Model Confidence Sets for Forecasting Models*

Peter Reinhard Hansen
Brown University
Department of Economics, Box B
Email: Peter_Hansen@brown.edu

Asger Lunde
The Aarhus School of Business
Department of Information Science
Email: alunde@asb.dk

James M. Nason
Federal Reserve Bank of Atlanta
Research Department
Email: jim.nason@atl.frb.org

Preliminary Version: March, 2004

Abstract

The paper introduces the *model confidence set* (MCS) and applies it to the selection of forecasting models. A MCS is a set of models that is constructed such that it will contain the 'best' forecasting model, given a level of confidence. Thus, a MCS is analogous to a confidence interval for a parameter. The MCS acknowledges the limitations of the data, such that uninformative data yields a MCS with many models, whereas informative data yields a 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. Although the MCS contains only a few models in the first subsample, there is little information in the second post-1984 subsample, which results in a large MCS. Yet, the random walk forecast is not contained in the MCS for either of the samples. This shows that the random walk forecast is inferior to principal component-based inflation forecasts.

JEL Classification: C12, C19, C44, C52, and C53.

Keywords: Model Confidence Set, Forecasting, Model Selection, Multiple Comparisons.

*The authors thank seminar participants at UCSD and CalTech for valuable comments. The first two authors are grateful to Financial support from the Danish Research Agency, grant no. 24-00-0363, and the first author wishes to thank the Federal Reserve Bank of Atlanta for its support and hospitality during his visit. The views in this paper should not be attributed either to the Federal Reserve Bank of Atlanta, the Federal Reserve System, or any of its Staff.

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 model – 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 $\mathcal{M}^*$ that consists of the 'best' model(s) from a collection of models, $\mathcal{M}_0$, where 'best' is defined in terms of some criterion that is user-specified. The MCS procedure yields a model confidence set, $\widehat{\mathcal{M}}^*$, which is a set of models that is constructed such that it will contain the best models with a given level of confidence. The MCS is constructed from 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 contains 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 (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 their sample performance that is specific to the criterion function that is employed, and the 'best' model is unlikely to be the same 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, α, is employed in all tests, which asymptotically guarantees that $P(\mathcal{M}^* \subset \widehat{\mathcal{M}}^*_{1-\alpha}) \geq 1-\alpha$, and in the case where $\mathcal{M}^*$ consists of one object we have the stronger results that $\lim_{n\to\infty} P(\mathcal{M}^* = \widehat{\mathcal{M}}^*_{1-\alpha}) = 1$. 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 $i \in \widehat{\mathcal{M}}^*_{1-\alpha}$, 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 that involves a sequential use of trace-tests is commonly used to select the number of cointegration relations within the vector autoregressive model, see Johansen (1988). The way that the MCS procedure determines the number of superior models is analogous to the way that 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 are based on simple t-statistics, because these 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 (2003) compare selection procedures that are based on information criteria and out-of-sample evaluation. Granger, King, and White (1995) argues 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 particularly 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 a critical views on this issue, see Clements (2002) and Inoue and Kilian (2002).

1.3 Bootstrap Implementation and Simulation Results

We propose a bootstrap implementation of 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 though simulation experiments. The results are very encouraging as 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 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 reasonable 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) present evidence this is not the case because 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 first subsample covers a period with large changes in the rate of inflation, this sample is expected to be relatively informative about which model might be the best forecasting models. Indeed, the MCS consists only of a few models, so the MCS proves to be effective at weeding out the inferior forecasts. The second subsample is a period with low and relatively stable inflation, and this sample contains relatively little information about which of the forecasting models that might be the best forecasts. In spite of the relatively low degree of information, 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) 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, $\mathcal{M}_0$, that contains a finite number of objects (forecasting models) that are indexed by $i = 1, \ldots, m$. 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 model i in period t as $L_{i,t}$.

We define the relative performance variables

$$d_{ij,t} \equiv L_{i,t} - L_{j,t}, \qquad \text{for all} \qquad i, j \in \mathcal{M}_0,$$

and make the following assumption.

Assumption 1 $\{d_{ij,t}\}_{i,j \in \mathcal{M}_0}$ *is strictly stationary and* $E|d_{ij,t}| < \infty$ *for all* $i, j \in \mathcal{M}_0$.

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 satisfies Assumption 1. Thus, the assumption allows for structural breaks and other aspects 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 $d_{ij,t}$.

Definition 1 *Let Assumption 1 hold. The set of superior objects is defined by*

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

In the following we let $\mathcal{M}^\dagger$ denote the complement to $\mathcal{M}^*$, i.e. $\mathcal{M}^\dagger \equiv \{i \in \mathcal{M}_0 : E(d_{ij,t}) > 0$ 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:

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

where $\mathcal{M} \subset \mathcal{M}_0$. We denote the alternative hypothesis ($E(d_{ij,t}) \neq 0$ 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 both $\mathcal{M}^*$ and $\mathcal{M}^\dagger$.

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 $\mathcal{M} \subset \mathcal{M}_0$,[1] 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.

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' we define the $\widehat{\mathcal{M}}^*_{1-\alpha} = \mathcal{M}$, otherwise we use $e_{\mathcal{M}}$ to eliminate objects 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 that is presented below shows that the term 'confidence set' is appropriate in this context, provided that the equivalence test and the elimination rule satisfies the following assumption.

Assumption 2 *For any $\mathcal{M} \subset \mathcal{M}_0$ we assume the following about $(\delta_{\mathcal{M}}, e_{\mathcal{M}})$: (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) $\lim_{n\to\infty} P(e_{\mathcal{M}} \in \mathcal{M}^* | H_{A,\mathcal{M}}) = 0$.*

Assumption 2 is standard. (a) requires the asymptotic level not to exceed α; (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 2, it holds that (i) $\lim_{n\to\infty} P(\mathcal{M}^* \subset \widehat{\mathcal{M}}^*_{1-\alpha}) \geq 1-\alpha$, and (ii) $\lim_{n\to\infty} P(i^\dagger \in \widehat{\mathcal{M}}^*_{1-\alpha}) = 0$ for all $i^\dagger \in \mathcal{M}^\dagger$.*

Proof. To prove (i) we consider the event that $i^* \in \mathcal{M}^*$ is eliminated from $\mathcal{M}$. From Assumption 2.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 $n \to \infty$, and Assumption 2.a shows that $P(\delta_{\mathcal{M}} = 1, e_{\mathcal{M}} = i^* | H_{0,\mathcal{M}}) = P(\delta_{\mathcal{M}} = 1 | H_{0,\mathcal{M}}) \leq \alpha$. To prove (ii),

[1] $\delta_{\mathcal{M}} = 0$ and $\delta_{\mathcal{M}} = 1$ correspond to the cases where $H_{0,\mathcal{M}}$ are 'accepted' and 'rejected' respectively.

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 $\mathcal{M} \cap \mathcal{M}^\dagger \neq \emptyset$. On the other hand, Assumption 2.*b* ensures that models will be eliminated as long as the null hypothesis is false. ■

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 $\widehat{\mathcal{M}}^*_{1-\alpha}$ is determined after the first 'acceptance'. Thus, a single null hypothesis is accepted, such that the familywise (Type I) error rate is bounded by the level of the test that was accepted! On the other hand, the MCS procedure exploits the fact that the power converges to unity as the sample size increases, such that all inferior models are (eventually) eliminated. In the event that the test lacks power the MCS procedure may result in a Type II error, in the sense that the MCS is 'too large' and will contain models from $\mathcal{M}^\dagger$. 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.

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

Corollary 2 *Suppose that Assumption 2 holds and that $\mathcal{M}^*$ is a singleton, $\mathcal{M}^* = \{i^*\}$. Then* $\lim_{n\to\infty} P(\mathcal{M}^* = \widehat{\mathcal{M}}^*_{1-\alpha}) = 1$.

Proof. This follows because i^* will be the last surviving model with probability one. Thus, 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 order, for simplicity, the elements $\mathcal{M}_0 = \{i_{(1)}, \ldots, i_{(m_0)}\}$ such that $i_{(k)} = e_{\mathcal{M}}$ for $\mathcal{M} = \{i_{(1)}, \ldots, i_{(k)}\}$. Thus, $i_{(m_0)} = e_{\mathcal{M}_0}$ is the first model to be eliminated in the event that $\delta_{\mathcal{M}_0}$ is rejected, $i_{(m_0-1)}$ the next model, etc.

Definition 3 (MCS *p*-values) *Let $p(k)$ be the p-value of $\delta_{\mathcal{M}_{(k)}}$ where $\mathcal{M}_{(k)} = \{i_{(1)}, \ldots, i_{(k)}\}$ for $k = 2, \ldots, m_0$, and use the convention $p(1) \equiv 1$. The MCS p-value of $j = i_{(j)} \in \mathcal{M}_0$ is defined by $\hat{p}_j \equiv \max_{k \geq i_{(j)}} p(k)$.*

The MCS p-values are convenient because they make it easy to determine whether a particular object is in $\widehat{\mathcal{M}}^*_{1-\alpha}$.

Theorem 3 *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 $i \in \mathcal{M}_0$.*

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, $\delta_{(k)}, \ldots, \delta_{(m_0)}$, are all rejected at significance level α. Hence, the first accepted hypothesis (if any) occurs after $i = e_{i_{(k)}}$ has been eliminated. This proves that $\hat{p}_i < \alpha$ implies $i \notin \widehat{\mathcal{M}}^*_{1-\alpha}$. Suppose now that $\hat{p}_i \geq \alpha$. Then for some $k \geq i$ it holds that $\hat{p}_k \geq \alpha$, such that $H_{0,\mathcal{M}_{(k)}}$ is accepted at significance level α. Similarly we conclude that $\hat{p}_i \geq \alpha$ implies that $i \in \widehat{\mathcal{M}}^*_{1-\alpha}$, which completes the proof. ■

2.2 Equivalence Tests and Elimination Rules

Now we consider specific equivalence tests and an elimination rule that satisfy Assumption 2. We shall make the following assumption that is sufficiently strong, such that the tests can be implemented by bootstrap methods.

Assumption 3 *For some $r > 2$ and $\delta > 0$ it holds that $E|d_{ij,t}|^{r+\delta} < \infty$ for all $i, j \in \mathcal{M}_0$, and that $\{d_{ij,t}\}_{i,j \in \mathcal{M}_0}$ is α-mixing of order $-r/(r-2)$.*

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 ι, 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 1, 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 given Assumption 3 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 $\Sigma \equiv \lim_{n \to \infty} \mathrm{var}(n^{1/2} \bar{X})$.*

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 α-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

$$T_Q \equiv n\bar{X}'\hat{\Sigma}^{\#}\bar{X} \qquad \text{and} \qquad T_F \equiv \frac{n-q}{q(n-1)}T_Q,$$

where $\hat{\Sigma}$ is some consistent estimator of Σ, $q \equiv \text{rank}(\hat{\Sigma})$, and $\hat{\Sigma}^{\#}$ denotes the Moore-Penrose inverse of $\hat{\Sigma}$.[2, 3] Here q denotes the effective number of *contrasts* under $H_{0,\mathcal{M}}$, and since $\hat{\Sigma} \xrightarrow{p} \Sigma$ (by assumption) it follows that $T_Q \xrightarrow{d} \chi^2_{(q)}$ and $T_F \xrightarrow{d} F_{(q,n-q)}$, where $\chi^2_{(q)}$ denotes the χ^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, $H_{A,\mathcal{M}}$, T_Q and T_F diverge to infinity with probability one. Thus, the test $\delta_{\mathcal{M}}$ will meet the requirements of Assumption 2, 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, Σ. Such tests can be constructed from the t-statistics

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

where we have defined $\bar{d}_{ij} \equiv n^{-1}\sum_{t=1}^{n} d_{ij,t}$ and $\bar{d}_{i\cdot} \equiv m^{-1}\sum_{j\in\mathcal{M}} \bar{d}_{ij}$, and $\widehat{\text{var}}(\bar{d}_{ij})$ and $\widehat{\text{var}}(\bar{d}_{i\cdot})$ denote estimates of $\text{var}(\bar{d}_{ij})$ and $\text{var}(\bar{d}_{i\cdot})$ respectively. The variable $\bar{d}_{ij}$ measures the sample loss differential between model i and j, whereas $\bar{d}_{i\cdot}$ 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 $\bar{d}_{i\cdot} = (\bar{L}_i - \bar{L}_\cdot)$, where $\bar{L}_i \equiv n^{-1}\sum_{t=1}^{n} L_{i,t}$ and $\bar{L}_\cdot \equiv m^{-1}\sum_{i\in\mathcal{M}} \bar{L}_i$.

The null hypothesis, $H_{0,\mathcal{M}}$, is equivalent to $E(\bar{d}_{i\cdot}) = 0$ for all $i \in \mathcal{M}$, (and equivalent to $E(\bar{d}_{ij}) = 0$ for all $i,j \in \mathcal{M}$ by definition). Test statistics, such as $T_D \equiv \sum_{i\in\mathcal{M}} t_{i\cdot}^2$, $T_R \equiv \max_{i,j\in\mathcal{M}} |t_{ij}|$, and $T_{SQ} = \sum_{i,j\in\mathcal{M}} t_{ij}^2$, can be used to test the hypothesis $H_{0,\mathcal{M}}$. The subscripts refer to *deviation* (from common average), *range*, and *semi-quadratic*, respectively. This paper

[2] 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 $\iota_\perp$.

[3] Under the additional assumption that $\{d_{ij,t}\}_{i,j\in\mathcal{M}}$ is uncorrelated (across t), we can use $\hat{\Sigma} = n^{-1}\sum_{t=1}^{n}(X_t - \bar{X})(X_t - \bar{X})'$, 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).

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.[4]

The asymptotic distributions of the test statistics, T_D, T_R, and T_{SQ}, are non-standard because they depend on nuisance parameters. However, this poses no obstacle 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 (2001), and Hansen (2003b), and Clark and McCracken (2003).

Besides an equivalence test, we need an elimination rule, $e_{\mathcal{M}}$, that meets the requirement of Assumption 2. When the test statistic, T_D, is used, the natural elimination rule is $e_{\mathcal{M}} \equiv \arg\max_i t_{i\cdot}$, 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, $e_{\mathcal{M}}$ 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 $e_{\mathcal{M}} = \arg\max_i t_{i\cdot}$, satisfies the necessary requirements of Assumption 2.

Lemma 5 *Suppose that Assumptions 1 and 3 hold and define* $\bar{Z} = (\bar{d}_{1\cdot}, \ldots, \bar{d}_{m\cdot})'$. *Then*

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

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

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

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

[4] 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}.

Theorem 6 *Suppose that Assumptions 1 and 3 hold, and let* $\hat{\omega}_i^2 \equiv \widehat{\text{var}}(n^{1/2}\bar{d}_{i\cdot}) = n\,\widehat{\text{var}}(\bar{d}_{i\cdot}) \xrightarrow{p} \omega_i^2$*, where* ω_i^2*,* $i = 1, \ldots, m$ *are the diagonal elements of* Ω*. Under* $H_{0,\mathcal{M}}$*, it holds that* $T_D \xrightarrow{d} F_\varrho$ *and under the alternative hypothesis,* $H_{A,\mathcal{M}}$*, it holds that* $T_D \to \infty$ *in probability.*

Proof. Let $D \equiv \text{diag}(\omega_1^2, \ldots, \omega_m^2)$ and $\hat{D} \equiv \text{diag}(\hat{\omega}_1^2, \ldots, \hat{\omega}_m^2)$. From Lemma 5 it follows that $\xi_n = (\xi_{1,n}, \ldots, \xi_{m,n})' \equiv D^{-1/2} n^{1/2} \bar{Z} \xrightarrow{d} N_m(0, \varrho)$, since $\varrho = D^{-1/2}\Omega D^{-1/2}$. From $t_{i\cdot} = \bar{d}_{i\cdot}/\sqrt{\widehat{\text{var}}(\bar{d}_{i\cdot})} = n^{1/2}\bar{d}_{i\cdot}/\hat{\omega}_i = \xi_{in}\frac{\omega_i}{\hat{\omega}_i}$ it now follows that $T_D = \sum_{i\in\mathcal{M}} t_{i\cdot}^2 = \bar{Z}'\hat{D}^{-1}\bar{Z} = \xi_n'(D\hat{D}^{-1})\xi_n \xrightarrow{d} F_\varrho$, since $D\hat{D}^{-1} \xrightarrow{p} I_m$ and $\xi_n'\xi_n \xrightarrow{d} F_\varrho$. Under the alternative hypothesis there exist an $j \in \mathcal{M}$, such that $\bar{d}_{j\cdot} \xrightarrow{p} c \neq 0$. Thus, $t_{j\cdot}^2$ 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 ϱ is 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 ϱ.

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 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.

2.3 MCS for Forecasting Models

In this subsection we consider 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 Assumptions 1 and 3 do 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), 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 $\delta_{\mathcal{M}}$. Given a proper choice for $\delta_{\mathcal{M}}$ and $e_{\mathcal{M}}$ the general (sequential) testing principle that is used to generate the MCS remains. However, in this paper we will not pursue this important generalization further.

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) 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 (2001).

The MCS has several advantages over the tests for superior predictive ability (SPA), see White (2000b) (reality check) and Hansen (2001). The tests for SPA are designed to address whether a particular forecasting model (the benchmark) is significantly outperformed by any model from a competing set of models. Whereas a SPA-test requires a benchmark to be specified,

the MCS procedure is 'benchmark free'. This is particular convenient in applications without a natural 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 the SPA tests are ill-suited for this problem. When the SPA test is rejected, there is little guidance about which set of models that are the possible best models, because the SPA test only identifies one model as significantly better.[5] We are 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 may also apply to several other models. The repeated use of SPA-tests where all models are used as the benchmark one-by-one is not a valid statistical procedure, unless the individual tests are adjusted for the number of tests that are made. The MCS procedure does not suffer from such problems because it implicitly controls the familywise error rate.

Finally, the null hypothesis when testing for SPA is a composite hypothesis that are defined from several inequality constraints. Since it is unclear how many of these inequalities bind under the null hypothesis, it becomes difficult to control the Type I error rate. For this reason, a SPA test can be quite conservative and have low power, see Hansen (2003a). In comparison, the MCS procedure is based on a sequence of hypotheses tests that only involves equalities, which avoids the composite hypothesis testing problem.

3.1 Bayesian Interpretation

The MCS procedure is entirely based on frequentist principles, but resembles some Bayesian procedures. By specifying a prior over the models in $\mathcal{M}_0$, a Bayesian procedure can derive posterior probabilities for each of the models. These can be used to construct a Bayesian confidence set by including the models with the largest posteriors until the posteriors add up to at least $1-\alpha$. Because MCS relies entirely on sample information, it bypasses the need to place priors on the models and their parameters. Note that the Bayesian notion of assigning probabilities to models is not meaningful in a frequentist setting. Instead, our probability statement is associated with

[5] Romano and Wolf (2003) impoves upon the SPA test and are able to identify a set of models that significantly dominate the benchmark. However, these models may be significantly different in terms of their performance. Thus this set has no direct relation to the MCS.

the MCS, which is a random data-dependent set of models. Therefore, it is meaningful to say that the best model can be found in the MCS with a certain probability.

4 Simulation Experiments

We consider two designs that are based on the m-dimensional vector,

$$\mu = \frac{\lambda}{\sqrt{n}}(0, \tfrac{1}{m-1}, \ldots, \tfrac{m-2}{m-1}, 1)',$$

that defines the relative performances, as we ensure that $E(d_{ij,t}) = \mu_i - \mu_j$. So our simulations are such that $\mathcal{M}^*$ consists of a single element, unless $\lambda = 0$, in which case we have $\mathcal{M}^* = \mathcal{M}_0$. The covariance structure is primarily defined by

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

Design I: In this design we define the (vector of) loss variables to be

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

and $\varepsilon_t \sim iid\, N(0,1)$. 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 $\text{var}(a_t) = \left(\exp[\varphi/(1-\varphi^2)] - 1\right)$. 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.

For our simulations we select $\lambda = 0, 5, 10, 20$, $\rho = 0.00, 0.50, 0.75, 0.95$, $\varphi = 0.0, 0.5, 0.8$, $m = 10, 40, 100$ with $2,500$ repetitions. We use the block-bootstrap using blocks with length $l = 2$ and our results are based on $B = 1,000$ resamples. Finally, we use $n = 250$ as the sample size, because this is in the order of magnitude that is common in empirical studies of macro economic variables.

We report two statistics from our simulation experiment. One is 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. So our simulation results are based on $\alpha = 10\%$.

[Table 1 about here]

The results reported in Table 1 show that the properties of the MCS procedure are as could be expected.[6] The frequency that the best models are contained in the MCS is virtually 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 (as λ increases). Further we also note that a strong correlation makes it easier to separate inferior models from superior model. 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 ρ. Thus a stronger correlation (holding the individual variances fixed) is associated with more information that makes is easier to separate good from bad models. Finally, the effect of heteroskedasticity are relatively small, but the heteroskedasticity does appear to add power to the MCS procedure, as the average number of models in $\widehat{\mathcal{M}}^*_{90\%}$ tends to fall as φ is increased.

When $\lambda > 0$ we have a situation where $\mathcal{M}^*$ only contains one model. So the consistency result of Corollary 2 applies in this case, and we do indeed observe that $\widehat{\mathcal{M}}^* = \mathcal{M}^*$ in a large number or simulations. For example when both of our statistics equal one, it shows that $\widehat{\mathcal{M}}^* = \mathcal{M}^*$ is all our simulations for that particular configuration.

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

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

such that $E(L_{it}) = 1/\sqrt{2} + \mu_i$, and $\text{var}(L_{it}) = 1 + 2\sqrt{2}\mu_i$. So like in Design I we have that $E(d_{ij,t}) = \mu_i - \mu_j$, whereas the variance of L_{it} is not increasing in μ_i, making the worst performing models the most volatile models.

We simulate this design with the same configurations as those for Design I, except that the parameter, φ, is not used in this design.

[Table 2 about here]

The results for Design II are reported in Table 2. Here 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 less than 90% in some cases where $\lambda = 0$, which indicates a minor small-sample size distortion (the simulation result are based on $n = 250$). The small sample distortion is likely caused by the non-Gaussian distribution that $\bar{d}_{ij}$ has in this design. Again, we see that MCS becomes better at separating the inferior models from the superior model as

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

$E(d_{ij,t})$ increases (as λ increases), and an increased correlation also adds power to the MCS procedure, as was the case in Design I.

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, π_t, as the CPI-U, all items ($PUNEW$) and the personal consumption expenditure implicit price deflator ($GMDC$). Their Phillips curve is

$$\pi_{t+h} - \pi_t \quad = \quad \phi + \beta(\mathbf{L})u_t + \gamma(\mathbf{L})(1-\mathbf{L})\pi_t + e_{t+h} \tag{3}$$

where u_t is the unemployment rate, $\mathbf{L}$ is the lag polynomial operator (e.g., $\mathbf{L}u_t = u_{t-1}$), 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.[7] 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 pth-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

[7] See Stock and Watson (1999) for details about their modeling strategy and data set.

transformation of the macro variable x_t in forecasting equation

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

Stock and Watson study the properties of their inflation forecasting models on the subsamples 1970:M1 - 1983:M12 and 1983:M1 - 1996:M9. The former subsample contains the great inflation of the 1970s 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.

[Table 3 about here]

Our Table 3 matches table 2 of Stock and Watson (1999, pp. 303-304).[8] 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.[9] 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, $\widehat{\mathcal{M}}^*_{90\%}$.[10] Only four of the 18 forecasting models *fail* to enter into $\widehat{\mathcal{M}}^*_{90\%}$ 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*-$\widehat{\mathcal{M}}^*_{90\%}$ are driven by macro variables related either to the labor market or to real economic activity. The labor market variables are *lpnag*, 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).

[8] 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.

[9] The "gap" is a one-sided Hodrick and Prescott (1997) filter of the relevant variable. See Stock and Watson (1999, p. 301) for details.

[10] Members of $\mathcal{M}^*_{1-\alpha}$ are listed by their MCS p-values being greater than or equal to α.

Three specifications of forecasting equation (4) that are in $PUNEW$-$\widehat{\mathcal{M}}^{*}_{90\%}$ include three real quantity variables. These models employ the variables *gmpyq*, real personal income, and *msmtq*, real manufacturing and trade, total, are embraced by $PUNEW$-$\widehat{\mathcal{M}}^{*}_{90\%}$. The former variable is the only variable that is included in $GMDC$-$\widehat{\mathcal{M}}^{*}_{90\%}$. The only "gap" specification that ends up in $PUNEW$-$\widehat{\mathcal{M}}^{*}_{90\%}$ is *hsbp*, (the natural log of) building permits for new private housing starts. These variables can be construed as signals 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 the Phillips curve models "cannot beat a random walk", a result that is reminiscent of famous work by Meese and Rogoff (1983). We find conclusive 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 forecast 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 $\widehat{\mathcal{M}}^*_{90\%}$ 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. However, our results suggest that a Phillips curve (4) specification tied to macro indicators other than unemployment yield better out-of-sample forecasts.

6 Summary and Concluding Remarks

In this paper, we have introduced the model confidence sets procedure (MCS). We discussed the relation of the MCS to other approaches to model selection and multiple comparisons, and we have established the asymptotic theory for the MCS. Further, we outlined a simple and convenient bootstrap method for the implementation of the MCS procedure and have presented Monte Carlo experiments that revealed good small sample properties of the MCS procedure.

As an empirical illustration of the MCS procedure, we applied the MCS to the forecasts of inflation of Stock and Watson (1999). Our results showed 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, however we also emphasized that the information content of the data matters for the conclusions 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 1984:M1 - 1996:M9 subsample is limited in comparison because the MCS procedure lets in almost any model that Stock and Watson consider. The upshot is that 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.

In applications of the MCS procedure it is important to understand the principle of the procedure. The MCS is constructed such that inference about the 'best' follows the conventional meaning of the word 'significance'. While the MCS will only contain the best model(s) asymp-

totically, it may contain several poor models in finite samples, and this aspect should be kept in mind 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. Thus *a model remains in the MCS until proven inferior*. This is the reason that some models in the MCS may not be good for forecasting, in much the same way that someone who is not convicted in court need not be innocent.[11]

The MCS has a wide variety of uses. For example, Hansen, Lunde, and Nason (2003) have used the MCS procedure to select the best volatility models. We apply the MCS to the problem of choosing the best forecasting models. Our empirical example employs the MCS procedure to revisit the Stock and Watson (1999) Phillips curve forecasting exercise. Our results reveal the MCS points to Phillips curve models in our search for the extracts the best set of forecasts of inflation. 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. Given the large number of forecasting problems economists face at central banks and other parts of government, in finance the markets, and in the academic setting, the MCS procedure faces a rich set of problems to study.

A Bootstrap Procedure

The bootstrap implementation.

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 $d_{i\cdot,t} = m^{-1} \sum_{j \in \mathcal{M}_0} d_{ij,t}$, $i = 1, \ldots, m$, 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 $b = 1, \ldots, B$:

[11] In future research, it would be interesting study the proportion of models in $\widehat{\mathcal{M}}^*_{1-\alpha}$ 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 (2003) for an application to comparisons of forecasting models.

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 ξ's.

(c) Save the full matrix of bootstrap indexes.

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 $\bar{L}_{i,\cdot} \equiv \frac{1}{n}\sum_{t=1}^{n} L_{i,t}$, $i = 1,\ldots,m$.

(b) The corresponding bootstrap variables are now given by

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

and calculate the bootstrap sample averages, $\bar{L}^*_{b,i} \equiv \frac{1}{n}\sum_{t=1}^{n} L^*_{b,i,t}$. The only variables that need to be stored are $\bar{L}_i$ and $\bar{\zeta}^*_{b,i} \equiv \bar{L}^*_{b,i} - \bar{L}_i$, 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

$$\bar{L}_{\cdot} \equiv \frac{1}{m}\sum_{i=1}^{m} \bar{L}_i, \qquad \zeta^*_{b,\cdot} = \frac{1}{m}\sum_{i=1}^{m} \zeta^*_{b,i}, \qquad \text{and} \qquad \widehat{\text{var}}(\bar{d}_{i\cdot}) \equiv \frac{1}{B}\sum_{b=1}^{B} (\zeta^*_{b,i} - \zeta^*_{b,\cdot})^2.$$

Now define $t_{i\cdot} \equiv \bar{d}_{i\cdot}/\sqrt{\widehat{\text{var}}(\bar{d}_{i\cdot})}$ and calculate the test statistic $T_D = \frac{1}{m}\sum_{i=1}^{m} t^2_{i\cdot}$.

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

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

where $t^*_{b,i\cdot} \equiv (\zeta^*_{b,i} - \zeta^*_{b,\cdot})/\sqrt{\widehat{\text{var}}(\bar{d}_{i\cdot})}$.

(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 α is the level the test, then $H_{\mathcal{M},0}$ is rejected and $e_{\mathcal{M}} \equiv \arg\max_i t_{i\cdot}$ is eliminated from $\mathcal{M}$.

(e) The steps in 3.(a)-(e) are repeated until first 'acceptance', and the resulting set of models is denoted $\widehat{\mathcal{M}}^*_{1-\alpha}$ and and referred to as the $(1-\alpha)$ MCS.

A.1 Justification of bootstrap implementation

Let $Z_t = (d_{1\cdot,t}, \ldots, d_{m\cdot,t})'$ then by Lemma 5 we have that $n^{1/2}(\bar{Z} - \psi) \xrightarrow{d} N_m(0, \Omega)$, where $\bar{Z} = \sum_{t=1}^n Z_t$. Let $Z^*_{b,t}$, $t = 1, \ldots, n$ and $b = 1, \ldots, B$ be generated by the stationary bootstrap of Politis and Romano (1994). Since Z_t has the properties of Assumption 3, it follows from Goncalves and de Jong (2003) that $n^{1/2}(\bar{Z}^*_b - \bar{Z}) \xrightarrow{d} N_m(0, \Omega)$ and $\hat{\Omega} \equiv n/B \sum_{b=1}^B (\bar{Z}^*_b - \bar{Z})(\bar{Z}^*_b - \bar{Z})' \xrightarrow{p} \Omega$. Now it follows that

$$\zeta^*_{b,i} - \zeta^*_{b,\cdot} = \bar{L}^*_{b,i} - \bar{L}_i - \frac{1}{m}\sum_{i=1}^m (\bar{L}^*_{b,i} - \bar{L}_i) = (\bar{L}^*_{b,i} - \bar{L}^*_{b,\cdot}) - (\bar{L}_i - \bar{L}_\cdot) = \bar{d}^*_{b,i\cdot} - \bar{d}_{i\cdot},$$

such that the diagonal elements of $\hat{\Omega}$ are given by

$$n/B \sum_{b=1}^B (\bar{Z}^*_{b,i} - \bar{Z}_i)^2 = n/B \sum_{b=1}^B (\bar{d}^*_{b,i\cdot} - \bar{d}_{i\cdot})^2 = \frac{n}{B}\sum_{b=1}^B (\zeta^*_{b,i} - \zeta^*_{b,\cdot})^2 = \widehat{\mathrm{var}}(n^{1/2}\bar{d}_{i\cdot}).$$

So under the null hypothesis the distribution of T_D is approximated by that of

$$\begin{aligned} n^{1/2}(\bar{Z}^*_b - \bar{Z})'\hat{D}n^{1/2}(\bar{Z}^*_b - \bar{Z}) &= (\bar{Z}^*_b - \bar{Z})'\mathrm{diag}(\widehat{\mathrm{var}}(\bar{d}_{1\cdot}), \ldots, \widehat{\mathrm{var}}(\bar{d}_{1\cdot}))(\bar{Z}^*_b - \bar{Z}) \\ &= \sum_{i=1}^m \frac{(\bar{d}^*_{b,i\cdot} - \bar{d}_{i\cdot})^2}{\widehat{\mathrm{var}}(\bar{d}_{i\cdot})} = \sum_{i=1}^m \frac{(\zeta^*_{b,i} - \zeta^*_{b,\cdot})^2}{\widehat{\mathrm{var}}(\bar{d}_{i\cdot})} = \sum_{i=1}^m (t^*_{b,i\cdot})^2 = T^*_{b,D}. \end{aligned}$$

References

Anderson, T. W. (1984): *An Introduction to Multivariate Statistical Analysis*. John Wiley and Sons, New York, 2nd edn.

Atkeson, A., and L. E. Ohanian (2001): "Are Phillips Curves Useful for Forecasting Inflation?," *Federal Reserve Bank of Minneapolis Quarterly Review*, 25.

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

Chong, Y. Y., and D. F. Hendry (1986): "Econometric Evaluation of Linear Macroeconomic Models," *Review of Economic Studies*, 53, 671–690.

Clark, T. E., and M. W. McCracken (2001): "Tests of Equal Forecast Accuracy and Encompassing for Nested Models," *Journal of Econometrics*, 105, 85–110.

——— (2003): "Evaluating Long Horizon Forecasts," Manuscript. Federal Reserve Bank of Kansas City.

Clements, M. P. (2002): "Why Forecast Performance Does Not Help Us Choose a Model," Mimeo. University of Warwick.

Corradi, V., and N. R. Swanson (2001): "Out-of-Sample Tests for Granger Causality," *Macroeconomic Dynamics*, 5, 598–620.

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

Doornik, J. A. (2001): *Ox: An Object-Orientated Matrix Programming Langguage*. Timberlake Consultants Ltd., London, 4 edn.

Engle, R. F., and S. J. Brown (1985): "Model Selection for Forecasting," *Journal of Computation in Statistics*.

Giacomini, R., and H. White (2003): "Tests of Conditional Predictive Ability," Boston College Working Paper 572.

Goncalves, S., and R. de Jong (2003): "Consistency of the Stationary Bootstrap under Weak Moment Conditions," *Economics Letters*, 81, 273–278.

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

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.

Gupta, S. S., and S. Panchapakesan (1979): *Multiple Decision Procedures*. John Wiley & Sons, New York.

Hansen, P. R. (2001): "A Test for Superior Predictive Ability," Brown University, Economics Working Paper. 2001-06
`http://www.econ.brown.edu/fac/Peter_Hansen`.

——— (2003a): "Asymptotic Tests of Composite Hypotheses," Brown University, Department of Economics Working Paper, 2003-09.
`http://www.econ.brown.edu/fac/Peter_Hansen`.

——— (2003b): "Regression Analysis with Many Specifications: A Bootstrap Method to Robust Inference," Mimeo
`http://www.econ.brown.edu/fac/Peter_Hansen`.

Hansen, P. R., A. Lunde, and J. M. Nason (2003): "Choosing the Best Volatility Models: The Model Confidence Set Approach," *Oxford Bulletin of Economics and Statistics*, 65, 839–861.

Harvey, D., and P. Newbold (2000): "Tests for Multiple Forecast Encompassing," *Journal of Applied Econometrics*, 15, 471–482.

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.

Horrace, W. C., and P. Schmidt (2000): "Multiple Comparisons with the Best, with Economic Applications," *Journal of Applied Econometrics*, 15, 1–26.

Hsu, J. C. (1996): *Multiple Comparisons*. Chapman & Hall/CRC, Boca Ranton, Florida.

Inoue, A., and L. Kilian (2002): "In-Sample or Out-of-Sample Tests of Predictability: Which One Should We Use?," *Manuscript*.

——— (2003): "On the Selection of Forecasting Models," European Central Bank Working Paper no. 214.

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

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

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.

McCracken, M. W., and S. Sapp (2003): "Evaluating the Predictability of Exchange Rates Using Long Horizon Regressions: Mind Your p's and q's!," Manuscript. University of Missouri-Columbia.

Meese, R., and K. Rogoff (1983): "Exchange Rate Models of the Seventies. Do They Fit Outof Sample?," *Journal of International Economics*, 14, 3–24.

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

Pantula, S. G. (1989): "Testing for Unit Roots in Time Series Data," *Econometric Theory*, 5, 256–71.

Politis, D. N., and J. P. Romano (1994): "The Stationary Bootstrap," *Journal of the American Statistical Association*, 89, 1303–1313.

ROMANO, J. P., AND M. WOLF (2003): "Stepwise Multiple Testing as Formalized Data Snooping," Mimeo.

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

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

STAIGER, D., J. H. STOCK, AND M. W. WATSON (1997): "The NAIRU, Unemployment, and Monetary Policy," *Journal of Economic Perspectives*, 11.

STOCK, J. H., AND M. W. WATSON (1999): "Forecasting Inflation," *Journal of Monetary Economics*, 44, 293–335.

——— (2002a): "Forecasting Using Principal Components From a Large Number of Predictors," *Journal of the American Statistical Association*, 97, 1167–1179.

——— (2002b): "Macroeconomic Forecasting Using Diffusion Indexes," *Journal of Business and Economic Statistics*, 20, 147–162.

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

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

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

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

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

WHITE, H. (2000a): *Asymptotic Theory for Econometricians*. Academic Press, San Diego, revised edn.

——— (2000b): "A Reality Check for Data Snooping," *Econometrica*, 68, 1097–1126.

Table 1: Simulation Design I.

	$m = 10$				$m = 40$				$m = 100$			
Panel A: $\varphi = 0$												
Frequency at which $\mathcal{M}^* \subset \widehat{\mathcal{M}}^*_{90\%}$ (size)												
$\rho =$	0	0.5	0.75	0.95	0	0.5	0.75	0.95	0	0.5	0.75	0.95
$\lambda = 0$	0.884	0.904	0.886	0.894	0.906	0.888	0.901	0.901	0.923	0.925	0.907	0.923
$\lambda = 5$	0.993	0.994	0.995	0.999	0.997	0.997	0.993	0.991	0.998	0.996	0.997	0.993
$\lambda = 10$	0.994	0.997	1.000	1.000	0.993	0.994	0.993	0.996	0.997	0.998	0.994	0.991
$\lambda = 20$	1.000	1.000	1.000	1.000	0.991	0.994	0.992	0.999	0.992	0.994	0.993	0.993
$\lambda = 40$	1.000	1.000	1.000	1.000	0.994	0.996	0.999	1.000	0.991	0.991	0.995	0.998
Average number of elements in $\widehat{\mathcal{M}}^*_{90\%}$ (power)												
$\lambda = 0$	9.806	9.834	9.816	9.817	39.84	39.81	39.84	39.83	99.87	99.86	99.83	99.86
$\lambda = 5$	5.936	4.284	3.088	1.530	24.51	17.62	12.60	5.767	59.26	42.42	30.70	14.36
$\lambda = 10$	3.089	2.224	1.663	1.031	12.63	9.019	6.501	2.962	30.72	22.08	15.72	7.278
$\lambda = 20$	1.650	1.280	1.064	1.000	6.506	4.651	3.305	1.629	15.88	11.49	8.195	3.744
$\lambda = 40$	1.074	1.004	1.000	1.000	3.291	2.412	1.784	1.054	8.112	5.796	4.216	1.970
Panel B: $\varphi = 0.5$												
Frequency at which $\mathcal{M}^* \subset \widehat{\mathcal{M}}^*_{90\%}$ (size)												
$\rho =$	0	0.5	0.75	0.95	0	0.5	0.75	0.95	0	0.5	0.75	0.95
$\lambda = 0$	0.927	0.920	0.906	0.927	0.957	0.960	0.962	0.964	0.992	0.989	0.992	0.990
$\lambda = 5$	0.990	0.993	0.997	1.000	0.999	0.995	0.995	0.991	0.998	0.999	0.999	0.996
$\lambda = 10$	0.996	0.999	1.000	1.000	0.994	0.994	0.994	0.996	0.999	0.996	0.995	0.994
$\lambda = 20$	1.000	1.000	1.000	1.000	0.994	0.995	0.993	1.000	0.997	0.995	0.992	0.992
$\lambda = 40$	1.000	1.000	1.000	1.000	0.997	0.997	0.999	1.000	0.995	0.994	0.995	1.000
Average number of elements in $\widehat{\mathcal{M}}^*_{90\%}$ (power)												
$\lambda = 0$	9.886	9.889	9.856	9.886	39.94	39.93	39.94	39.95	99.99	99.99	99.99	99.98
$\lambda = 5$	5.814	4.162	2.948	1.500	24.56	17.72	12.63	5.774	61.77	44.24	31.46	14.32
$\lambda = 10$	2.958	2.195	1.614	1.052	12.63	8.975	6.416	2.931	31.61	22.29	15.98	7.213
$\lambda = 20$	1.650	1.290	1.075	1.000	6.432	4.583	3.212	1.578	16.03	11.33	8.079	3.605
$\lambda = 40$	1.067	1.011	1.001	1.000	3.272	2.321	1.730	1.064	7.994	5.696	4.066	1.926
Panel C: $\varphi = 0.8$												
Frequency at which $\mathcal{M}^* \subset \widehat{\mathcal{M}}^*_{90\%}$ (size)												
$\rho =$	0	0.5	0.75	0.95	0	0.5	0.75	0.95	0	0.5	0.75	0.95
$\lambda = 0$	0.955	0.962	0.953	0.967	0.994	0.994	0.995	0.994	1.000	1.000	1.000	1.000
$\lambda = 5$	0.996	0.999	1.000	1.000	0.998	0.996	0.997	0.997	1.000	0.999	0.999	0.998
$\lambda = 10$	0.999	1.000	1.000	1.000	0.997	0.996	0.997	1.000	0.998	0.997	0.997	0.995
$\lambda = 20$	1.000	1.000	1.000	1.000	0.995	0.998	0.998	1.000	0.996	0.996	0.997	0.999
$\lambda = 40$	1.000	1.000	1.000	1.000	0.999	1.000	1.000	1.000	0.996	0.998	0.998	1.000
Average number of elements in $\widehat{\mathcal{M}}^*_{90\%}$ (power)												
$\lambda = 0$	9.934	9.940	9.931	9.956	39.99	39.99	39.99	39.99	100.0	100.0	100.00	100.0
$\lambda = 5$	4.259	3.148	2.315	1.306	18.94	13.72	9.959	4.441	48.15	35.33	25.14	11.62
$\lambda = 10$	2.330	1.741	1.414	1.058	9.850	6.975	4.944	2.269	25.66	17.87	12.75	5.614
$\lambda = 20$	1.389	1.198	1.078	1.014	4.914	3.535	2.504	1.373	12.56	8.992	6.422	2.833
$\lambda = 40$	1.076	1.022	1.009	1.003	2.511	1.870	1.450	1.081	6.459	4.309	3.122	1.558

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.

Table 2: Simulation Design II.

	$m = 10$				$m = 40$				$m = 100$			
Panel A: $n = 250$												
Frequency at which $\mathcal{M}^* \subset \widehat{\mathcal{M}}^*_{90\%}$ (size)												
$\rho =$	0	0.5	0.75	0.95	0	0.5	0.75	0.95	0	0.5	0.75	0.95
$\lambda = 0$	0.874	0.880	0.882	0.888	0.833	0.879	0.914	0.927	0.833	0.905	0.945	0.957
$\lambda = 5$	0.991	0.993	0.994	0.996	0.992	0.993	0.996	0.994	0.992	0.998	0.995	0.998
$\lambda = 10$	0.992	0.995	0.998	1.000	0.994	0.994	0.991	0.996	0.995	0.997	0.998	0.994
$\lambda = 20$	0.999	0.999	1.000	1.000	0.990	0.990	0.995	0.996	0.988	0.994	0.993	0.996
$\lambda = 40$	1.000	1.000	1.000	1.000	0.996	0.996	0.997	1.000	0.991	0.987	0.990	0.994
Average number of elements in $\widehat{\mathcal{M}}^*_{90\%}$ (power)												
$\lambda = 0$	9.738	9.770	9.789	9.809	39.58	39.73	39.82	39.87	99.49	99.75	99.87	99.93
$\lambda = 5$	6.530	5.873	4.661	2.750	25.58	23.20	18.43	10.17	61.07	55.93	44.34	24.65
$\lambda = 10$	3.517	3.072	2.483	1.604	13.83	12.12	9.665	5.314	32.51	29.04	23.24	12.72
$\lambda = 20$	1.846	1.659	1.416	1.110	7.293	6.376	5.080	2.940	17.02	15.10	12.03	6.629
$\lambda = 40$	1.143	1.094	1.032	1.002	3.794	3.280	2.624	1.674	9.125	7.863	6.332	3.527
Panel B: $n = 1000$												
Frequency at which $\mathcal{M}^* \subset \widehat{\mathcal{M}}^*_{90\%}$ (size)												
$\rho =$	0	0.5	0.75	0.95	0	0.5	0.75	0.95	0	0.5	0.75	0.95
$\lambda = 0$	0.888	0.891	0.892	0.906	0.894	0.894	0.912	0.912	0.890	0.903	0.915	0.930
$\lambda = 5$	0.990	0.996	0.994	0.997	0.998	0.995	0.998	0.994	0.997	0.998	0.998	0.997
$\lambda = 10$	0.992	0.995	0.998	0.999	0.993	0.993	0.992	0.993	0.998	0.996	0.997	0.996
$\lambda = 20$	0.998	0.998	0.999	1.000	0.993	0.993	0.991	0.998	0.996	0.994	0.993	0.996
$\lambda = 40$	1.000	1.000	1.000	1.000	0.993	0.994	0.991	1.000	0.995	0.994	0.992	0.994
Average number of elements in $\widehat{\mathcal{M}}^*_{1-\alpha}$												
$\lambda = 0$	9.792	9.809	9.824	9.842	39.80	39.80	39.83	39.84	99.73	99.79	99.83	99.87
$\lambda = 5$	7.962	7.364	6.050	3.232	32.43	29.55	23.97	12.68	79.18	72.06	58.09	30.42
$\lambda = 10$	4.600	4.021	3.159	1.754	18.50	16.14	12.63	6.644	44.13	38.44	30.21	15.91
$\lambda = 20$	2.370	2.122	1.730	1.151	9.584	8.427	6.576	3.524	23.07	20.28	15.83	8.177
$\lambda = 40$	1.358	1.246	1.096	1.001	5.005	4.352	3.393	1.923	12.15	10.52	8.277	4.346

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.

Table 3: MCS p-values for Stock and Watson JME (1999) table 2.

		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		3.290	0.001	2.140	0.008	2.208	0.015	1.751	0.000
uniar	-	2.675	0.002	1.360	$0.778^{\star}$	1.941	0.033	1.082	$0.413^{\star}$
'Gaps' specifications									
dtip	DT	2.519	0.013	1.310	$0.781^{\star}$	1.913	0.074	1.043	$0.549^{\star}$
dtgmpyq	DT	2.644	0.001	1.446	$0.101^{\star}$	2.067	0.006	1.103	$0.239^{\star}$
dtmsmtq	DT	2.341	0.089	1.280	$0.848^{\star}$	1.844	0.083	1.007	$0.969^{\star}$
dtlpnag	DT	2.482	0.029	1.323	$0.778^{\star}$	2.024	0.020	1.012	$0.969^{\star}$
ipxmca	LV	2.373	0.066	1.264	$1.000^{\star}$	1.887	0.083	1.026	$0.969^{\star}$
hsbp	LN	2.205	$0.682^{\star}$	1.392	$0.663^{\star}$	1.829	0.083	0.993	$1.000^{\star}$
lhmu25	LV	2.433	0.052	1.401	$0.402^{\star}$	1.937	0.041	1.055	$0.763^{\star}$
First difference specifications									
ip	DLN	2.384	0.060	1.429	$0.244^{\star}$	1.819	0.083	1.115	0.064
gmpyq	DLN	2.233	$0.653^{\star}$	1.532	0.039	1.565	$1.000^{\star}$	1.149	$0.129^{\star}$
msmtq	DLN	2.169	$1.000^{\star}$	1.353	$0.774^{\star}$	1.778	0.083	1.062	$0.491^{\star}$
lpnag	DLN	2.308	$0.124^{\star}$	1.317	$0.781^{\star}$	1.809	0.083	1.009	$0.969^{\star}$
dipxmca	DLV	2.355	0.066	1.456	0.068	1.839	0.083	1.128	0.035
dhsbp	DLN	2.701	0.004	1.405	$0.496^{\star}$	1.969	0.021	1.077	$0.450^{\star}$
dlhmu25	DLV	2.352	0.080	1.474	0.026	1.878	0.083	1.103	0.095
dlhur	DLV	2.321	$0.153^{\star}$	1.451	$0.139^{\star}$	1.843	0.083	1.088	$0.316^{\star}$
Phillips curve									
LHUR		2.387	0.060	1.371	$0.582^{\star}$	1.939	0.059	1.050	$0.602^{\star}$

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 $\star$ are those in $\widehat{\mathcal{M}}^{*}_{90\%}$. 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.

	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	3.290	0.001	2.140	0.020	2.208	0.013	1.751	0.004
Univariate	2.675	0.001	1.360	$0.617^{\star}$	1.941	0.025	1.082	$0.199^{\star}$
Panel A. All indicators								
Mul. factors	2.158	$0.222^{\star}$	1.291	$0.998^{\star}$	1.894	$0.132^{\star}$	0.964	$0.631^{\star}$
1 factor	2.069	$0.610^{\star}$	1.274	$1.000^{\star}$	1.692	$1.000^{\star}$	1.002	$0.602^{\star}$
Comb. mean	2.439	0.002	1.289	$0.999^{\star}$	1.853	$0.162^{\star}$	1.036	$0.543^{\star}$
Comb. median	2.550	0.002	1.316	$0.981^{\star}$	1.895	$0.115^{\star}$	1.063	$0.442^{\star}$
Comb. ridge reg.	2.209	0.023	1.280	$0.999^{\star}$	1.842	$0.193^{\star}$	1.019	$0.543^{\star}$
Panel B. Real activity indicators								
Mul. factors	2.019	$1.000^{\star}$	1.357	$0.775^{\star}$	1.792	$0.202^{\star}$	0.946	$1.000^{\star}$
1 factor	2.079	$0.610^{\star}$	1.281	$0.999^{\star}$	1.753	$0.234^{\star}$	1.017	$0.543^{\star}$
Comb. mean	2.346	0.004	1.284	$0.999^{\star}$	1.807	$0.202^{\star}$	1.020	$0.543^{\star}$
Comb. median	2.381	0.002	1.299	$0.994^{\star}$	1.831	$0.193^{\star}$	1.036	$0.506^{\star}$
Comb. ridge reg.	2.192	0.084	1.298	$0.994^{\star}$	1.773	$0.234^{\star}$	1.022	$0.543^{\star}$
Panel C. Interest rates								
Mul. factors	2.585	0.002	1.495	$0.205^{\star}$	1.976	0.036	1.173	0.069
1 factor	2.524	0.004	1.495	$0.117^{\star}$	2.038	0.007	1.077	$0.356^{\star}$
Comb. mean	2.424	0.008	1.341	$0.883^{\star}$	1.900	$0.100^{\star}$	1.079	$0.184^{\star}$
Comb. median	2.513	0.002	1.336	$0.941^{\star}$	1.912	0.055	1.078	$0.300^{\star}$
Comb. ridge reg.	2.432	0.008	1.368	$0.454^{\star}$	1.943	0.029	1.123	$0.106^{\star}$
Panel D. Money								
Mul. factors	2.679	0.001	1.360	$0.462^{\star}$	1.933	0.032	1.080	$0.218^{\star}$
1 factor	2.679	0.001	1.360	$0.544^{\star}$	1.933	0.047	1.080	$0.254^{\star}$
Comb. mean	2.664	0.001	1.350	$0.700^{\star}$	1.964	0.020	1.066	$0.486^{\star}$
Comb. median	2.670	0.001	1.348	$0.789^{\star}$	1.954	0.021	1.070	$0.409^{\star}$
Comb. ridge reg.	2.638	0.001	1.385	$0.390^{\star}$	1.934	0.074	1.121	$0.151^{\star}$
Phillips curve								
LHUR	2.387	0.004	1.371	$0.437^{\star}$	1.939	0.060	1.050	$0.543^{\star}$

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 $\star$ are those in $\widehat{\mathcal{M}}^{*}_{90\%}$. The results of this table correspond to those of S&W (1999, table 4).