Hansen, Peter Reinhard; Lunde, Asger; Nason, James M.

Working Paper
Model confidence sets for forecasting models


Provided in Cooperation with:
Federal Reserve Bank of Atlanta

Suggested Citation: Hansen, Peter Reinhard; Lunde, Asger; Nason, James M. (2005) : Model confidence sets for forecasting models, Working Paper, No. 2005-07, Federal Reserve Bank of Atlanta, Atlanta, GA

This Version is available at:
http://hdl.handle.net/10419/100950
Model Confidence Sets for Forecasting Models

Peter Reinhard Hansen, Asger Lunde, and James M. Nason

Working Paper 2005-7
March 2005
Model Confidence Sets for Forecasting Models

Peter Reinhard Hansen, Asger Lunde, and James M. Nason

Working Paper 2005-7
March 2005

Abstract: The paper introduces the model confidence set (MCS) and applies it to the selection of forecasting models. An MCS is a set of models that is constructed so that it will contain the “best” forecasting model, given a level of confidence. Thus, an MCS is analogous to a confidence interval for a parameter. The MCS acknowledges the limitations of the data so that uninformative data yield an MCS with many models, whereas informative data yield an MCS with only a few models. We revisit the empirical application in Stock and Watson (1999) and apply the MCS procedure to their set of inflation forecasts. In the first pre-1984 subsample we obtain an MCS that contains only a few models, notably versions of the Solow-Gordon Phillips curve. On the other hand, the second post-1984 subsample contains little information and results in a large MCS. Yet, the random walk forecast is not contained in the MCS for either of the samples. This outcome shows that the random walk forecast is inferior to inflation forecasts based on Phillips curve-like relationships.

JEL classification: C12, C19, C44, C52, C53

Key words: model confidence set, forecasting, model selection, multiple comparisons

The authors gratefully thank Barbara Rossi, Michael Wolf, Jim Stock, and seminar participants at UCSD, CalTech, University of Copenhagen, Cass, Queen Mary, Universitat Pompeu Fabra, and the 2004 NBER Summer Institute for valuable comments. The first two authors are grateful to financial support from the Danish Research Agency, grant no. 24-00-0363, and the last author wishes to thank the Federal Reserve Bank of Atlanta for its support and hospitality during his visit. The views expressed here are the authors’ and not necessarily those of the Federal Reserve Bank of Atlanta or the Federal Reserve System. Any remaining errors are the authors’ responsibility.

Please address questions regarding content to Peter Reinhard Hansen, Department of Economics, Stanford University, 579 Serra Mall, Stanford, Calif., 94305-6072, 650-725-1869, peter.hansen@stanford.edu.

Federal Reserve Bank of Atlanta working papers, including revised versions, are available on the Atlanta Fed’s Web site at www.frbatlanta.org. Click “Publications” and then “Working Papers.” Use the WebScriber Service (at www.frbatlanta.org) to receive e-mail notifications about new papers.
1 Introduction

Which is the ‘best’ forecasting model? This question is onerous for most data to answer, especially when the set of competing models is large. Many applications will not yield a single model that significantly dominates all competitors because the data is not sufficiently informative to give a 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 is guaranteed to contain the ‘best’ forecasting model, given a pre-specified level of confidence.

The objective of the model confidence set (MCS) procedure is to determine the $\mathcal{M}^*$ 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 set of models constructed to contain the best models with a given level of confidence. The models in $\widehat{\mathcal{M}}^*$ are evaluated using sample information about the relative performances of the models in $\mathcal{M}_0$. Thus, the MCS is a random data-dependent set of models that includes the best forecasting model(s), as a standard confidence interval covers the 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 makes it difficult to distinguish between models and may result in a MCS that contains several (or possibly all the) 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’, i.e., $\mathcal{M}^*$ may contain more than a single model.

The contributions of this paper can be summarized as follows: First, we introduce the model confidence set and derive its theoretical properties. Second, we propose a practical implementation of the MCS procedure that is based on bootstrap methods. 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 revisit the empirical application in Stock and Watson (1999) and apply the MCS procedure to their set of inflation forecasts.
1.1 Theory of Model Confidence Sets

We do not treat ‘models’ as sacred objects, nor 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 ‘model’ loosely. It can refer to a forecasting model, method, or rule that need not involve any modelling of data. The MCS procedure is not specific to comparisons of forecasting models. It can also be used to seek the ‘best’ among more general objects. For example, one could construct a MCS for a set of different ‘treatments’ by comparing sample estimates of the corresponding treatment effects.

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 of objects \( \mathcal{M} = \mathcal{M}_0 \). If \( \delta_{\mathcal{M}} \) is rejected, there is evidence that the models in \( \mathcal{M} \) are not equally ‘good’ and \( e_{\mathcal{M}} \) is used to eliminate an object with poor sample performance from \( \mathcal{M} \). This procedure is repeated until \( \delta_{\mathcal{M}} \) is ‘accepted’, and the MCS is now defined by the set of ‘surviving’ models. The same significance level, \( \alpha \), is employed in all tests, which asymptotically guarantees that \( P(\mathcal{M}^* \subset \hat{\mathcal{M}}_{1-\alpha}^*) \geq 1 - \alpha \), and in the case where \( \mathcal{M}^* \) consists of one object we have the stronger results that \( \lim_{n \to \infty} P(\mathcal{M}^* = \hat{\mathcal{M}}_{1-\alpha}^*) = 1 \). The MCS procedure also yields \( p \)-values for each of the models. For a given model \( i \in \mathcal{M}_0 \), the MCS \( p \)-value, \( \hat{p}_i \), is the threshold at which \( i \in \hat{\mathcal{M}}_{1-\alpha}^* \), if and only if \( \hat{p}_i \geq \alpha \). Thus, a model with a small MCS \( p \)-value makes it unlikely that model \( i \) is one of the ‘best’ models (is a member of \( \mathcal{M}^* \)).

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 of Anderson (1984). This procedure involves a sequence of trace-tests, which 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. We discuss this issue and related testing procedures in Section 3.

1.2 Forecasting Models

The focus of this paper is the multiple comparisons of forecasting models. In this context, the equivalence test amounts to a test for equal predictive ability (EPA), such as those by Diebold and Mariano (1995) and West (1996). The natural extension of these tests to the comparison of multiple forecasting models
leads to quadratic-form tests, such as that of West and Cho (1995). These tests require an estimate of a covariance matrix that has a dimension that is proportional to the number of models. Estimation of this covariance matrix can be difficult when the number of models in \( \mathcal{M}_0 \) is large. For this reason, we consider alternative tests that combine simple \( t \)-statistics, which do not require an estimate of the covariance matrix.

Several papers have studied the problem of selecting the best forecasting model from a set of competing models. For example, Engle and Brown (1985) compare selection procedures that are based on six information criteria and two testing procedures (‘general-to-specific’ and ‘specific-to-general’), Sin and White (1996) analyze information criteria for possibly misspecified models, and Inoue and Kilian (2005) compare selection procedures that are based on information criteria and out-of-sample evaluation. Granger, King, and White (1995) argue 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 defines the null hypothesis. Instead, all models are treated equally in terms of their sample performance, and in the context of forecasting models, these are evaluated through their out-of-sample predictive ability. We make no attempt to justify that forecasts should be evaluated in terms of their out-of-sample predictive ability. For views on this issue, see Clements (2002), Clements and Hendry (2003), and Inoue and Kilian (2004). We not address issues that relate to parameter estimation (in the context of forecasting). A thorough analysis of this problem would obscure our main objective, which is to lay out the key ideas of the MCS and make it operative.

### 1.3 Bootstrap Implementation and Simulation Results

We propose a bootstrap implementation of the MCS procedure that is very 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 as the best model does end up in the MCS at the appropri-
ate frequency, and the MCS procedure does have power to weed out all the poor models when the data contains sufficient information.

1.4 Empirical Analysis of Inflation Forecasts

We apply the MCS to the problem of forecasting inflation. The tradition of the Phillips curve suggests it remains a useful vehicle for this task. Stock and Watson (1999) make the case that a reasonably specified Phillips curve is the best tool for forecasting inflation; also see Gordon (1997), Staiger, Stock, and Watson (1997), and Stock and Watson (2003). Atkeson and Ohanian (2001) conclude that this is not the case because they find it is difficult for any of the Phillips curves they study to beat a random walk in out-of-sample point prediction.

Our empirical analysis is based on the same data as Stock and Watson (1999), and we partition the evaluation period in the same two subsamples as did Stock and Watson (1999). The main advantage of the MCS procedure in this context is that it allows us to make statements about significance that are valid, in the traditional sense. This property is difficult to achieve using the traditional approach of making multiple pairwise comparisons. The problem is particularly severe when the comparisons are made with reference to a benchmark that is selected based on information from the same set of data.

There are several interesting results of our analysis. Since the earlier subsample covers a period with persistent and volatile inflation, this sample is expected to be relatively informative about which model(s) might be the best forecasting model(s). Indeed, the MCS consists only of a few models, so the MCS proves to be effective at weeding out 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 might deliver the best forecasts. In spite of the relatively low degree of information in the more recent subsample, we are able to conclude that the simple random walk forecast is indeed inferior to other forecasts, and the performance difference is significant. This conclusion can be made because the random walk forecasts never ends up in the MCS. Although we cannot point to a single models as the ‘significantly best’ forecasting models, we do find that the index-based forecasts of Stock and Watson (2002a, 2002b) generally perform quite well.

The paper is organized as follows. Section 2 presents the theoretical framework for the MCS. Some theoretical aspects of the MCS procedure are discussed in Section 3, where we emphasize similarities and differences to some existing methods for multiple comparisons (of forecasting models). The results
of simulation experiments are discussed in Section 4. The next section applies the MCS to the problem of inflation forecasting and reports the outcome. Section 6 concludes.

2 Theory for General Confidence Set

In this section, we discuss the theory of model confidence sets for general objects. Our leading example concerns the comparison of forecasting model. 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, \( \mathcal{C}_0 \), that contains a finite number of objects (forecasting models) that are indexed by \( i = 1, \ldots, m_0 \). The objects are evaluated over the sample \( t = 1, \ldots, n \), in terms of a loss function and we denote the loss that is associated with object \( i \) in period \( t \) as \( L_{i,t} \).

We define the relative performance variables

\[
d_{ij,t} = L_{i,t} - L_{j,t}, \quad \text{for all } i, j \in \mathcal{C}_0,
\]

and throughout we assume that \( E(d_{ij,t}) \) is finite and does not depend on \( t \), for all \( i, j \in \mathcal{M}_0 \).

**Definition 1** The set of superior objects is defined by

\[
\mathcal{M}^* \equiv \{ i \in \mathcal{M}_0 : E(d_{ij,t}) \leq 0 \quad \text{for all } j \in \mathcal{M}_0 \}.
\]

In the following, we let \( \mathcal{M}^\dagger \) denote the complement to \( \mathcal{M}^* \), i.e. \( \mathcal{M}^\dagger \equiv \{ i \in \mathcal{M}_0 : E(d_{ij,t}) > 0 \quad \text{for some } j \in \mathcal{M}_0 \} \) and we use \( i^* \) and \( i^\dagger \) to represent typical elements of \( \mathcal{M}^* \) and \( \mathcal{M}^\dagger \), respectively.

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:

\[
H_{0,\mathcal{M}} : E(d_{ij,t}) = 0 \quad \text{for all } i, j \in \mathcal{M}, \tag{1}
\]

where \( \mathcal{M} \subset \mathcal{M}_0 \). We denote the alternative hypothesis \( (E(d_{ij,t}) \neq 0 \quad \text{for some } i, j \in \mathcal{M}) \) by \( H_{A,\mathcal{M}} \).

Note that \( H_{0,\mathcal{M}^*} \) is always true given our definition of \( \mathcal{M}^* \), whereas \( H_{0,\mathcal{M}} \) is always false if \( \mathcal{M} \) contains elements from \( \mathcal{M}^* \) and \( \mathcal{M}^\dagger \).

Note that there might be an even better model outside the set of ‘candidate models’, \( \mathcal{M}_0 \). Although the quest for the ‘best of all models’ is an interesting problem, it is a difficult one. One aspect of this

\[1\text{In the situation where a point forecast, } \hat{Y}_{i,t}, \text{ of } Y_t \text{ is evaluated in terms of a loss function, } L, \text{ we define } L_{i,t} = L(Y_t, \hat{Y}_{i,t}).\]
problem that is important to acknowledge is that statements about models outside $\mathcal{M}_0$ hinge on untestable assumptions, unless one has sample information about these models. If one has sample information about additional models one can, in principle, include these models in $\mathcal{M}_0$ and derive the MCS for the larger set of candidate models.

As stated in the introduction, the MCS procedure is based on an equivalence test, $\delta_M$, and an elimination rule, $e_M$. The equivalence test, $\delta_M$, is used to test the hypothesis $H_{0,M}$ for any $\mathcal{M} \subset \mathcal{M}_0$, and $e_M$ identifies the object of $\mathcal{M}$ that is to be removed from $\mathcal{M}$, in the event that $H_{0,M}$ is rejected.

**Definition 2 (MCS Algorithm)** Step 0: Initially set $\mathcal{M} = \mathcal{M}_0$. Step 1: Test $H_{0,M}$ using $\delta_M$ at level $\alpha$.

Step 2: If $H_{0,M}$ is ‘accepted’ we define the $\widehat{\mathcal{M}}_{1-\alpha}^*$ as $\mathcal{M}$, otherwise we use $e_M$ to eliminate an object from $\mathcal{M}$ and repeat the procedure beginning with 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 shows below that the term ‘confidence set’ is appropriate in this context, provided that the equivalence test and the elimination rule satisfies the following assumption.

**Assumption 1** For any $\mathcal{M} \subset \mathcal{M}_0$ we assume the following about $(\delta_M, e_M)$: (a) $\limsup_{n \to \infty} P(\delta_M = 1|H_{0,M}) \leq \alpha$; (b) $\lim_{n \to \infty} P(\delta_M = 1|H_{A,M}) = 1$; and (c) $\lim_{n \to \infty} P(e_M \in \mathcal{M}^*|H_{A,M}) = 0$.

Assumption 1 is standard. (a) requires the asymptotic level not to exceed $\alpha$; (b) requires the asymptotic power to be one; whereas (c) requires that a superior object $i^* \in \mathcal{M}^*$ is not 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) $\lim_{n \to \infty} P(\mathcal{M}^* \subset \widehat{\mathcal{M}}_{1-\alpha}^*) \geq 1 - \alpha$, and (ii) $\lim_{n \to \infty} P(i^* \in \widehat{\mathcal{M}}_{1-\alpha}^*) = 0$ for all $i^* \in \mathcal{M}^*$.

**Proof.** To prove (i) we consider the event that $i^* \in \mathcal{M}^*$ is eliminated from $\mathcal{M}$. From Assumption 1.c it follows that $P(\delta_M = 1, e_M = i^*|H_{A,M}) \leq P(e_M = i^*|H_{A,M}) \to 0$ as $n \to \infty$, and Assumption 1.a shows that $P(\delta_M = 1, e_M = i^*|H_{0,M}) = P(\delta_M = 1|H_{0,M}) \leq \alpha$. To prove (ii), we first note that $\lim_{n \to \infty} P(e_M = i^*|H_{A,M}) = 0$ such that only poor models will be eliminated (asymptotically) as long as $\mathcal{M} \cap \mathcal{M}^\dagger \neq \emptyset$. On the other hand, Assumption 1.b ensures that models will be eliminated as long as the null hypothesis is false. ■

---

2We let $\delta_M = 0$ and $\delta_M = 1$ correspond to the cases where $H_{0,M}$ are ‘accepted’ and ‘rejected’ respectively.
Consider first the situation where the data contains 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 because the lack of power is tied to the lack of information in the data. It is only appropriate that the MCS is large when the data does not contain sufficient information to tell the good and bad models apart.

Consider now the situation where the data is informative. In this situation, the equivalence test is powerful and will reject all false hypotheses, while the elimination rule will not question any of the superior models 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. So the asymptotic familywise (Type I) error rate is bounded by the level that is used in all tests.

Econometricians often worry about the properties of sequential testing procedures, because these can ‘build-up’ Type I errors, and result in unfortunate properties, 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 2** Suppose that Assumption 1 holds and that \( \mathcal{M}^* \) is a singleton, \( \mathcal{M}^* = \{i^*\} \). Then \( \lim_{n \to \infty} P(\mathcal{M}^* = \hat{\mathcal{M}}_{1-n}^*) = 1 \).

**Proof.** This follows because \( i^* \) will be the last surviving model with probability one. So this model is never eliminated asymptotically.

### 2.1 MCS p-Values

Next, we introduce the MCS \( p \)-values. Let \( m_0 \) denote the number of elements in \( \mathcal{M}_0 \), and suppose for simplicity that the elements of \( \mathcal{M}_0 = \{1, \ldots, m_0\} \) are ordered such that \( k = e_{M_{(k)}} \) where \( \mathcal{M}_{(k)} = [k, k+1, \ldots, m_0], k = 1, \ldots, m_0 \). Hence \( e_{M_0} = e_{M_{(1)}} = 1 \) is the first model to be eliminated in the event that \( H_{0, \mathcal{M}_{(1)}} \) is rejected, \( e_{M_{(2)}} = 2 \) is the next model, etc.
**Model Confidence Sets**

**Definition 3 (MCS p-values)** Let \( p(k) \) be the p-value of the hypothesis \( H_{0,M(k)} \), with the convention: \( p(m_0) \equiv 1 \). The MCS p-value for model \( i \in \mathcal{M}_0 \) is defined by \( \hat{p}_i \equiv \max_{k \leq i} p(k) \).

The following table describes how the MCS p-values are defined and how they relate to the p-values of the individual tests, \( p(k), k = 1, \ldots, m_0 \).

<table>
<thead>
<tr>
<th>( e_i )</th>
<th>p-value of ( H_{0,M(k)} )</th>
<th>MCS p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( p(1) = 0.01 )</td>
<td>( \hat{p}_1 = 0.01 )</td>
</tr>
<tr>
<td>2</td>
<td>( p(2) = 0.04 )</td>
<td>( \hat{p}_2 = 0.04 )</td>
</tr>
<tr>
<td>3</td>
<td>( p(3) = 0.02 )</td>
<td>( \hat{p}_3 = 0.04 )</td>
</tr>
<tr>
<td>4</td>
<td>( p(4) = 0.03 )</td>
<td>( \hat{p}_4 = 0.04 )</td>
</tr>
<tr>
<td>5</td>
<td>( p(5) = 0.07 )</td>
<td>( \hat{p}_5 = 0.07 )</td>
</tr>
<tr>
<td>6</td>
<td>( p(6) = 0.07 )</td>
<td>( \hat{p}_6 = 0.07 )</td>
</tr>
<tr>
<td>7</td>
<td>( p(7) = 0.11 )</td>
<td>( \hat{p}_7 = 0.11 )</td>
</tr>
<tr>
<td>8</td>
<td>( p(8) = 0.25 )</td>
<td>( \hat{p}_8 = 0.25 )</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
</tr>
<tr>
<td>( m_0 )</td>
<td>( p(m_0) \equiv 1.00 )</td>
<td>( \hat{p}_{m_0} = 1.00 )</td>
</tr>
</tbody>
</table>

The MCS p-values are convenient because they make it easy to determine whether a particular object is in \( \mathcal{M}^*_\alpha \). Thus, the MCS p-values are an effective way of conveying the information in the data.

**Theorem 3** The MCS p-value, \( \hat{p}_i \), is such that \( i \in \mathcal{M}^*_\alpha \) if and only if \( \hat{p}_i \geq \alpha \), for any \( i \in \mathcal{M}_0 \).

**Proof.** Suppose that \( \hat{p}_i < \alpha \) and let \( i(k) \equiv i \). Since \( \hat{p}_i = \max_{j \geq k} p(j) \) it follows that the tests, \( \delta_{i(k)}, \ldots, \delta_{(m_0)} \), are all rejected at significance level \( \alpha \). Hence, the first accepted hypothesis (if any) occurs after \( i = e_{i(k)} \) has been eliminated. This proves that \( \hat{p}_i < \alpha \) implies \( i \notin \mathcal{M}^*_\alpha \). Suppose now that \( \hat{p}_i \geq \alpha \). Then for some \( k \geq i \) it holds that \( \hat{p}_k \geq \alpha \), such that \( H_{0,M(k)} \) is accepted at significance level \( \alpha \). Similarly we conclude that \( \hat{p}_i \geq \alpha \) implies that \( i \in \mathcal{M}^*_\alpha \), which completes the proof. \( \blacksquare \)

The interpretation of a MCS p-value is analogous to that of a classical p-value. So the MCS p-value cannot be interpreted as the probability that a particular model is the best model, analogous to the fact that a classical p-value does not give the probability that a particular hypothesis is true. The MCS is a random subset of models, that contains \( \mathcal{M}^* \) with a certain probability, and the probability interpretation of a MCS p-value is tied to the random nature of the MCS. The analogy to the classical setting is that of a \((1 - \alpha)\) confidence interval that contains the ‘true’ parameter with a probability no less than \(1 - \alpha\).
2.2 Equivalence Tests and Elimination Rules

Now we consider specific equivalence tests and an elimination rule that satisfy Assumption 1. We shall make the following assumption that is sufficiently strong, such that the tests can be implemented by bootstrap methods.

Assumption 2 For some \( r > 2 \) and \( \gamma > 0 \) it holds that \( E|d_{ij,t}|^{r+\gamma} < \infty \) for all \( i, j \in \mathcal{M}_0 \), and that \( \{d_{ij,t}\}_{i,j \in \mathcal{M}_0} \) is strictly stationary and \( \alpha \)-mixing of order \( -r/(r - 2) \).

It is worthwhile to note that \( \{L_{i,t}\} \) is not required to be stationary or ‘well-behaved’ as long as the relative performance variables that are used for the comparisons satisfy Assumption 2. Thus, the assumption allows for structural breaks and other features that may cause \( \{L_{i,t}\} \) to be non-stationary, as long as all objects of \( \mathcal{M}_0 \) are affected in a ‘similar’ way that preserves the stationarity of \( d_{ij,t} \).

Let \( \mathcal{M} \) be some subset of \( \mathcal{M}_0 \) and let \( m \) be the number of models in \( \mathcal{M} = \{i_1, \ldots, i_m\} \). We define the vector of loss-variables, \( L_t \equiv (L_{i_1,t}, \ldots, L_{i_m,t})' \), \( t = 1, \ldots, n \), and its sample average, \( \bar{L} \equiv n^{-1} \sum_{t=1}^{n} L_t \), and we let \( \iota \equiv (1, \ldots, 1)' \) be the column vector where all \( m \) entries equal one.

The orthogonal complement to \( \iota \), is an \( m \times (m - 1) \) matrix, \( \iota_\perp \), that has full column rank and satisfies \( \iota_\perp' \iota = 0 \) (an 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 4 Given Assumption 2, let \( X_t \equiv \iota_\perp' L_t \) and define \( \mu \equiv E(X_t) \). The null hypothesis \( H_{0,\mathcal{M}} \) is equivalent to \( \mu = 0 \) and it holds that \( n^{1/2}(\bar{X} - \mu) \xrightarrow{d} N(0, \Sigma) \), where \( \bar{X} \equiv n^{-1} \sum_{t=1}^{n} X_t \) and \( \Sigma \equiv \lim_{n \to \infty} \text{var}(n^{1/2} \bar{X}) \).

Proof. First note that \( X_t = \iota_\perp' L_t \) can be written as a linear combination of \( d_{ij,t}, i, j \in \mathcal{M}_0 \), since \( \iota_\perp' \iota = 0 \). Thus \( H_{0,\mathcal{M}} \) is given by \( \mu = 0 \), and the asymptotic normality follows by the central limit theorem for \( \alpha \)-mixing processes, see e.g. White (2000a).

Lemma 4 shows that \( H_{0,\mathcal{M}} \) can be tested using traditional quadratic-form tests, such as those that are based on the test statistics

\[
T_Q \equiv n \bar{X}' \hat{\Sigma}^* \bar{X} \quad \text{and} \quad T_F \equiv \frac{n - q}{q(n - 1)} T_Q.
\]
where \( \hat{\Sigma} \) is some consistent estimator of \( \Sigma, q \equiv \text{rank}(\hat{\Sigma}) \), and \( \hat{\Sigma}^\dagger \) denotes the Moore-Penrose inverse of \( \hat{\Sigma} \).\(^3\)\(^4\) Here \( q \) denotes the effective number of \textit{contrasts} (the number of linearly independent comparisons) under \( H_{0,M} \), and since \( \hat{\Sigma} \rightarrow \Sigma \) (by assumption) it follows that \( T_Q \overset{d}{\rightarrow} \chi^2_q \) and \( T_F \overset{d}{\rightarrow} F_{(q,n-q)} \), where \( \chi^2_q \) denotes the \( \chi^2 \)-distribution with \( q \) degrees of freedom and \( F_{(q,n-q)} \) is the \( F \)-distribution with \( (q, n-q) \) degrees of freedom (\( F_{q,\infty} = \chi^2_q/q \) in the limit). Under the alternative hypothesis, \( H_{A,M} \), \( T_Q \) and \( T_F \) diverge to infinity with probability one. Thus, the test \( \delta_M \) will meet the requirements of Assumption 1, when constructed from either of the statistics \( T_Q \) or \( T_F \).

An empirical problem arises when the number of elements, \( m \), become large relative to the sample size, \( n \). In this case, it is useful to consider alternative tests that do not require an estimate of the \((m-1) \times (m-1)\) covariance matrix, \( \Sigma \). Such tests can be constructed from the \( t \)-statistics

\[
t_{ij} = \frac{\hat{d}_{ij}}{\sqrt{\text{var}(\hat{d}_{ij})}} \quad \text{and} \quad t_i = \frac{\hat{d}_i}{\sqrt{\text{var}(\hat{d}_i)}} \quad \text{for } i, j \in \mathcal{M},
\]

where we have defined \( \hat{d}_{ij} \equiv n^{-1} \sum_{t=1}^n d_{ij,t} \) and \( \hat{d}_i \equiv m^{-1} \sum_{j \in \mathcal{M}} \hat{d}_{ij} \), and \( \text{var}(\hat{d}_{ij}) \) and \( \text{var}(\hat{d}_i) \) denote estimates of \( \text{var}(\tilde{d}_{ij}) \) and \( \text{var}(\tilde{d}_i) \) respectively. The variable \( \hat{d}_{ij} \) measures the sample loss differential between model \( i \) and \( j \), whereas \( \hat{d}_i \) is a contrast of model \( i \)'s sample loss to that of the average across all models. The latter can be seen from the identity \( \hat{d}_i = (\tilde{L}_i - \tilde{L}) \), where \( \tilde{L}_i \equiv n^{-1} \sum_{t=1}^n L_{i,t} \) and \( \tilde{L} \equiv m^{-1} \sum_{i \in \mathcal{M}} L_i \).

The null hypothesis, \( H_{0,M} \), is equivalent to \( E(\hat{d}_{ij}) = 0 \) for all \( i \in \mathcal{M} \), (and equivalent to \( E(\tilde{d}_{ij}) = 0 \) for all \( i, j \in \mathcal{M} \) by definition). Test statistics, such as \( T_D \equiv \sum_{i \in \mathcal{M}} t_{i}^2 \), \( T_R \equiv \max_{i,j \in \mathcal{M}} |t_{ij}| \), and \( T_{SQ} \equiv \sum_{i,j \in \mathcal{M}} t_{ij}^2 \), can be used to test the hypothesis \( H_{0,M} \). The subscripts refer to \textit{deviation} (from common average), \textit{range}, and \textit{semi-quadratic}, respectively. This paper focuses on the test statistic, \( T_D \), because it involves the fewest pairwise comparisons, \( m \), as opposed to the \( m(m-1) \) comparisons that \( T_R \) and \( T_{SQ} \) involve.\(^5\)

The asymptotic distributions of the test statistics, \( T_D \), \( T_R \), and \( T_{SQ} \), are non-standard because they depend on nuisance parameters (under both the null and the alternative). However, this poses no obstacle

---

\(^3\)Note that the matrix \( \iota_{\perp} \) is not fully identified (the space spanned by the columns of \( \iota_{\perp} \) is). However, this does not create any problems for the tests that are based on \( T_Q \) and \( T_F \), because these statistics are invariant to the choice for \( \iota_{\perp} \).

\(^4\)Under the additional assumption that \( \{d_{ij,t}\}_{i,j \in \mathcal{M}} \) is uncorrelated (across \( t \)), we can use \( \hat{\Sigma} = n^{-1} \sum_{t=1}^n (X_t - \tilde{X})(X_t - \tilde{X})' \), whereas in the case with autocorrelation one can use a robust estimator such as that of Newey and West (1987). The test based on \( T_Q \) in combination with (asymptotic) critical values from \( \chi^2_q \), was first used by West and Cho (1995).

\(^5\)The MCS procedure of the present paper has been applied to volatility models by Hansen, Lunde, and Nason (2003) who also provide some simulation results for the statistics \( T_R \) and \( T_{SQ} \).
as their distributions are easily estimated using bootstrap methods that implicitly solve the nuisance parameter problem. This feature of the bootstrap has previously been used in this context by Killian (1999), White (2000b), Hansen (2005), and Hansen (2003b), and Clark and McCracken (2003).

Besides an equivalence test, we need an elimination rule, $e_M$, that meets the requirement of Assumption 1. When the test statistic, $T_D$, is used, the natural elimination rule is $e_M \equiv \arg \max_i t_i$, because it removes the model that contributes most to the test statistic, $T_D$, among the models with a sample performance that is worse than the average across models. In fact, $e_M$ selects the object that has the largest standardized excess loss, relative to the average across all models in $\mathcal{M}$.

Next, we derive some intermediate results that are used to prove that the MCS, which is based on $T_D$ and $e_M = \arg \max_i t_i$, satisfies the necessary requirements of Assumption 1.

**Lemma 5** Suppose that Assumption 2 holds and define $\tilde{Z} = (\tilde{a}_1, \ldots, \tilde{a}_m)'$. Then

$$n^{1/2}(\tilde{Z} - \psi) \overset{d}{\rightarrow} N_m(0, \Omega), \quad \text{as } n \rightarrow \infty,$$

(2)

where $\psi \equiv E(\tilde{Z})$ and $\Omega \equiv \lim_{n \rightarrow \infty} \text{var}(n^{1/2}\tilde{Z})$. The null hypothesis, $H_{0,M}$, corresponds to $\psi = 0$.

**Proof.** From the identity $\tilde{a}_i = \tilde{L}_i - \tilde{L} = \tilde{L}_i - m^{-1} \sum_{j \in \mathcal{M}} \tilde{L}_j = m^{-1} \sum_{j \in \mathcal{M}} (\tilde{L}_i - \tilde{L}_j) = m^{-1} \sum_{j \in \mathcal{M}} \tilde{a}_{ij}$, we see that the elements of $\tilde{Z}$ are linear transformations of $\tilde{X}$ from Lemma 4. Thus for some $(m - 1) \times m$ matrix $G$ we have $\tilde{Z} = G' \tilde{X}$, and the result now follows, where $\psi = G' \mu$ and $\Omega = G' \Sigma G$. (The $m \times m$ covariance matrix, $\Omega$, has reduced rank, as rank($\Omega) \leq m - 1$.)

In the following, we let $\varrho$ denote the $m \times m$ correlation matrix that is implied by the covariance matrix, $\Omega$, of Lemma 5. Further, from a vector of random variables, $\xi \sim N_m(0, \varrho)$, we let $F_{\varrho}$ denote the distribution of $\xi'\xi$. Note that each of the elements of $\xi$ have a standard normal distribution, so that $F_{\varrho} = \chi^2_{(m)}$ for the special case where $\varrho = I_m$. This follows from the fact that $\xi_i^2 \sim \text{iid} \chi^2_{(1)}$ when $\varrho = I$ such that $\xi'\xi = \sum_{i=1}^m \xi_i^2 \sim \chi^2_{(m)} = F_{\varrho}$.

**Theorem 6** Let Assumption 2 hold and suppose that $\hat{\omega}_i^2 \equiv \text{var}(n^{1/2}\tilde{a}_i) = n \text{var}(\tilde{a}_i) \overset{p}{\rightarrow} \omega_i^2$, where $\omega_i^2$, $i = 1, \ldots, m$ are the diagonal elements of $\Omega$. Under $H_{0,M}$, it holds that $T_D \overset{d}{\rightarrow} F_{\varrho}$ and under the alternative hypothesis, $H_{A,M}$, it holds that $T_D \rightarrow \infty$ in probability.

**Proof.** Let $D \equiv \text{diag}(\omega_1^2, \ldots, \omega_m^2)$ and $\tilde{D} \equiv \text{diag}(\hat{\omega}_1^2, \ldots, \hat{\omega}_m^2)$. From Lemma 5 it follows that $\xi_n = (\xi_{1,n}, \ldots, \xi_{m,n})' \equiv D^{-1/2}n^{1/2}Z \overset{d}{\rightarrow} N_m(0, \varrho)$, since $\varrho = D^{-1/2}\Omega D^{-1/2}$. From $t_i = \tilde{a}_i/\sqrt{\text{var}(\tilde{a}_i)} = n^{1/2}\tilde{a}_i/\hat{\omega}_i = \xi_{i,n} \overset{w}{\rightarrow}$ it now follows that $T_D = \sum_{i \in \mathcal{M}} t_i^2 = \tilde{Z}' \tilde{D}^{-1} \tilde{Z} = \xi_n'(D \tilde{D}^{-1})\xi_n \overset{d}{\rightarrow} F_{\varrho}$, since
Model Confidence Sets

\[ D \tilde{D}^{-1} \xrightarrow{D} I_m \text{ and } \xi_n^d \xi_n^d \xrightarrow{d} F_\varrho. \] Under the alternative hypothesis there exist an \( j \in \mathcal{M} \), such that \( \tilde{d}_j \xrightarrow{p} c \neq 0 \). Thus, \( t_j^2 \) diverges at rate \( n^{1/2} \) and \( T_D \) at rate \( n \) in probability. ■

Theorem 6 shows that the asymptotic distribution of \( T_D \) depends on the correlation matrix, which makes \( \varrho \) a nuisance parameter in this testing problem. Nonetheless, as we have discussed earlier, we can solve this nuisance parameter problem by using bootstrap methods. Our bootstrap implementation produces a consistent estimate of \( T_D \)’s distribution for all values of \( \varrho \).

2.3 MCS for Forecasting Models

This subsection considers some issues that are relevant when the MCS procedure is applied to out-of-sample evaluation of forecasting models.

Parameter estimation can play an important role in the evaluation and comparison of forecasting models. Specifically, when nested models are being compared and the parameters estimated using certain estimation schemes, the limit distribution of our test statistic need not be Gaussian, see West and McCracken (1998) and Clark and McCracken (2001). The problem is that Assumption 2 does not hold in this instance. Some of these problems can be avoided by using a rolling window (of the sample) for the parameter estimation, which is the approach taken by Giacomini and White (2003). 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. E.g. modifications are needed in the case with nested models, see e.g. Chong and Hendry (1986), Harvey and Newbold (2000), Chao, Corradi, and Swanson (2001), and Clark and McCracken (2001). The key modification that is needed to accommodate the case with nested models, is to make a proper choice for \( \delta_M \). Given a proper choice for \( \delta_M \) and \( e_M \) the general (sequential) testing principle that is used to generate the MCS remains. However, in this paper we will not pursue this important generalization, because it would obscure our main objective, which is to lay out the key ideas of the MCS.

3 Relation to Some Existing Empirical Procedures

In the introduction, we discussed the relation between the MCS procedure and trace-test procedure that is used to select the number of cointegration relations, see Johansen (1988). The underlying testing
principle that both the MCS procedure and the trace-test procedure is based on, is known as intersection-union testing (IUT) that was formalized by Berger (1982). See also Pantula (1989) who applied IUT to select the lag-length and order of integration in univariate autoregressive processes.

Another way to cast the MCS problem is as a multiple comparisons problem. Problems of multiple comparisons have a long history in the statistics literature, see Gupta and Panchapakesan (1979), Hsu (1996), Dudoit, Shaffer, and Boldrick (2003) and Lehmann and Romano (2005, chapter 9) and references therein. Result from this literature has recently been adopted in the econometrics literature. One problem is that of multiple comparisons with best, where objects are compared to that with the ‘best’ sample performance. Statistical procedures for multiple comparisons with best are discussed and applied to economic applications by Horrace and Schmidt (2000). Another related problem is the case where the benchmark, to which all objects are compared, is selected independent of the data used for the comparison. This problem is known as multiple comparisons with control, and this is the testing problem that arises in the reality check for data snooping by White (2000b) and the test for superior predictive ability (SPA) by Hansen (2005).

The MCS has several advantages over the reality check of White (2000b) and the test for SPA of Hansen (2005). The reality check and the SPA tests are designed to address whether a particular forecasting model (e.g., a benchmark model) is significantly outperformed by any member of a competing set of models. Unlike these tests the MCS procedure does not require a benchmark to be specified, which is very useful in many applications without an objective benchmark. In the situation where there is a natural benchmark, the MCS procedure can still address the same objective as that of the SPA tests. This is done by observing whether the designated benchmark is in the MCS or not, where the latter correspond to a rejection of the null hypothesis that is tested by a SPA test.

The MCS procedure has the advantage that it can be used for model selection, whereas SPA tests are ill-suited for this problem. Since a rejection of the SPA test only identifies one model as significantly better, the SPA gives little guidance about which models reside in the set of ‘best’ models. 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 one-by-one comparisons. The

---

6Romano and Wolf (2003) improves upon the SPA test and are able to identify a set of models that significantly dominate the benchmark. However, these models may differ significantly in terms of their performance. Thus, this set has no direct relation to the MCS.
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 the individual SPA-tests need to be adjusted for the number of tests that are made, in order to control the family-wise Type I error rate. For example, if the level in each of the SPA-tests is $\alpha/m$ (rather than $\alpha$) the resulting set of ‘surviving’ benchmarks is a MCS with coverage $(1 - \alpha)$. However, there is a substantial loss of power associated with the use of a small level in each of the individual tests, and this highlights a major pitfall of sequential SPA-tests.

A second drawback of constructing a MCS from SPA-tests, is that the null hypothesis of a SPA-test is a composite hypothesis. The null is defined by several inequality constraints and the asymptotic distribution of the SPA test-statistic depends on the number of binding inequalities. This creates a nuisance parameter problem that makes it difficult to control the Type I error rate, and this results in an additional loss of power, see Hansen (2003a). In comparison, the MCS procedure is based on a sequence of hypotheses tests that only involve equalities, whereby composite hypothesis testing is avoided.

### 3.1 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 which 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 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 MCS, which is a random data-dependent set of models, and it is therefore meaningful to state that the best model can be found in the MCS with a certain probability. (The Bayesian notion of assigning probabilities to models is not meaningful in a frequentist setting.) Another advantage of our frequentist approach is that it does not require extraordinary computational demands, unlike the synthetic data creation that Bayesian Markov chain-Monte Carlo methods rely on. Thus, we argue our MCS procedure is
(i) econometrically meaningful because it selects the best model(s) with a certain probability and (ii) economically useful because it is straightforward to implement.

4 Simulation Experiments

This section reports on Monte Carlo experiments that show the MCS to be properly sized and possess good power. We consider two designs that are based on the \( m \)-dimensional vector,

\[
\mu = \frac{\lambda}{\sqrt{n}} (0, \frac{1}{m-1}, \ldots, \frac{1}{m-1}, 1)^t,
\]

that defines the relative performances, \( E(d_{ij,t}) = \mu_i - \mu_j \). 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 driven by

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

where \( \rho \) controls the degree of correlation between alternatives.

**Design I:** Define the (vector of) loss variables to be

\[
L_t \equiv \mu + \frac{a_t}{\sqrt{E(a_t^2)}} X_t, \quad \text{where } a_t = \exp(y_t), \ y_t = \frac{-\varphi}{2(1 + \varphi)} + \varphi y_{t-1} + \sqrt{\varphi} \epsilon_t,
\]

and \( \epsilon_t \sim iid \ N(0, 1) \). This implies that \( E(y_t) = -\varphi/[2(1 - \varphi^2)] \) and \( \text{var}(y_t) = \varphi/(1 - \varphi^2) \), such that \( E(a_t) = \exp[E(y_t) + \text{var}(y_t)/2] = \exp[0] = 1 \), and \( \text{var}(a_t) = (\exp[\varphi/(1 - \varphi^2)] - 1) \). Further \( E(a_t^2) = \text{var}(a_t) + 1 = \exp(\varphi/(1 - \varphi^2)) \) such that \( \text{var}(L_t) = 1 \). Note that \( \varphi = 0 \) corresponds to homoskedastic errors and \( \varphi > 0 \) corresponds to (GARCH-type) heteroskedastic errors.

The simulations employ 2500 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 \( B = 1000 \) resamples. The size of a synthetic sample is \( n = 250 \). This approximates the sample size on which many model selection exercises are based in macroeconomics.

We report two statistics from our simulation experiment based on \( \alpha = 10\% \). One is the frequency at which \( \hat{\mathcal{M}}_{90\%} \) contains \( \mathcal{M}^* \) and the other is the average number of models in \( \hat{\mathcal{M}}_{90\%} \). The former shows the ‘size’ properties of the MCS procedure and the latter is informative about the ‘power’ of the procedure.

[Table 1 about here]
Table 1 presents simulation results that show 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 \(E(d_{ij,t})\) increases (e.g., \(\lambda\) increases).

Note also that a larger correlation makes it easier to separate inferior models from superior models. This is not surprising because \(\text{var}(d_{ij,t}) = \text{var}(L_{it}) + \text{var}(L_{jt}) - 2\text{cov}(L_{it}, L_{jt}) = 2(1 - \rho)\), which is decreasing in \(\rho\). 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 \(\hat{\mathcal{M}}_{90\%}^*\) tends to fall as \(\varphi\) increases.

The consistency result of Corollary 2 applies if \(\lambda > 0\), because \(\mathcal{M}^*\) contains only one model in this case. According to Table 1, the MCS matches this theoretical prediction because \(\hat{\mathcal{M}}^* = \mathcal{M}^*\) in a large number of simulations, in particular when \(\lambda\) and \(\rho\) are large. This follows from the fact that the simulation experiments yield size and power statistics equal to (or nearly equal to) one. The former statistic implies \(\mathcal{M}^* \subset \hat{\mathcal{M}}^*\) (in all the synthetic samples), while the latter indicates that \(\hat{\mathcal{M}}^*\) is reduced to a single model in all the synthetic samples.

**Design II:** (MSE-type loss). In this design, we generate the individual loss variables

\[
L_{it} = \left(2^{-\frac{1}{4}} X_{it} + \sqrt{\mu_i}\right)^2, \quad i = 1, \ldots, m \quad \text{and} \quad t = 1, \ldots, n,
\]

such that \(E(L_{it}) = 1/\sqrt{2} + \mu_i\), and \(\text{var}(L_{it}) = 1 + 2\sqrt{2}\mu_i\). Thus, the variance of \(L_{it}\) increases in \(\mu_i\) in Design II which causes the worst performing models to be the most volatile models. From the expected loss it follows that \(E(d_{ij,t}) = \mu_i - \mu_j\) such that the two designs are identical in terms of expected relative loss.

The results for Design II are found in Table 2. The frequency at which the best models are contained in the MCS is somewhat smaller than it was the case for Design I. In fact, the estimated frequency is...

---

7 All the calculations made in this paper are based on software written by the authors using the Ox language of Doornik (2001).

8 This follows from the fact that \(\sqrt{2}L_{it} = \left(X_{it} + \frac{2}{\sqrt{2}} \sqrt{\mu_i}\right)^2\) has a non-central \(\chi^2\)-distribution with parameters \((1, \sqrt{2}\mu_i)\), since \(X_{it} \sim N(0, 1)\). Thus \(E(\sqrt{2}L_{it}) = 1 + \sqrt{2}\mu_i\) and \(\text{var}(\sqrt{2}L_{it}) = 2 + 4\sqrt{2}\mu_i\).
Model Confidence Sets

less than 90% in over half the cases when $\lambda = 0$. This suggests a minor small sample size distortion (the simulation results are based on $n = 250$). The small sample distortion is likely caused by the non-Gaussian distribution of $\hat{d}_{ij}$ in this design. Once again, the MCS becomes better at separating inferior from superior models as $E(d_{ij,i})$ increases (as $\lambda$ increases), and a larger correlation ($\rho$ rises) also adds power to the MCS procedure, as was the case in Design I.

**Design III:** This design has $L_t \sim iid \ N_{10}(\mu, \Sigma)$, where the covariance matrix has the following structure,

$$
\Sigma_{ij} = \rho^{|i-j|}, \quad \text{for } \rho = 0.00, 0.50, 0.75, \text{ and } 0.95.
$$

For the mean vector we use $\mu = (0, \ldots, 0, \frac{1}{10}, \ldots, \frac{1}{10})'$ where $m_0 = 10$, such that the number of zeros in $\mu$ defines the number of elements in $M^*$. We report results when $M^*$ consists of single model and five models. The objective of this design is to evaluate the performance of the $T_D$ test by comparing it to the quadratic $T_Q$ test, as the latter is an optimal test in an asymptotic single-hypothesis sense. In our implementation we use $\chi^2(m)$ critical values for evaluating $T_Q$.

The simulation results are presented in Figures 1 and 2. The left panels display the frequency at which $\hat{M}_{90\%}^*$ contains $M^*$ (‘size’) at various sample sizes. The right panels present the average number of models in $\hat{M}_{90\%}^*$ (‘power’). The different rows of panels correspond to different values of $\rho$. Figure 1 contains the results for the case where $M^*$ contains a single model. The left panels indicate that the best model is almost always contained in the MCS, which agrees with Corollary 2. The right panels show the ‘power’ of the $T_D$ and $T_Q$ tests. These tests are comparable for small values of $\rho$, but diverge as $\rho$ becomes larger. As $\rho$ rises, the quadratic test, $T_Q$, has the edge in terms of power. This power advantage disappears when $M^*$ contains five models, as can be seen from the right panels of Figure 2. In this case, the $T_D$ test has the edge because it tends not to over reject at small sample sizes, as the $T_Q$ test does. Figure 2 also reveals that the $T_D$ test dominates the $T_Q$ test in terms of size considerations. Thus, the inference we draw from the simulation results are quite encouraging for the $T_D$ test, especially because it can be invoked in the situation where $m_0$ is large.
5 Empirical Application to US Inflation Forecasts: Stock & Watson (JME, 1999) Revisited

This section revisits the Stock and Watson (1999) empirical application that pairwise compares a large number of inflation forecasting models, including several that have a Phillips curve type specification.

The Stock and Watson (1999) forecasting inflation-data set measure inflation, \( \pi_t \), as the CPI-U, all items \( (PUNEW) \) and the personal consumption expenditure implicit price deflator \( (GMDC) \). Their Phillips curve is

\[
\pi_{t+h} - \pi_t = \phi + \beta(L)u_t + \gamma(L)(1-L)\pi_t + e_{t+h}
\]

(3)

where \( u_t \) is the unemployment rate, \( L \) is the lag polynomial operator (e.g., \( Lu_t = u_{t-1} \)), and \( e_{t+h} \) is the long-horizon inflation forecast innovation. Note that the natural rate hypothesis is not imposed on this Phillips curve (3) and that inflation as a regressor variable is in its first difference. Besides the Phillips curve (3), Stock and Watson forecast inflation with a range of models, where unemployment is replaced with different macro variables that are labeled \( x_t \).\(^9\) The entire sample runs from 1959:M1 to 1997:M9. The first observation used in the regressions is 1960:M2, and the period over which simulated out-of-sample forecasts are computed and compared is 1970:M1 through 1996:M9.

We compute the MCS across all of the Stock and Watson inflation forecasting models. This includes the Phillips curve model and the models that run through all of the macro variables that Stock and Watson consider, a random walk model, and a univariate \( p \)-th order autoregressive model, AR\((p)\). Stock and Watson also present results with bivariate and multivariate forecast combinations and with indicator variables constructed using principal component decompositions. Our analysis employs their complete collection of models and variables.

Tables 3-4 consist (of the level) of the root mean square error (RMSE) and MCS \( p \)-values of the Stock and Watson forecast inflation models. The first column of tables 3-4 also lists the transformation of the macro variable \( x_t \) in the forecasting equation

\[
\pi_{t+h} - \pi_t = \phi + \beta(L)x_t + \gamma(L)(1-L)\pi_t + e_{t+h}.
\]

(4)


\(^9\)See Stock and Watson (1999) for details about their modeling strategy and data set.
and the substantial disinflation of the early 1980s. Inflation does not exhibit this behavior in the later subsample. Rather than an increase and then decrease in inflation, an important feature of the latter 1980s and first-half of the 1990s is a disinflation at the lower frequencies.

Our Table 3 matches table 2 of Stock and Watson (1999, pp. 303-304). The RMSEs and the p-values for the Phillips curve forecasting model (3) appear in the bottom row of our Table 3. The results for the random walk and AR(\(p\)) are the first two rows of the table, respectively. The rest of the rows of Table 3 are the ‘gap’ and ‘first difference’ specifications of Stock and Watson’s aggregate activity variables. There is a total of 18 models.

A glance at Table 3 reveals that the MCS of subsamples 1970:M1 - 1983:M12 and 1984:M1 - 1996:M9 are strikingly different. The MCS of the former subsample contains only five forecasting models for \(PUNEW\) and just one model for \(GMDC\) at the 90 percent level, \(\hat{\mathcal{M}}_{90\%}^*\). Only four of the 18 forecasting models fail to enter into \(\hat{\mathcal{M}}_{90\%}^*\) either for \(PUNEW\) or for \(GMDC\) based on the 1984:M1 - 1996:M9 subsample. Thus, the earlier sample possesses useful information to tell the forecast apart, whereas the later sample is less informative.

Another intriguing feature of Table 3 is the models that reside in the MCS of the 1970:M1 - 1983:M12 subsample. The five models that are in \(PUNEW-\hat{\mathcal{M}}_{90\%}^*\) are driven by macro variables related either to the labor market or to real economic activity. The labor market variables are \(lpng\), the first difference of employees on nonagricultural payrolls, and \(dlhur\), the first difference of the unemployment rate, all workers 16 years and older. Thus, there is labor market information that is important for predicting inflation. This is consistent with traditional Keynesian measures of aggregate demand; see Solow (1976).

Three specifications of forecasting equation (4) that are in \(PUNEW-\hat{\mathcal{M}}_{90\%}^*\) include three real quantity variables. These models employ the variables \(gmppyq\), real personal income, and \(msmtq\), real manufacturing and trade, total, are embraced by \(PUNEW-\hat{\mathcal{M}}_{90\%}^*\). The former variable is the only variable that is included in \(GMDC-\hat{\mathcal{M}}_{90\%}^*\). The only ‘gap’ specification that ends up in \(PUNEW-\hat{\mathcal{M}}_{90\%}^*\) is \(hsbp\), (the natural log of) building permits for new private housing starts. These variables can be construed as signals

---

10 In this paper we present tables that corresponds to tables 2 and 4 of Stock and Watson (1999). The tables with results that corresponds to Stock and Watson (1999, tables 3, 5 and 6) are available upon request.


12 Members of \(\hat{\mathcal{M}}_{1-\alpha}^*\) are listed by their MCS p-values being greater than or equal to \(\alpha\).
Model Confidence Sets

about the anticipated path either of real aggregate demand or real aggregate supply.

The last inference we draw from Table 3 is a rejection of the random walk forecasting model for \( {PUNEW} \) and \( {GMDC} \). This is contrary to Atkeson and Ohanian (2001). They report that Phillips curve models ‘cannot beat a random walk’, a result that is reminiscent of famous work on exchange rate forecasting by Meese and Rogoff (1983). We find strong evidence that the random walk inflation forecasts are inferior to other inflation forecasting model specifications. The MCS \( p \)-values for the random walk forecasting model are all very small (all are less than 0.015), which is consistent with table 3 of Stock and Watson (1999). Thus, we agree with Stock and Watson that the Phillips curve is a device that helps to forecast inflation.

[Table 4 about here]

Table 4 generates MCSs of inflation forecasting models using multivariate forecasting techniques, which replicates Table 4 of Stock and Watson (1999, pp. 318-319). They combine a large set of inflation forecasts from an array of 168 models using sample means, sample medians, and ridge estimation to produce these forecast weighting schemes. The other multivariate forecasting approach depends on principal components of the 168 macro-predictors. 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) argues 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.

Multivariate forecasting of inflation yields results consistent with those of our Table 3. The earlier subsample contains information that enables the MCS to distinguish between competing specifications, unlike the latter subsample. Table 4 shows that all specifications, but the random walk model, is covered by the MCS during the 1984:M1 - 1996:M9 subsamples. Thus, we continue to find that the random walk model forecasts poorly on the 1970:M1 - 1983:M12 and 1984:M1 - 1996:M9 subsamples, relative to other models. This is the case for both measures of inflation (\( PUNEW \) and \( GMDC \)), see of Table 4.

Only the multi-factor and one-factor specifications for ‘all indicators’ and ‘real activity indicators’ appear in the MCS of \( PUNEW \) at the 90 percent level in the 1970:M1 - 1983:M12 subsample. Table 4 shows that the MCS of \( GMDC \) is larger in this case, as the \( \hat{M}_{90\%} \) contains the entire collection of specifications for ‘all indicators’ and ‘real activity indicators’, as well as the combined-mean-forecast for ‘interest rates’. Since the multiple and one-factor specifications for ‘all indicator’ and ‘real activity
indicator’ appear in the MCSs across inflation measures and subsamples, we have further evidence that the Phillips curve is a useful tool for inflation forecasting.

6 Summary and Concluding Remarks

This paper introduces the model confidence sets procedure (MCS), relates it to other approaches of model selection and multiple comparisons, and establishes the asymptotic theory of the MCS. A simple and convenient bootstrap method for the implementation of the MCS procedure is also outlined in the paper. The paper employs our bootstrap method to generate Monte Carlo experiments that reveal the MCS procedure 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. This is a key feature of the MCS in practice. The MCS procedure operates on a metric that discards a model only if it is found to be significantly inferior to another model in the set being considered, given the finite sample at hand. Since a model remains in the MCS until proven inferior, all models in the MCS may not be judged good forecasting tools.\textsuperscript{13}

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 the MCS procedure yields a set of models that summarizes key sample information.

The paper applies the MCS procedure to the inflation forecasting problem of Stock and Watson (1999), as an empirical example. Results show that the MCS procedure provides a powerful tool for evaluating competing inflation forecasts. We agree with Stock and Watson that the Phillips curve yields good inflation forecasts, but we emphasize that the information content of the data matters for the inferences that can be drawn. The great inflation-disinflation subsample of 1970:M1 - 1983:M12 has movements in inflation and macro variables that allows the MCS procedure to make 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 con-

\textsuperscript{13}In future research, it would be interesting to study the proportion of models in \( \hat{\mathcal{M}}_{1-\alpha} \) that are members of \( \mathcal{M}^* \). This issue is related to the false discovery rate and the \( q \)-value theory of Storey (2002). See McCracken and Sapp (2005) for an application to comparisons of forecasting models.
Model Confidence Sets

sider. However, the random walk model of inflation never resides in the MCS in either the earlier or the later periods. Nonetheless, the question of what constitutes the best inflation forecasting model for the last 35 years of U.S. data remains unanswered. We pursue this task in future research.

The MCS has a wide variety of potential uses, beyond inflation forecasting. For example, Hansen, Lunde, and Nason (2003) have used the MCS procedure to determine the set of superior volatility models in an analysis of IBM return 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.

On the theoretical front there are important generalizations that need be addressed in future research. Of particular importance are the issues that arise when forecasts are based on estimated parameters. We leave this for future research.

A Bootstrap Procedure

This appendix describes the bootstrap implementation of the MCS procedure.

1. (Bootstrap indexes for resampling)

This is the first step because we need to use common random numbers for the bootstrap resamples in each iteration of the sequential test.

(a) Choose the block-length bootstrap parameter, $l$. The optimal choice for $l$ is tied to the persistence in $d_{i,t} = m^{-1} \sum_{j \in M_0} d_{i,t-j}$, $i = 1, \ldots, m$, which is difficult to estimate precisely when $m$ is large. Instead one can use different choices for $l$, and verify that the result is not sensitive to the choice.

(b) Generate $B$ bootstrap resamples of $\{1, \ldots, n\}$. I.e., for $b = 1, \ldots, B$:
   i. Choose $\xi_{b_1} \sim U\{1, \ldots, n\}$ and set $(\tau_{b,1}, \ldots, \tau_{b,l}) = (\xi_{b_1}, \xi_{b_1} + 1, \ldots, \xi_{b_1} + l - 1)$, with the convention $n + i = i$ for $i \geq 1$.
   ii. Choose $\xi_{b_2} \sim U\{1, \ldots, n\}$ and set $(\tau_{b,l+1}, \ldots, \tau_{b,2l}) = (\xi_{b_2}, \xi_{b_2} + 1, \ldots, \xi_{b_2} + l - 1)$.
   iii. Continue until a sample size of $n$, is constructed.
   iv. This is repeated for all resamples $b = 1, \ldots, B$, using independent draws of the $\xi$'s.

(c) Save the full matrix of bootstrap indexes.
Alternatively one can use a different bootstrap scheme, such as the stationary bootstrap of Politis and Romano (1994).

2. **(Sample and Bootstrap Statistics)**

(a) For each model and each point in time we evaluate the performance to obtain the variables $L_{i,t}$, for $i = 1, \ldots, m$, and $t = 1, \ldots, n$. These variables are used to calculate the sample averages for each model $\bar{L}_i = \frac{1}{n} \sum_{t=1}^{n} L_{i,t}, i = 1, \ldots, m$.

(b) The corresponding bootstrap variables are now given by

$$L^*_{b,i,t} = L_{i,t_b}, \quad \text{for} \quad b = 1, \ldots, B, \quad i = 1, \ldots, m, \quad \text{and} \quad t = 1, \ldots, n.$$ 

and calculate the bootstrap sample averages, $\bar{L}^*_b \equiv \frac{1}{n} \sum_{t=1}^{n} L^*_{b,i,t}$. The only variables that need to be stored are $\bar{L}_i$ and $\bar{ζ}^*_b \equiv \bar{L}^*_b - \bar{L}_i$, as all required statistics can be calculated from these two variables.

3. **(Sequential Testing)** Initialize by setting $\mathcal{M} = \mathcal{M}_0$.

(a) Let $m$ denote the number of elements in $\mathcal{M}$, and calculate

$$\bar{L}_i \equiv \frac{1}{m} \sum_{i=1}^{m} L_{i,t}, \quad \bar{ζ}^*_b \equiv \frac{1}{m} \sum_{i=1}^{m} ζ^*_b,i, \quad \text{and} \quad \hat{\text{var}}(\bar{d}_i) \equiv \frac{1}{B} \sum_{b=1}^{B} (ζ^*_b,i - ζ^*_b,·)^2.$$ 

Alternatively one can define $\hat{\text{var}}(·)$ to be its analytical value under the bootstrap scheme that is employed.

Now define $t_i \equiv \bar{d}_i / \sqrt{\hat{\text{var}}(\bar{d}_i)}$ and calculate the test statistic $T_D = \frac{1}{m} \sum_{i=1}^{m} t_i^2$.

(b) The bootstrap estimate of $T_D$’s distribution is given by empirical distribution of

$$T^*_{D,b} = \frac{1}{m} \sum_{i=1}^{m} t^2_{b,i}, \quad \text{for} \quad b = 1, \ldots, B,$$

where $t^*_{b,i} \equiv (ζ^*_b,i - ζ^*_b,·) / \sqrt{\hat{\text{var}}(\bar{d}_i)}$.

(c) The $p$-value of $H_{\mathcal{M},0}$ is given by

$$\hat{p}(m) \equiv \frac{1}{B} \sum_{b=1}^{B} 1_{\{T_D > T^*_{D,b}\}},$$

where $1_{\{\cdot\}}$ is the indicator function.
(d) If \( \hat{p}(m) < \alpha \), where \( \alpha \) is the level the test, then \( H_{\mathcal{M},0} \) is rejected and \( e_{\mathcal{M}} \equiv \arg \max_t t_i \) is eliminated from \( \mathcal{M} \).

(e) The steps in 3.(a)-(d) are repeated until first ‘acceptance’, and the resulting set of models is denoted \( \hat{\mathcal{M}}_{T_{-a}} \) and referred to as the \((1 - \alpha)\) MCS.

### A.1 Justification of bootstrap implementation

Let \( Z_t = (d_{1,t}, \ldots, d_{m,t})' \) then by Lemma 5 we have that \( n^{1/2}(\hat{Z} - \psi) \rightarrow^d N_m(0, \Omega) \), where \( \hat{Z} = \sum_{t=1}^n Z_t \). The bootstrap variables \( \{Z_{b,t}\} \) are generated such that \( n^{1/2}(\hat{Z}_b - \hat{Z}) \rightarrow^d N_m(0, \Omega) \), where the covariance matrix can be estimated by its analytical form under the bootstrap scheme, \( \hat{\Omega}_b \) say, where \( \hat{\Omega}_b \) is consistent for \( \Omega \) as \( n \rightarrow \infty \). Alternatively, \( \Omega \) can be estimated directly from the resamples by \( \hat{\Omega}_{n,B} \equiv n/B \sum_{b=1}^B (\hat{Z}_b - \hat{Z})(\hat{Z}_b - \hat{Z})' \), where \( \hat{\Omega}_{n,B} \rightarrow^p \hat{\Omega}_n \) as \( B \rightarrow \infty \) by the law of large numbers.

Our implementation is based on \( \hat{\Omega}_{n,B} \), and the identity

\[
\zeta_{b,i} - \zeta_{b,i} = \hat{L}_{b,i} - \hat{L}_i - \frac{1}{m} \sum_{i=1}^m (\hat{L}_{b,i} - \hat{L}_i) = (\hat{L}_{b,i} - \hat{L}_i) - (\hat{L}_i - \hat{L}) = \hat{d}_{b,i} - \bar{d}_i,
\]

that shows that the diagonal elements of \( \hat{\Omega} \) are given by

\[
n/B \sum_{b=1}^B (\hat{Z}_{b,i} - \hat{Z}_i)^2 = n/B \sum_{b=1}^B (\hat{d}_{b,i} - \bar{d}_i)^2 = n/B \sum_{b=1}^B (\zeta_{b,i} - \zeta_{b,i})^2 = \text{var}(n^{1/2}\bar{d}_i).
\]

Under the null hypothesis, the distribution of \( T_D \) is approximated by that of

\[
n^{1/2}(\hat{Z}_b - \hat{Z})' \hat{D} n^{1/2}(\hat{Z}_b - \hat{Z}) = (\hat{Z}_b - \hat{Z})' \text{diag}(\hat{\text{var}}(d_1), \ldots, \hat{\text{var}}(d_i)) (\hat{Z}_b - \hat{Z})
\]

\[
= \sum_{i=1}^m (\hat{d}_{b,i} - \bar{d}_i)^2 / \text{var}(d_i) = \sum_{i=1}^m (\zeta_{b,i} - \zeta_{b,i})^2 / \text{var}(d_i) = \sum_{i=1}^m (t_{b,i})^2
\]

\(= T_{b,D} \).

### References


Model Confidence Sets


## Model Confidence Sets

### Table 1: Simulation Design I.

<table>
<thead>
<tr>
<th></th>
<th>$m = 10$</th>
<th>$m = 40$</th>
<th>$m = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Panel A: $\varphi = 0$</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frequency at which $\mathcal{M}^+ \subset \hat{\mathcal{M}}_{95%}$ (size)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rho = 0$</td>
<td>0.884</td>
<td>0.904</td>
<td>0.886</td>
</tr>
<tr>
<td>$\lambda = 0$</td>
<td>0.993</td>
<td>0.994</td>
<td>0.995</td>
</tr>
<tr>
<td>$\lambda = 10$</td>
<td>0.994</td>
<td>0.997</td>
<td>1.000</td>
</tr>
<tr>
<td>$\lambda = 20$</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>$\lambda = 40$</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td><strong>Average number of elements in $\hat{\mathcal{M}}_{95%}$ (power)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda = 0$</td>
<td>9.086</td>
<td>9.834</td>
<td>9.816</td>
</tr>
<tr>
<td>$\lambda = 5$</td>
<td>5.936</td>
<td>4.284</td>
<td>3.088</td>
</tr>
<tr>
<td>$\lambda = 10$</td>
<td>3.089</td>
<td>2.224</td>
<td>1.663</td>
</tr>
<tr>
<td>$\lambda = 20$</td>
<td>1.650</td>
<td>1.280</td>
<td>1.064</td>
</tr>
<tr>
<td>$\lambda = 40$</td>
<td>1.074</td>
<td>1.004</td>
<td>1.000</td>
</tr>
<tr>
<td><strong>Panel B: $\varphi = 0.5$</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frequency at which $\mathcal{M}^+ \subset \hat{\mathcal{M}}_{95%}$ (size)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rho = 0$</td>
<td>0.927</td>
<td>0.920</td>
<td>0.906</td>
</tr>
<tr>
<td>$\lambda = 0$</td>
<td>0.990</td>
<td>0.993</td>
<td>0.997</td>
</tr>
<tr>
<td>$\lambda = 10$</td>
<td>0.996</td>
<td>0.999</td>
<td>1.000</td>
</tr>
<tr>
<td>$\lambda = 20$</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>$\lambda = 40$</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td><strong>Average number of elements in $\hat{\mathcal{M}}^+_{95%}$ (power)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda = 0$</td>
<td>9.868</td>
<td>9.889</td>
<td>9.856</td>
</tr>
<tr>
<td>$\lambda = 5$</td>
<td>5.814</td>
<td>4.162</td>
<td>2.948</td>
</tr>
<tr>
<td>$\lambda = 10$</td>
<td>2.958</td>
<td>2.195</td>
<td>1.614</td>
</tr>
<tr>
<td>$\lambda = 20$</td>
<td>1.650</td>
<td>1.290</td>
<td>1.075</td>
</tr>
<tr>
<td>$\lambda = 40$</td>
<td>1.067</td>
<td>1.011</td>
<td>1.001</td>
</tr>
<tr>
<td><strong>Panel C: $\varphi = 0.8$</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frequency at which $\mathcal{M}^+ \subset \hat{\mathcal{M}}_{95%}$ (size)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rho = 0$</td>
<td>0.955</td>
<td>0.962</td>
<td>0.953</td>
</tr>
<tr>
<td>$\lambda = 0$</td>
<td>0.996</td>
<td>0.999</td>
<td>1.000</td>
</tr>
<tr>
<td>$\lambda = 10$</td>
<td>0.999</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>$\lambda = 20$</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>$\lambda = 40$</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td><strong>Average number of elements in $\hat{\mathcal{M}}^+_{95%}$ (power)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda = 0$</td>
<td>9.934</td>
<td>9.940</td>
<td>9.931</td>
</tr>
<tr>
<td>$\lambda = 5$</td>
<td>4.259</td>
<td>3.148</td>
<td>2.315</td>
</tr>
<tr>
<td>$\lambda = 10$</td>
<td>2.330</td>
<td>1.741</td>
<td>1.414</td>
</tr>
<tr>
<td>$\lambda = 20$</td>
<td>1.389</td>
<td>1.198</td>
<td>1.075</td>
</tr>
<tr>
<td>$\lambda = 40$</td>
<td>1.076</td>
<td>1.022</td>
<td>1.009</td>
</tr>
</tbody>
</table>

The two statistics are the frequency at which $\hat{\mathcal{M}}_{95\%}$ contains $\mathcal{M}^+$ and the other is the average number of models in $\hat{\mathcal{M}}_{95\%}$. The former shows the ‘size’ properties of the MCS procedure and the latter is informative about the ‘power’ of the procedure.
Model Confidence Sets

Table 2: Simulation Design II.

<table>
<thead>
<tr>
<th>m</th>
<th>m = 10</th>
<th>m = 40</th>
<th>m = 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho = 0$</td>
<td>0.874</td>
<td>0.833</td>
<td>0.833</td>
</tr>
<tr>
<td>$\lambda = 5$</td>
<td>0.991</td>
<td>0.992</td>
<td>0.992</td>
</tr>
<tr>
<td>$\lambda = 10$</td>
<td>0.992</td>
<td>0.994</td>
<td>0.995</td>
</tr>
<tr>
<td>$\lambda = 20$</td>
<td>0.999</td>
<td>0.990</td>
<td>0.988</td>
</tr>
<tr>
<td>$\lambda = 40$</td>
<td>1.000</td>
<td>0.996</td>
<td>0.991</td>
</tr>
</tbody>
</table>

Panel A: $n = 250$

Frequency at which $M^* \subset \tilde{M}_{90\%}$ (size)

Average number of elements in $\tilde{M}_{90\%} (power)$

Panel B: $n = 1000$

Frequency at which $M^* \subset \tilde{M}_{90\%}$ (size)

Average number of elements in $\tilde{M}_{1-\alpha}$

The two statistics are the frequency at which $\tilde{M}_{90\%}$ contains $M^*$ and the other is the average number of models in $\tilde{M}_{90\%}$. The former shows the ‘size’ properties of the MCS procedure and the latter is informative about the ‘power’ of the procedure.
### Table 3: MCS p-values for Stock and Watson JME (1999) table 2.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No change</td>
<td></td>
<td>3.290</td>
<td>0.001</td>
<td>2.140</td>
<td>0.008</td>
<td>2.208</td>
<td>0.015</td>
<td>1.751</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>uniar</td>
<td>-</td>
<td>2.675</td>
<td>0.002</td>
<td>1.360</td>
<td>0.778*</td>
<td>1.941</td>
<td>0.033</td>
<td>1.082</td>
<td>0.413*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

'Gaps’ specifications

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>dip</td>
<td>DT</td>
<td>2.519</td>
<td>0.013</td>
<td>1.310</td>
<td>0.781*</td>
<td>1.913</td>
<td>0.074</td>
<td>1.043</td>
<td>0.549*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dsgmpyq</td>
<td>DT</td>
<td>2.644</td>
<td>0.001</td>
<td>1.446</td>
<td>0.101*</td>
<td>2.067</td>
<td>0.006</td>
<td>1.103</td>
<td>0.239*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dtsmsmtq</td>
<td>DT</td>
<td>2.341</td>
<td>0.089</td>
<td>1.280</td>
<td>0.848*</td>
<td>1.844</td>
<td>0.083</td>
<td>1.007</td>
<td>0.969*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dtlnpnag</td>
<td>DT</td>
<td>2.482</td>
<td>0.029</td>
<td>1.323</td>
<td>0.778*</td>
<td>2.024</td>
<td>0.020</td>
<td>1.012</td>
<td>0.969*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ipxmca</td>
<td>LV</td>
<td>2.373</td>
<td>0.066</td>
<td>1.264</td>
<td>1.000*</td>
<td>1.887</td>
<td>0.083</td>
<td>1.026</td>
<td>0.969*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hspb</td>
<td>LN</td>
<td>2.205</td>
<td>0.682*</td>
<td>1.392</td>
<td>0.663*</td>
<td>1.829</td>
<td>0.083</td>
<td>0.993</td>
<td>1.000*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lhm25</td>
<td>LV</td>
<td>2.433</td>
<td>0.052</td>
<td>1.401</td>
<td>0.402*</td>
<td>1.937</td>
<td>0.041</td>
<td>1.055</td>
<td>0.763*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

First difference specifications

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ip</td>
<td>DLN</td>
<td>2.384</td>
<td>0.060</td>
<td>1.429</td>
<td>0.244*</td>
<td>1.819</td>
<td>0.083</td>
<td>1.115</td>
<td>0.064</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gmpyq</td>
<td>DLN</td>
<td>2.233</td>
<td>0.653*</td>
<td>1.532</td>
<td>0.039</td>
<td>1.565</td>
<td>1.000*</td>
<td>1.149</td>
<td>0.129*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>msmtq</td>
<td>DLN</td>
<td>2.169</td>
<td>1.000*</td>
<td>1.353</td>
<td>0.774*</td>
<td>1.778</td>
<td>0.083</td>
<td>1.062</td>
<td>0.491*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lnpag</td>
<td>DLN</td>
<td>2.308</td>
<td>0.124*</td>
<td>1.317</td>
<td>0.781*</td>
<td>1.809</td>
<td>0.083</td>
<td>1.099</td>
<td>0.969*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dipxmca</td>
<td>DLV</td>
<td>2.355</td>
<td>0.066</td>
<td>1.456</td>
<td>0.068</td>
<td>1.839</td>
<td>0.083</td>
<td>1.128</td>
<td>0.035</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dhspb</td>
<td>DLN</td>
<td>2.701</td>
<td>0.004</td>
<td>1.405</td>
<td>0.496*</td>
<td>1.969</td>
<td>0.021</td>
<td>1.077</td>
<td>0.450*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dlhmu25</td>
<td>DLV</td>
<td>2.352</td>
<td>0.080</td>
<td>1.474</td>
<td>0.026</td>
<td>1.878</td>
<td>0.083</td>
<td>1.103</td>
<td>0.095</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dlhur</td>
<td>DLV</td>
<td>2.321</td>
<td>0.153*</td>
<td>1.451</td>
<td>0.139*</td>
<td>1.843</td>
<td>0.083</td>
<td>1.088</td>
<td>0.316*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Phillips curve

<table>
<thead>
<tr>
<th>Variable</th>
<th>Trans</th>
<th>1970-1983</th>
<th>RMSE</th>
<th>$p_{MCS}$</th>
<th>1984-1996</th>
<th>RMSE</th>
<th>$p_{MCS}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>LHUR</td>
<td></td>
<td>2.387</td>
<td>0.060</td>
<td>1.371</td>
<td>0.582*</td>
<td>1.939</td>
<td>0.059</td>
</tr>
</tbody>
</table>

The Table report the RMSE’s and the MCS p-values for the different forecasting models for US inflation. The p-values that are marked with an * are those in $\hat{M}_{\alpha_{y/9}}$. The results of this table correspond to those of S&W (1999, table 2).
Table 4: MCS $p$-values for Stock and Watson JME (1999) Table 4.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE  $p_{MCS}$</td>
<td>RMSE  $p_{MCS}$</td>
<td>RMSE  $p_{MCS}$</td>
<td>RMSE  $p_{MCS}$</td>
</tr>
<tr>
<td>No change</td>
<td>3.290 0.001</td>
<td>2.140 0.020</td>
<td>2.208 0.013</td>
<td>1.751 0.004</td>
</tr>
<tr>
<td>Univariate</td>
<td>2.675 0.001</td>
<td>1.360 0.617*</td>
<td>1.941 0.025</td>
<td>1.082 0.199*</td>
</tr>
</tbody>
</table>

**Panel A. All indicators**

| Mul. factors | 2.158 0.222* | 1.291 0.998* | 1.894 0.132* | 0.964 0.631* |
| 1 factor     | 2.069 0.610* | 1.274 1.000* | 1.692 1.000* | 1.002 0.602* |
| Comb. mean   | 2.439 0.002  | 1.289 0.999* | 1.853 0.162* | 1.036 0.543* |
| Comb. median | 2.550 0.002  | 1.316 0.981* | 1.895 0.115* | 1.063 0.442* |
| Comb. ridge reg. | 2.209 0.023 | 1.280 0.999* | 1.842 0.193* | 1.019 0.543* |

**Panel B. Real activity indicators**

| Mul. factors | 2.019 1.000* | 1.357 0.775* | 1.792 0.202* | 0.946 1.000* |
| 1 factor     | 2.079 0.610* | 1.281 0.999* | 1.753 0.234* | 1.017 0.543* |
| Comb. mean   | 2.346 0.004  | 1.284 0.999* | 1.807 0.202* | 1.020 0.543* |
| Comb. median | 2.381 0.002  | 1.299 0.994* | 1.831 0.193* | 1.036 0.506* |
| Comb. ridge reg. | 2.192 0.084 | 1.298 0.994* | 1.773 0.234* | 1.022 0.543* |

**Panel C. Interest rates**

| Mul. factors | 2.585 0.002  | 1.495 0.205* | 1.976 0.036  | 1.173 0.069 |
| 1 factor     | 2.524 0.004  | 1.495 0.117* | 2.038 0.007  | 1.077 0.356* |
| Comb. mean   | 2.424 0.008  | 1.341 0.883* | 1.900 0.100* | 1.079 0.184* |
| Comb. median | 2.513 0.002  | 1.336 0.941* | 1.912 0.055  | 1.078 0.300* |
| Comb. ridge reg. | 2.432 0.008 | 1.368 0.454* | 1.943 0.029  | 1.123 0.106* |

**Panel D. Money**

| Mul. factors | 2.679 0.001  | 1.360 0.462* | 1.933 0.032  | 1.080 0.218* |
| 1 factor     | 2.679 0.001  | 1.360 0.544* | 1.933 0.047  | 1.080 0.254* |
| Comb. mean   | 2.664 0.001  | 1.350 0.700* | 1.964 0.020  | 1.066 0.486* |
| Comb. median | 2.670 0.001  | 1.348 0.789* | 1.954 0.021  | 1.070 0.409* |
| Comb. ridge reg. | 2.638 0.001 | 1.385 0.390* | 1.934 0.074  | 1.121 0.151* |

Phillips curve

| LHUR        | 2.387 0.004  | 1.371 0.437* | 1.939 0.060  | 1.050 0.543* |

The Table report the RMSE’s and the MCS $p$-values for the different forecasting models for US inflation. The $p$-values that are marked with an * are those in $\hat{\mathcal{M}}_{90\%}$. The results of this table correspond to those of S&W (1999, table 4).
Model Confidence Sets

Figure 1: Design III with one good model. The left panels report the frequency at which $\mathcal{M}^*$ is contained in $\hat{\mathcal{M}}^*_{90\%}$ (size properties) and the right panels report the average number of models in $\hat{\mathcal{M}}^*_{90\%}$ (power properties). The rows correspond to different values of the correlation parameter $\rho$. 

\[ \rho = 0 \]
\[ \rho = 0.5 \]
\[ \rho = 0.75 \]
\[ \rho = 0.95 \]
Figure 2: Design III with five good model. The left panels report the frequency at which $\mathcal{M}^*$ is contained in $\hat{\mathcal{M}}^*_{90\%}$ (size properties) and the right panels report the average number of models in $\hat{\mathcal{M}}^*_{90\%}$ (power properties). The rows corresponds to different values of the correlation parameter $\rho$. 