Econometrica, Vol. 79, No. 2 (March, 2011), 453–497

THE MODEL CONFIDENCE SET

BY PETER R. HANSEN, ASGER LUNDE, AND JAMES M. NASON[1]

This paper introduces the *model confidence set* (MCS) and applies it to the selection of models. A MCS is a set of models that is constructed such that it will contain the *best* model with a given level of confidence. The MCS is in this sense analogous to a confidence interval for a parameter. The MCS acknowledges the limitations of the data, such that uninformative data yield a MCS with many models, whereas informative data yield a MCS with only a few models. The MCS procedure does not assume that a particular model is the true model; in fact, the MCS procedure can be used to compare more general objects, beyond the comparison of models. We apply the MCS procedure to two empirical problems. First, we revisit the inflation forecasting problem posed by Stock and Watson (1999), and compute the MCS for their set of inflation forecasts. Second, we compare a number of Taylor rule regressions and determine the MCS of the best regression in terms of in-sample likelihood criteria.

KEYWORDS: Model confidence set, model selection, forecasting, multiple comparisons.

1. INTRODUCTION

ECONOMETRICIANS OFTEN FACE a situation where several models or methods are available for a particular empirical problem. A relevant question is, "Which is the best?" This question is onerous for most data to answer, especially when the set of competing alternatives is large. Many applications will not yield a single model that significantly dominates all competitors because the data are not sufficiently informative to give an unequivocal answer to this question. Nonetheless, it is possible to reduce the set of models to a smaller set of models—a model confidence set—that contains the best model with a given level of confidence.

The objective of the model confidence set (MCS) procedure is to determine the set of models, $\mathcal{M}^*$, that consists of the best model(s) from a collection of models, $\mathcal{M}^0$, where *best* is defined in terms of a criterion that is user-specified. The MCS procedure yields a model confidence set, $\widehat{\mathcal{M}}^*$, that is a collection of models built to contain the best models with a given level of confidence. The process of winnowing models out of $\mathcal{M}^0$ relies on sample information about

[1] The authors thank Joe Romano, Barbara Rossi, Jim Stock, Michael Wolf, and seminar participants at several institutions and the NBER Summer Institute for valuable comments, and Thomas Trimbur for sharing his code for the Baxter–King filter. The Ox language of Doornik (2006) was used to perform the calculations reported here. The first two authors are grateful for financial support from the Danish Research Agency, Grant 24-00-0363, and thank the Federal Reserve Bank of Atlanta for its support and hospitality during several visits. The views in this paper should not be attributed to either the Federal Reserve Bank of Philadelphia or the Federal Reserve System, or any of its staff. The Center for Research in Econometric Analysis of Time Series (CREATES) is a research center at Aarhus University funded by the Danish National Research Foundation.

© 2011 The Econometric Society DOI: 10.3982/ECTA5771

the relative performances of the models in $\mathcal{M}^0$. This sample information drives the MCS to create a random data-dependent set of models, $\widehat{\mathcal{M}}^*$. The set $\widehat{\mathcal{M}}^*$ includes the best model(s) with a certain probability in the same sense that a confidence interval covers a population parameter.

An attractive feature of the MCS approach is that it acknowledges the limitations of the data. Informative data will result in a MCS that contains only the best model. Less informative data make it difficult to distinguish between models and may result in a MCS that contains several (or possibly all) models. Thus, the MCS differs from extant model selection criteria that choose a single model without regard to the information content of the data. Another advantage is that the MCS procedure makes it possible to make statements about significance that are valid in the traditional sense—a property that is not satisfied by the commonly used approach of reporting p-values from multiple pairwise comparisons. Another attractive feature of the MCS procedure is that it allows for the possibility that more than one model can be the best, in which case $\mathcal{M}^*$ contains more than a single model.

The contributions of this paper can be summarized as follows: First, we introduce a model confidence set procedure and establish its theoretical properties. Second, we propose a practical bootstrap implementation of the MCS procedure for a set of problems that includes comparisons of forecasting models evaluated out of sample and regression models evaluated in sample. This implementation is particularly useful when the number of objects to be compared is large. Third, the finite sample properties of the bootstrap MCS procedure are analyzed in simulation studies. Fourth, we apply the MCS procedure to two empirical applications. We revisit the out-of-sample prediction problem of Stock and Watson (1999) and construct MCSs for their inflation forecasts. We also build a MCS for Taylor rule regressions using three likelihood criteria that include the Akaike information criterion (AIC) and Bayesian information criterion (BIC).

1.1. *Theory of Model Confidence Sets*

We do not treat *models* as sacred objects; neither do we assume that a particular model represents the true data generating process. Models are evaluated in terms of a user-specified criterion function. Consequently, the "best" model is unlikely to be replicated for all criteria. Also, we use the term "models" loosely. It can refer to econometric models, competing forecasts, or alternatives that need not involve any modelling of data, such as trading rules. So the MCS procedure is not specific to comparisons of models. For example, one could construct a MCS for a set of different "treatments" by comparing sample estimates of the corresponding treatment effects or construct a MCS for trading rules with the best Sharpe ratio.

A MCS is constructed from a collection of competing objects, $\mathcal{M}^0$, and a criterion for evaluating these objects empirically. The MCS procedure is based on an *equivalence test*, $\delta_{\mathcal{M}}$, and an *elimination rule*, $e_{\mathcal{M}}$. The equivalence test is applied to the set $\mathcal{M} = \mathcal{M}^0$. If $\delta_{\mathcal{M}}$ is rejected, there is evidence that the objects in $\mathcal{M}$ are not equally "good" and $e_{\mathcal{M}}$ is used to eliminate from $\mathcal{M}$ an object with poor sample performance. This procedure is repeated until $\delta_{\mathcal{M}}$ is "accepted" and the MCS is now defined by the set of "surviving" objects. By using the same significance level, α, in all tests, the procedure guarantees that $\lim_{n\to\infty} P(\mathcal{M}^* \subset \widehat{\mathcal{M}}^*_{1-\alpha}) \geq 1 - \alpha$; in the case where $\mathcal{M}^*$ consists of one object, we have the stronger result that $\lim_{n\to\infty} P(\mathcal{M}^* = \widehat{\mathcal{M}}^*_{1-\alpha}) = 1$. The MCS procedure also yields p-values for each of the objects. For a given object, $i \in \mathcal{M}^0$, the MCS p-value, $\hat{p}_i$, is the threshold at which $i \in \widehat{\mathcal{M}}^*_{1-\alpha}$ if and only if $\hat{p}_i \geq \alpha$. Thus, an object with a small MCS p-value makes it unlikely that it is one of the best alternatives in $\mathcal{M}^0$.

The idea behind the sequential testing procedure that we use to construct the MCS may be recognized by readers who are familiar with the trace-test procedure for selecting the rank of a matrix. This procedure involves a sequence of trace tests (see Anderson (1984)), and is commonly used to select the number of cointegration relations within a vector autoregressive model (see Johansen (1988)). The MCS procedure determines the number of superior models in the same way the trace test is used to select the number of cointegration relations. A key difference is that the trace-test procedure has a natural ordering in which the hypotheses are to be tested, whereas the MCS procedure requires a carefully chosen elimination rule to define the sequence of tests. We discuss this issue and related testing procedures in Section 4.

1.2. *Bootstrap Implementation and Simulation Results*

We propose a bootstrap implementation of the MCS procedure that is convenient when the number of models is large. The bootstrap implementation is simple to use in practice and avoids the need to estimate a high-dimensional covariance matrix. White (2000b) is the source of many of the ideas that underlies our bootstrap implementation.

We study the properties of our bootstrap implementation of the MCS procedure through simulation experiments. The results are very encouraging because the best model does end up in the MCS at the appropriate frequency and the MCS procedure does have power to weed out all the poor models when the data contain sufficient information.

1.3. *Empirical Analysis of Inflation Forecasts and Taylor Rules*

We apply the MCS to two empirical problems. First, the MCS is used to study the inflation forecasting problem. The choice of an inflation forecasting model is an especially important issue for central banks, treasuries, and private sector agents. The 50-plus year tradition of the Phillips curve suggests it remains an effective vehicle for the task of inflation forecasting. Stock and

Watson (1999) made the case that "a reasonably specified Phillips curve is the best tool for forecasting inflation"; also see Gordon (1997), Staiger, Stock, and Watson (1997b), and Stock and Watson (2003). Atkeson and Ohanian (2001) concluded that this is not the case because they found it is difficult for any of the Phillips curves they studied to beat a simple no-change forecast in out-of-sample point prediction.

Our first empirical application is based on the Stock and Watson (1999) data set. Several interesting results come out of our analysis. We partition the evaluation period in the same two subsamples as did Stock and Watson (1999). The earlier subsample covers a period with persistent and volatile inflation: this sample is expected to be relatively informative about which models might be the best forecasting models. Indeed, the MCS consists of relatively few models, so the MCS proves to be effective at purging the inferior forecasts. The later subsample is a period in which inflation is relatively smooth and exhibits little volatility. This yields a sample that contains relatively little information about which of the models deliver the best forecasts. However, Stock and Watson (1999) reported that a no-change forecast, which uses last month's inflation rate as the point forecast, is inferior in both subsamples. In spite of the relatively low degree of information in the more recent subsample, we are able to conclude that this no-change forecast is indeed inferior to other forecasts. We come to this conclusion because the Stock and Watson no-change forecast never ends up in the MCS. Next, we add the no-change forecast employed by Atkeson and Ohanian (2001) to the comparison. Their forecast uses the past year's inflation rate as the point prediction rather than month-over-month inflation. This turns out to matter for the second subsample, because the no-change (year) forecast has the smallest mean square prediction error (MSPE) of all forecasts. This enables us to reconcile Stock and Watson (1999) with Atkeson and Ohanian (2001) by showing that their different definitions of the benchmark forecast—no-change (month) and no-change (year), respectively—explain the different conclusions they reach about these forecasts.

Our second empirical example shows that the MCS approach is a useful tool for in-sample evaluation of regression models. This example applies the MCS to choosing from a set of competing (nominal) interest rate rule regressions on a quarterly U.S. sample that runs from 1979 through 2006. These regressions fall into the class of interest rate rules promoted by Taylor (1993). His (Taylor's) rule forms the basis of a class of monetary policy rules that gauge the success of monetary policy at keeping inflation low and the real economy close to trend. The MCS does not reveal which Taylor rule regressions best describe the actual U.S. monetary policy; neither does it identify the best policy rule. Rather the MCS selects the Taylor rule regressions that have the best empirical fit of the U.S. federal funds rate in this sample period, where the "best fit" is defined by different likelihood criteria.

The MCS procedure begins with 25 regression models. We include a pure first-order autoregression, AR(1), of the federal funds rate in the initial MCS.

The remaining 24 models are Taylor rule regressions that contain different combinations of lagged inflation, lags of various definitions of real economic activity (i.e., the output gap, the unemployment rate gap, or real marginal cost), and in some cases the lagged federal funds rate.

It seems that there is limited information in our U.S. sample for the MCS procedure to narrow the set of Taylor rule regressions. The one exception is that the MCS only holds regressions that admit the lagged interest rate. This includes the pure AR(1). The reason is that the time-series properties of the federal funds rate is well explained by its own lag. Thus, the lagged federal funds rate appears to dominate lags of inflation and the real activity variables for explaining the current funds rate. There is some solace for advocates of interest rate rules, because under one likelihood criterion, the MCS often tosses out Taylor rule regression lacking in lags of inflation. Nonetheless, the MCS indicates that the data are consistent with either lags of the output gap, the unemployment rate gap, or real marginal cost playing the role of the real activity variables in the Taylor rule regression. This is not a surprising result. Measurement of gap and marginal cost variables remain an unresolved issue for macroeconometrics; for example, see Orphanides and Van Norden (2002) and Staiger, Stock, and Watson (1997a). It is also true that monetary policymakers rely on sophisticated information sets that cannot be spanned by a few aggregate variables (see Bernanke and Boivin (2003)). The upshot is that the sample used to calculate the MCS has difficulties extracting useful information to separate the pure AR(1) from Taylor rule regressions that include the lagged federal funds rate.

1.4. *Outline of Paper*

The paper is organized as follows. We present the theoretical framework of the MCS in Section 2. Section 3 outlines practical bootstrap methods to implement the MCS. Multiple model comparison methods related to the MCS are discussed in Section 4. Section 5 reports the results of simulation experiments. The MCS is applied to two empirical examples in Section 6. Section 7 concludes. The Supplemental Material (Hansen, Lunde, and Nason (2011)) provides detailed description of our bootstrap implementation and some tables that substantiate the results presented in the simulation and empirical section.

2. GENERAL THEORY FOR MODEL CONFIDENCE SET

In this section, we discuss the theory of model confidence sets for a general set of alternatives. Our leading example concerns the comparison of empirical models, such as forecasting models. Nevertheless, we do not make specific references to models in the first part of this section, in which we lay out the general theory.

We consider a set, $\mathcal{M}^0$, that contains a finite number of objects that are indexed by $i = 1, \dots, m_0$. The objects are evaluated in terms of a loss function and we denote the loss that is associated with object i in period t as $L_{i,t}$, $t = 1, \dots, n$. For example, in the situation where a point forecast $\hat{Y}_{i,t}$ of Y_t is evaluated in terms of a loss function L, we define $L_{i,t} = L(Y_t, \hat{Y}_{i,t})$.

Define the relative performance variables

$$d_{ij,t} \equiv L_{i,t} - L_{j,t}, \quad \text{for all } i, j \in \mathcal{M}^0.$$

This paper assumes that $\mu_{ij} \equiv \mathrm{E}(d_{ij,t})$ is finite and does not depend on t for all $i, j \in \mathcal{M}^0$. We rank alternatives in terms of expected loss, so that alternative i is preferred to alternative j if $\mu_{ij} < 0$.

DEFINITION 1: The set of superior objects is defined by

$$\mathcal{M}^* \equiv \{i \in \mathcal{M}^0 : \mu_{ij} \leq 0 \text{ for all } j \in \mathcal{M}^0\}.$$

The objective of the MCS procedure is to determine $\mathcal{M}^*$. This is done through a sequence of significance tests, where objects that are found to be significantly inferior to other elements of $\mathcal{M}^0$ are eliminated. The hypotheses that are being tested take the form

$$(1) \qquad H_{0,\mathcal{M}} : \mu_{ij} = 0 \quad \text{for all } i, j \in \mathcal{M},$$

where $\mathcal{M} \subset \mathcal{M}^0$. We denote the alternative hypothesis, $\mu_{ij} \neq 0$ for some $i, j \in \mathcal{M}$, by $H_{A,\mathcal{M}}$. Note that $H_{0,\mathcal{M}^*}$ is true given our definition of $\mathcal{M}^*$, whereas $H_{0,\mathcal{M}}$ is false if $\mathcal{M}$ contains elements from $\mathcal{M}^*$ and its complement, $\mathcal{M}^0 \setminus \mathcal{M}^*$. Naturally, the MCS is specific to a set of candidate models, $\mathcal{M}^0$, and therefore silent about the relative merits of objects that are not included in $\mathcal{M}^0$.

We define a model confidence set to be any subset of $\mathcal{M}^0$ that contains all of $\mathcal{M}^*$ with a given probability (its coverage probability). The challenge is to design a procedure that produces a set with the proper coverage probability. The next subsection introduces a generic MCS procedure that meets this requirement. This MCS procedure is constructed from an equivalence test and an elimination rule that are assumed to have certain properties. Next, Section 3 presents feasible tests and elimination rules that can be used for specific problems, such as comparing out-of-sample forecasts and in-sample regression models.

2.1. *The MCS Algorithm and Its Properties*

As stated in the Introduction, the MCS procedure is based on an *equivalence test*, $\delta_{\mathcal{M}}$, and an *elimination rule*, $e_{\mathcal{M}}$. The equivalence test, $\delta_{\mathcal{M}}$, is used to test the hypothesis $H_{0,\mathcal{M}}$ for any $\mathcal{M} \subset \mathcal{M}^0$, and $e_{\mathcal{M}}$ identifies the object of $\mathcal{M}$ that is to be removed from $\mathcal{M}$ in the event that $H_{0,\mathcal{M}}$ is rejected. As a convention, we let $\delta_{\mathcal{M}} = 0$ and $\delta_{\mathcal{M}} = 1$ correspond to the cases where $H_{0,\mathcal{M}}$ are accepted and rejected, respectively.

DEFINITION 2—MCS Algorithm:
Step 0. Initially set $\mathcal{M} = \mathcal{M}^0$.
Step 1. Test $H_{0,\mathcal{M}}$ using $\delta_{\mathcal{M}}$ at level α.
Step 2. If $H_{0,\mathcal{M}}$ is accepted, define $\widehat{\mathcal{M}}^*_{1-\alpha} = \mathcal{M}$; otherwise, use $e_{\mathcal{M}}$ to eliminate an object from $\mathcal{M}$ and repeat the procedure from Step 1.

The set $\widehat{\mathcal{M}}^*_{1-\alpha}$, which consists of the set of surviving objects (those that survived all tests without being eliminated), is referred to as the *model confidence set*. Theorem 1, which is stated below, shows that the term "confidence set" is appropriate in this context, provided that the equivalence test and the elimination rule satisfy the following assumption.

ASSUMPTION 1: *For any* $\mathcal{M} \subset \mathcal{M}^0$, *we assume about* $(\delta_{\mathcal{M}}, e_{\mathcal{M}})$ *that* (a) $\limsup_{n\to\infty} P(\delta_{\mathcal{M}} = 1 | H_{0,\mathcal{M}}) \leq \alpha$, (b) $\lim_{n\to\infty} P(\delta_{\mathcal{M}} = 1 | H_{A,\mathcal{M}}) = 1$, *and* (c) $\lim_{n\to\infty} P(e_{\mathcal{M}} \in \mathcal{M}^* | H_{A,\mathcal{M}}) = 0$.

The conditions that Assumption 1 states for $\delta_{\mathcal{M}}$ are standard requirements for hypothesis tests. Assumption 1(a) requires the asymptotic level not exceed α and Assumption 1(b) requires the asymptotic power be 1, whereas Assumption 1(c) requires that a superior object $i^* \in \mathcal{M}^*$ not be eliminated (as $n \to \infty$) as long as there are inferior models in $\mathcal{M}$.

THEOREM 1—Properties of MCS: *Given Assumption* 1, *it holds that* (i) $\liminf_{n\to\infty} P(\mathcal{M}^* \subset \widehat{\mathcal{M}}^*_{1-\alpha}) \geq 1 - \alpha$ *and* (ii) $\lim_{n\to\infty} P(i \in \widehat{\mathcal{M}}^*_{1-\alpha}) = 0$ *for all* $i \notin \mathcal{M}^*$.

PROOF: Let $i^* \in \mathcal{M}^*$. To prove (i) we consider the event that i^* is eliminated from $\mathcal{M}$. From Assumption 1(c) it follows that $P(\delta_{\mathcal{M}} = 1, e_{\mathcal{M}} = i^* | H_{A,\mathcal{M}}) \leq P(e_{\mathcal{M}} = i^* | H_{A,\mathcal{M}}) \to 0$ as $n \to \infty$. So the probability that a good model is eliminated when $\mathcal{M}$ contains poor models vanishes as $n \to \infty$. Next, Assumption 1(a) shows that $\limsup_{n\to\infty} P(\delta_{\mathcal{M}} = 1, e_{\mathcal{M}} = i^* | H_{0,\mathcal{M}}) = \limsup_{n\to\infty} P(\delta_{\mathcal{M}} = 1 | H_{0,\mathcal{M}}) \leq \alpha$ such that the probability that i^* is eliminated when all models in $\mathcal{M}$ are good models is asymptotically bounded by α. To prove (ii), we first note that $\lim_{n\to\infty} P(e_{\mathcal{M}} = i^* | H_{A,\mathcal{M}}) = 0$ such that only poor models will be eliminated (asymptotically) as long as $\mathcal{M} \nsubseteq \mathcal{M}^*$. On the other hand, Assumption 1(b) ensures that models will be eliminated as long as the null hypothesis is false. *Q.E.D.*

Consider first the situation where the data contain little information such that the equivalence test lacks power and the elimination rule may question a superior model prior to the elimination of all inferior models. The lack of power causes the procedure to terminate too early (on average), and the MCS will contain a large number of models, including several inferior models. We view this as a strength of the MCS procedure. Since lack of power is tied to the lack of information in the data, the MCS should be large when there is insufficient information to distinguish good and bad models.

In the situation where the data are informative, the equivalence test is powerful and will reject all false hypotheses. Moreover, the elimination rule will not question any superior model until all inferior models have been eliminated. (This situation is guaranteed asymptotically.) The result is that the first time a superior model is questioned by the elimination rule is when the equivalence test is applied to $\mathcal{M}^*$. Thus, the probability that one (or more) superior model is eliminated is bounded (asymptotically) by the size of the test! Note that additional superior models may be eliminated in subsequent tests, but these tests will only be performed if $H_{0,\mathcal{M}^*}$ is rejected. Thus, the asymptotic familywise error rate (FWE), which is the probability of making one or more false rejections, is bounded by the level that is used in all tests.

Sequential testing is key for building a MCS. However, econometricians often worry about the properties of a sequential testing procedure, because it can "accumulate" Type I errors with unfortunate consequences (see, e.g., Leeb and Pötscher (2003)). The MCS procedure does not suffer from this problem because the sequential testing is halted when the first hypothesis is accepted.

When there is only a single model in $\mathcal{M}^*$ (one best model), we obtain a stronger result.

COROLLARY 1: *Suppose that Assumption 1 holds and that $\mathcal{M}^*$ is a singleton. Then* $\lim_{n\to\infty} P(\mathcal{M}^* = \widehat{\mathcal{M}}^*_{1-\alpha}) = 1$.

PROOF: When $\mathcal{M}^*$ is a singleton, $\mathcal{M}^* = \{i^*\}$, then it follows from Theorem 1 that i^* will be the last surviving element with probability approaching 1 as $n \to \infty$. The result now follows, because the last surviving element is never eliminated. *Q.E.D.*

2.2. *Coherency Between Test and Elimination Rule*

The previous asymptotic results do not rely on any direct connection between the hypothesis test, $\delta_{\mathcal{M}}$, and the elimination rule, $e_{\mathcal{M}}$. Nonetheless when the MCS is implemented in finite samples, there is an advantage to the hypothesis test and elimination rule being coherent. The next theorem establishes a finite sample version of the result in Theorem 1(i) when there is a certain coherency between the hypothesis test and the elimination rule.

THEOREM 2: *Suppose that $P(\delta_{\mathcal{M}} = 1, e_{\mathcal{M}} \in \mathcal{M}^*) \leq \alpha$. Then we have*

$$P(\mathcal{M}^* \subset \widehat{\mathcal{M}}^*_{1-\alpha}) \geq 1 - \alpha.$$

PROOF: We only need to consider the first instance that $e_{\mathcal{M}} \in \mathcal{M}^*$, because all preceding tests will not eliminate elements that are in $\mathcal{M}^*$. Regardless of the null hypothesis being true or false, we have $P(\delta_{\mathcal{M}} = 1, e_{\mathcal{M}} \in \mathcal{M}^*) \leq \alpha$. So it follows that α bounds the probability that an element from $\mathcal{M}^*$ is eliminated. Additional elements from $\mathcal{M}^*$ may be eliminated in subsequent tests, but these test will only be undertaken if all preceding tests are rejected. So we conclude that $P(\mathcal{M}^* \subset \widehat{\mathcal{M}}^*_{1-\alpha}) \geq 1 - \alpha$. *Q.E.D.*

The property that $P(\delta_{\mathcal{M}} = 1, e_{\mathcal{M}} \in \mathcal{M}^*) \leq \alpha$ holds under both the null hypothesis and the alternative hypothesis is key for the result in Theorem 2. For a test with the correct size, we have $P(\delta_{\mathcal{M}} = 1 | H_{0,\mathcal{M}}) \leq \alpha$, which implies $P(\delta_{\mathcal{M}} = 1, e_{\mathcal{M}} \in \mathcal{M}^* | H_{0,\mathcal{M}}) \leq \alpha$. The additional condition, $P(\delta_{\mathcal{M}} = 1, e_{\mathcal{M}} \in \mathcal{M}^* | H_{A,\mathcal{M}}) \leq \alpha$, ensures that a rejection, $\delta_{\mathcal{M}} = 1$, can be taken as significant evidence that $e_{\mathcal{M}}$ is not in $\mathcal{M}^*$.

In practice, hypothesis tests often rely on asymptotic results that cannot guarantee $P(\delta_{\mathcal{M}} = 1, e_{\mathcal{M}} \in \mathcal{M}^*) \leq \alpha$ holds in finite samples. We provide a definition of coherency between a test and an elimination rule that is useful in situations where testing is grounded on asymptotic distributions. In what follows, we use P_0 to denote the probability measure that arises via imposing the null hypothesis via the transformation $d_{ij,t} \mapsto d_{ij,t} - \mu_{ij}$. Thus P is the true probability measure, whereas P_0 is a simple transformation of P that satisfies the null hypothesis.

DEFINITION 3: There is said to be *coherency* between test and elimination rule when

$$P(\delta_{\mathcal{M}} = 1, e_{\mathcal{M}} \in \mathcal{M}^*) \leq P_0(\delta_{\mathcal{M}} = 1).$$

The coherency in conjunction with an asymptotic control of the Type I error, $\limsup_{n\to\infty} P_0(\delta_{\mathcal{M}} = 1) \leq \alpha$, translates into an asymptotic version of the assumption we made in Theorem 2. Coherency places restrictions on the combinations of tests and elimination rules we can employ. These restrictions go beyond those imposed by the asymptotic conditions we formulated in Assumption 1. In fact, coherency serves to curb the reliance on asymptotic properties so as to avoid perverse outcomes in finite samples that could result from absurd combinations of test and elimination rule. Coherency prevents us from adopting the most powerful test of the hypothesis $H_{0,\mathcal{M}}$ in some situations. The reason is that tests do not necessarily identify a single element as the cause for the rejection. A good analogy is found in the standard regression model, where an F-test may reject the joint hypothesis that all regression coefficients are zero, even though all t-statistics are insignificant.[2]

In our bootstrap implementations of the MCS procedure, we adopt the required coherency between the test and the elimination rule.

[2] Another analogy is that it is easier to conclude that a murder has taken place than it is to determine who committed the murder.

2.3. *MCS p-Values*

In this section we introduce the notion of MCS p-values. The elimination rule, $e_{\mathcal{M}}$, defines a sequence of (random) sets $\mathcal{M}^0 = \mathcal{M}_1 \supset \mathcal{M}_2 \supset \cdots \supset \mathcal{M}_{m_0}$, where $\mathcal{M}_i = \{e_{\mathcal{M}_i}, \ldots, e_{\mathcal{M}_{m_0}}\}$ and m_0 is the number of elements in $\mathcal{M}^0$. So $e_{\mathcal{M}^0} = e_{\mathcal{M}_1}$ is the first element to be eliminated in the event that $H_{0,\mathcal{M}_1}$, is rejected, $e_{\mathcal{M}_2}$ is the second element to be eliminated, and so forth.

DEFINITION 4—MCS p-Values: Let $P_{H_{0,\mathcal{M}_i}}$ denote the p-value associated with the null hypothesis $H_{0,\mathcal{M}_i}$, with the convention that $P_{H_{0,\mathcal{M}_{m_0}}} \equiv 1$. The MCS p-value for model $e_{\mathcal{M}_j} \in \mathcal{M}^0$ is defined by $\hat{p}_{e_{\mathcal{M}_j}} \equiv \max_{i \leq j} P_{H_{0,\mathcal{M}_i}}$.

The advantage of this definition of MCS p-values will be evident from Theorem 3 which is stated below. Since $\mathcal{M}_{m_0}$ consists of a single model, the null hypothesis, $H_{0,\mathcal{M}_{m_0}}$, simply states that the last surviving model is as good as itself, making the convention $P_{H_{0,\mathcal{M}_{m_0}}} \equiv 1$ logical.

Table I illustrates how MCS p-values are computed and how they relate to p-values of the individual tests $P_{H_{0,\mathcal{M}_i}}$, $i = 1, \ldots, m_0$. The MCS p-values are convenient because they make it easy to determine whether a particular object is in $\widehat{\mathcal{M}}^*_{1-\alpha}$ for any α. Thus, the MCS p-values are an effective way to convey the information in the data.

THEOREM 3: *Let the elements of $\mathcal{M}^0$ be indexed by $i = 1, \ldots, m_0$. The MCS p-value, $\hat{p}_i$, is such that $i \in \widehat{\mathcal{M}}^*_{1-\alpha}$ if and only if $\hat{p}_i \geq \alpha$ for any $i \in \mathcal{M}^0$.*

TABLE I

COMPUTATION OF MCS p-VALUES[a]

Elimination Rule	p-Value for $H_{0,\mathcal{M}_k}$	MCS p-Value
$e_{\mathcal{M}_1}$	$P_{H_{0,\mathcal{M}_1}} = 0.01$	$\hat{p}_{e_{\mathcal{M}_1}} = 0.01$
$e_{\mathcal{M}_2}$	$P_{H_{0,\mathcal{M}_2}} = 0.04$	$\hat{p}_{e_{\mathcal{M}_2}} = 0.04$
$e_{\mathcal{M}_3}$	$P_{H_{0,\mathcal{M}_3}} = 0.02$	$\hat{p}_{e_{\mathcal{M}_3}} = 0.04$
$e_{\mathcal{M}_4}$	$P_{H_{0,\mathcal{M}_4}} = 0.03$	$\hat{p}_{e_{\mathcal{M}_4}} = 0.04$
$e_{\mathcal{M}_5}$	$P_{H_{0,\mathcal{M}_5}} = 0.07$	$\hat{p}_{e_{\mathcal{M}_5}} = 0.07$
$e_{\mathcal{M}_6}$	$P_{H_{0,\mathcal{M}_6}} = 0.04$	$\hat{p}_{e_{\mathcal{M}_6}} = 0.07$
$e_{\mathcal{M}_7}$	$P_{H_{0,\mathcal{M}_7}} = 0.11$	$\hat{p}_{e_{\mathcal{M}_7}} = 0.11$
$e_{\mathcal{M}_8}$	$P_{H_{0,\mathcal{M}_8}} = 0.25$	$\hat{p}_{e_{\mathcal{M}_8}} = 0.25$
$\vdots$	$\vdots$	$\vdots$
$e_{\mathcal{M}_{(m_0)}}$	$P_{H_{0,\mathcal{M}_{m_0}}} \equiv 1.00$	$\hat{p}_{e_{\mathcal{M}_{m_0}}} = 1.00$

[a]Note that MCS p-values for some models do not coincide with the p-values for the corresponding null hypotheses. For example, the MCS p-value for $e_{\mathcal{M}_3}$ (the third model to be eliminated) exceeds the p-value for $H_{0,\mathcal{M}_3}$, because the p-value associated with $H_{0,\mathcal{M}_2}$—a null hypothesis tested prior to $H_{0,\mathcal{M}_3}$—is larger.

PROOF: Suppose that $\hat{p}_i < \alpha$ and determine the k for which $i = e_{\mathcal{M}_k}$. Since $\hat{p}_i = \hat{p}_{e_{\mathcal{M}_k}} = \max_{j \le k} P_{H_{0,\mathcal{M}_j}}$, it follows that $H_{0,\mathcal{M}_1}, \dots, H_{0,\mathcal{M}_k}$ are all rejected at significance level α. Hence, the first accepted hypothesis (if any) occurs after $i = e_{\mathcal{M}_k}$ has been eliminated. So $\hat{p}_i < \alpha$ implies $i \notin \widehat{\mathcal{M}}^*_{1-\alpha}$. Suppose now that $\hat{p}_i \ge \alpha$. Then for some $j \le k$, we have $P_{H_{0,\mathcal{M}_j}} \ge \alpha$, in which case $H_{0,\mathcal{M}_j}$ is accepted at significance level α that terminates the MCS procedure before the elimination rule gets to $e_{\mathcal{M}_k} = i$. So $\hat{p}_i \ge \alpha$ implies $i \in \widehat{\mathcal{M}}^*_{1-\alpha}$. This completes the proof. *Q.E.D.*

The interpretation of a MCS p-value is analogous to that of a classical p-value. The analogy is to a $(1-\alpha)$ confidence interval that contains the "true" parameter with a probability no less than $1-\alpha$. The MCS p-value also cannot be interpreted as the probability that a particular model is the best model, exactly as a classical p-value is not the probability that the null hypothesis is true. Rather, the probability interpretation of a MCS p-value is tied to the random nature of the MCS because the MCS is a *random* subset of models that contains $\mathcal{M}^*$ with a certain probability.

3. BOOTSTRAP IMPLEMENTATION

3.1. *Equivalence Tests and Elimination Rules*

Now we consider specific equivalence tests and an elimination rule that satisfy Assumption 1. The following assumption is sufficiently strong to enable us to implement the MCS procedure with bootstrap methods.

ASSUMPTION 2: *For some $r > 2$ and $\gamma > 0$, it holds that $\mathrm{E}|d_{ij,t}|^{r+\gamma} < \infty$ for all $i, j \in \mathcal{M}^0$ and that $\{d_{ij,t}\}_{i,j\in\mathcal{M}^0}$ is strictly stationary with $\mathrm{var}(d_{ij,t}) > 0$ and α-mixing of order $-r/(r-2)$.*

Assumption 2 places restrictions on the relative performance variables, $\{d_{ij,t}\}$, not directly on the loss variables $\{L_{i,t}\}$. For example, a loss function need not be stationary as long as the loss differentials, $\{d_{ij,t}\}$, $i, j \in \mathcal{M}^0$, satisfy Assumption 2. The assumption allows for some types of structural breaks and other features that can create nonstationary $\{L_{i,t}\}$ as long as all objects in $\mathcal{M}^0$ are affected in a similar way that preserves the stationarity of $\{d_{ij,t}\}$.

3.1.1. *Quadratic-Form Test*

Let $\mathcal{M}$ be some subset of $\mathcal{M}^0$ and let m be the number of models in $\mathcal{M} = \{i_1, \dots, i_m\}$. We define the vector of loss variables $L_t \equiv (L_{i_1,t}, \dots, L_{i_m,t})'$, $t = 1, \dots, n$, and its sample average $\bar{L} \equiv n^{-1}\sum_{t=1}^{n} L_t$, and we let $\iota \equiv (1, \dots, 1)'$ be the column vector where all m entries equal 1. The orthogonal complement to ι is an $m \times (m-1)$ matrix, $\iota_\perp$ that has full column rank and satisfies $\iota_\perp' \iota = 0$

(a vector of zeros). The $m-1$-dimensional vector $X_t \equiv \iota_\perp' L_t$ can be viewed as $m-1$ contrasts, because each element of X_t is a linear combination of $d_{ij,t}$, $i, j \in \mathcal{M}$, which has mean zero under the null hypothesis.

LEMMA 1: *Given Assumption 2, let $X_t \equiv \iota_\perp' L_t$ and define $\theta \equiv \mathrm{E}(X_t)$. The null hypothesis $H_{0,\mathcal{M}}$ is equivalent to $\theta = 0$ and it holds that $n^{1/2}(\bar{X} - \theta) \overset{d}{\to} N(0, \Sigma)$, where $\bar{X} \equiv n^{-1}\sum_{t=1}^{n} X_t$ and $\Sigma \equiv \lim_{n\to\infty} \mathrm{var}(n^{1/2}\bar{X})$.*

PROOF: Note that $X_t = \iota_\perp' L_t$ can be written as a linear combination of $d_{ij,t}$, $i, j \in \mathcal{M}^0$, because $\iota_\perp' \iota = 0$. Thus $H_{0,\mathcal{M}}$ is given by $\theta = 0$ and the asymptotic normality follows by the central limit theorem for α-mixing processes (see, e.g., White (2000a)). *Q.E.D.*

Lemma 1 shows that $H_{0,\mathcal{M}}$ can be tested using traditional quadratic-form statistics. An example is $T_Q \equiv n\bar{X}'\hat{\Sigma}^{\#}\bar{X}$, where $\hat{\Sigma}$ is some consistent estimator of Σ and $\hat{\Sigma}^{\#}$ denotes the Moore–Penrose inverse of $\hat{\Sigma}$.[3] The rank $q \equiv \mathrm{rank}(\hat{\Sigma})$ represents the effective number of *contrasts* (the number of linearly independent comparisons) under $H_{0,\mathcal{M}}$. Since $\hat{\Sigma} \overset{p}{\to} \Sigma$ (by assumption), it follows that $T_Q \overset{d}{\to} \chi^2_{(q)}$, where $\chi^2_{(q)}$ denotes the χ^2 distribution with q degrees of freedom. Under the alternative hypothesis, T_Q diverge to infinity with probability 1. So the test $\delta_{\mathcal{M}}$ will meet the requirements of Assumption 1 when constructed from T_Q. Although the matrix $\iota_\perp$ is not fully identified by the requirements $\iota_\perp'\iota = 0$ and $\det(\iota_\perp'\iota_\perp) \neq 0$ (but the subspace spanned by the columns of $\iota_\perp$ is), there is no problem because the statistic T_Q is invariant to the choice for $\iota_\perp$.

A rejection of the null hypothesis based on the quadratic-form test need not identify an inferior alternative because a large value of T_Q can stem from several $\bar{d}_{ij}$ being slightly different from zero. To achieve the required coherence between test and elimination rule, additional testing is needed. Specifically, one needs to test all subhypotheses of any rejected hypothesis, unless the subhypothesis is nested in an accepted hypothesis, before further elimination is justified. The underlying principle is known as the *closed testing procedure* (see Lehmann and Romano (2005, pp. 366–367)).

When m is large relative to the sample size, n, reliable estimates of Σ are difficult to obtain, because the number of elements of Σ to be estimated are of order m^2. It is convenient to use a test statistic that does not require an explicit estimate of Σ in this case. We consider test statistics that resolve this issue in the next section.

[3] Under the additional assumption that $\{d_{ij,t}\}_{i,j\in\mathcal{M}}$ is uncorrelated (across t), we can use $\hat{\Sigma} = n^{-1}\sum_{t=1}^{n}(X_t - \bar{X})(X_t - \bar{X})'$. Otherwise, we need a robust estimator along the lines of Newey and West (1987). In the context of comparing forecasts, West and Cho (1995) were the first investigators to use the test statistic T_Q. They based their test on (asymptotic) critical values from $\chi^2_{(m-1)}$.

3.1.2. *Tests Constructed From t-Statistics*

This section develops two tests that are based on multiple t-statistics. This approach has two advantages. First, it bypasses the need for an explicit estimate of Σ. Second, the multiple t-statistic approach simplifies the construction of an elimination rule that satisfies the notion of coherency formulated in Definition 3.

Define the relative sample loss statistics $\bar{d}_{ij} \equiv n^{-1}\sum_{t=1}^{n} d_{ij,t}$ and $\bar{d}_{i\cdot} \equiv m^{-1}\sum_{j\in\mathcal{M}} \bar{d}_{ij}$. Here $\bar{d}_{ij}$ measures the relative sample loss between the ith and jth models, while $\bar{d}_{i\cdot}$ is the sample loss of the ith model relative to the average across models in $\mathcal{M}$. The latter can be seen from the identity $\bar{d}_{i\cdot} = (\bar{L}_i - \bar{L}_{\cdot})$, where $\bar{L}_i \equiv n^{-1}\sum_{t=1}^{n} L_{i,t}$ and $\bar{L}_{\cdot} \equiv m^{-1}\sum_{i\in\mathcal{M}} \bar{L}_i$. From these statistics, we construct the t-statistics

$$t_{ij} = \frac{\bar{d}_{ij}}{\sqrt{\widehat{\operatorname{var}}(\bar{d}_{ij})}} \quad \text{and} \quad t_{i\cdot} = \frac{\bar{d}_{i\cdot}}{\sqrt{\widehat{\operatorname{var}}(\bar{d}_{i\cdot})}} \quad \text{for} \quad i, j \in \mathcal{M},$$

where $\widehat{\operatorname{var}}(\bar{d}_{ij})$ and $\widehat{\operatorname{var}}(\bar{d}_{i\cdot})$ denote estimates of $\operatorname{var}(\bar{d}_{ij})$ and $\operatorname{var}(\bar{d}_{i\cdot})$, respectively. The first statistic, t_{ij}, is used in the well known test for comparing two forecasts; see Diebold and Mariano (1995) and West (1996). The t-statistics t_{ij} and $t_{i\cdot}$ are associated with the null hypothesis that $H_{ij}: \mu_{ij} = 0$ and $H_{i\cdot}: \mu_{i\cdot} = 0$, where $\mu_{i\cdot} = \mathrm{E}(\bar{d}_{i\cdot})$. These statistics form the basis of tests of the hypothesis $H_{0,\mathcal{M}}$. We take advantages of the equivalence between $H_{0,\mathcal{M}}$, $\{H_{ij}$ for all $i, j \in \mathcal{M}\}$, and $\{H_{i\cdot}$ for all $i \in \mathcal{M}\}$. With $\mathcal{M} = \{i_1, \dots, i_m\}$ the equivalence follows from

$$\begin{aligned} \mu_{i_1} = \cdots = \mu_{i_m} \quad &\Leftrightarrow \quad \mu_{ij} = 0 \quad \text{for all } i, j \in \mathcal{M} \\ &\Leftrightarrow \quad \mu_{i\cdot} = 0 \quad \text{for all } i \in \mathcal{M}. \end{aligned}$$

Moreover, the equivalence extends to $\{\mu_{i\cdot} \le 0$ for all $i \in \mathcal{M}\}$ as well as $\{|\mu_{ij}| \le 0$ for all $i, j \in \mathcal{M}\}$, and these two formulations of the null hypothesis map naturally into the test statistics

$$T_{\max,\mathcal{M}} = \max_{i\in\mathcal{M}} t_{i\cdot} \quad \text{and} \quad T_{R,\mathcal{M}} \equiv \max_{i,j\in\mathcal{M}} |t_{ij}|,$$

which are available to test the hypothesis $H_{0,\mathcal{M}}$.[4] The asymptotic distributions of these test statistics are nonstandard because they depend on nuisance parameters (under the null and the alternative). However, the nuisance parameters pose few obstacles, as the relevant distributions can be estimated with bootstrap methods that implicitly deal with the nuisance parameter problem.

[4] An earlier version of this paper has results for the test statistics $T_D = \sum_{j=1}^{n} t_{i\cdot}^2$ and T_Q.

This feature of the bootstrap has previously been used in this context by Kilian (1999), White (2000b), Hansen (2003b, 2005), and Clark and McCracken (2005).

Characterization of the MCS procedure needs an elimination rule, $e_{\mathcal{M}}$, that meets the requirements of Assumption 1(c) and the coherency of Definition 3. For the test statistic $T_{\max,\mathcal{M}}$, the natural elimination rule is $e_{\max,\mathcal{M}} \equiv \arg\max_{i\in\mathcal{M}} t_{i\cdot}$ because a rejection of the null hypothesis identifies the hypothesis $\mu_{j\cdot} = 0$ as false for $j = e_{\max,\mathcal{M}}$. In this case the elimination rule removes the model that contributes most to the test statistic. This model has the largest standardized excess loss relative to the average across all models in $\mathcal{M}$. With the other test statistic, $T_{R,\mathcal{M}}$, the natural elimination rule is $e_{R,\mathcal{M}} = \arg\max_{i\in\mathcal{M}} \sup_{j\in\mathcal{M}} t_{ij}$ because this model is such that $t_{e_{R,\mathcal{M}}j} = T_{R,\mathcal{M}}$ for some $j \in \mathcal{M}$. These combinations of test and elimination rule will satisfy the required coherency.

PROPOSITION 1: *Let $\delta_{\max,\mathcal{M}}$ and $\delta_{R,\mathcal{M}}$ denote the tests based on the statistics $T_{\max,\mathcal{M}}$ and $T_{R,\mathcal{M}}$, respectively. Then $(\delta_{\max,\mathcal{M}}, e_{\max,\mathcal{M}})$ and $(\delta_{R,\mathcal{M}}, e_{R,\mathcal{M}})$ satisfy the coherency of Definition 3.*

PROOF: Let T_i denote either $t_{i\cdot}$ or $\max_{j\in\mathcal{M}} t_{ij}$, and note that the test statistics $T_{\max,\mathcal{M}}$ and $T_{R,\mathcal{M}}$ are both of the form $T = \max_{i\in\mathcal{M}} T_i$. Let P_0 be as defined in Section 2.2. From the definitions of $t_{i\cdot}$ and t_{ij}, we have for $i \in \mathcal{M}^*$ the first-order stochastic dominance result $P_0(\max_{i\in\mathcal{M}'} T_i > x) \geq P(\max_{i\in\mathcal{M}'} T_i > x)$ for any $\mathcal{M}' \subset \mathcal{M}^*$ and all $x \in \mathbb{R}$. The coherency now follows from

$$\begin{aligned}
&P(T > c, e_{\mathcal{M}} = i \text{ for some } i \in \mathcal{M}^*) \\
&\quad = P(T > c, T = T_i \text{ for some } i \in \mathcal{M}^*) \\
&\quad = P\Big(\max_{i\in\mathcal{M}\cap\mathcal{M}^*} T_i > c, T_i \geq T_j \text{ for all } j \in \mathcal{M}\Big) \leq P\Big(\max_{i\in\mathcal{M}\cap\mathcal{M}^*} T_i > c\Big) \\
&\quad \leq P_0\Big(\max_{i\in\mathcal{M}\cap\mathcal{M}^*} T_i > c\Big) \leq P_0\Big(\max_{i\in\mathcal{M}} T_i > c\Big) = P_0(T > c).
\end{aligned}$$

This completes the proof. *Q.E.D.*

Next, we establish two intermediate results that underpin the bootstrap implementation of the MCS.

LEMMA 2: *Suppose that Assumption 2 holds and define $\bar{Z} = (\bar{d}_{1\cdot}, \ldots, \bar{d}_{m\cdot})'$. Then*

$$(2) \qquad n^{1/2}(\bar{Z} - \psi) \xrightarrow{d} N_m(0, \Omega) \quad \text{as} \quad n \to \infty,$$

where $\psi \equiv \mathrm{E}(\bar{Z})$ and $\Omega \equiv \lim_{n\to\infty} \mathrm{var}(n^{1/2}\bar{Z})$, and the null hypothesis $H_{0,\mathcal{M}}$ is equivalent to: $\psi = 0$.

PROOF: From the identity $\bar{d}_{i\cdot} = \bar{L}_i - \bar{L}_{\cdot} = \bar{L}_i - m^{-1}\sum_{j\in\mathcal{M}}\bar{L}_j = m^{-1} \times \sum_{j\in\mathcal{M}}(\bar{L}_i - \bar{L}_j) = m^{-1}\sum_{j\in\mathcal{M}}\bar{d}_{ij}$, we see that the elements of $\bar{Z}$ are linear transformations of $\bar{X}$ from Lemma 1. Thus for some $(m-1)\times m$ matrix G, we have $\bar{Z} = G'\bar{X}$ and the result now follows, where $\psi = G'\theta$ and $\Omega = G'\Sigma G$. (The $m \times m$ covariance matrix Ω has reduced rank, as $\operatorname{rank}(\Omega) \leq m-1$.) *Q.E.D.*

In the following discussion, we let ϱ denote the $m\times m$ correlation matrix that is implied by the covariance matrix Ω of Lemma 2. Further, given the vector of random variables $\xi \sim N_m(0,\varrho)$, we let F_ϱ denote the distribution of $\max_i \xi_i$.

THEOREM 4: *Let Assumption 2 hold and suppose that* $\hat{\omega}_i^2 \equiv \widehat{\operatorname{var}}(n^{1/2}\bar{d}_{i\cdot}) = n\widehat{\operatorname{var}}(\bar{d}_{i\cdot}) \xrightarrow{p} \omega_i^2$, *where* ω_i^2, $i = 1,\dots,m$, *are the diagonal elements of* Ω. *Under* $H_{0,\mathcal{M}}$, *we have* $T_{\max,\mathcal{M}} \xrightarrow{d} F_\varrho$, *and under the alternative hypothesis* $H_{A,\mathcal{M}}$, *we have* $T_{\max,\mathcal{M}} \to \infty$ *in probability. Moreover, under the alternative hypothesis, we have* $T_{\max,\mathcal{M}} = t_{j\cdot}$, *where* $j = e_{\max,\mathcal{M}} \notin \mathcal{M}^*$ *for n sufficiently large.*

PROOF: Let $D \equiv \operatorname{diag}(\omega_1^2,\dots,\omega_m^2)$ and $\hat{D} \equiv \operatorname{diag}(\hat{\omega}_1^2,\dots,\hat{\omega}_m^2)$. From Lemma 2 it follows that $\xi_n = (\xi_{1,n},\dots,\xi_{m,n})' \equiv D^{-1/2}n^{1/2}\bar{Z} \xrightarrow{d} N_m(0,\varrho)$, since $\varrho = D^{-1/2}\Omega D^{-1/2}$. From $t_{i\cdot} = \bar{d}_{i\cdot}/\sqrt{\widehat{\operatorname{var}}(\bar{d}_{i\cdot})} = n^{1/2}\bar{d}_{i\cdot}/\hat{\omega}_i = \xi_{i,n}\frac{\omega_i}{\hat{\omega}_i}$, it now follows that $T_{\max,\mathcal{M}} = \max_i t_{i\cdot} = \max_i(\hat{D}^{-1/2}n^{1/2}\bar{Z})_i \xrightarrow{d} F_\varrho$. Under the alternative hypothesis, we have $\bar{d}_{j\cdot} \xrightarrow{p} \mu_{j\cdot} > 0$ for any $j \notin \mathcal{M}^*$, so that both $t_{j\cdot}$ and $T_{\max,\mathcal{M}}$ diverge to infinity at rate $n^{1/2}$ in probability. Moreover, it follows that $e_{\max,\mathcal{M}} \notin \mathcal{M}^*$ for n sufficiently large. *Q.E.D.*

Theorem 4 shows that the asymptotic distribution of $T_{\max,\mathcal{M}}$ depends on the correlation matrix ϱ. Nonetheless, as discussed earlier, bootstrap methods can be employed to deal with this nuisance parameter problem. Thus, we construct a test of $H_{0,\mathcal{M}}$ by comparing the test statistic $T_{\max,\mathcal{M}}$ to an estimate of the 95% quantile, say, of its limit distribution under the null hypothesis. Although the quantile may depend on ϱ, our bootstrap implementation leads to an asymptotically valid test because the bootstrap consistently estimates the desired quantile. A detailed description of our bootstrap implementation is available in a separate appendix (Hansen, Lunde, and Nason (2011)).

Theorem 4 formulates results for the situation where the MCS is constructed with $T_{\max,\mathcal{M}}$ and $e_{\max,\mathcal{M}} = \arg\max_i t_{i\cdot}$. Similar results hold for the MCS that is constructed from $T_{R,\mathcal{M}}$ and $e_{R,\mathcal{M}}$. The arguments are almost identical to those used for Theorem 4.

3.2. *MCS for Regression Models*

This section shows how to construct the MCS for regression models using likelihood-based criteria. Information criteria, such as the AIC and BIC, are special cases for building a MCS of regression models. The MCS approach departs from standard practice where the AIC and BIC select a single model, but are silent about the uncertainty associated with this selection.[5] Thus, the MCS procedure yields valuable additional information about the uncertainty surrounding model selection. In Section 6.2, application of the MCS procedure in sample to Taylor rule regressions indicates this uncertainty can be substantial.

Although we focus on regression models for simplicity, it will be evident that the MCS procedure laid out in this setting can be adapted to more complex models, such as the type of models analyzed in Sin and White (1996).

3.2.1. *Framework and Assumptions*

Consider the family of regression models $Y_t = \beta_j' X_{j,t} + \varepsilon_{j,t}$, $t = 1, \dots, n$, where $X_{j,t}$ is a subset of the variables in X_t for $j = 1, \dots, m_0$. The set of regression models, $\mathcal{M}^0$, may consist of nested, nonnested, and overlapping specifications.

Throughout we assume that the pair (Y_t, X_t') is strictly stationary and satisfies Assumption 1 in Goncalves and White (2005). This justifies our use of the moving-block bootstrap to implement our resampling procedure. The framework of Goncalves and White (2005) permits weak serial dependence in (Y_t, X_t'), which is important for many applications.

The population parameters for each of the models are defined by $\beta_{0j} = [\mathrm{E}(X_{j,t} X_{j,t}')]^{-1} \mathrm{E}(X_{j,t} Y_t)$ and $\sigma_{0j}^2 = \mathrm{E}(\varepsilon_{j,t}^2)$, where $\varepsilon_{j,t} = Y_t - \beta_{0j}' X_{j,t}$, $t = 1, \dots, n$. Furthermore, the Gaussian quasi-log-likelihood function is, apart from a constant, given by

$$\ell(\beta_j, \sigma_j^2) = -\frac{n}{2} \log \sigma_j^2 - \sigma_j^{-2} \frac{1}{2} \sum_{t=1}^{n} (Y_t - \beta_j' X_{j,t})^2.$$

3.2.2. *MCS by Kullback–Leibler Divergence*

One way to define the best regression model is in terms of the Kullback–Leibler information criterion (KLIC) (see, e.g., Sin and White (1996)). This is equivalent to ranking the models in terms of the expected value of the quasi-log-likelihood function when evaluated at their respective population parameters, that is, $\mathrm{E}[\ell(\beta_{0j}, \sigma_{0j}^2)]$. It is convenient to define

$$Q(\mathcal{Z}, \theta_j) = -2\ell(\beta_j, \sigma_j^2) = n \log \sigma_j^2 + \sum_{t=1}^{n} \frac{(Y_t - \beta_j' X_{j,t})^2}{\sigma_j^2},$$

[5] The same point applies to the Autometrics procedure; see Doornik (2009) and references therein. Autometrics is constructed from a collection of tests and decision rules but does not control a familywise error rate, and the set of models that Autometrics seeks to identify is not defined from a single criterion, such as the Kullback–Leibler information criterion.

where θ_j can be viewed as a high-dimensional vector that is restricted by the parameter space $\Theta_j \subset \Theta$ that defines the jth regression model. The population parameters are here given by $\theta_{0j} = \arg\min_{\theta \in \Theta_j} \mathrm{E}[Q(\mathcal{Z}, \theta)]$, $j = 1, \ldots, m_0$, and the best model is defined by $\min_j \mathrm{E}[Q(\mathcal{Z}, \theta_{0j})]$. In the notation of the MCS framework, the KLIC leads to

$$\mathcal{M}^*_{\mathrm{KLIC}} = \Big\{ j : \mathrm{E}[Q(\mathcal{Z}, \theta_{0j})] = \min_i \mathrm{E}[Q(\mathcal{Z}, \theta_{0i})] \Big\},$$

which (as always) permits the existence of more than one best model.[6] The extension to other criteria, such as the AIC and the BIC, is straightforward. For instance, the set of best models in terms of the AIC is given by $\mathcal{M}^*_{\mathrm{AIC}} = \{j : \mathrm{E}[Q(\mathcal{Z}, \theta_{0j}) + 2k_j] = \min_i \mathrm{E}[Q(\mathcal{Z}, \theta_{0i}) + 2k_i]\}$, where k_j is the degrees of freedom in the jth model.

The likelihood framework enables us to construct either $\widehat{\mathcal{M}}^*_{\mathrm{KLIC}}$ or $\widehat{\mathcal{M}}^*_{\mathrm{AIC}}$ by drawing on the theory of quasi-maximum-likelihood estimation (see, e.g., White (1994)). Since the family of regression models is linear, the quasi-maximum-likelihood estimators are standard, $\hat{\beta}_j = (\sum_{t=1}^n X_{j,t} X'_{j,t})^{-1} \times \sum_{t=1}^n X_{j,t} Y_t$, and $\hat{\sigma}_j^2 = n^{-1} \sum_{t=1}^n \hat{\varepsilon}_{j,t}^2$, where $\hat{\varepsilon}_{j,t} = Y_t - \hat{\beta}'_j X_{j,t}$. We have

$$\begin{aligned} &Q(\mathcal{Z}, \hat{\theta}_j) - Q(\mathcal{Z}, \theta_{0j}) \\ &\quad = n \Bigg\{ (\log \sigma_{0j}^2 - \log \hat{\sigma}_j^2) + \Bigg(n^{-1} \sum_{t=1}^n \varepsilon_{j,t}^2 / \sigma_{0j}^2 - 1 \Bigg) \Bigg\}, \end{aligned}$$

which is the quasi-likelihood ratio (QLR) statistic for the null hypothesis, $H_0 : \theta = \theta_{0j}$.

In the event that the jth model is correctly specified, it is well known that the limit distribution of $Q(\mathcal{Z}, \hat{\theta}_j) - Q(\mathcal{Z}, \theta_{0j})$ is $\chi^2_{(k_j)}$, where the degrees of freedom, k_j, is given by the dimension of $\theta_{0j} = (\beta'_{0j}, \sigma_{0j}^2)'$. In the present multimodel setup, it is unlikely that all models are correctly specified. More generally, the limit distribution of the QLR statistic has the form, $\sum_{i=1}^{k_j} \lambda_{i,j} Z_{i,j}^2$, where $\lambda_{1,j}, \ldots, \lambda_{k_j,j}$ are the eigenvalues of $\mathcal{I}_j^{-1} \mathcal{J}_j$ and $Z_{1,j}, \ldots, Z_{k_j,j} \sim \text{i.i.d.} N(0, 1)$. The information matrices $\mathcal{I}_j$ and $\mathcal{J}_j$ are those associated with the jth model,

[6] In the present situation, we have $\mathrm{E}[Q(\mathcal{Z}_j, \theta_{0j})] \propto \sigma_{0j}^2$. The implication is that the error variance, σ_{0j}^2, induces the same ranking as KLIC, so that $\mathcal{M}^*_{\mathrm{KLIC}} = \{j : \sigma_{0j}^2 = \min_{j'} \sigma_{0j'}^2\}$.

$\mathcal{I}_j = \mathrm{diag}(\sigma_{0j}^{-2}\mathrm{E}(X_{j,t}X_{j,t}'), \frac{1}{2}\sigma_{0j}^{-4})$ and

$$\mathcal{J}_j = \mathrm{E}\begin{pmatrix} \sigma_{0j}^{-4}n^{-1}\sum_{s,t=1}^{n} X_{j,s}\varepsilon_{j,s}\varepsilon_{j,t}X_{j,t}' & \frac{1}{2}\sigma_{0j}^{-6}n^{-1}\sum_{s,t=1}^{n} X_{j,s}\varepsilon_{j,s}\varepsilon_{j,t}^2 \\ \bullet & \frac{1}{4}\sigma_{0j}^{-8}n^{-1}\sum_{s,t=1}^{n}(\varepsilon_{j,s}^2\varepsilon_{j,t}^2 - \sigma_{0j}^4) \end{pmatrix}.$$

The effective degrees of freedom, $k_j^\star$, is defined by the mean of the QLR limit distribution:

$$\begin{aligned} k_j^\star &= \lambda_{1,j} + \cdots + \lambda_{k_j,j} = \mathrm{tr}\{\mathcal{I}_j^{-1}\mathcal{J}_j\} \\ &= \mathrm{tr}\left\{[\mathrm{E}(X_{j,t}X_{j,t}')]^{-1}\sigma_{0j}^{-2}n^{-1}\sum_{s,t=1}^{n}\mathrm{E}(X_{j,s}\varepsilon_{j,s}X_{j,t}'\varepsilon_{j,t})\right\} \\ &\quad + n^{-1}\frac{1}{2}\sum_{s,t=1}^{n}\mathrm{E}\left(\frac{\varepsilon_{j,s}^2\varepsilon_{j,t}^2}{\sigma_{0j}^4} - 1\right). \end{aligned}$$

The previous expression points to estimating $k_j^\star$ with heteroskedasticity and autocorrelation consistent (HAC) type estimators that account for the autocorrelation in $\{X_{j,t}\varepsilon_{j,t}\}$ and $\{\varepsilon_{j,t}^2\}$ (e.g., Newey and West (1987) and Andrews (1991)). Below we use a simple bootstrap estimate of $k_j^\star$, which is also employed in our simulations and our empirical Taylor rule regression application.

The effective degrees of freedom in the context of misspecified models was first derived by Takeuchi (1976). He proposed a modified AIC, sometimes referred to as the Takeuchi information criterion (TIC), which computes the penalty with the effective degrees of freedom rather than the number of parameters as is used by the AIC; see also Sin and White (1996) and Hong and Preston (2008). We use the notation AIC* and BIC* to denote the information criteria that are defined by substituting the effective degrees of freedom $k_j^\star$ for k_j in the AIC and BIC, respectively. In this case, our AIC* is identical to the TIC proposed by Takeuchi (1976).

3.2.3. *The MCS Procedure*

The MCS procedure can be implemented by the moving-block bootstrap applied to the pair (Y_t, X_t); see Goncalves and White (2005). We compute resamples $\mathcal{Z}_b^* = (Y_{b,t}^*, X_{b,t}^*)_{t=1}^n$ for $b = 1, \dots, B$, which equates the original point estimate, $\hat{\theta}_j$, to the population parameter in the jth model under the bootstrap scheme.

The literature has proposed several bootstrap estimators of the effective degrees of freedom, $k_j^\star = \mathrm{E}[Q(\mathcal{Z}, \theta_{0j}) - Q(\mathcal{Z}, \hat{\theta}_j)]$; see, for example, Efron

(1983, 1986) and Cavanaugh and Shumway (1997). These and additional estimators are analyzed and compared in Shibata (1997). We adopt the estimator for $k^{\star}_j$ that is labelled B_3 in Shibata (1997). In the regression context, this estimator takes the form

$$\begin{aligned}\hat{k}^{\star}_j &= B^{-1}\sum_{b=1}^{B} Q(\mathcal{Z}^*_b, \hat{\theta}_j) - Q(\mathcal{Z}^*_b, \hat{\theta}^*_{b,j}) \\ &= B^{-1}\sum_{b=1}^{B}\left\{ n\log\frac{\hat{\sigma}^2_j}{\hat{\sigma}^{*2}_{b,j}} + \frac{\sum_{t=1}^{n}(\varepsilon^*_{b,j,t})^2}{\hat{\sigma}^2_j} - n\right\},\end{aligned}$$

where $\varepsilon^*_{b,j,t} = Y^*_{b,t} - \hat{\beta}'_j X^*_{b,j,t}$, $\hat{\varepsilon}^*_{b,j,t} = Y^*_{b,t} - \hat{\beta}^{*\prime}_{b,j} X^*_{b,j,t}$, and $\hat{\sigma}^{*2}_{b,j} = n^{-1}\sum_{t=1}^{n}(\hat{\varepsilon}^*_{b,j,t})^2$. This is an estimate of the expected overfit that results from maximization of the likelihood function. For a correctly specified model, we have $k^{\star}_j = k_j$, so we would expect $\hat{k}^{\star}_j \approx k_j$ when the jth model is correctly specified. This is indeed what we find in our simulations; see Section 5.2.

Given an estimate of the effective degrees of freedom $\hat{k}^{\star}_j$, compute the AIC* statistic $Q(\mathcal{Z}, \hat{\theta}_j) + \hat{k}^{\star}_j$, which is centered about $\mathrm{E}\{Q(\mathcal{Z}, \theta_{0j})\}$. The null hypothesis $H_{0,\mathcal{M}}$ states that $\mathrm{E}[Q(\mathcal{Z}, \theta_{0i}) - Q(\mathcal{Z}, \theta_{0j})] = 0$ for all $i, j \in \mathcal{M}$. This motivates the range statistic

$$T_{R,\mathcal{M}} = \max_{i,j\in\mathcal{M}}\left|[Q(\mathcal{Z}, \hat{\theta}_i) + \hat{k}^{\star}_i] - [Q(\mathcal{Z}, \hat{\theta}_j) + \hat{k}^{\star}_j]\right|$$

and the elimination rule $e_{\mathcal{M}} = \arg\max_{j\in\mathcal{M}}[Q(\mathcal{Z}, \hat{\theta}_j) + \hat{k}^{\star}_j]$. This elimination rule removes the model with the largest bias adjusted residual variance. Our test statistic, $T_{R,\mathcal{M}}$, is a range statistic over recentered QLR statistics computed for all pairs of models in $\mathcal{M}$. In the special case with independent and identically distributed (i.i.d.) data and just two models in $\mathcal{M}$, we could simply adopt the QLR test of Vuong (1989) as our equivalence test.

Next, we estimate the distribution of $T_{R,\mathcal{M}}$ under the null hypothesis. The estimate is calculated with methods similar to those used in White (2000b) and Hansen (2005). The joint distribution of

$$\begin{aligned}&\big(Q(\mathcal{Z}, \hat{\theta}_1) + k^{\star}_1 - \mathrm{E}[Q(\mathcal{Z}, \theta_{01})], \dots, \\ &\quad Q(\mathcal{Z}, \hat{\theta}_{m_0}) + k^{\star}_{m_0} - \mathrm{E}\big[Q(\mathcal{Z}, \theta_{0m_0})\big]\big)\end{aligned}$$

is estimated by the empirical distribution of

$$\big\{Q(\mathcal{Z}^*_b, \hat{\theta}^*_{b,1}) + \hat{k}^{\star}_1 - Q(\mathcal{Z}, \hat{\theta}_1), \dots, Q(\mathcal{Z}^*_b, \hat{\theta}^*_{b,m_0}) + \hat{k}^{\star}_{m_0} - Q(\mathcal{Z}, \hat{\theta}_{m_0})\big\} \tag{3}$$

for $b = 1, \dots, B$, because $Q(\mathcal{Z}, \hat{\theta}_j)$ plays the role of $\mathrm{E}[Q(\mathcal{Z}, \theta_{0j})]$ under the resampling scheme. These bootstrap statistics are relatively easy to compute because the structure of the likelihood function is

$$Q(\mathcal{Z}_b^*, \hat{\theta}_{b,j}^*) - Q(\mathcal{Z}, \hat{\theta}_j) = n(\log \hat{\sigma}_{b,j}^{*2} + 1) - n(\log \hat{\sigma}_j^2 + 1) = n \log \frac{\hat{\sigma}_{b,j}^{*2}}{\hat{\sigma}_j^2},$$

where $\hat{\sigma}_{b,j}^{*2} = n^{-1} \sum_{t=1}^{n} (Y_{b,t}^* - \hat{\beta}_{b,j}^{*\prime} X_{b,j,t}^*)^2$. For each of the bootstrap resamples, we compute the test statistic

$$T_{b,R,\mathcal{M}}^* = \max_{i,j \in \mathcal{M}} \big| \{Q(\mathcal{Z}_b^*, \hat{\theta}_{b,i}^*) + \hat{k}_i^\star - Q(\mathcal{Z}, \hat{\theta}_i)\} - \{Q(\mathcal{Z}_b^*, \hat{\theta}_{b,j}^*) + \hat{k}_j^\star - Q(\mathcal{Z}, \hat{\theta}_j)\} \big|.$$

The p-value for the hypothesis test with which we are concerned is computed by

$$p_{\mathcal{M}} = B^{-1} \sum_{b=1}^{B} 1_{\{T_{b,R,\mathcal{M}}^* \geq T_{R,\mathcal{M}}\}}.$$

The empirical distribution of $n^{-1/2} T_{b,R,\mathcal{M}}^*$ yields a conservative estimate of the distribution of $n^{-1/2} T_{R,\mathcal{M}}$ as $n, B \to \infty$. The conservative nature of this estimate refers to the p-value, $p_{\mathcal{M}}$, being conservative in situations where the comparisons involve nested models. We discuss this issue at some length in the next subsection.

It is also straightforward to construct the MCS using either the AIC, the BIC, the AIC*, or the BIC*. The relevant test statistic has the form

$$T_{R,\mathcal{M}} = \max_{i,j \in \mathcal{M}} \big| [Q(\mathcal{Z}, \hat{\theta}_i) + c_i] - [Q(\mathcal{Z}, \hat{\theta}_j) + c_j] \big|,$$

where $c_j = 2k_j$ for the AIC, $c_j = \log(n) k_j$ for the BIC, $c_j = 2\hat{k}_j^\star$ for the AIC*, and $c_j = \log(n)\hat{k}_j^\star$ for the BIC*. The computation of the resampled test statistics, $T_{b,R,\mathcal{M}}^*$, is identical for the three criteria. The reason is that the location shift c_j has no effect on the bootstrap statistics once the null hypothesis is imposed. Under the null hypothesis, we recenter the bootstrap statistics about zero and this offsets the location shift $c_i - c_j$.

3.2.4. *Issues Related to the Comparison of Nested Models*

When two models are nested, the null hypothesis used with KLIC, $\mathrm{E}[Q(\mathcal{Z}, \theta_{0i})] = \mathrm{E}[Q(\mathcal{Z}, \theta_{0j})]$, has the strong implication that $Q(\mathcal{Z}, \theta_{0i}) = Q(\mathcal{Z}, \theta_{0j})$ a.e. (almost everywhere), and this causes the limit distribution of the quasi-likelihood ratio statistic, $Q(\mathcal{Z}, \hat{\theta}_i) - Q(\mathcal{Z}, \hat{\theta}_j)$, to differ for nested or nonnested comparisons (see Vuong (1989)). This property of nested comparisons can be imposed on the bootstrap resamples by replacing $Q(\mathcal{Z}, \hat{\theta}_j)$ with $Q(\mathcal{Z}^*, \hat{\theta}_j)$, because the latter is the bootstrap variant of $Q(\mathcal{Z}, \theta_{0j})$. The MCS procedure can be adapted so that different bootstrap schemes are used for nested and nonnested comparisons, and imposing the stronger null hypothesis $Q(\mathcal{Z}, \theta_{0i}) = Q(\mathcal{Z}, \theta_{0j})$ a.e. may improve the power of the procedure. The key difference is that the null hypothesis with KLIC has $Q(\mathcal{Z}, \hat{\theta}_i) - Q(\mathcal{Z}, \hat{\theta}_j) = O_p(1)$ for nested comparisons and $Q(\mathcal{Z}, \hat{\theta}_i) - Q(\mathcal{Z}, \hat{\theta}_j) = O_p(n^{1/2})$ for nonnested comparisons. Our bootstrap implementation is such that $\{Q(\mathcal{Z}_b^*, \hat{\theta}_{b,i}^*) + \hat{k}_i^\star - Q(\mathcal{Z}, \hat{\theta}_i)\} - \{Q(\mathcal{Z}_b^*, \hat{\theta}_{b,j}^*) + \hat{k}_j^\star - Q(\mathcal{Z}, \hat{\theta}_j)\}$ is $O_p(n^{1/2})$, whether the comparison involves nested or nonnested models, which causes the bootstrap critical values to be conservative. Under the alternative, $Q(\mathcal{Z}, \hat{\theta}_i) - Q(\mathcal{Z}, \hat{\theta}_j)$ diverges at rate n for nested and nonnested comparisons, so the bootstrap testing procedure is consistent in both cases.

Since nested and nonnested comparisons result in different rates of convergence and different limit distributions, there are better ways to construct an adaptive procedure than through the test statistic $T_{R,\mathcal{M}}$, for instance, by combining the p-values for the individual subhypotheses. We shall not pursue such an adaptive bootstrap implementation in this paper. It is, however, important to note that the issue with nested models is only relevant for KLIC because the underlying null hypotheses of other criteria, including AIC* and BIC*, do not imply $Q(\mathcal{Z}, \theta_{0i}) = Q(\mathcal{Z}, \theta_{0j})$ a.e. for nested models.

4. RELATION TO EXISTING MULTIPLE COMPARISONS METHODS

The Introduction discussed the relationship between the MCS and the trace test used to select the number of cointegration relations (see Johansen (1988)). The MCS and the trace test share an underlying testing principle known as *intersection–union testing* (IUT). Berger (1982) was responsible for formalizing the IUT, while Pantula (1989) applied the IUT to the problem of selecting the lag length and order of integration in univariate autoregressive processes.

Another way to cast the MCS problem is as a multiple comparisons problem. The multiple comparisons problem has a long history in the statistics literature; see Gupta and Panchapakesan (1979), Hsu (1996), Dudoit, Shaffer, and Boldrick (2003), and Lehmann and Romano (2005, Chap. 9) and references therein. Results from this literature have recently been adopted in the econometrics literature. One problem is that of *multiple comparisons with best*, where objects are compared to those with the best sample performance. Statistical procedures for multiple comparisons with best are discussed and applied to economic problems in Horrace and Schmidt (2000). Shimodaira (1998) used a variant of Gupta's subset selection (see Gupta and Panchapakesan (1979)) to construct a set of models that he terms a model confidence set. His procedure is specific to a ranking of models in terms of $\mathrm{E}(\mathrm{AIC}_j)$, and his framework is different from ours in a number of ways. For instance, his preferred set of models does not control the FWE. He also invoked a Gaussian approximation that rules out comparisons of nested models.

Our MCS employs a sequential testing procedure that mimics step-down procedures for multiple hypothesis testing; see, for example, Dudoit, Shaffer, and Boldrick (2003), Lehmann and Romano (2005, Chap. 9), or Romano, Shaikh, and Wolf (2008). Our definition of MCS p-values implies the monotonicity, $\hat{p}_{e_{\mathcal{M}_1}} \leq \hat{p}_{e_{\mathcal{M}_2}} \leq \cdots \leq \hat{p}_{e_{\mathcal{M}_{m_0}}}$ that is key for the result of Theorem 3. This monotonicity is also a feature of the so-called *step-down Holm adjusted p-values*.

4.1. *Relationship to Tests for Superior Predictive Ability*

Another related problem is the case where the benchmark, to which all objects are compared, is selected independently of the data used for the comparison. This problem is known as *multiple comparisons with control*. In the context of forecast comparisons, this is the problem that arises when testing for *superior predictive ability* (SPA); see White (2000b), Hansen (2005), and Romano and Wolf (2005).

The MCS has several advantages over tests for superior predictive ability. The *reality check for data snooping* of White (2000b) and the SPA test of Hansen (2005) are designed to address whether a particular benchmark is significantly outperformed by any of the alternatives used in the comparison. Unlike these tests, the MCS procedure does not require a benchmark to be specified, which is very useful in applications without an obvious benchmark. In the situation where there is a natural benchmark, the MCS procedure can still address the same objective as the SPA tests. This is done by observing whether the designated benchmark is in the MCS, where the latter corresponds to a rejection of the null hypothesis that is relevant for a SPA test.

The MCS procedure has the advantage that it can be employed for model selection, whereas a SPA test is ill-suited for this problem. A rejection of the SPA test only identifies one or more models as significantly better than the benchmark.[7] Thus, the SPA test offers little guidance about which models reside in $\mathcal{M}^*$. We are also faced with a similar problem in the event that the null hypothesis is not rejected by the SPA test. In this case, the benchmark may be the best model, but this label may also be applied to other models. This issue can be resolved if all models serve as the benchmark in a series of comparisons. The result is a sequence of SPA tests that define the MCS to be the set of "benchmark" models that are found not to be significantly inferior to the alternatives. However, the level of individual SPA tests needs to be adjusted

[7] Romano and Wolf (2005) improved on the reality check by identifying the entire set of alternatives that significantly dominate the benchmark. This set of models is specific to the choice of benchmark and has, therefore, no direct relation to the MCS.

for the number of tests that are computed to control the FWE. For example, if the level in each of the SPA tests is α/m, the Bonferroni bound states that the resulting set of surviving benchmarks is a MCS with coverage $(1-\alpha)$. Nonetheless, there is a substantial loss of power associated with the small level applied to the individual tests. The loss of power highlights a major pitfall of sequential SPA tests.

Another drawback of constructing a MCS from SPA-tests is that the null of a SPA test is a composite hypothesis. The null is defined by several inequality constraints which affect the asymptotic distribution of the SPA test statistic because it depends on the number of binding inequalities. The binding inequality constraints create a nuisance parameter problem. This makes it difficult to control the Type I error rate, inducing an additional loss of power; see Hansen (2003a). In comparison, the MCS procedure is based on a sequence of hypothesis tests that only involve equalities, which avoids composite hypothesis testing.

4.2. *Related Sequential Testing Procedures for Model Selection*

This subsection considers some relevant aspects of out-of-sample evaluation of forecasting models and how the MCS procedure relates to these issues.

Several papers have studied the problem of selecting the best forecasting model from a set of competing models. For example, Engle and Brown (1985) compared selection procedures that are based on six information criteria and two testing procedures (general-to-specific and specific-to-general), Sin and White (1996) analyzed information criteria for possibly misspecified models, and Inoue and Kilian (2006) compared selection procedures that are based on information criteria and out-of-sample evaluation. Granger, King, and White (1995) argued that the general-to-specific selection procedure is based on an incorrect use of hypothesis testing, because the model chosen to be the null hypothesis in a pairwise comparison is unfairly favored. This is problematic when the data set under investigation does not contain much information, which makes it difficult to distinguish between models. The MCS procedure does not assume that a particular model is the true model; neither is the null hypothesis defined by a single model. Instead, all models are treated equally in the comparison and only evaluated on out-of-sample predictive ability.

4.3. *Aspects of Parameter Uncertainty and Forecasting*

Parameter estimation can play an important role in the evaluation and comparison of forecasting models. Specifically, when the comparison of nested models relies on parameters that are estimated using certain estimation schemes, the limit distribution of our test statistics need not be Gaussian; see West and McCracken (1998) and Clark and McCracken (2001). In the present context, there will be cases that do not fulfil Assumption 2. Some of these problems can be avoided by using a rolling window for parameter estimation, known as the *rolling scheme*. This is the approach taken by Giacomini and White (2006). Alternatively one can estimate the parameters once (using data that are dated prior to the evaluation period) and then compare the forecasts *conditional on these parameter estimates*. However, the MCS should be applied with caution when forecasts are based on estimated parameters because our assumptions need not hold in this case. As a result, modifications are needed in the case with nested models; see Chong and Hendry (1986), Harvey and Newbold (2000), Chao, Corradi, and Swanson (2001), and Clark and McCracken (2001) among others. The key modification that is needed to accommodate the case with nested models is to adopt a test with a proper size. With proper choices for $\delta_{\mathcal{M}}$ and $e_{\mathcal{M}}$, the general theory for the MCS procedure remains. However, in this paper we will not pursue this extension because it would obscure our main objective, which is to lay out the key ideas of the MCS.

4.4. *Bayesian Interpretation*

The MCS procedure is based on frequentist principles, but resembles some aspects of Bayesian model selection techniques. By specifying a prior over the models in $\mathcal{M}^0$, a Bayesian procedure would produce a posterior distribution for each model, conditional on the actual data. This approach to MCS construction includes those models with the largest posteriors that sum at least to $1-\alpha$. If the Bayesian were also to choose models by minimizing the "risk" associated with the loss attributed to each model, the MCS would be a Bayes decision procedure with respect to the model posteriors. Note that the Bayesian and frequentist MCSs rely on the metric under which loss is calculated and depend on sample information.

We argue that our approach to the MCS and its bootstrap implementation compares favorably to Bayesian methods of model selection. One advantage of the frequentist approach is that it avoids having to place priors on the elements of $\mathcal{M}^0$ (and their parameters). Our probability statement is associated with the random data-dependent set of models that is the MCS. It therefore is meaningful to state that the best model can be found in the MCS with a certain probability. The MCS also places moderate computational demands on the researcher, unlike the synthetic data creation methods on which Bayesian Markov chain Monte Carlo methods rely.

5. SIMULATION RESULTS

This section reports on Monte Carlo experiments that show the MCS to be properly sized and possess good power in various simulation designs.

5.1. *Simulation Experiment I*

We consider two designs that are based on the m-dimensional vector $\theta = (0, \frac{1}{m-1}, \dots, \frac{m-2}{m-1}, 1)'\lambda/\sqrt{n}$ that defines the relative performances $\mu_{ij} = \mathrm{E}(d_{ij,t}) = \theta_i - \theta_j$. The experimental design ensures that $\mathcal{M}^*$ consists of a single element, unless $\lambda = 0$, in which case we have $\mathcal{M}^* = \mathcal{M}^0$. The stochastic nature of the simulation is primarily driven by

$$X_t \sim \text{i.i.d.}\, N_m(0, \Sigma), \quad \text{where}$$
$$\Sigma_{ij} = \begin{cases} 1 & \text{for } i = j, \\ \rho & \text{for } i \neq j, \text{for some } 0 \leq \rho \leq 1, \end{cases}$$

where ρ controls the degree of correlation between alternatives.

DESIGN I.A—Symmetric Distributed Loss: Define the (vector of) loss variables to be

$$L_t \equiv \theta + \frac{a_t}{\sqrt{\mathrm{E}(a_t^2)}} X_t, \quad \text{where}$$
$$a_t = \exp(y_t), \quad y_t = \frac{-\varphi}{2(1+\varphi)} + \varphi y_{t-1} + \sqrt{\varphi}\varepsilon_t,$$

and $\varepsilon_t \sim \text{i.i.d.}\, N(0, 1)$. This implies that $\mathrm{E}(y_t) = -\varphi/\{2(1 - \varphi^2)\}$ and $\mathrm{var}(y_t) = \varphi/(1-\varphi^2)$ such that $\mathrm{E}(a_t) = \exp\{\mathrm{E}(y_t)+\mathrm{var}(y_t)/2\} = \exp\{0\} = 1$ and $\mathrm{var}(a_t) = (\exp\{\varphi/(1 - \varphi^2)\} - 1)$. Furthermore, $\mathrm{E}(a_t^2) = \mathrm{var}(a_t) + 1 = \exp\{\varphi/(1 - \varphi^2)\}$ such that $\mathrm{var}(L_t) = 1$. Note that $\varphi = 0$ corresponds to homoskedastic errors and $\varphi > 0$ corresponds to (generalized autoregressive conditional heteroskedasticity) (GARCH type) heteroskedastic errors.

The simulations employ 2,500 repetitions, where $\lambda = 0, 5, 10, 20$, $\rho = 0.00, 0.50, 0.75, 0.95$, $\varphi = 0.0, 0.5, 0.8$, and $m = 10, 40, 100$. We use the block bootstrap, in which blocks have length $l = 2$, and results are based on $B = 1{,}000$ resamples. The size of a synthetic sample is $n = 250$. This approximates sample sizes often available for model selection exercises in macroeconomics.

We report two statistics from our simulation experiment based on $\alpha = 10\%$: one is the frequency at which $\widehat{\mathcal{M}}^*_{90\%}$ contains $\mathcal{M}^*$; the other is the average number of models in $\widehat{\mathcal{M}}^*_{90\%}$. The former shows the size properties of the MCS procedure; the latter is informative about the power of the procedure.

Table II presents simulation results that show that the small sample properties of the MCS procedure closely match its theoretical predictions. The frequency that the best models are contained in the MCS is almost always greater than $(1 - \alpha)$, and the MCS becomes better at separating the inferior models from the superior model, as the μ_{ij}s become more disperse (e.g., as λ increases). Note also that a larger correlation makes it easier to separate inferior models from superior model. This is not surprising because

TABLE II

SIMULATION DESIGN I.A[a]

	$m = 10$				$m = 40$				$m = 100$			
λ $\rho =$	0	0.5	0.75	0.95	0	0.5	0.75	0.95	0	0.5	0.75	0.95
Panel A: $\varphi = 0$												
Frequency at which $\mathcal{M}^* \subset \widehat{\mathcal{M}}^*_{90\%}$ (size)												
0	0.885	0.898	0.884	0.885	0.882	0.882	0.877	0.880	0.880	0.870	0.877	0.875
5	0.990	0.988	0.991	1.000	0.980	0.979	0.976	0.984	0.975	0.976	0.975	0.976
10	0.994	0.998	0.999	1.000	0.978	0.983	0.985	0.993	0.973	0.975	0.974	0.980
20	0.998	1.000	1.000	1.000	0.988	0.981	0.991	1.000	0.975	0.978	0.986	0.992
40	1.000	1.000	1.000	1.000	0.992	0.996	0.998	1.000	0.981	0.984	0.990	0.998
Average number of elements in $\widehat{\mathcal{M}}^*_{90\%}$ (power)												
0	9.614	9.658	9.646	9.632	38.68	38.78	38.91	38.82	97.02	96.84	97.11	97.20
5	6.498	4.693	3.239	1.544	25.30	18.79	13.35	6.382	59.87	43.92	32.51	15.04
10	3.346	2.390	1.732	1.027	13.59	9.829	7.142	3.266	32.32	23.04	16.97	7.902
20	1.702	1.307	1.062	1.000	7.060	5.010	3.617	1.674	17.03	12.40	8.785	4.049
40	1.072	1.005	1.000	1.000	3.572	2.597	1.840	1.052	8.778	6.375	4.521	2.083
Panel B: $\varphi = 0.5$												
Frequency at which $\mathcal{M}^* \subset \widehat{\mathcal{M}}^*_{90\%}$ (size)												
0	0.908	0.897	0.905	0.894	0.911	0.907	0.910	0.916	0.925	0.918	0.909	0.913
5	0.985	0.990	0.995	1.000	0.971	0.976	0.977	0.987	0.974	0.974	0.973	0.973
10	0.992	0.999	1.000	1.000	0.978	0.985	0.982	0.995	0.975	0.969	0.983	0.984
20	0.999	1.000	1.000	1.000	0.988	0.989	0.988	1.000	0.979	0.976	0.981	0.992
40	1.000	1.000	1.000	1.000	0.996	0.996	1.000	1.000	0.980	0.982	0.991	0.999
Average number of elements in $\widehat{\mathcal{M}}^*_{90\%}$ (power)												
0	9.660	9.664	9.664	9.649	38.97	38.93	39.03	39.05	98.35	98.05	97.94	97.73
5	6.076	4.497	3.213	1.564	24.33	17.72	13.13	6.112	57.84	41.60	30.35	14.54
10	3.188	2.278	1.680	1.035	12.95	9.268	6.791	3.136	30.54	22.30	16.56	7.510
20	1.700	1.274	1.069	1.000	6.819	4.883	3.563	1.659	16.04	11.56	8.430	3.894
40	1.085	1.008	1.000	1.000	3.506	2.517	1.811	1.061	8.339	6.166	4.360	2.034
Panel C: $\varphi = 0.8$												
Frequency at which $\mathcal{M}^* \subset \widehat{\mathcal{M}}^*_{90\%}$ (size)												
0	0.931	0.940	0.939	0.947	0.963	0.968	0.958	0.962	0.970	0.975	0.969	0.972
5	0.990	0.997	0.998	1.000	0.977	0.980	0.989	0.993	0.970	0.975	0.976	0.981
10	0.998	1.000	1.000	1.000	0.984	0.987	0.992	0.998	0.982	0.976	0.974	0.991
20	1.000	1.000	1.000	1.000	0.990	0.993	0.996	1.000	0.982	0.982	0.992	0.998
40	1.000	1.000	1.000	1.000	0.999	1.000	1.000	1.000	0.988	0.994	0.996	1.000
Average number of elements in $\widehat{\mathcal{M}}^*_{90\%}$ (power)												
0	9.739	9.814	9.794	9.799	39.61	39.61	39.53	39.55	99.00	99.44	99.15	99.43
5	4.301	3.318	2.386	1.322	16.26	12.31	9.118	4.401	39.69	28.13	20.56	10.12
10	2.424	1.864	1.419	1.062	9.133	6.643	4.727	2.349	20.72	14.77	11.26	5.470
20	1.455	1.220	1.092	1.010	4.770	3.520	2.535	1.454	11.15	8.014	5.948	2.840
40	1.098	1.037	1.011	1.003	2.645	1.967	1.490	1.081	5.932	4.356	3.248	1.645

[a] The two statistics are the frequency at which $\widehat{\mathcal{M}}^*_{90\%}$ contains $\mathcal{M}^*$ and the other is the average number of models in $\widehat{\mathcal{M}}^*_{90\%}$. The former shows the ‘size’ properties of the MCS procedure and the latter is informative about the ‘power’ of the procedure.

$\text{var}(d_{ij,t}) = \text{var}(L_{it}) + \text{var}(L_{jt}) - 2\,\text{cov}(L_{it}, L_{jt}) = 2(1-\rho)$, which is decreasing in ρ. Thus, a larger correlation (holding the individual variances fixed) is associated with more information that allows the MCS to separate good from bad models. Finally, the effects of heteroskedasticity are relatively small, but heteroskedasticity does appear to add power to the MCS procedure. The average number of models in $\widehat{\mathcal{M}}^*_{90\%}$ tends to fall as φ increases.

Corollary 1 has a consistency result that applies when $\lambda > 0$. The implication is that only one model enters $\mathcal{M}^*$ under this restriction. Table II shows that $\mathcal{M}^*$ often contains only one model given $\lambda > 0$. The MCS matches this theoretical prediction in Table II because $\widehat{\mathcal{M}}^*_{90\%} = \mathcal{M}^*$ in a large number of simulations. This equality holds especially when λ and ρ are large. These are also the simulation experiments that yield size and power statistics equal (or nearly equal) to 1. With size close to 1 or equal to 1, observe that $\mathcal{M}^* \subset \widehat{\mathcal{M}}^*_{90\%}$ (in all the synthetic samples). On the other hand, $\widehat{\mathcal{M}}^*_{90\%}$ is reduced to a single model (in all the synthetic samples) when power is close to 1 or equal to 1.

DESIGN I.B—Dependent Loss: This design sets $L_t \sim \text{i.i.d.}\, N_{10}(\theta, \Sigma)$, where the covariance matrix has the structure $\Sigma_{ij} = \rho^{|i-j|}$ for $\rho = 0, 0.5$, and 0.75. The mean vector takes the form $\theta = (0, \dots, 0, \frac{1}{5}, \dots, \frac{1}{5})'$ so that the number of zero elements in θ defines the number of elements in $\mathcal{M}^*$. We report simulation results for the case where $m_0 = 10$ and $\mathcal{M}^*$ consists of either one, two, or five models.

The simulation results are presented in Figure 1. The left panels display the frequency at which $\widehat{\mathcal{M}}^*_{90\%}$ contains $\mathcal{M}^*$ (size) at various sample sizes. The right panels present the average number of models in $\widehat{\mathcal{M}}^*_{90\%}$ (power). The two upper panels contain the results for the case where $\mathcal{M}^*$ is a single model. The upper-left panel indicates that the best model is almost always contained in the MCS. This agrees with Corollary 1, which states that $\widehat{\mathcal{M}}^*_{1-\alpha} \xrightarrow{p} \mathcal{M}^*$ as $n \to \infty$, whenever $\mathcal{M}^*$ consists of a single model. The upper-right panel illustrates the power of the procedure based on $T_{\max,\mathcal{M}} = \max_{i\in\mathcal{M}} t_{i\cdot}$. We note that it takes about 800 observations to weed out the 9 inferior models in this design. The MCS procedure is barely affected by the correlation parameter ρ, but we note that a larger ρ results in a small loss in power. In the lower-left panel, we see that the frequency at which $\mathcal{M}^*$ is contained in $\widehat{\mathcal{M}}^*_{90\%}$ is reasonably close to 90% except for the very short sample sizes. From the middle-right and lower-right panels, we see that it takes about 500 observations to remove all the poor models.

The middle-right and lower-right panels illustrate another aspect of the MCS procedure. For large sample sizes, we note that the average number of models in $\widehat{\mathcal{M}}^*_{90\%}$ falls below the number of models in $\mathcal{M}^*$. The explanation is simple. After all poor models have been eliminated, as occurs with probability approaching 1 as $n \to \infty$, there is a positive probability that $H_{0,\mathcal{M}^*}$ is rejected,

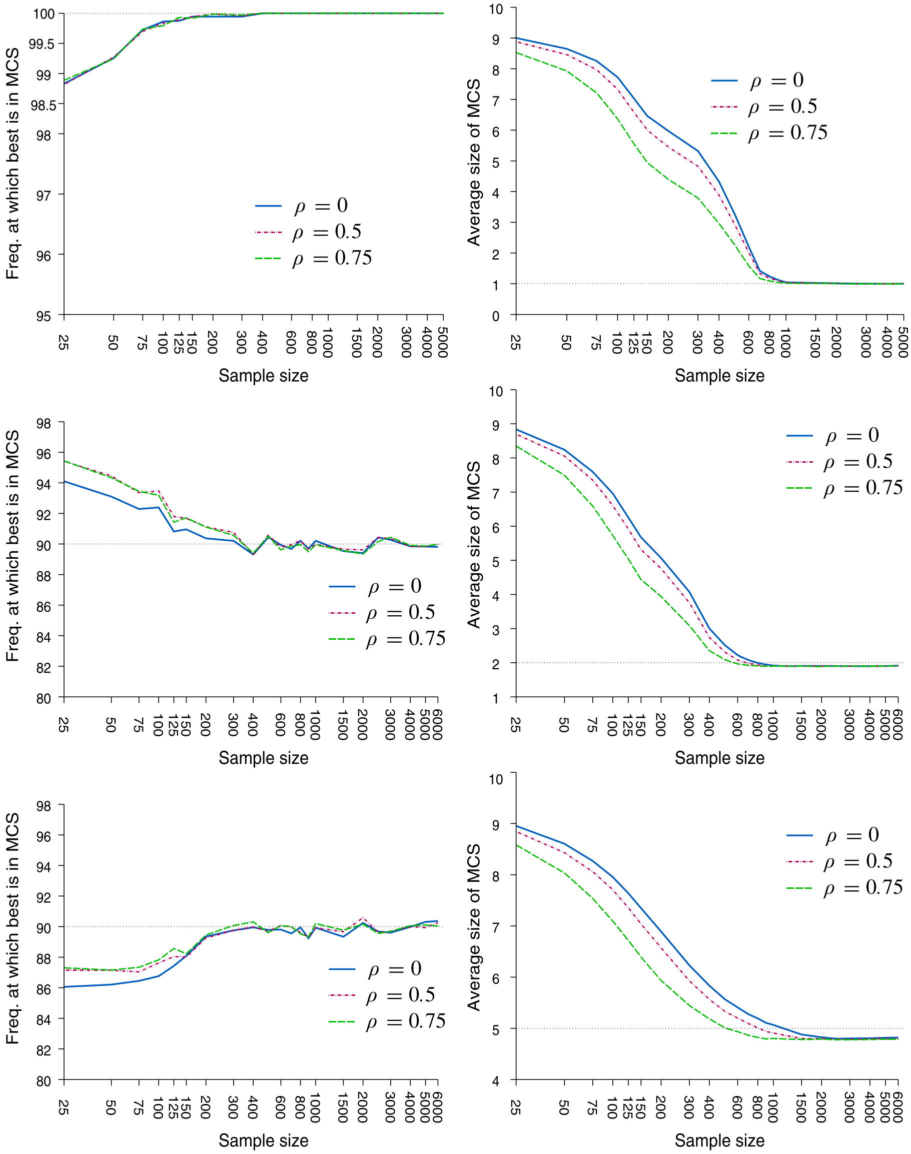

FIGURE 1.—Simulation Design I.B with 10 alternatives and 1, 2, or 5 elements in $\mathcal{M}^*$. The left panels report the frequency at which $\mathcal{M}^*$ is contained in $\widehat{\mathcal{M}}^*_{90\%}$ (size properties) and the right panels report the average number of models in $\widehat{\mathcal{M}}^*_{90\%}$ (power properties).

which causes the MCS procedure to eliminate a good model. Thus, the inferences we draw from the simulation results are quite encouraging for the $T_{\max,\mathcal{M}}$ test.

5.2. *Simulation Experiment II: Regression Models*

Next we study the properties of the MCS procedure in the context of in-sample evaluation of regression models as we laid out in Section 3.2. We consider a setup with six potential regressors, $X_t = (X_{1,t}, \dots, X_{6,t})'$, that are distributed as

$$X_t \sim \text{i.i.d.}\, N_6(0, \Sigma), \quad \text{where}$$
$$\Sigma_{ij} = \begin{cases} 1 & \text{for } i = j, \\ \rho & \text{for } i \neq j, \text{ for some } 0 \leq \rho < 1, \end{cases}$$

where ρ measures the degree of dependence between the regressors. We define the dependent variable by $Y_t = \mu + \beta X_{1,t} + \sqrt{1-\beta^2}\varepsilon_t$, where $\varepsilon_t \sim \text{i.i.d.}\, N(0,1)$. In addition to the six variables in X_t, we include a constant, $X_{0,t} = 1$, in all regression models. The set of regressions being estimated is given by the 12 regression models that are listed in each of the panels in Table III.

We report simulation results based on 10,000 repetitions, using a design with an $R^2 = 50\%$ (i.e., $\beta^2 = 0.5$) and either $\rho = 0.3$ or $\rho = 0.9$.[8] For the number of bootstrap resamples, we use $B = 1{,}000$. Since $X_{0,t} = 1$ is included in all regression models, the relevant MCS statistics are invariant to the actual value for μ, so we set $\mu = 0$ in our simulations.

The definition of $\mathcal{M}^*$ will depend on the criterion. With KLIC, the set of best models is given by the set of regression models that includes X_1. The reason is that KLIC does not favor parsimonious models, unlike the AIC* and BIC*. With these two criteria, $\mathcal{M}^*$ is defined to be the most parsimonious regression model that includes X_1. The models in $\mathcal{M}^*$ are identified by the shaded regions in Table III.

Our simulation results are reported in Table III. The average value of $Q(\mathcal{Z}_j, \hat{\theta}_j)$ is given in the first pair of data columns, followed by the average estimate of the effective degrees of freedom, $\hat{k}^\star$. The Gaussian setup is such that all models are correctly specified, so the effective degrees of freedom is simply the number of free parameters, which is the number of regressors plus 1 for σ_j^2. Table III shows that the average value of $\hat{k}_j^\star$ is very close to the number of free parameters in the jth regression model. The last three pairs of columns report the frequency that each of the models are in $\widehat{\mathcal{M}}^*_{90\%}$. We want large numbers inside the shaded region and small numbers outside the shaded region. The results are intuitive. As the sample size increases from 50 to 100 and then to 500, the MCS procedure becomes better at eliminating the models that do not reside in $\mathcal{M}^*$. With a sample size of $n = 500$, the consistent criterion, BIC*,

[8] Simulation results for $\beta^2 = 0.1$ and 0.9 are available in a separate appendix; see Hansen, Lunde, and Nason (2011).

TABLE III

SIMULATION EXPERIMENT II[a]

	$Q(\mathcal{Z}_j, \hat{\theta}_j)$		$\hat{k}^\star$		KLIC		AIC* (TIC)		BIC*	
$\rho =$	0.3	0.9	0.3	0.9	0.3	0.9	0.3	0.9	0.3	0.9
Panel A: $n = 50$										
X_0	48.1	48.1	1.99	2.00	0.058	0.038	0.085	0.070	0.118	0.124
X_0, X_1	12.4	12.4	3.02	3.02	0.998	0.999	1.000	1.000	1.000	1.000
$X_0, \dots, X_2$	11.3	11.3	4.08	4.08	0.998	0.999	0.962	0.999	0.566	0.940
$X_0, \dots, X_3$	10.2	10.2	5.18	5.18	0.999	0.999	0.940	0.998	0.469	0.912
$X_0, \dots, X_4$	9.09	9.04	6.32	6.32	1.000	1.000	0.905	0.997	0.367	0.803
$X_0, \dots, X_5$	7.95	7.88	7.50	7.50	1.000	1.000	0.867	0.994	0.279	0.598
$X_0, \dots, X_6$	6.77	6.69	8.73	8.74	1.000	1.000	0.806	0.990	0.203	0.400
X_0, X_2	44.7	21.0	3.02	3.02	0.086	0.905	0.100	0.935	0.099	0.877
X_0, X_2, X_3	42.3	18.1	4.08	4.08	0.106	0.948	0.107	0.949	0.077	0.806
$X_0, X_2, \dots, X_4$	40.4	16.3	5.18	5.18	0.120	0.958	0.105	0.938	0.054	0.665
$X_0, X_2, \dots, X_5$	38.8	14.8	6.32	6.32	0.132	0.962	0.100	0.913	0.036	0.501
$X_0, X_2, \dots, X_6$	37.2	13.4	7.50	7.51	0.145	0.964	0.094	0.869	0.022	0.348
Panel B: $n = 100$										
X_0	98.0	98.1	1.99	1.99	0.000	0.000	0.000	0.000	0.000	0.000
X_0, X_1	27.6	27.8	3.00	3.00	0.998	1.000	1.000	1.000	1.000	1.000
$X_0, \dots, X_2$	26.6	26.7	4.03	4.03	0.999	1.000	0.959	0.982	0.402	0.675
$X_0, \dots, X_3$	25.5	25.7	5.07	5.06	0.999	1.000	0.939	0.975	0.276	0.619
$X_0, \dots, X_4$	24.4	24.6	6.12	6.12	1.000	1.000	0.908	0.960	0.174	0.545
$X_0, \dots, X_5$	23.4	23.6	7.19	7.18	1.000	1.000	0.864	0.942	0.101	0.390
$X_0, \dots, X_6$	22.3	22.5	8.28	8.27	1.000	1.000	0.800	0.920	0.059	0.238
X_0, X_2	92.4	45.1	3.00	3.01	0.000	0.548	0.000	0.585	0.000	0.490
X_0, X_2, X_3	88.8	40.4	4.03	4.03	0.000	0.691	0.000	0.666	0.000	0.443
$X_0, X_2, \dots, X_4$	86.1	38.1	5.07	5.07	0.000	0.736	0.000	0.675	0.000	0.338
$X_0, X_2, \dots, X_5$	83.9	36.3	6.12	6.12	0.000	0.759	0.000	0.655	0.000	0.236
$X_0, X_2, \dots, X_6$	82.0	34.8	7.19	7.19	0.001	0.772	0.000	0.631	0.000	0.143
Panel C: $n = 500$										
X_0	498	498	2.00	2.00	0.000	0.000	0.000	0.000	0.000	0.000
X_0, X_1	151	151	3.00	3.00	0.999	0.999	1.000	1.000	1.000	1.000
$X_0, \dots, X_2$	150	150	4.00	4.00	0.999	0.999	0.958	0.960	0.207	0.206
$X_0, \dots, X_3$	149	149	5.01	5.01	0.999	1.000	0.938	0.938	0.100	0.099
$X_0, \dots, X_4$	148	148	6.02	6.01	1.000	1.000	0.907	0.901	0.044	0.042
$X_0, \dots, X_5$	147	147	7.03	7.02	1.000	1.000	0.858	0.852	0.020	0.017
$X_0, \dots, X_6$	145	146	8.04	8.03	1.000	1.000	0.790	0.792	0.006	0.008
X_0, X_2	474	238	3.00	3.00	0.000	0.000	0.000	0.000	0.000	0.000
X_0, X_2, X_3	460	219	4.00	4.00	0.000	0.002	0.000	0.002	0.000	0.002
$X_0, X_2, \dots, X_4$	451	211	5.01	5.01	0.000	0.004	0.000	0.004	0.000	0.001
$X_0, X_2, \dots, X_5$	444	206	6.02	6.01	0.000	0.006	0.000	0.006	0.000	0.001
$X_0, X_2, \dots, X_6$	439	203	7.03	7.02	0.000	0.008	0.000	0.007	0.000	0.000

[a] The average value of the maximized log-likelihood function multiplied by −2 is reported in the first two data columns. The next pair of columns has the average of the effective degrees of freedom. The last three pairs of columns report the frequency that a particular regression model is in the $\widehat{\mathcal{M}}^*_{90\%}$ for each of the three criteria: KLIC, AIC*, and BIC*.

has reduced the MCS to the single best model in the majority of simulations. This is not true for the AIC* criterion. Although it tends to settle on more parsimonious models than the KLIC, the AIC* has a penalty that makes it possible for an overparameterized model to have the best AIC*. The bootstrap testing procedure is conservative when the comparisons involve nested models under KLIC; see our discussion in the last paragraph of Section 3.2. This explains that both Type I and Type II errors are close to zero when $n = 500$, an ideal outcome that is not guaranteed when $\mathcal{M}^*_{\text{KLIC}}$ includes nonnested models.[9]

6. EMPIRICAL APPLICATIONS

6.1. *U.S. Inflation Forecasts: Stock and Watson (1999) Revisited*

This section revisits the Stock and Watson (1999) study of the best out-of-sample predictors of inflation. Their empirical application consists of pairwise comparisons of a large number of inflation forecasting models. The set of inflation forecasting models includes several that have a Phillips curve interpretation, along with autoregressive and a no-change (month-over-month) forecast. We extend their set of forecasts by adding a second no-change (12-months-over-12-months) forecast that was used in Atkeson and Ohanian (2001).

Stock and Watson (1999) measured inflation, π_t, as either the CPI-U, all items (PUNEW), or the headline personal consumption expenditure implicit price deflator (GMDC).[10] The relevant Phillips curve is

$$\pi_{t+h} - \pi_t = \phi + \beta(\mathbf{L})u_t + \gamma(\mathbf{L})(1 - \mathbf{L})\pi_t + e_{t+h}, \tag{4}$$

where u_t is the unemployment rate, $\mathbf{L}$ is the lag polynomial operator, and e_{t+h} is the long-horizon inflation forecast innovation. Note that the natural rate hypothesis is not imposed on the Phillips curve (4) and that inflation as a regressor is in its first difference. Stock and Watson also forecasted inflation with (4) where the unemployment rate u_t is replaced with different macrovariables.

The entire sample runs from 1959:M1 to 1997:M9. Following Stock and Watson, we study the properties of their forecasting models on the pre- and post-1984 subsamples of 1970:M1–1983:M12 and 1984:M1–1996:M9.[11] The former subsample contains the great inflation of the 1970s and the rapid disinflation of the early 1980s. Inflation does not exhibit this volatile behavior in the post-1984 subsample. We follow Stock and Watson so as to replicate their inflation

[9] In an unreported simulation study where $\mathcal{M}^*_{\text{KLIC}}$ was designed to include nonnested models, we found the frequency by which $\mathcal{M}^*_{\text{KLIC}} \subset \widehat{\mathcal{M}}^*_{90\%}$ converges to 90%.

[10] The data for this applications was downloaded from Mark Watson's web page. We refer the interested reader to Stock and Watson (1999) for details about the data and model specifications.

[11] Stock and Watson split their sample at the end of 1983 to account for structural change in inflation dynamics. This structural break is ignored when estimating the Phillips curve model (4) and the alternative inflation forecasting equations. This is justified by Stock and Watson because the impact of the 1984 structural break on their estimated Phillips curve coefficients is small.

forecasts. However, our MCS bootstrap implementation, which is described in Section 3, relies on an assumption that $d_{ij,t}$ is stationary. This is not plausible when the parameters are estimated with a recursive estimation scheme, as was used by Stock and Watson (1999). We avoid this problem by following Giacomini and White (2006) and present empirical results that are based on parameters estimated over a rolling window with a fixed number of observations.[12] Regressions are estimated on data that begin no earlier than 1960:M2, although lagged regressors impinge on observations back to 1959:M1.

We compute the MCS across all of the Stock and Watson inflation forecasting models. This includes the Phillips curve model (4), the inflation forecasting equation that runs through all of the macrovariables considered by Stock and Watson, a univariate autoregressive model, and two no-change forecasts. The first no-change forecast is the past month's inflation rate; the second no-change forecast uses the past year's inflation rate as its forecast. The former matches the no-change forecast in Stock and Watson (1999) and the latter matches the no-change forecast in Atkeson and Ohanian (2001). Stock and Watson also presented results for forecast combinations and forecasts based on principal component indicator variables.[13]

Tables IV and V report (the level of) the root mean square error (RMSE) and MCS p-values for each of the inflation forecasting models. The second column of Table IV also lists the transformation of the macrovariable employed by the forecasting equation.

Our Table IV matches the results reported in Stock and Watson (1999, Table 2). The initial model space $\mathcal{M}^0$ is filled with a total of 19 models. The results for the two no-change forecasts and the AR(p) are the first three rows of Table IV. The RMSEs and the p-values for the Phillips curve forecasting model (4) appear in the bottom row of our Table IV. The rest of the rows of Table IV are the "gap" and "first difference" specifications of Stock and Watson's aggregate activity variables that appear in place of u_t in inflation forecasting equation (4). The gap variables are computed with a one-sided Hodrick and Prescott (1997) filter; see Stock and Watson (1999, p. 301) for details.[14]

A glance at Table IV reveals that the MCS of subsamples 1970:M1–1983:M12 and 1984:M1–1996:M9 are strikingly different for both inflation series, PUNEW and GMDC. The MCS of the pre-1984 subsample places seven

[12] The corresponding empirical results that are based on parameters that are estimated with the recursive scheme, as was used in Stock and Watson (1999), are available in a separate appendix; see Hansen, Lunde, and Nason (2011). Although our assumption does not justify the recursive estimation scheme, it produces pseudo-MCS results that are very similar to those obtained under the rolling window estimation scheme.

[13] See Stock and Watson (1999) for details about their modelling strategy, forecasting procedures, and data set.

[14] The MCS p-values are computed using a block size of $l = 12$ in the bootstrap implementation. The MCS p-values are qualitatively similar when computed with $l = 6$ and $l = 9$. These are reported in a separate appendix; see Hansen, Lunde, and Nason (2011).

TABLE IV

MCS FOR SIMPLE REGRESSION-BASED INFLATION FORECASTS[a]

		PUNEW				GMDC			
		1970–1983		1984–1996		1970–1983		1984–1996	
Variable	Trans	RMSE	p_{MCS}	RMSE	p_{MCS}	RMSE	p_{MCS}	RMSE	p_{MCS}
No change (month)		3.290	0.001	2.140	0.122 *	2.208	0.042	1.751	0.113*
No change (year)	–	2.798	0.006	1.207	1.00**	2.100	0.109*	0.888	1.00**
uniar	–	2.802	0.004	1.330	0.736**	2.026	0.145*	1.070	0.411**
Gap specifications									
dtip	DT	2.597	0.059	1.475	0.651**	2.103	0.095	1.050	0.411**
dtgmpyq	DT	2.751	0.020	1.691	0.299**	2.090	0.157*	1.125	0.317**
dtmsmtq	DT	2.202	0.872**	1.704	0.477**	1.806	0.464**	1.046	0.411**
dtlpnag	DT	2.591	0.068	1.433	0.694**	2.132	0.075	1.026	0.411**
ipxmca	LV	2.609	0.034	1.318	0.736**	2.040	0.261**	1.034	0.411**
hsbp	LN	2.114	1.00**	1.582	0.579**	1.967	0.364**	1.034	0.411**
lhmu25	LV	2.968	0.006	1.439	0.651**	2.231	0.061	1.040	0.411**
First difference specifications									
ip	DLN	2.344	0.306**	1.393	0.736**	1.946	0.298**	1.058	0.411**
gmpyq	DLN	2.306	0.842**	1.524	0.421**	1.709	1.00**	1.158	0.317**
msmtq	DLN	2.158	0.872**	1.391	0.736**	1.857	0.464**	1.066	0.411**
lpnag	DLN	2.408	0.430**	1.341	0.736**	1.940	0.298**	1.027	0.411**
dipxmca	DLV	2.379	0.139*	1.353	0.736**	1.903	0.446**	1.041	0.411**
dhsbp	DLN	2.850	0.003	1.456	0.665**	2.076	0.075	1.070	0.411**
dlhmu25	DLV	2.383	0.169*	1.440	0.579**	2.035	0.102*	1.065	0.411**
dlhur	DLV	2.296	0.631**	1.429	0.691**	1.904	0.330**	1.067	0.411**
Phillips curve									
lhur		2.637	0.034	1.388	0.736**	2.076	0.098	1.162	0.325**

[a] RMSEs and MCS p-values for the different forecasts. The forecasts in $\widehat{\mathcal{M}}^*_{90\%}$ and $\widehat{\mathcal{M}}^*_{75\%}$ are identified by one and two asterisks, respectively.

forecasting models in PUNEW-$\widehat{\mathcal{M}}^*_{75\%}$ and nine models in GMDC-$\widehat{\mathcal{M}}^*_{75\%}$. For the post-1984 subsample, all but one model ends up in $\widehat{\mathcal{M}}^*_{75\%}$ for both PUNEW and GMDC. The only model that is consistently kicked out of these MCSs is the monthly no-change forecast, which uses last month's inflation rate as its forecast.

Another intriguing feature of Table IV is the inflation forecasting models that reside in the MCS when faced with the 1970:M1–1983:M12 subsample. The seven models that are in PUNEW-$\widehat{\mathcal{M}}^*_{75\%}$ are driven by macrovariables related either to real economic activity (e.g., manufacturing and trade, and building permits) or to the labor market. The labor market variables are lpnag (employees on nonagricultural payrolls) and dlhur (first difference of the unemployment rate, all workers 16 years and older). Thus, there is labor mar-

TABLE V

MCS RESULTS FOR SHRINKAGE-TYPE INFLATION FORECASTS[a]

	PUNEW				GMDC			
	1970–1983		1984–1996		1970–1983		1984–1996	
Variable	RMSE	p_{MCS}	RMSE	p_{MCS}	RMSE	p_{MCS}	RMSE	p_{MCS}
No change (month)	3.290	0.006	2.140	0.000	2.208	0.006	1.751	0.000
No change (year)	2.798	0.020	1.207	1.00**	2.100	0.120*	0.888	1.00**
Univariate	2.802	0.012	1.330	0.718**	2.026	0.046	1.070	0.378**
Panel A. All indicators								
Mul. factors	2.367	0.266**	1.407	0.069	2.105	0.088	1.013	0.570**
1 factor	2.106	1.00**	1.351	0.186*	1.746	1.00**	1.038	0.570**
Comb. mean	2.423	0.093	1.269	0.869**	1.880	0.585**	1.030	0.570**
Comb. median	2.585	0.030	1.294	0.869**	1.939	0.323**	1.055	0.530**
Comb. ridge reg.	2.121	0.975**	1.318	0.869**	1.918	0.518**	1.013	0.570**
Panel B. Real activity indicators								
Mul. factors	2.245	0.768**	1.416	0.022	1.959	0.323**	0.990	0.570**
1 factor	2.115	0.975**	1.347	0.358**	1.774	0.720**	1.041	0.570**
Comb. mean	2.284	0.615**	1.263	0.869**	1.827	0.698**	1.012	0.570**
Comb. median	2.329	0.495**	1.284	0.869**	1.854	0.647**	1.038	0.553**
Comb. ridge reg.	2.160	0.953**	1.326	0.855**	1.888	0.518**	1.013	0.570**
Panel C. Interest rates								
Mul. factors	2.828	0.019	1.512	0.005	2.215	0.008	1.294	0.008
1 factor	2.776	0.030	1.463	0.003	2.111	0.007	1.102	0.161*
Comb. mean	2.474	0.092	1.349	0.123*	1.935	0.323**	1.060	0.522**
Comb. median	2.567	0.077	1.377	0.034	1.974	0.290**	1.066	0.418**
Comb. ridge reg.	2.436	0.164*	1.372	0.069	1.962	0.216*	1.052	0.530**
Panel D. Money								
Mul. factors	2.801	0.015	1.340	0.597**	2.028	0.020	1.075	0.057
1 factor	2.805	0.013	1.352	0.186*	2.027	0.031	1.104	0.026
Comb. mean	2.742	0.019	1.390	0.022	2.033	0.012	1.088	0.015
Comb. median	2.752	0.019	1.340	0.386**	2.032	0.008	1.077	0.095
Comb. ridge reg.	2.721	0.019	1.446	0.007	2.013	0.088	1.088	0.010
Phillips curve								
LHUR	2.637	0.030	1.388	0.022	2.076	0.031	1.162	0.423**

[a] RMSEs and MCS p-values for the different forecasts. The forecasts in $\widehat{\mathcal{M}}^*_{90\%}$ and $\widehat{\mathcal{M}}^*_{75\%}$ are identified by one and two asterisks, respectively.

ket information that is important for predicting inflation during the pre-1984 subsample. This result is consistent with traditional Keynesian measures of aggregate demand.

Table IV also shows that there are two levels and five first difference specifications of the forecasting equation that consistently appear in $\widehat{\mathcal{M}}^*_{75\%}$ using the 1970:M1–1983:M12 subsample. On this subsample, only msmtq (total real manufacturing and trade) is consistently embraced by PUNEW- and GMDC-$\widehat{\mathcal{M}}^*_{75\%}$

whether in levels or first differences. In summary, we interpret these variables as signals about the anticipated path of either real aggregate demand or real aggregate supply that helps to predict inflation out of sample in the pre-1984 subsample.

There are several more inferences to draw from Table IV. These concern the two types of no-change forecasts whose predictive accuracy is strikingly different. The no-change (month) forecast fails to appear in $\widehat{\mathcal{M}}^*_{75\%}$ either on the pre-1984 or on the post-1984 subsamples, whereas the no-change (year) forecast finds its way into $\widehat{\mathcal{M}}^*_{75\%}$ for the post-1984 subsample, but not the 1970:M1–1983:M12 subsample. These results are especially of interest because the no-change (year) forecast yields the best inflation forecasts on the 1984:M1–1996:M9 subsample for both PUNEW and GMDC. These empirical results for the no-change inflation forecasts are interesting because they reconcile the results of Stock and Watson (1999) with those of Atkeson and Ohanian (2001). Stock and Watson (1999, p. 327) found that "[T]he conventionally specified Phillips curve, based on the unemployment rate, was found to perform reasonably well. Its forecasts are better than univariate forecasting models (both autoregressions and random walk models)." In contrast, Atkeson and Ohanian (2001, p. 10) concluded that "economists have not produced a version of the Phillips curve that makes more accurate inflation forecasts than those from a naive model that presumes inflation over the next four quarters will be equal to inflation over the last four quarters." The source of the disagreement is that Stock and Watson and Atkeson and Ohanian studied different no-change inflation forecasts. The no-change forecast Stock and Watson (1999) deployed is last month's inflation rate, whereas the no-change forecasts in Atkeson and Ohanian (2001) is the past year's inflation rate.

We agree with Stock and Watson that the Phillips curve is a device that yields better forecasts of inflation in the pre-1984 period. The relevant $\widehat{\mathcal{M}}^*_{75\%}$ do not include either of the no-change forecasts for PUNEW and GMDC. However, for the post-1984 sample, we observe that no-change (year) forecast has the smallest sample loss of all forecasts, which supports the conclusion of Atkeson and Ohanian (2001).

Table V generates MCSs using factor models and forecast combination methods that replicate the set of forecasts in Stock and Watson (1999, Table 4). They combined a large set of inflation forecasts from an array of 168 models using sample means, sample medians, and ridge estimation to produce forecast weighting schemes. The other forecasting approach depends on principal components of the 168 macropredictors. The idea is that there exists an underlying factor or factors (e.g., real aggregate demand, financial conditions) that summarize the information of a large set of predictors. For example, Solow (1976) argued that a motivation for the Phillips curves of the 1960s and 1970s was that unemployment captured, albeit imperfectly, the true unobserved state of real aggregate demand.

The factor models and forecast combination methods produce inflation forecasts that are, in general, better than those in Table IV. The forecasts constructed from "All indicators" and "Real activity indicators" in Panels A and B do particularly well across the board. Interestingly, the best forecast during the 1970:M1–1983:M12 subsample is the one-factor "All indicators" model, while the second best is the one-factor "Real activity indicators" model. Most of the forecasts constructed from the "Money" variables do not find their way into the MCSs.

Despite the better predictive accuracy produced by factor models and forecast combinations, during the post-1984 period the best forecast is the no-change (year) forecast.

6.2. *Likelihood-Based Comparison of Taylor-Rule Models*

Monetary policy is often evaluated with the Taylor (1993) rule. A Taylor rule summarizes the objectives and constraints that define monetary policy by mapping (implicitly) from this decision problem to the path of the short-term nominal interest rate. A canonical monetary policy loss function penalizes the decision maker for inflation volatility against its target and output volatility around its trend. The mapping generates a Taylor rule that the interest rate responds to inflation and output deviations from trend. Thus, Taylor rules measure ex post the success monetary policy has had at meeting the goals of keeping inflation close to target and output at trend. Articles by Taylor (1999), Clarida, Galí, and Gertler (2000), and Orphanides (2003) are leading examples of using Taylor rules to evaluate actual monetary policy, while McCallum (1999) provided an introduction for consumers of monetary policy rules.

This section shows how the MCS can be used to evaluate which Taylor rule regression best approximates the underlying data generating process. We posit the general Taylor rule regression

$$R_t = (1-\rho)\left[\gamma_0 + \sum_{j=1}^{p_\pi} \gamma_{\pi,j}\pi_{t-j} + \sum_{j=1}^{p_y} \gamma_{y,j}y_{t-j}\right] + \rho R_{t-1} + v_t, \tag{5}$$

where R_t denotes the short-term nominal interest rate, π_t is inflation, y_t equals deviations of output from trend (i.e., the output gap), and the error term, v_t, is assumed to be a martingale difference process. The Taylor principle is satisfied if $\sum_{j=1}^{p_\pi} \gamma_{\pi,j}$ exceeds 1 because a 1% rise in the sum of p_π lags of inflation indicates that R_t should rise by more than 100 basis points. The monetary policy response to real side fluctuations is given by $\sum_{j=1}^{p_y} \gamma_{y,j}$ on the p_y lags of the output gap. The intercept γ_0 is the equilibrium steady state real rate plus the target inflation rate (weighted by $1-\sum_{j=1}^{p_\pi} \gamma_{\pi,j}$). The Taylor rule regression (5) includes lagged interest, R_{t-1}, which may be interpreted as interest rate smoothing by the central bank. Alternatively, the lagged interest rate could be interpreted as

TABLE VI

TAYLOR RULE REGRESSION DATA SET[a]

	Observable	Construction
Dependent variable		
R_t: Interest rate	Effective Fed Funds Rate (EFFR), $R_{\text{fed funds},t}$	Temporally aggregate daily return (annual rate) to quarterly, $R_t = 100 \times \ln[1 + R_{\text{fed funds},t}/100]$
Independent variables		
π_t: Inflation	Implicit GDP deflator, P_t, seasonally adjusted (SA)	$\pi_t = 400 \times \ln[P_t/P_{t-1}]$
y_t: Output gap	$\ln Q_t - \text{trend}\, Q_t$, i.e., transitory component of output, where Q_t is real GDP in billions of chained 2000 $, SA at annual rates	Apply Hodrick–Prescott filter to $\ln Q_t$
ur_t: Unemployment rate gap	$\text{UR}_t - \text{trend}\,\text{UR}_t$, i.e., transitory component of UR_t, where UR_t is the is the civilian unemployment rate, SA	Temporally aggregate monthly to quarterly frequency to get UR_t. Apply Baxter–King filter to UR_t
rulc_t: Real unit labor costs	The cointegrating residual of nominal ULC_t $(= \text{LS}_t - \text{LS}_t)$ and $\ln P_t$. LS_t is labor share, i.e., log of compensation per hour in the nonfarm business sector; LP_t is labor productivity, i.e., log of output per hour of all persons nonfarm business sector	$\text{rulc}_t = \text{LS}_t - \text{LP}_t - \hat{a}_0 - \hat{a}_1 t$ $-\hat{a}_2 \ln P_t$

[a] The effective federal funds rate is obtained from H.15 Selected Interest Rates in Federal Reserve Statistical Releases. The implicit price deflator, real GDP, the unemployment rate, compensation per hour, and output per hour of all persons are constructed by the Bureau of Economic Analysis and are available at the FRED Data Bank at the Federal Reserve Bank of St. Louis. The sample period is 1979:Q1–2006:Q4. The data are drawn from data available online from the Board of Governors and FRED at the Federal Reserve Bank of St. Louis.

a proxy for other determinants of the interest rate that are not captured by the regression (5). Note also that the Taylor rule regression (5) avoids issues that arise in the estimation of simultaneous equation systems because contemporaneous inflation, π_t, and the output gap, y_t, are not regressors, only lags of these variables are. In this case, structural interpretations have to be applied to the Taylor rule regression (5) with care.

The Taylor rule regression (5) is estimated by ordinary least squares on a U.S. sample that runs from 1979:Q1 to 2006:Q4. Table VI provides details about the data used to estimate the Taylor rule regression.[15] The (effective) federal funds rate defines the Taylor rule policy rate R_t. The growth rate of the im-

[15] We have generated results on a shorter post-1984 sample. Omitting the volatile 1979–1983 period from the analysis does not substantially change our results, beyond the loss of information that one would expect with a shorter sample. These results are available in a separate appendix (Hansen, Lunde, and Nason (2011)).

plicit gross domestic product (GDP) deflator is our measure of inflation, π_t. The cyclical component of the Hodrick and Prescott (1997) filter is applied to real GDP to obtain estimates of the output gap, y_t. We also employ two real activity variables to fill out the model space and to act as alternatives to the output gap. These real activity variables are the Baxter and King (1999) filtered unemployment rate gap, ur_t, and the Nason and Smith (2008) measure of real unit labor costs, rulc_t. We compute the Baxter–King ur_t using the maximum likelihood–Kalman filter methods of Harvey and Trimbur (2003).

The model space consists of 25 specifications. The model space is built by setting ρ to zero or estimating it ($p_\pi = 1$ or 2, $p_y = 1$ or 2) and equating y_t with the output gap, or replacing it with either the unemployment rate gap or real unit labor costs. We add to these 24 ($= 2 \times 2 \times 3 \times 2$) regressions a pure AR(1) model of the effective federal funds rate.

TABLE VII

MCS FOR TAYLOR RULES: 1979:Q1–2006:Q4[a]

Model Specification			$Q(\mathcal{Z}_j, \hat{\theta}_j)$	$\hat{k}^\star$	KLIC	AIC*	BIC*
R_{t-1}			93.15	13.74	106.89 (0.30)**	120.63 (0.47)**	157.99 (0.63)**
	π_{t-1}	y_{t-1}	284.82	11.44	296.25 (0.00)	307.69 (0.00)	338.79 (0.00)
	$\pi_{t-j},\ _{j=1,2}$	$y_{t-j},\ _{j=1,2}$	258.95	14.66	273.61 (0.00)	288.28 (0.01)	328.14 (0.01)
	π_{t-1}	ur_{t-1}	289.65	10.20	299.84 (0.00)	310.04 (0.00)	337.75 (0.00)
	$\pi_{t-j},\ _{j=1,2}$	$\text{ur}_{t-j},\ _{j=1,2}$	268.90	12.82	281.72 (0.00)	294.53 (0.00)	329.37 (0.01)
	π_{t-1}	rulc_{t-1}	289.99	9.89	299.88 (0.00)	309.77 (0.00)	336.67 (0.01)
	$\pi_{t-j},\ _{j=1,2}$	$\text{rulc}_{t-j},\ _{j=1,2}$	266.07	12.12	278.19 (0.00)	290.31 (0.01)	323.26 (0.01)
	y_{t-1}	ur_{t-1}	387.45	17.04	404.49 (0.00)	421.54 (0.00)	467.86 (0.00)
	$y_{t-j},\ _{j=1,2}$	$\text{ur}_{t-j},\ _{j=1,2}$	385.86	23.42	409.28 (0.00)	432.69 (0.00)	496.35 (0.00)
	y_{t-1}	rulc_{t-1}	386.47	14.92	401.39 (0.00)	416.32 (0.00)	456.89 (0.00)
	$y_{t-j},\ _{j=1,2}$	$\text{rulc}_{t-j},\ _{j=1,2}$	385.43	19.44	404.87 (0.00)	424.31 (0.00)	477.16 (0.00)
	ur_{t-1}	rulc_{t-1}	386.21	15.41	401.62 (0.00)	417.02 (0.00)	458.90 (0.00)
	$\text{ur}_{t-j},\ _{j=1,2}$	$\text{rulc}_{t-j},\ _{j=1,2}$	384.82	19.86	404.68 (0.00)	424.54 (0.00)	478.52 (0.00)
R_{t-1}	π_{t-1}	y_{t-1}	68.57	17.71	86.28 (0.86)**	103.98 (1.00)**	152.12 (0.64)**
R_{t-1}	$\pi_{t-j},\ _{j=1,2}$	$y_{t-j},\ _{j=1,2}$	62.11	22.11	84.22 (1.00)**	106.32 (0.93)**	166.43 (0.41)**
R_{t-1}	π_{t-1}	ur_{t-1}	77.57	16.32	93.89 (0.72)**	110.22 (0.89)**	154.60 (0.64)**
R_{t-1}	$\pi_{t-j},\ _{j=1,2}$	$\text{ur}_{t-j},\ _{j=1,2}$	73.27	18.79	92.07 (0.80)**	110.86 (0.89)**	161.95 (0.57)**
R_{t-1}	π_{t-1}	rulc_{t-1}	72.80	16.06	88.86 (0.86)**	104.92 (0.93)**	148.58 (1.00)**
R_{t-1}	$\pi_{t-j},\ _{j=1,2}$	$\text{rulc}_{t-j},\ _{j=1,2}$	69.21	19.26	88.47 (0.86)**	107.73 (0.92)**	160.09 (0.58)**
R_{t-1}	y_{t-1}	ur_{t-1}	86.16	19.16	105.33 (0.33)**	124.49 (0.38)**	176.59 (0.16)*
R_{t-1}	$y_{t-j},\ _{j=1,2}$	$\text{ur}_{t-j},\ _{j=1,2}$	85.51	24.32	109.83 (0.28)**	134.16 (0.18)*	200.28 (0.02)
R_{t-1}	y_{t-1}	rulc_{t-1}	89.42	18.92	108.35 (0.29)**	127.27 (0.31)**	178.72 (0.15)*
R_{t-1}	$y_{t-j},\ _{j=1,2}$	$\text{rulc}_{t-j},\ _{j=1,2}$	88.11	22.42	110.53 (0.28)**	132.94 (0.20)*	193.88 (0.03)
R_{t-1}	ur_{t-1}	rulc_{t-1}	87.42	18.07	105.49 (0.33)**	123.55 (0.38)**	172.66 (0.21)*
R_{t-1}	$\text{ur}_{t-j},\ _{j=1,2}$	$\text{rulc}_{t-j},\ _{j=1,2}$	85.93	21.32	107.25 (0.30)**	128.56 (0.28)**	186.51 (0.06)

[a] We report the maximized log-likelihood function (multiplied by -2), the effective degress of freedom, and the three criteria KLIC, AIC*, and BIC* along with the corresponding MCS p-values. The regression models in $\widehat{\mathcal{M}}^*_{90\%}$ and $\widehat{\mathcal{M}}^*_{75\%}$ are identified by one and two asterisks, respectively. See the text and Table VI for variable mnemonics and definitions.

TABLE VIII

REGRESSION MODELS IN $\widehat{\mathcal{M}}^*_{90\%}$-KLIC[a]

γ_0	ρ	$\gamma_{\pi,1}$	$\gamma_{\pi,2}$	$\gamma_{y,1}$	$\gamma_{y,2}$	$\gamma_{\text{ur},1}$	$\gamma_{\text{ur},2}$	$\gamma_{\text{rulc},1}$	$\gamma_{\text{rulc},2}$
5.29	0.96								
(2.50)	(30.1)								
0.12	0.84	1.87		1.20					
(0.13)	(17.0)	(7.01)		(2.17)					
0.00	0.80	0.77	1.14	1.50	−0.39				
(0.00)	(12.1)	(2.58)	(4.76)	(1.25)	(0.33)				
0.82	0.86	1.60				1.58			
(0.67)	(16.8)	(4.85)				(0.25)			
0.64	0.83	0.68	0.97			5.90	−6.56		
(0.56)	(12.9)	(1.77)	(2.85)			(0.68)	(1.16)		
0.37	0.87	1.76						−0.81	
(0.30)	(17.0)	(5.38)						(1.56)	
0.39	0.84	0.76	0.99					−0.18	−0.55
(0.35)	(12.9)	(2.12)	(3.55)					(0.23)	(0.68)
5.63	0.97			4.89		45.9			
(2.20)	(37.3)			(1.05)		(0.79)			
5.56	0.97			6.42	−1.71	60.7	−22.9		
(2.12)	(32.3)			(0.58)	(0.19)	(0.66)	(0.42)		
5.33	0.97			1.04				−2.47	
(2.22)	(35.5)			(0.32)				(0.79)	
5.42	0.97			8.37	−8.05			2.52	−5.43
(2.22)	(32.6)			(0.64)	(0.56)			(0.75)	(0.96)
5.35	0.97					30.9		−3.62	
(2.02)	(37.8)					(0.63)		(1.04)	
5.43	0.97					52.5	−25.6	−1.18	−2.74
(2.10)	(34.2)					(0.64)	(0.54)	(0.30)	(0.85)

[a] Parameter estimates with t-statistics (in absolute values) in parentheses. The shaded area identifies the models in $\widehat{\mathcal{M}}^*_{75\%}$-BIC*.

We present results of applying the MCS and likelihood-based criteria to the choice of the best Taylor rule regression (5) and AR(1) regressions in Tables VII and VIII. Table VII reports $Q(\mathcal{Z}_j, \hat{\theta}_j)$ (the log-likelihood function multiplied by -2), the bootstrap estimate of the effective degrees of freedom, $\hat{k}^\star$, and the realizations of the three empirical criteria, KLIC, AIC*, and BIC*. The numbers surrounded by parentheses in columns headed KLIC, AIC*, and BIC* are the MCS p-values, and an asterisk identifies the specifications that enter $\widehat{\mathcal{M}}^*_{90\%}$. Table VIII lists estimates of the regression models that are in $\widehat{\mathcal{M}}^*_{90\%}$ along with their corresponding t-statistics in parentheses.

The t-statistics are based on robust standard errors following Newey and West (1987).

Table VII shows that the MCS procedure selects 10–13 of the 25 possible regressions depending on the information criteria. The lagged nominal rate R_{t-1} is the one regressor common to the regressions that enter $\widehat{\mathcal{M}}^*_{90\%}$ for the KLIC, AIC*, and BIC*. Besides the AR(1), $\widehat{\mathcal{M}}^*_{90\%}$ consists of the six Taylor rule specifications that nest the AR(1). Under the KLIC and AIC*, the Taylor rule regressions include all one or two lag combinations of π_t, y_t, ur_t, and rulc_t. The BIC produces a smaller $\widehat{\mathcal{M}}^*_{90\%}$ because it ejects the two lag Taylor rule specifications that exclude lagged π_t. Thus, the Taylor rule regression–MCS example finds that the BIC tends to settle on more parsimonious models. This is to be expected, given its larger penalty on model complexity.

The AR(1) falls into $\widehat{\mathcal{M}}^*_{90\%}$ under the KLIC, AIC*, and BIC*. Although the first line of Table VII shows that the AR(1) has the largest $Q(\mathcal{Z}_j, \hat{\theta}_j)$ of the regressions covered by $\widehat{\mathcal{M}}^*_{90\%}$, the MCS recruits the AR(1) because it has a relatively small estimate of the effective degrees of freedom, $\hat{k}^\star$. It is important to keep in mind that estimates of the effective degrees of freedom are larger than the number of free parameters in each of the models. This reflects the fact that the Gaussian model is misspecified. For example, the conventional AIC penalty (that doubles the number of free parameters) is misleading in the context of misspecified models; see Takeuchi (1976), Sin and White (1996), and Hong and Preston (2008).

It is somewhat disappointing that the MCS procedure yields as many as 13 models in $\widehat{\mathcal{M}}^*_{90\%}$. The reason is that the data lack the information to resolve precisely which Taylor rule specification is best in terms of Kullback–Leibler discrepancy. The large set of models is also an outcome of the strict requirements that characterize the MCS. The MCS procedure is designed to control the familywise error rate (FWE), which is the probability of making one or more false rejections. We will be able to trim $\widehat{\mathcal{M}}^*$ further if we relax the control of the FWE, but that will affect the interpretation of $\widehat{\mathcal{M}}^*_{1-\alpha}$. For instance, if we control the probability of making k or more false rejections, k-FWE (see, e.g., Romano, Shaikh, and Wolf (2008)), additional models can be eliminated. The drawback of k-FWE and other alternative controls is that the MCS looses its key property, which is to contain the best models with probability $1-\alpha$.

Table VIII provides information about the regressions in $\widehat{\mathcal{M}}^*_{90\%}$-KLIC. The shaded area identifies the models in $\widehat{\mathcal{M}}^*_{75\%}$-BIC*. First, note that the estimated Taylor rules always satisfy the Taylor principle (i.e., $\hat{\gamma}_{\pi,1} > 1$ or $\hat{\gamma}_{\pi,1} + \hat{\gamma}_{\pi,2} > 1$). The coefficients associated with real activity variables have insignificant t-statistics in most cases. Only the first lag of the output gap produces a positive coefficient with a t-ratio above 2 in the first Taylor rule regression listed in Table VIII. Moreover, the statistically insignificant coefficients for the unemployment rate gap and real unit labor costs variables often have counterintuitive signs. Finally, the estimates of ρ are between 0.83 and 0.87 in the Taylor rule regressions that include a lag of π_t, which suggests interest rate smoothing.[16]

The fact that the MCS cannot settle on a single specification is not a surprising result. Monetary policymakers almost surely rely on a more complex information set than can be summarized by a simple model. Furthermore, any real activity variable is an imperfect measure of the underlying state of the economy, and there are important and unresolved issues regarding the measurement of gap and marginal cost variables that translate into uncertainty about the proper definitions of the real activity variables.

7. SUMMARY AND CONCLUDING REMARKS

This paper introduces the model confidence set (MCS) procedure, relates it to other approaches of model selection and multiple comparisons, and establishes the asymptotic theory of the MCS. The MCS is constructed from a hypothesis test, $\delta_{\mathcal{M}}$, and an elimination rule, $e_{\mathcal{M}}$. We defined coherency between test and elimination rule, and stressed the importance of this concept for the finite sample properties of the MCS. We also outlined simple and convenient bootstrap methods for the implementation of the MCS procedure. The paper employs Monte Carlo experiments to study the MCS procedure that reveal it has good small sample properties.

It is important to understand the principle of the MCS procedure in applications. The MCS is constructed such that inference about the "best" follows the conventional meaning of the word "significance." Although the MCS will contain only the best model(s) asymptotically, it may contain several poor models in finite samples. A key feature of the MCS procedure is that a model is discarded only if it is found to be significantly inferior to another model. Models remain in the MCS until proven inferior, which has the implication that not all models in the MCS may be judged good models.[17]

An important advantage of the MCS, compared to other selection procedures, is that the MCS acknowledges the limits to the informational content of the data. Rather than selecting a single model without regard to degree of information, the MCS procedure yields a set of models that summarizes key sample information.

We applied the MCS procedure to the inflation forecasting problem of Stock and Watson (1999). Results show that the MCS procedure provides a powerful tool for evaluating competing inflation forecasts. We emphasize that the information content of the data matters for the inferences that can be drawn. The

[16] We have also estimated Taylor rule regressions with moving average (MA) errors, as an alternative to using R_{t-1} as a regressor. The empirical fit of models with MA errors is, in all cases, inferior to the Taylor rule regressions that include R_{t-1}.

[17] The proportion of models in $\widehat{\mathcal{M}}^*_{1-\alpha}$ that are members of $\mathcal{M}^*$ can be related to the *false discovery rate* and the q-value theory of Storey (2002). See McCracken and Sapp (2005) for an application that compares forecasting models. See also Romano, Shaikh, and Wolf (2008).

great inflation–disinflation subsample of 1970:M1–1983:M12 has movements in inflation and macrovariables that allow the MCS procedure to make relatively sharp choices across the relevant models. The information content of the less persistent, less volatile 1984:M1–1996:M9 subsample is limited in comparison because the MCS procedure lets in almost any model that Stock and Watson considered. A key exception is the no-change (month) forecast that uses last month's inflation rate as a predictor of future inflation. This no-change forecast never resides in the MCS in either the earlier or the later periods. A likely explanation is that month-to-month inflation is a noisy measure of core inflation. This view is supported by the fact that a second no-change (year) forecast, which employs a year-over-year inflation rate as the forecast, is a better forecast. This result enables us to reconcile the empirical results in Stock and Watson (1999) with those of Atkeson and Ohanian (2001). Nonetheless, the question of what constitutes the best inflation forecasting model for the last 35 years of U.S. data remains unanswered because the data provide insufficient information to distinguish between good and bad models.

This paper also constructs a MCS for Taylor rule regressions based on three likelihood criteria. Such interest rate rules are often used to evaluate the success of monetary policy, but this is not our intent for the MCS. Instead, we study the MCS that selects the best fitting Taylor rule regressions under either a quasi-likelihood criterion, the AIC, or the BIC using the effective degrees of freedom. The competing Taylor rule regressions consist of different combinations of lags of inflation, lags of three different real activity variables, and the lagged federal funds rate. Besides these Taylor rule regressions, the MCS must also contend with a first-order autoregression of the federal funds rate. The regressions are estimated on a 1979:Q1–2006:Q4 sample of U.S. data. Under the three likelihood criteria, the MCS settles on Taylor rule regressions that satisfy the Taylor principle, include all three competing real activity variables, and add the lagged federal funds rate. Furthermore, we find that the first-order autoregression also enters the MCS. Thus, the U.S. data lack the information to resolve precisely which Taylor rule specification best describes the data.

Given the large number of forecasting problems economists face at central banks and other parts of government, in financial markets, and other settings, the MCS procedure faces a rich set of problems to study. Furthermore, the MCS has a wide variety of potential uses beyond forecast comparisons and regression models. We leave this work for future research.

REFERENCES

ANDERSON, T. W. (1984): *An Introduction to Multivariate Statistical Analysis* (Second Ed.). New York: Wiley. [455]

ANDREWS, D. W. K. (1991): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59, 817–858. [470]

ATKESON, A., AND L. E. OHANIAN (2001): "Are Phillips Curves Useful for Forecasting Inflation?" *Federal Reserve Bank of Minneapolis Quarterly Review*, 25, 2–11. [456,483,484,487,494]

BAXTER, M., AND R. G. KING (1999): "Measuring Business Cycles: Approximate Bandpass Filters for Economic Time Series," *Review of Economics and Statistics*, 81, 575–593. [490]

BERGER, R. L. (1982): "Multiparameter Hypothesis Testing and Acceptance Sampling," *Technometrics*, 24, 295–300. [473]

BERNANKE, B. S., AND J. BOIVIN (2003): "Monetary Policy in a Data-Rich Environment," *Journal of Monetary Economics*, 50, 525–546. [457]

CAVANAUGH, J. E., AND R. H. SHUMWAY (1997): "A Bootstrap Variant of AIC for State-Space Model Selection," *Statistica Sinica*, 7, 473–496. [471]

CHAO, J. C., V. CORRADI, AND N. R. SWANSON (2001): "An Out of Sample Test for Granger Causality," *Macroeconomic Dynamics*, 5, 598–620. [476]

CHONG, Y. Y., AND D. F. HENDRY (1986): "Econometric Evaluation of Linear Macroeconomic Models," *Review of Economic Studies*, 53, 671–690. [476]

CLARIDA, R., J. GALÍ, AND M. GERTLER (2000): "Monetary Policy Rules and Macroeconomic Stability: Evidence and Some Theory," *Quarterly Journal of Economics*, 115, 147–180. [488]

CLARK, T. E., AND M. W. MCCRACKEN (2001): "Tests of Equal Forecast Accuracy and Encompassing for Nested Models," *Journal of Econometrics*, 105, 85–110. [475,476]

——— (2005): "Evaluating Direct Multi-Step Forecasts," *Econometric Reviews*, 24, 369–404. [466]

DIEBOLD, F. X., AND R. S. MARIANO (1995): "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 13, 253–263. [465]

DOORNIK, J. A. (2009): "Autometrics," in *The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry*, ed. by N. Shephard and J. L. Castle. New York: Oxford University Press, 88–121. [468]

——— (2006): *Ox: An Object-Orientated Matrix Programming Language* (Fifth Ed.). London: Timberlake Consultants Ltd. [453]

DUDOIT, S., J. P. SHAFFER, AND J. C. BOLDRICK (2003): "Multiple Hypothesis Testing in Microarray Experiments," *Statistical Science*, 18, 71–103. [473,474]

EFRON, B. (1983): "Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation," *Journal of the American Statistical Association*, 78, 316–331. [470,471]

——— (1986): "How Biased Is the Apparent Error Rate of a Prediction Rule?" *Journal of the American Statistical Association*, 81, 461–470. [471]

ENGLE, R. F., AND S. J. BROWN (1985): "Model Selection for Forecasting," *Journal of Computation in Statistics*, 51, 341–365. [475]

GIACOMINI, R., AND H. WHITE (2006): "Tests of Conditional Predictive Ability," *Econometrica*, 74, 1545–1578. [476,484]

GONCALVES, S., AND H. WHITE (2005): "Bootstrap Standard Error Estimates for Linear Regression," *Journal of the American Statistical Association*, 100, 970–979. [468,470]

GORDON, R. J. (1997): "The Time-Varying NAIRU and Its Implications for Economic Policy," *Journal of Economic Perspectives*, 11, 11–32. [456]

GRANGER, C. W. J., M. L. KING, AND H. WHITE (1995): "Comments on Testing Economic Theories and the Use of Model Selection Criteria," *Journal of Econometrics*, 67, 173–187. [475]

GUPTA, S. S., AND S. PANCHAPAKESAN (1979): *Multiple Decision Procedures*. New York: Wiley. [473]

HANSEN, P. R. (2003a): "Asymptotic Tests of Composite Hypotheses," Working Paper 03-09, Brown University Economics. Available at http://ssrn.com/abstract=399761. [475]

——— (2003b): "Regression Analysis With Many Specifications: A Bootstrap Method to Robust Inference," Mimeo, Stanford University. [466]

——— (2005): "A Test for Superior Predictive Ability," *Journal of Business & Economic Statistics*, 23, 365–380. [466,471,474]

HANSEN, P. R., A. LUNDE, AND J. M. NASON (2011): "Supplement to 'The Model Confidence Set'," *Econometrica Supplemental Material*, 79, http://www.econometricsociety.org/ecta/Supmat/5771_tables.pdf; http://www.econometricsociety.org/ecta/Supmat/5771_data and programs.zip. [457,467,481,484,489]

HARVEY, A. C., AND T. M. TRIMBUR (2003): "General Model-Based Filters for Extracting Cycles and Trends in Economic Time Series," *Review of Economics and Statistics*, 85, 244–255. [490]

HARVEY, D., AND P. NEWBOLD (2000): "Tests for Multiple Forecast Encompassing," *Journal of Applied Econometrics*, 15, 471–482. [476]

HODRICK, R. J., AND E. C. PRESCOTT (1997): "Postwar U.S. Business Cycles: An Empirical Investigation," *Journal of Money, Credit, and Banking Economy*, 29, 1–16. [484,490]

HONG, H., AND B. PRESTON (2008): "Bayesian Averaging, Prediction and Nonnested Model Selection," Working Paper W14284, NBER. [470,492]

HORRACE, W. C., AND P. SCHMIDT (2000): "Multiple Comparisons With the Best, With Economic Applications," *Journal of Applied Econometrics*, 15, 1–26. [473]

HSU, J. C. (1996): *Multiple Comparisons*. Boca Raton, FL: Chapman & Hall/CRC. [473]

INOUE, A., AND L. KILIAN (2006): "On the Selection of Forecasting Models," *Journal of Econometrics*, 130, 273–306. [475]

JOHANSEN, S. (1988): "Statistical Analysis of Cointegration Vectors," *Journal of Economic Dynamics and Control*, 12, 231–254. [455,473]

KILIAN, L. (1999): "Exchange Rates and Monetary Fundamentals: What Do We Learn From Long Horizon Regressions?" *Journal of Applied Econometrics*, 14, 491–510. [466]

LEEB, H., AND B. PÖTSCHER (2003): "The Finite-Sample Distribution of Post-Model-Selection Estimators, and Uniform versus Non-Uniform Approximations," *Econometric Theory*, 19, 100–142. [460]

LEHMANN, E. L., AND J. P. ROMANO (2005): *Testing Statistical Hypotheses* (Third Ed.). New York: Wiley. [464,473,474]

MCCALLUM, B. T. (1999): "Issues in the Design of Monetary Policy Rules," in *Handbook of Macroeconomics*, Vol. 1C, ed. by J. B. Taylor and M. Woodford. Amsterdam: North-Holland, 1483–1530. [488]

MCCRACKEN, M. W., AND S. SAPP (2005): "Evaluating the Predictability of Exchange Rates Using Long Horizon Regressions: Mind Your p's and q's!" *Journal of Money, Credit, and Banking*, 37, 473–494. [493]

NASON, J. M., AND G. W. SMITH (2008): "Identifying the New Keynesian Phillips Curve," *Journal of Applied Econometrics*, 23, 525–551. [490]

NEWEY, W., AND K. WEST (1987): "A Simple Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708. [464,470,492]

ORPHANIDES, A. (2003): "Historical Monetary Policy Analysis and the Taylor Rule," *Journal of Monetary Economics*, 50, 983–1022. [488]

ORPHANIDES, A., AND S. VAN NORDEN (2002): "The Unreliability of Output-Gap Estimates in Real Time," *Review of Economics and Statistics*, 84, 569–583. [457]

PANTULA, S. G. (1989): "Testing for Unit Roots in Time Series Data," *Econometric Theory*, 5, 256–271. [473]

ROMANO, J. P., AND M. WOLF (2005): "Stepwise Multiple Testing as Formalized Data Snooping," *Econometrica*, 73, 1237–1282. [474]

ROMANO, J. P., A. M. SHAIKH, AND M. WOLF (2008): "Formalized Data Snooping Based on Generalized Error Rates," *Econometric Theory*, 24, 404–447. [474,492,493]

SHIBATA, R. (1997): "Bootstrap Estimate of Kullback–Leibler Information for Model Selection," *Statistica Sinica*, 7, 375–394. [471]

SHIMODAIRA, H. (1998): "An Application of Multiple Comparison Techniques to Model Selection," *Annals of the Institute of Statistical Mathematics*, 50, 1–13. [473]

SIN, C.-Y., AND H. WHITE (1996): "Information Criteria for Selecting Possibly Misspecified Parametric Models," *Journal of Econometrics*, 71, 207–225. [468,470,475,492]

SOLOW, R. M. (1976): "Down the Phillips Curve With Gun and Camera," in *Inflation, Trade, and Taxes*, ed. by D. A. Belsley, E. J. Kane, P. A. Samuelson, and R. M. Solow. Columbus, OH: Ohio State University Press. [487]

STAIGER, D., J. H. STOCK, AND M. W. WATSON (1997a): "How Precise Are Estimates of the Natural Rate of Unemployment?" in *Reducing Inflation: Motivation and Strategy*, ed. by C. Romer and D. Romer. Chicago: University of Chicago Press, 195–242. [457]

——— (1997b): "The NAIRU, Unemployment, and Monetary Policy," *Journal of Economic Perspectives*, 11, 33–49. [456]

STOCK, J. H., AND M. W. WATSON (1999): "Forecasting Inflation," *Journal of Monetary Economics*, 44, 293–335. [453-456,483,484,487,493,494]

——— (2003): "Forecasting Output and Inflation: The Role of Asset Prices," *Journal of Economic Literature*, 61, 788–829. [456]

STOREY, J. D. (2002): "A Direct Approach to False Discovery Rates," *Journal of the Royal Statistical Society, Ser. B*, 64, 479–498. [493]

TAKEUCHI, K. (1976): "Distribution of Informational Statistics and a Criterion of Model Fitting," *Suri-Kagaku (Mathematical Sciences)*, 153, 12–18. (In Japanese.) [470,492]

TAYLOR, J. B. (1993): "Discretion versus Policy Rules in Practice," *Carnegie–Rochester Conference Series on Public Policy*, 39, 195–214. [456,488]

——— (1999): "A Historical Analysis of Monetary Policy Rules," in *Monetary Policy Rules*, ed. by J. B. Taylor. Chicago: University of Chicago Press, 319–341. [488]

VUONG, Q. H. (1989): "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses," *Econometrica*, 57, 307–333. [471,473]

WEST, K. D. (1996): "Asymptotic Inference About Predictive Ability," *Econometrica*, 64, 1067–1084. [465]

WEST, K. D., AND D. CHO (1995): "The Predictive Ability of Several Models of Exchange Rate Volatility," *Journal of Econometrics*, 69, 367–391. [464]

WEST, K. D., AND M. W. MCCRACKEN (1998): "Regression Based Tests of Predictive Ability," *International Economic Review*, 39, 817–840. [475]

WHITE, H. (1994): *Estimation, Inference and Specification Analysis*. Cambridge: Cambridge University Press. [469]

——— (2000a): *Asymptotic Theory for Econometricians* (Revised Ed.). San Diego: Academic Press. [464]

——— (2000b): "A Reality Check for Data Snooping," *Econometrica*, 68, 1097–1126. [455,466, 471,474]

Dept. of Economics, Stanford University, 579 Serra Mall, Stanford, CA 94305-6072, U.S.A. and CREATES; peter.hansen@stanford.edu,

School of Economics and Management, Aarhus University, Bartholins Allé 10, Aarhus, Denmark and CREATES; alunde@econ.au.dk,

and

Federal Reserve Bank of Philadelphia, Ten Independence Mall, Philadelphia, PA 19106-1574, U.S.A.; Jim.Nason@phil.frb.org.

Manuscript received March, 2005; final revision received March, 2010.