear time series models". Review of Economics and Statistics 85, 1048–1062.

Hueng, C.J. (1999). "Money demand in an open-economy shopping-time model: An out-of-sample-prediction application to Canada". Journal of Economics and Business 51, 489–503.

Hueng, C.J., Wong, K.F. (2000). "Predictive abilities of inflation-forecasting models using real time data". Working Paper No 00-10-02, The University of Alabama.

Ing, C.-K. (2003). "Multistep prediction in autoregressive processes". Econometric Theory 19, 254–279.

Inoue, A., Kilian, L. (2004a). "In-sample or out-of-sample tests of predictability: Which one should we use?". Econometric Reviews. In press.

Inoue, A., Kilian, L. (2004b). "On the selection of forecasting models". Manuscript, University of Michigan.

Leitch, G., Tanner, J.E. (1991). "Economic forecast evaluation: Profits versus the conventional error measures". American Economic Review 81, 580–590.

Lettau, M., Ludvigson, S. (2001). "Consumption, aggregate wealth, and expected stock returns". Journal and Finance 56, 815–849.

Marcellino, M., Stock, J.H., Watson, M.W. (2004). "A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time". Manuscript, Princeton University.

Mark, N. (1995). "Exchange rates and fundamentals: Evidence on long-horizon predictability". American Economic Review 85, 201–218.

McCracken, M.W. (2000). "Robust out of sample inference". Journal of Econometrics 99, 195–223.

McCracken, M.W. (2004). "Asymptotics for out of sample tests of causality". Manuscript, University of Missouri.

McCracken, M.W., Sapp, S. (2003). "Evaluating the predictability of exchange rates using long horizon regressions". Journal of Money, Credit and Banking. In press.

Meese, R.A., Rogoff, K. (1983). "Empirical exchange rate models of the seventies: Do they fit out of sample?". Journal of International Economics 14, 3–24.

Meese, R.A., Rogoff, K. (1988). "Was it real? The exchange rate – interest differential over the modern floating rate period". Journal of Finance 43, 933–948.

Mizrach, B. (1995). "Forecast comparison in L_2". Manuscript, Rutgers University.

Morgan, W.A. (1939). "A test for significance of the difference between two variances in a sample from a normal bivariate population". Biometrika 31, 13–19.

Newey, W.K., West, K.D. (1987). "A simple, positive semidefinite, heteroskedasticity and autocorrelation consistent covariance matrix". Econometrica 55, 703–708.

Newey, W.K., West, K.D. (1994). "Automatic lag selection in covariance matrix estimation". Review of Economic Studies 61, 631–654.

Pagan, A.R., Hall, A.D. (1983). "Diagnostic tests as residual analysis". Econometric Reviews 2, 159–218.

Politis, D.N., Romano, J.P. (1994). "The stationary bootstrap". Journal of the American Statistical Association 89, 1301–1313.

Romano, J.P., Wolf, M. (2003). "Stepwise multiple testing as formalize data snooping". Manuscript, Stanford University.

Rossi, B. (2003). "Testing long-horizon predictive ability with high persistence the Meese–Rogoff puzzle". International Economic Review. In press.

Sarno, L., Thornton, D.L., Valente, G. (2005). "Federal funds rate prediction". Journal of Money, Credit and Banking. In press.

Shintani, M. (2004). "Nonlinear analysis of business cycles using diffusion indexes: Applications to Japan and the US". Journal of Money, Credit and Banking. In press.

Stock, J.H., Watson, M.W. (1999). "Forecasting inflation". Journal of Monetary Economics 44, 293–335.

Stock, J.H., Watson, M.W. (2002). "Macroeconomic forecasting using diffusion indexes". Journal of Business and Economic Statistics 20, 147–162.

Storey, J.D. (2002). "A direct approach to false discovery rates". Journal of the Royal Statistical Society, Series B 64, 479–498.

West, K.D. (1996). "Asymptotic inference about predictive ability". Econometrica 64, 1067–1084.

West, K.D. (2001). "Tests of forecast encompassing when forecasts depend on estimated regression parameters". Journal of Business and Economic Statistics 19, 29–33.

West, K.D., Cho, D. (1995). "The predictive ability of several models of exchange rate volatility". Journal of Econometrics 69, 367–391.

West, K.D., Edison, H.J., Cho, D. (1993). "A utility based comparison of some models of exchange rate volatility". Journal of International Economics 35, 23–46.

West, K.D., McCracken, M.W. (1998). "Regression based tests of predictive ability". International Economic Review 39, 817–840.

White, H. (1984). Asymptotic Theory for Econometricians. Academic Press, New York.

White, H. (2000). "A reality check for data snooping". Econometrica 68, 1097–1126.

Wilson, E.B. (1934). "The periodogram of American business activity". The Quarterly Journal of Economics 48, 375–417.

Wooldridge, J.M. (1990). "A unified approach to robust, regression-based specification tests". Econometric Theory 6, 17–43.