Power Loss in Predictive Performance Tests

Daniel Indacochea*
Department of Economics
University of Toronto

January 2020

Most recent version here.

Abstract

Out-of-sample predictive performance tests were originally intended for forecast comparison and remain valid for that purpose. However, the last two decades has seen a boom in research on the uses of out-of-sample tests for purposes besides forecast comparison. The wisdom of using such out-of-sample tests when full sample alternatives are available has come under scrutiny in recent years. Using Monte Carlo simulations, this paper demonstrates that a recently proposed out-of-sample procedure for model selection indeed performs markedly worse compared to its full sample counterpart in terms of statistical size and power. This paper also reproduces an applied example whereby the conclusions are sensitive to the choice of out-of-sample versus full sample procedure.

Keywords: *Forecasting; Model Selection; Bootstrap Method; Model Misspecification.*
JEL Codes: C12, C15.

* I thank participants at the 32nd Meeting of Canadian Economic Study Group in Guelph for valuable comments. Financial support from OGS is gratefully acknowledged. Any errors are my own. E-mail: daniel.indacochea@mail.utoronto.ca.

Contents

1 Introduction **2**

2 Background **3**

2.1 Diebold-Mariano Test: Model-Invariant Forecast Comparisons 3

2.2 WCM Tests: Model-Variant Forecast Comparisons 4

2.3 Vuong two-step tests: In-sample 6

2.4 Clark-McCracken two-step tests: Out-of-sample 10

3 Monte Carlo Evidence **11**

3.1 Monte Carlo Experimental Design 11

3.2 Simulation Results 13

3.2.1 Case 1: H_0^ω true (valid, equally-close models) 13

3.2.2 Case 2: $H_0 \backslash H_0^\omega$ true (misspecified, equally-close models) 15

3.2.3 Case 3: $H_0 \backslash H_0^\omega$ true (misspecified, equally-close models with reduced power) 16

3.2.4 Case 4: H_f true (misspecified models, but M1 closer in the KLIC sense) 16

4 Application to GDP growth **19**

5 Conclusion **20**

6 References **22**

1 Introduction

Diebold and Mariano (1995) developed a predictive performance test—now referred to as the **DM-test** in the literature—which allows economic practitioners to test if a set of forecasts were *significantly* more accurate than another. West (1996) proved the asymptotic distribution of the DM-test applied to forecasts from *estimated* models to be asymptotically standard normal. Following this result, there was a proliferation of research on the uses of out-of-sample DM-type tests for various purposes besides forecast comparison, such as parameter restriction (Avramov 2002, Clark and McCracken 2001), lag selection (Chao, Corradi and Swanson 2001), and non-nested model selection (Clark and McCracken 2014). However, the wisdom of using such out-of-sample tests for purposes other than forecast comparison has come under scrutiny in recent years (see Diebold 2013). In particular, Diebold (2013) argues that such tests discard valuable information by not making use of the full sample. Moreover, out-of-sample approaches use recursive testing schemes which can be difficult to implement and often have inferior statistical power.

This paper puts the criticism made by Diebold (2013) to practice. In particular, we compare in-sample alternatives for non-nested hypothesis testing of two possibly misspecified models developed by Vuong (1989), and compare those to the out-of-sample tests developed by Clark and McCracken (2014). We find that the in-sample methods perform comparably well in terms of statistical size and dramatically better in terms of power, thereby corroborating the thesis of Diebold (2013). Moreover, this paper reproduces the applied example from Clark and McCracken (2014), and finds that some test conclusions change when using the full-sample Vuong test.[1]

This paper is organized as follows: Section 2 summarizes the key background literature, Section 3 outlines the Monte Carlo experiment used to compare the in-sample and out-of-sample alternatives and summarizes the key findings, Section 4 applies these findings to a real-world example, and Section 5 outlines future research and concludes.

[1] "Full-sample" is redundant here but is included for emphasis.

2 Background

2.1 Diebold-Mariano Test: Model-Invariant Forecast Comparisons

Economic forecasters often wish to compare the predictive accuracy of two or more sets of forecasts. One natural way to do so is to compare the **mean squared prediction error** (MSPE) of the vector of forecasts $\hat{\boldsymbol{f}}_i$ to the vector of realized values $\boldsymbol{f}$:

$$\text{MSPE}_i = \frac{1}{P}\sum_{t=1}^{P}(\hat{f}_{it} - f_t)^2, \tag{1}$$

where $i \in \{1, \ldots, M\}$, M is the number of competing forecast vectors, t is the time index, and P is the number of forecasted periods. Hence, this MSPE approach tells the practitioner which forecasts performed best in terms of average prediction error. The obvious drawback of such an approach, however, is that it says nothing about how much more accurate any set of forecasts were. This set the stage for Diebold and Mariano (1995), who developed a test for comparing predictive performance that allowed practitioners to say whether one set of forecasts were *significantly* more accurate than another. The statistic proposed by Diebold and Mariano (1995) is today commonly referred to as the **DM-statistic** in the literature, and is straightforward to compute:

$$\begin{aligned}
\hat{d}_{t,ij} &= g(e_{it}) - g(e_{jt}), \\
\bar{d}_{ij} &= \frac{1}{P}\sum_{t=1}^{P}\hat{d}_{t,ij}, \\
\hat{S}_{dd} &= \frac{1}{P-1}\sum_{t=1}^{P}\hat{d}_{t,ij} - \bar{d}_{ij}, \qquad (2) \\
\text{MSE-}t &= \sqrt{P}\frac{\bar{d}_{ij}}{\sqrt{\hat{S}_{dd}}}, \qquad (3)
\end{aligned}$$

where $g(\cdot)$ is the loss function—usually chosen to be quadratic or absolute—and MSE-t is the DM-statistic. The null hypothesis of equal forecast accuracy can therefore be stated as

$$\mathrm{H}_0 : \mathbf{E}[d_{t,ij}] = 0. \tag{4}$$

Diebold and Mariano (1995) showed that MSE-$t \xrightarrow{D} N(0,1)$ under some stationarity assumptions about the forecast error loss differential $\hat{d}_{t,ij}$. This last point is crucial; the DM test does not make *any* assumptions about the models used to generate the forecasts themselves, and is thereby invariant to models used to generate those forecasts. Subsequent research, however, went in a decidedly different direction.

2.2 WCM Tests: Model-Variant Forecast Comparisons

As Diebold (2013) notes, subsequent research on comparative out-of-sample performance focused not on using DM-type tests for comparing *forecasts*, but on comparing econometric *models* themselves. This modeling paradigm has its origins with West (1996) and Clark and McCracken (2001), and will be referred to broadly as "WCM" as per Diebold (2013).

Computing WCM out-of-samples statistics is relatively straightforward. For concreteness, suppose we have M candidate models of the form

$$\boldsymbol{y} = \boldsymbol{X}_i \boldsymbol{\beta}_i + \boldsymbol{u}_i, \; i \in \{1, \ldots, M\}, \tag{5}$$

where $\boldsymbol{y}$ is a $T \times 1$ regressand vector, $\boldsymbol{X}_i$ is a $T \times k_i$ matrix of regressors, $\boldsymbol{\beta}_i$ is a $k_i \times 1$ vector of parameters, and $\boldsymbol{u}_i$ is a $T \times 1$ vector of errors for model i. For each candidate model $i \in \{1, \ldots, M\}$, the out-of-sample statistics are computed as follows:

1. For a sample of size T, split the sample into R minimum observations used to estimate the model (by regression or otherwise), and P periods used for prediction.

2. Set $N = R$, the number of observations used in the regression.

3. Estimate $\boldsymbol{\beta}_i$ using the N periods.

4. Compute $\hat{y}_{t+1} = X_{i,t+1}\hat{\boldsymbol{\beta}}_i^N$, where $\hat{\boldsymbol{\beta}}_i^N$ is the estimate of $\boldsymbol{\beta}_i$ using N observations.

5. Compute $\hat{e}_{it} = \hat{y}_{t+1} - y_{t+1}$ and store it.

6. Repeat steps 1-5 for $N \in \{R+1, R+2, \ldots, R+P-1\}$ (Resulting in P **one-step-ahead predictions**).

7. Once steps 1-6 have been performed for all M models, compute MSE-t for models i, j as in Equation (3).

The above algorithm is by no means the only way to proceed. For example, one could compute prediction errors based on any horizon smaller than P instead of one-step-ahead predictions. Similarly, one could use a rolling window so that the number of observations used in the regression is fixed at R instead of recursively expanding N as outlined above. However, the one-step-ahead approach with a recursively expanding window is what is most commonly used in practice and will therefore be the focus of this paper.

As Diebold (2013) highlights, a litany of complications arise under the WCM paradigm, as statistical inference must account for the following:

1. Nesting structure. (see Definitions 1-3.)
2. Functional form. (*e.g.* linear or non-linear.)
3. Distribution of error terms.
4. Estimation scheme. (Recurive, fixed, or rolling window.)
5. Choice of estimation method.
6. Asymptotics. (WCM tests are sensitive to $\underset{T\to\infty}{\text{plim}}\, P/R$. See Figure 1.)

2.3 Vuong two-step tests: In-sample

Here we outline the theory and application of the two-step test developed by Vuong (1989), which is the in-sample basis of Clark and McCracken (2014)—the two of which will be compared in Section 3. The purpose of Vuong (1989) is to provide a general framework for selecting among candidate models

$$\boldsymbol{F}_\theta = \{f(\boldsymbol{y}|\boldsymbol{X}_1;\boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^{k_1}\}, \tag{6}$$

$$\boldsymbol{G}_\gamma = \{g(\boldsymbol{y}|\boldsymbol{X}_2;\boldsymbol{\gamma}); \boldsymbol{\gamma} \in \Gamma \subseteq \mathbb{R}^{k_2}\}, \tag{7}$$

when the conditional density of the true **data generating process** (DGP) is $h^0(\boldsymbol{y}|\boldsymbol{Z})$, which is unobserved. Thus, this framework allows $\boldsymbol{F}_\theta$ and $\boldsymbol{G}_\gamma$ to be misspecified and either **nested**, **overlapping** or **strictly non-nested**.

Definition 1 (Nested models) *$\boldsymbol{F}_\theta$ is said to be nested in $\boldsymbol{G}_\gamma$ iff*
$\boldsymbol{F}_\theta \subseteq \boldsymbol{G}_\gamma$.

Definition 2 (Overlapping models) *$\boldsymbol{F}_\theta$ and $\boldsymbol{G}_\gamma$ are said to be overlapping iff*
i) $\boldsymbol{F}_\theta \cap \boldsymbol{G}_\gamma \neq \emptyset$.
ii) $\boldsymbol{F}_\theta \subsetneq \boldsymbol{G}_\gamma$ *and* $\boldsymbol{G}_\gamma \subsetneq \boldsymbol{F}_\theta$.

Definition 3 (Strictly non-nested models) *$\boldsymbol{F}_\theta$ and $\boldsymbol{G}_\gamma$ are said to be strictly non-nested iff*
$\boldsymbol{F}_\theta \cap \boldsymbol{G}_\gamma = \emptyset$.

In the context of classical linear regression models, two models are nested if one contains all of the regressors of the other, overlapping if the models share some regressors but also have regressors not in common, and strictly non-nested if the models have no regressors in common.

Although we do not observe $h^0(\boldsymbol{y}|\boldsymbol{Z})$ directly, we can still define a measure[2] of the "distance" between our candidate models $\boldsymbol{F}_\theta$ and $\boldsymbol{G}_\gamma$ and the true DGP using the **Kullback-Leibler**

[2] The Kullback-Leibler approximation is not a true measure in the strict mathematical sense as it does not satisfy the symmetry condition nor the triangle inequality (See Gourierroux and Monfort (1995), Volume I, pg. 13).

Information Criterion (KLIC)

$$\text{KLIC}(H^0_{\boldsymbol{y}|\boldsymbol{Z}}, \boldsymbol{F}_{\boldsymbol{\theta}}) = \mathbf{E}^0[\log h^0(\boldsymbol{y}|\boldsymbol{Z})] - \mathbf{E}^0[\log f(\boldsymbol{y}|\boldsymbol{X}_1; \boldsymbol{\theta}^*)], \tag{8}$$

where $H^0_{\boldsymbol{y}|\boldsymbol{Z}}$ is the true joint distribution of y_t and $\boldsymbol{Z}_t$ and $\boldsymbol{\theta}^*$ is the pseudo-true parameter which solves

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\text{argmin}}\ \mathbf{E}^0[\log h^0(\boldsymbol{y}|\boldsymbol{Z})] - \mathbf{E}^0[\log f(\boldsymbol{y}|\boldsymbol{X}_1; \boldsymbol{\theta})]. \tag{9}$$

Although we do not actually observe $h^0(\boldsymbol{y}|\boldsymbol{Z})$, we can still compute $\boldsymbol{\theta}^*$ identically as

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\text{argmax}}\ \mathbf{E}^0[\log f(\boldsymbol{y}|\boldsymbol{X}_1; \boldsymbol{\theta})],$$

because $h^0(\boldsymbol{y}|\boldsymbol{Z})$ is constant in $\boldsymbol{\theta}$. $\text{KLIC}(H^0_{\boldsymbol{y}|\boldsymbol{Z}}, \boldsymbol{G}_\gamma)$ is defined analogously to Equation (8). Thus, testing which candidate model is closer to the true DGP in the Kullback-Leibler sense is equivalent to testing:

$$\begin{aligned}
\text{H}_0 &: \mathbf{E}^0[\log f(\boldsymbol{y}|\boldsymbol{X}_1; \boldsymbol{\theta})] = \mathbf{E}^0[\log g(\boldsymbol{y}|\boldsymbol{X}_2; \boldsymbol{\gamma})] \\
\text{H}_f &: \mathbf{E}^0[\log f(\boldsymbol{y}|\boldsymbol{X}_1; \boldsymbol{\theta})] > \mathbf{E}^0[\log g(\boldsymbol{y}|\boldsymbol{X}_2; \boldsymbol{\gamma})] \\
\text{H}_g &: \mathbf{E}^0[\log f(\boldsymbol{y}|\boldsymbol{X}_1; \boldsymbol{\theta})] < \mathbf{E}^0[\log g(\boldsymbol{y}|\boldsymbol{X}_2; \boldsymbol{\gamma})].
\end{aligned}$$

Before we formally summarize Vuong's (1989) findings, we need three more definitions.

Definition 4 (Pseudo-true variance) $\omega^2_* = \text{Var}^0\left[\log \frac{f(\boldsymbol{y}|\boldsymbol{X}_1;\boldsymbol{\theta}^*)}{g(\boldsymbol{y}|\boldsymbol{X}_2;\boldsymbol{\gamma}^*)}\right]$

Definition 5 (Mixed-chi square distribution) *Let* $z_i \sim \text{N}(0,1)\ \forall i \in \{1, \ldots, m\}$*, and let* $\boldsymbol{\lambda}$ *be an* m*-vector with typical element* $\lambda_i \in \mathbb{R}$*, then* $\sum_{i=1}^m \lambda_i z_i^2$ *is said to follow a mixed chi-square distribution denoted* $\text{M}_m(\boldsymbol{\lambda})$*.*

Definition 6 (Asymptotic covariance matrix of MLE) *Let* $\hat{\theta}, \hat{\gamma}$ *denote the* **Maximum Likelihood** *(ML) estimators of* θ^*, γ^**, respectively. The asymptotic covariance matrix of*

these ML estimators is given by:

$$\boldsymbol{W} = \begin{bmatrix} -\boldsymbol{B}_f(\boldsymbol{\theta}^*)\boldsymbol{A}_f^{-1}(\boldsymbol{\theta}^*) & -\boldsymbol{B}_{fg}(\boldsymbol{\theta}^*, \boldsymbol{\gamma}^*)\boldsymbol{A}_g^{-1}(\boldsymbol{\gamma}^*) \\ -\boldsymbol{B}_{gf}(\boldsymbol{\gamma}^*, \boldsymbol{\theta}^*)\boldsymbol{A}_f^{-1}(\boldsymbol{\theta}^*) & -\boldsymbol{B}_g(\boldsymbol{\gamma}^*)\boldsymbol{A}_g^{-1}(\boldsymbol{\gamma}^*) \end{bmatrix}, \tag{10}$$

where

$$\begin{aligned} \boldsymbol{A}_f(\boldsymbol{\theta}^*) &= \mathbf{E}^0\left[\frac{\partial^2 \log f}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top}\right], \quad \boldsymbol{A}_g(\boldsymbol{\gamma}^*) = \mathbf{E}^0\left[\frac{\partial^2 \log g}{\partial \boldsymbol{\gamma}\partial \boldsymbol{\gamma}^\top}\right], \\ \boldsymbol{B}_f(\boldsymbol{\theta}^*) &= \mathbf{E}^0\left[\frac{\partial^2 \log f}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top}\right], \quad \boldsymbol{B}_g(\boldsymbol{\gamma}^*) = \mathbf{E}^0\left[\frac{\partial^2 \log g}{\partial \boldsymbol{\gamma}\partial \boldsymbol{\gamma}^\top}\right], \end{aligned}$$

$$\boldsymbol{B}_{fg}(\boldsymbol{\theta}^*, \boldsymbol{\gamma}^*) = \mathbf{E}^0\left[\frac{\partial \log f \partial \log g}{\partial \boldsymbol{\theta}\partial \boldsymbol{\gamma}^\top}\right] = \boldsymbol{B}_{fg}(\boldsymbol{\gamma}^*, \boldsymbol{\theta}^*)^\top.$$

When $\boldsymbol{F}_\theta$ and $\boldsymbol{G}_\gamma$ are overlapping, testing H_0 turns out to be a difficult problem. This is because there are two ways H_0 can be satisfied, which in turn implies the asymptotic distributions of whatever tests we propose depends precisely on how this null is satisfied. The two ways in which H_0 can be satisfied are $\omega_*^2 = 0$ and $\omega_*^2 \neq 0$. Let us denote the null where $\omega_*^2 = 0$ as H_0^ω, and the null where $\omega_*^2 \neq 0$ as $\mathrm{H}_0 \backslash \mathrm{H}_0^\omega$. Under some regularity conditions, Vuong (1989) proved the following:

$$\text{Under } \mathrm{H}_0{}^\omega : T\hat{\omega}^2 \xrightarrow{D} \mathrm{M}_{p+q}(\boldsymbol{\lambda}^*), \tag{11}$$

$$\text{Under } (\mathrm{H}_0{}^\omega)^c : T\hat{\omega}^2 \xrightarrow{a.s.} +\infty, \tag{12}$$

$$\text{Under } \mathrm{H}_0 \backslash \mathrm{H}_0^\omega : t_V = \frac{\mathrm{LR}_T(\hat{\theta}, \hat{\gamma})}{\sqrt{T\hat{\omega}^2}} \xrightarrow{D} \mathrm{N}(0,1), \tag{13}$$

$$\text{Under } \mathrm{H}_f : t_V = \frac{\mathrm{LR}_T(\hat{\theta}, \hat{\gamma})}{\sqrt{T\hat{\omega}^2}} \xrightarrow{a.s.} +\infty, \tag{14}$$

$$\text{Under } \mathrm{H}_g : t_V = \frac{\mathrm{LR}_T(\hat{\theta}, \hat{\gamma})}{\sqrt{T\hat{\omega}^2}} \xrightarrow{a.s.} -\infty, \tag{15}$$

where $\boldsymbol{\lambda}^*$ is the vector of eigenvalue of the asymptotic covariance matrix defined in Equation (10), and

$$\hat{\omega}^2 \equiv \frac{1}{T}\sum_{t=1}^{T}\left[\log \frac{f(y_t|\boldsymbol{X}_{1t};\hat{\theta})}{g(y_t|\boldsymbol{X}_{2t};\hat{\gamma})}\right]^2 - \left[\frac{1}{T}\sum_{t=1}^{T}\log \frac{f(y_t|\boldsymbol{X}_{1t};\hat{\theta})}{g(y_t|\boldsymbol{X}_{2t};\hat{\gamma})}\right]^2, \tag{16}$$

$$\mathrm{LR}_T(\hat{\theta}, \hat{\gamma}) \equiv \sum_{t=1}^{T}\log \frac{f(y_t|\boldsymbol{X}_{1t};\hat{\theta})}{g(y_t|\boldsymbol{X}_{2t};\hat{\boldsymbol{\gamma}})}. \tag{17}$$

To account for the possibility $\mathrm{H}_0{}^\omega$ is true with two overlapping, possibly misspecified models, Vuong (1989) proposes the following two-step procedure, which is henceforth referred to collectively as the **Vuong two-step test**:

1. **Vuong Variance test**: To test the overlapping null H_0^ω is true, we test if $T\hat{\omega}^2 = 0$, which we reject with $1 - \alpha_1\%$ confidence and move on to step 2 if it exceeds the $100 - \alpha_1$ percentile from $\mathrm{M}_{p+q}(\boldsymbol{\lambda}^*)$. Otherwise, we cannot discriminate between $\boldsymbol{F}_\theta$

and $\boldsymbol{G}_\gamma$ given the data, so we stop.

2. **Vuong *t*-test**: To test the nested null $\mathrm{H}_0 \backslash \mathrm{H}_0^\omega$ is true, we test if $t_V = T^{-1/2}\mathrm{LR}_T(\hat{\theta}, \hat{\gamma})/\sqrt{\hat{\omega}^2} = 0$, which we reject with $1 - \alpha_2\%$ confidence if it exceeds the $100 - \alpha_2/2$ critical value from the standard normal (thereby concluding $\boldsymbol{F}_\theta$ is closer in the KLIC sense), or falls below the $\alpha_2/2$ critical value from the standard normal (thereby concluding $\boldsymbol{G}_\gamma$ is closer in the KLIC sense). Otherwise, we conclude these models are nested.

Vuong (1989) showed that this two-step procedure is asymptotically conservative. That is, the size of the Vuong test under H_0 is asymptotically bounded by $\max(\alpha_1, \alpha_2)$.

2.4 Clark-McCracken two-step tests: Out-of-sample

Clark and McCracken (2014) developed an out-of-sample version of the Vuong two-step procedure that makes use of the wild bootstrap. Consider a special case of Equation (5) with $M = 2$ linear models,

$$\begin{aligned}
\mathrm{M1}: &\quad \boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{u}_1; \quad \mathbf{E}[\boldsymbol{u}_1|\boldsymbol{X}_1] = \mathbf{0}; \quad \mathbf{E}[\boldsymbol{u}_1\boldsymbol{u}_1^\top|\boldsymbol{X}_1] = \sigma_1^2\boldsymbol{I}_T, \\
\mathrm{M2}: &\quad \boldsymbol{y} = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}_2; \quad \mathbf{E}[\boldsymbol{u}_2|\boldsymbol{X}_2] = \mathbf{0}; \quad \mathbf{E}[\boldsymbol{u}_2\boldsymbol{u}_2^\top|\boldsymbol{X}_2] = \sigma_2^2\boldsymbol{I}_T,
\end{aligned}$$

where $\boldsymbol{I}_T$ denotes the $T \times T$ identity matrix, and $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ may contain a vector of lag dependent variables $\boldsymbol{y}_1$. Suppose further that the true DGP is given by:

$$\mathrm{M0}: \quad \boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta}_0 + \boldsymbol{u}_0; \quad \mathbf{E}[\boldsymbol{u}_0|\boldsymbol{Z}] = \mathbf{0}; \quad \mathbf{E}[\boldsymbol{u}_0\boldsymbol{u}_0^\top|\boldsymbol{Z}] = \sigma_0^2\boldsymbol{I}_T,$$

where $\boldsymbol{Z}$ may also contain a vector of lagged dependent variables $\boldsymbol{y}_1$. Clark and McCracken (2014) propose the following out-of-sample model selection procedure:

1. Compute $P\hat{S}_{dd}$ (see 2) and MSE-t (see 3) from the original sample.

2. Regress $\boldsymbol{y}$ on $\boldsymbol{X}_\cap \equiv \boldsymbol{X}_1 \cap \boldsymbol{X}_2$ and store the estimated parameter coefficients $\hat{\boldsymbol{\beta}}_\cap$.

3. Regress $\boldsymbol{y}$ on $\boldsymbol{X}_{\cup} \equiv \boldsymbol{X}_1 \cup \boldsymbol{X}_2$ and store the vector of residuals $\hat{\boldsymbol{u}}_{\cup}$.

4. Generate B wild bootstrap samples using the following procedure.

   (a) Generate a T-vector of residuals $\boldsymbol{\epsilon}_j^*$ with typical element $\epsilon_t^{*j} = \hat{u}_{\cup i} z_i^{*j}$, where $z_i^{*j} \overset{iid}{\sim} \mathrm{N}(0,1)\ \forall j \in \{1,\ldots,B\}$.

   (b) Generate a T-vector $\boldsymbol{y}_j^*$ with typical element $y_t^{*j} = \boldsymbol{X}_{\cap t}^{*j}\hat{\boldsymbol{\beta}}_{\cap} + \epsilon_t^{*j}\ \forall j \in \{1,\ldots,B\}$. Each $\boldsymbol{y}^{*j}$ is a bootstrap sample of size T. (Note that $\boldsymbol{X}_{\cap}^{*j} = \boldsymbol{X}_{\cap}\ \forall j \in \{1,\ldots,B\}$ if $\boldsymbol{X}_{\cap}$ does not contain a lag dependent variable.)

   (c) Compute $P\hat{S}_{dd}^{*j}$ just as with the original sample in step 1.

5. Perform the out of sample two-step procedure, which is henceforth referred to collectively as the **CM two-step test**:

   (a) **CM variance test** If $P\hat{S}_{dd}$ exceeds the $100 - \alpha_1$ percentile of the bootstrap distribution of $P\hat{S}_{dd}^{*j}$, move on to step 5.(b). Otherwise, we cannot distinguish between M1 and M2, so we stop.

   (b) **CM *t*-test**: Reject the null of equal forecast accuracy with $100 - \alpha_2\%$ confidence if MSE-t exceeds the $100 - \alpha_2/2$ critical value from the standard normal (thereby concluding M1 has more predictive content in the WCM sense), or falls below the $\alpha_2/2$ critical value from the standard normal (thereby concluding M2 has more predictive content in the WCM sense). Otherwise, we fail we to reject that these models have equal predictive performance.

3 Monte Carlo Evidence

3.1 Monte Carlo Experimental Design

We reproduce the Monte Carlo experimental design of Clark and McCracken (2014) to compare the performance of the *out-of-sample* CM two-step test outlined in Section 2.4 to the

full-sample Vuong two-step test outlined in Section 2.3. Each pseudo-observation consists of error terms drawn independently from the normal distribution and the autoregressive structure of the DGP. To resemble monthly and quarterly data, sample sizes (R, P) vary from $R \in \{50, 100, 200\}$ and $P \in \{20, 50, 100, 200\}$, hence there are $T = R + P$ total observations such that

$$T \in \{70, 100, 120, 150, 200, 220, 250, 300, 400\}$$

by construction. The experiments are intended to loosely mimic the empirical properties of GDP growth (denoted by the regressand $\boldsymbol{y}$), and the spread between 10-year and 1-year yields $\boldsymbol{x}_1$, and the spread between AAA and BBB corporate bonds $\boldsymbol{x}_2$:

$$y_{t+1} = 0.3y_t + b_1 x_{1,t} + b_2 x_{t,2} + u_{t+1}, \tag{18}$$

$$x_{i,t} = 0.9x_{t-1,i} + \epsilon_{t,i} \;\; \forall i \in \{1, 2\}, \tag{19}$$

where

$$\text{Var}\left(\begin{bmatrix} u_t \\ \epsilon_{t,1} \\ \epsilon_{t,2} \end{bmatrix}\right) = \begin{bmatrix} 10.0 & \cdot & \cdot \\ 0.0 & 0.1 & \cdot \\ 0.0 & 0.0 & 0.1 \end{bmatrix}. \tag{20}$$

The two candidate models under consideration are

$$\text{M1} : y_{t+1} = \alpha_0 + \alpha_1 y_t + \alpha_2 x_{t,1} + \alpha_3 x_{t-1,1} + u_{t+1,1}, \tag{21}$$

$$\text{M2} : y_{t+1} = \beta_0 + \beta_1 y_t + \beta_2 x_{t,2} + \beta_3 x_{t-1,2} + u_{t+1,2}. \tag{22}$$

Since these models have common regressors but neither model is a special case of the other, these models are overlapping. Four cases for of the true DGP (18) are considered:

Case 1: $b_1 = b_2 = 0$

Both models contain the null and are therefore valid. Under these conditions $T\hat{\omega}^2$ and $P\hat{S}_{dd}$ asymptotically follow mixed chi-square distributions while t_V and MSE-t asymptotically follow non-standard distributions.

Case 2: $b_1 = b_2 = -2$

Neither model contains the true DGP (18) and are therefore misspecified, but each model is equally close to the true DGP in the KLIC sense. Under these conditions $T\hat{\omega}^2$ and $P\hat{S}_{dd}$ are asymptotically infinite while t_V and MSE-t are asymptotically standard normal.

Case 3: $b_1 = b_2 = -0.7$

Analogous to Case 2, only now there should be a substantial reduction in power.

Case 4: $b_1 = -2, b_2 = 0$

Neither model contains the true DGP (18) and are therefore both misspecified, but M1 is closer to the true DGP in the KLIC sense. Under these conditions, all the test statistics are asymptotically infinite.

3.2 Simulation Results

Tables 1-4 summarize the simulation results. It should be noted that Clark and McCracken (2014) use 10,000 Monte Carlo replications and 499 bootstrap samples per replication. To save on simulation time, this study used 5,000 Monte Carlo replications and 399 bootstrap samples per replication, but the results are quantitatively similar and qualitatively identical. Table 4 has the most pertinent results, so Tables 1-3 are discussed tersely.

3.2.1 Case 1: $\mathbf{H_0^\omega}$ true (valid, equally-close models)

As seen in Table 1, both the CM two-step test and the Vuong-two step tests are conservative, with rejection rates well below $\max(\alpha_1, \alpha_2) = 0.10$. For the pre-test, both $P\hat{S}_{dd}$ and $T\hat{\omega}^2$ overreject slightly, though for small samples (R=50, P=20) $T\hat{\omega}^2$ performs slightly better. It is therefore left for future research to examine how severe these biases are with fewer

than $T = 70$ observations, which is a plausible scenario for practitioners using quarterly data limited in size because of structural breaks.

Table 1: Monte Carlo Rejection Rates: overlapping models ($\alpha_1 = \alpha_2 = 0.10$; $b_1 = b_2 = 0$)

R	P	CM Two-step	Vuong Two-step	MSE-t vs N(0, 1)	Voung-t vs N(0, 1)	$P\hat{S}_{dd}$	$T\hat{\omega}^2$
50	20	0.0141	0.0067	0.0974	0.0532	0.1452	0.1259
50	50	0.0074	0.0073	0.0598	0.0617	0.1234	0.1182
50	100	0.0042	0.0074	0.0402	0.0694	0.1046	0.1066
50	200	0.0017	0.0047	0.0176	0.0447	0.0987	0.1057
100	20	0.0110	0.0053	0.0967	0.0475	0.1138	0.1105
100	50	0.0090	0.0074	0.0787	0.0694	0.1144	0.1066
100	100	0.0722	0.0034	0.0722	0.0320	0.1080	0.1062
100	200	0.0053	0.0047	0.0537	0.0460	0.0993	0.1030
200	20	0.0098	0.0054	0.0945	0.0501	0.1037	0.1077
200	50	0.0088	0.0040	0.0800	0.0359	0.1100	0.1114
200	100	0.0075	0.0043	0.0735	0.0385	0.1020	0.1103
200	200	0.0076	0.0024	0.0701	0.0220	0.1084	0.1090

3.2.2 Case 2: $\mathbf{H_0 \backslash H_0^\omega}$ true (misspecified, equally-close models)

If we look at the two rightmost columns of Table 2, we see that $T\hat{\omega}^2$ performs unequivocally better than $P\hat{S}_{dd}$ in the pre-test, particularly in small samples. In particular, for $R = 50, P = 20$, we see that $T\hat{\omega}^2$ has power of 89.5%, compared to 72.6% of $P\hat{S}_{dd}$. If we look at the third and fourth columns, we see that both two-step procedures have comparable size.

Table 2: Monte Carlo Rejection Rates: overlapping models ($\alpha_1 = \alpha_2 = 0.10$; $b_1 = b_2 = -2$)

R	P	CM Two-step	Vuong Two-step	MSE-t vs N(0, 1)	Voung-t vs N(0, 1)	$P\hat{S}_{dd}$	$T\hat{\omega}^2$
50	20	0.0945	0.1022	0.1241	0.1142	0.7616	0.8953
50	50	0.1052	0.1144	0.1168	0.1167	0.9009	0.9807
50	100	0.1040	0.1136	0.1064	0.1138	0.9776	0.9986
50	200	0.0940	0.1037	0.0941	0.1037	0.9991	1.0000
100	20	0.1204	0.1144	0.1320	0.1151	0.9124	0.9942
100	50	0.1164	0.1176	0.1184	0.1177	0.9834	0.9994
100	100	0.1106	0.1158	0.1108	0.1158	0.9982	1.0000
100	200	0.1196	0.1144	0.1196	0.1144	1.0000	1.0000
200	20	0.1177	0.1140	0.1199	0.1140	0.9814	1.0000
200	50	0.1200	0.1164	0.1202	0.1164	0.9986	1.0000
200	100	0.1208	0.1195	0.1208	0.1195	1.0000	1.0000
200	200	0.1116	0.1253	0.1116	0.1253	1.0000	1.0000

3.2.3 Case 3: $\mathbf{H_0 \backslash H_0^\omega}$ true (misspecified, equally-close models with reduced power)

The results are analogous to Table 2 but now we see the expected reduction in power.

Table 3: Monte Carlo Rejection Rates: overlapping models ($\alpha_1 = \alpha_2 = 0.10$; $b_1 = b_2 = -0.7$)

R	P	CM Two-step	Vuong Two-step	MSE-t vs N(0,1)	Voung-t vs N(0,1)	$P\hat{S}_{dd}$	$T\hat{\omega}^2$
50	20	0.0300	0.0246	0.1067	0.0800	0.2810	0.3076
50	50	0.0281	0.0312	0.0844	0.0742	0.3335	0.4203
50	100	0.0260	0.0440	0.0646	0.0744	0.4022	0.5912
50	200	0.0320	0.0546	0.0570	0.0656	0.5614	0.8314
100	20	0.0382	0.0359	0.0977	0.0721	0.3909	0.4977
100	50	0.0454	0.0442	0.0958	0.0742	0.4740	0.5956
100	100	0.0538	0.0494	0.0917	0.0663	0.5866	0.7450
100	200	0.0538	0.0692	0.0740	0.0770	0.7268	0.8986
200	20	0.0598	0.0607	0.1039	0.0787	0.5756	0.7715
200	50	0.0666	0.0678	0.0979	0.0822	0.6800	0.8250
200	100	0.0755	0.0685	0.0954	0.0763	0.7915	0.8978
200	200	0.0794	0.0740	0.0888	0.0767	0.8942	0.9642

3.2.4 Case 4: $\mathbf{H_f}$ true (misspecified models, but M1 closer in the KLIC sense)

The results in Table 4 are dramatic. For $R = 50$, $P = 20$, the CM two-step test has 11.6% power while the Vuong two-step test has 29.6% power. Moreover, at $R = 50, P = 50$, the difference is even more striking: the CM two-step test has 23.9% power while the Vuong two-step test has 50.6% power. This means that for this experimental design, using the out-of-sample procedure discards over half of the available power in small- to medium-sized samples. To further illustrate just how severely the out-of-sample procedure reduces power, consider the power of the Vuong two-step test when $R = 50, P = 100$: 78.3%. By contrast, the CM two-step test needs $R = 50, P = 180$ to achieve this same level of power[3], for a total

[3] This value $P = 180$ was deduced with a search algorithm and is not shown in Table 4.

of 80 *additional* observations. In terms of quarterly data, this means the CM two-step test requires 20 additional years of data to achieve the same power performance as the full-sample Vuong test.

Table 4: Monte Carlo Rejection Rates: overlapping models ($\alpha_1 = \alpha_2 = 0.10$; $b_1 = -2, b_2 = 0$)

R	P	CM Two-step	Vuong Two-step	MSE-t vs N$(0,1)$	Voung-t vs N$(0,1)$	$P\hat{S}_{dd}$	$T\hat{\omega}^2$
50	20	0.1157	0.2958	0.1947	0.4205	0.5942	0.7034
50	50	0.2392	0.5063	0.3214	0.5761	0.7442	0.8789
50	100	0.4978	0.7826	0.5619	0.7999	0.8860	0.9784
50	200	0.8590	0.9742	0.8779	0.9747	0.9785	0.9995
100	20	0.1646	0.6339	0.2022	0.6766	0.8139	0.9369
100	50	0.3390	0.7764	0.3685	0.7958	0.9200	0.9756
100	100	0.5858	0.9160	0.5995	0.9189	0.9772	0.9968
100	200	0.8790	0.9938	0.8809	0.9940	0.9978	0.9998
200	20	0.2125	0.9470	0.2231	0.9479	0.9523	0.9990
200	50	0.3672	0.9734	0.3700	0.9738	0.9924	0.9996
200	100	0.6172	0.9898	0.6179	0.9900	0.9990	0.9998
200	200	0.8780	0.9993	0.8780	0.9993	1.0000	1.0000

Apart from severe power loss, another major drawback with the CM test is that its power properties vary with the choice of (R, P). As such, the practitioner has to divine which particular combination of (R, P) provides the best statistical size/power trade-off. By contrast, the Vuong test suffers no such drawback, because its power increases monotonically with sample size T and is invariant to the choice of (R, P), as illustrated in Figure 1.

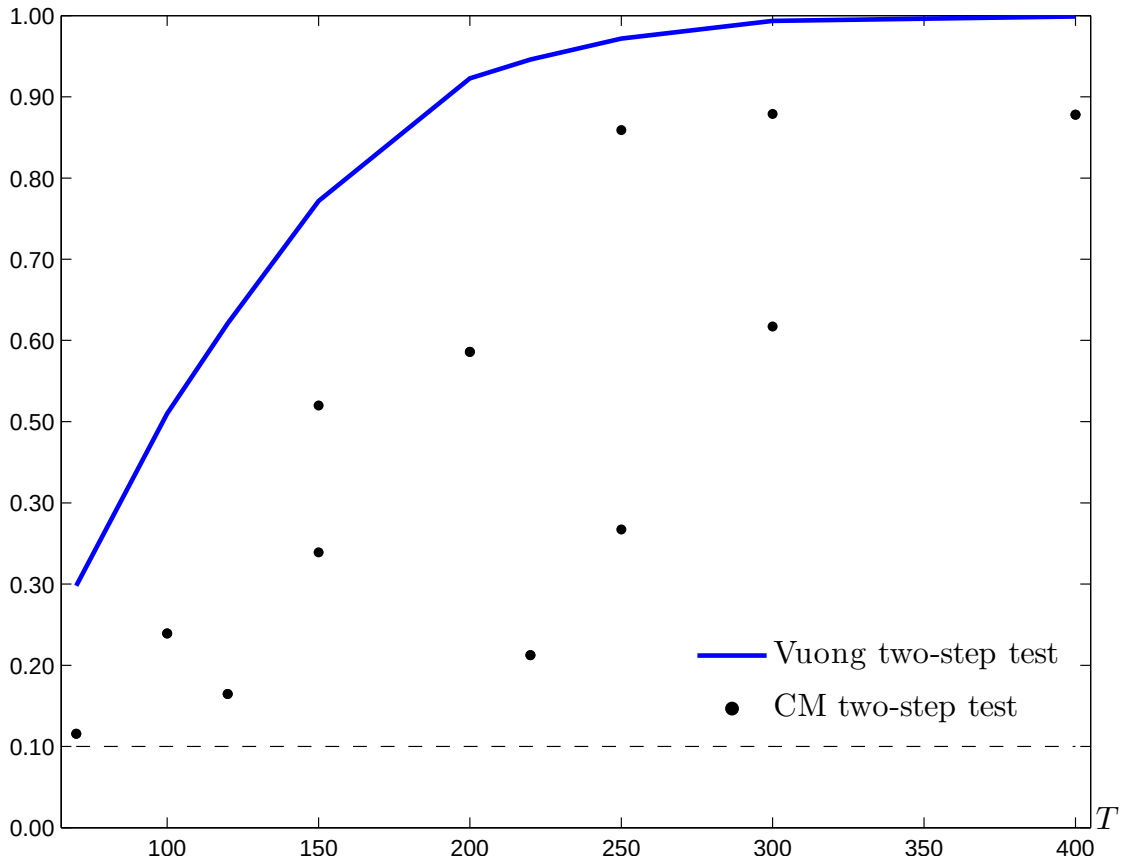

Figure 1: Vuong test rejection frequencies increase monotonically with sample size T, CM test rejection frequencies vary with sample split (R, P)

In one important respect, the comparison between the CM test and the Vuong test needs to be qualified. The CM test—as with all out-of-sample tests generally—hedges against **overfitting**, while the Vuong test used for these experiments does not. That said, Vuong (1989) proposed an Akaike-type variation of his test which penalizes for the number of regressors. Thus, further investigation is needed to see if the loss of power associated with CM two-step tests is as drastic in a more general model selection environment where specified models differ in the number of regressors they possess.

4 Application to GDP growth

This section reproduces the applied example in Clark and McCracken (2014) and assesses if any conclusions change when using the Vuong procedure. The competing forecast models are

$$\text{M1}: \Delta\log GDP_{t+1} = \alpha_0 + \alpha_1 \Delta\log GDP_t + \alpha_2 \Delta\log SP500_t + \alpha_3 \Delta\log SP500_{t-1} + u_{t+1,1} \tag{23}$$

$$\text{M2}: \Delta\log GDP_{t+1} = \beta_0 + \beta_1 \Delta\log GDP_t + \beta_2 spread_t + \beta_3 spread_{t-1} + u_{t+1,2}. \tag{24}$$

where $\log GDP_t$ denotes 400 times the log difference of real GDP (Seasonally Adjusted, Chained 2009\$), $\log SP500_t$ denotes the log difference in the real[4] S&P500 index, and $spread_t$ denotes the credit spread between Moody's AAA-rated and BBB-rated bonds. The CM two-step test is performed with one-quarter-ahead forecasts from 1985:Q12010:Q4. The forecasting models are estimated with a recursively expanding sample starting in 1960:Q1. The Vuong two-step test is computed on the full sample. Note that $\hat{S}_{dd}$ and $\hat{\omega}^2$ are reported without being multiplied by their respective samples sizes P and T because the bootstrapped statistics are invariant to scaling.

The results are summarized in Table 5. For the CM procedure, the results are identical to those reported in Clark and McCracken (2014). We see that we would reject H_0^ω with 90% confidence but fail to reject with 95% confidence. By contrast, the Vuong procedure rejects H_0^ω at 95% confidence. This discrepancy is consistent with our simulation results in Tables 2-4, which show that $\hat{\omega}^2$ has substantially more power than $\hat{S}_{dd}$. In the second step, both procedures fail to reject the null of equally-close models. However, it is worth noting that the sign of the MSE-t statistic is reversed compared to t_V (-0.1641 vs 0.7766). Thus, these two-step approaches each provide weak evidence against the null of equal accuracy but point

[4] S&P500 index is divided by the price index for personal consumption expenditures less food and energy.

in opposite directions.

Table 5: Results of application to forecasts of GDP growth

CM Two-Step Test	
MSPE_1	4.8999
MSPE_2	4.9782
$\bar{d}_{12}$	-0.0783
$\hat{S}_{dd}$	23.6640
90%, 95% bootstrap critical values for $\hat{S}_{dd}$	16.5925, 25.9368
MSE-t	-0.1641
90% Normal Critical Values	$-1.6449,\ 1.6449$
95% Normal Critical Values	$-1.9600,\ 1.9600$
Vuong Two-Step Test	
$\frac{1}{T}\text{LR}_T(\hat{\boldsymbol{\alpha}})$	-2.5609
$\frac{1}{T}\text{LR}_T(\hat{\boldsymbol{\beta}})$	-2.5447
$\frac{1}{T}\text{LR}_T(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$	-0.0161
$\hat{\omega}^2$	0.1630
90%, 95% bootstrap critical values for $\hat{\omega}^2$	$0.0303, 0.0393$
t_V	0.7766
90% Normal Critical Values	$-1.6449,\ 1.6449$
95% Normal Critical Values	$-1.9600,\ 1.9600$

5 Conclusion

This paper corroborates the critique made by Diebold (2013) about the use of out-of-sample tests for purposes beyond forecast comparison. In particular, the Monte Carlo experiments suggest that in the context of overlapping model selection, the out-of-sample tests developed by Clark and McCracken (2014) provide no improvement in terms of statistical size and lead to a substantial reduction in terms of power. With quarterly data, there are cases where this reduction in power can represent decades of wasted information about the true DGP.

Moreover, we saw that the power of the CM two-step test varies with how the sample is split, while the full-sample Vuong test suffers no such drawback. Finally, we examined an applied example where the choice of test leads to different conclusions about whether or not two models are truly overlapping in the sense of Clark and McCracken (2014). At the very least, these results should give the practitioner pause if using an out-of-sample procedure when a full-sample alternative is available.

To save on computation time, this paper has left for future research to examine how these results change if one decides to bootstrap the second step of these two-step procedures. Moreover, CM two-step tests—as with all out-of-sample tests generally—are able to account for overfitting. It would therefore be illuminating to see the Monte Carlo experiments used here extended to a more general model selection framework with a varying number of regressors among competing models. Finally, although the simulations used here made use of the wild bootstrap, the experimental design does not introduce heteroskedasticity into the pseudo-data. And since heteroskedasticity is a common feature in time series data, it would be fruitful to examine how its presence impacts the results presented here.

6 References

Avramov, Doron, "Stock Return Predictability and Model Uncertainty" (2002), Journal of Financial Economics

64, 423-458.

Chao, John, Valentina Corradi and Norman R. Swanson, "Out-Of-Sample Tests for Granger Causality" (2001),

Macroeconomic Dynamics 5, 598-620.

Clark, Todd E. and Michael W. McCracken, "Tests of Equal Forecast Accuracy for Overlapping Models" (2001),

Journal of Applied Econometrics 29, 415-430.

Clark, Todd E. and Michael W. McCracken, "Tests of Equal Forecast Accuracy and Encompassing for Nested

Models" (2001), Journal of Econometrics 105, 85-110.

Diebold, F.X. and Mariano, R.S., "Comparing Predictive Accuracy" (1995), Journal of Business and Economic

Statistics, 13, 253-263.

Gourieroux, C. and Monfort, A., "Statistics and Econometric Models, Volume I" (1995)

Vuong, Q. (1989), "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses," Econometrica, 57,

307-333.

West, Kenneth D., "Asymptotic Inference About Predictive Ability" (1996), Econometrica 64, 1067-1084.