The Model Confidence Set

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

The Model Confidence Set

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

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

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

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

The Model Confidence Set*

Peter R. Hansen^{a†}, **Asger Lunde**^{b}, and **James M. Nason**^{c}

^{a}*Stanford University, Department of Economics, 579 Serra Mall,*
Stanford, CA 94305-6072, USA & CREATES

^{b}*Aarhus School of Business, University of Aarhus, Department of Marketing and Statistics,*
Haslegaardsvej 10, 8210 Aarhus V, Denmark & CREATES

^{c}*Federal Reserve Bank of Atlanta, Research Department, 1000 Peachtree St., NE,*
Atlanta, GA 30309, USA

May 2009

Abstract

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

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

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

1 Introduction

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

The objective of the model confidence set (MCS) procedure is to determine the set of models, $\mathcal{M}^*$, that consists of the best model(s) from a collection of models, $\mathcal{M}^0$, where *best* is defined in terms of a criterion that is user-specified. The MCS procedure yields a model confidence set, $\widehat{\mathcal{M}}^*$, that is a collection of models built to contain the best models with a given level of confidence. The process of winnowing models out of $\mathcal{M}^0$ relies on sample information about the relative performances of the models in $\mathcal{M}^0$. This sample information drives the MCS to create a random data-dependent set of models, $\widehat{\mathcal{M}}^*$. The set $\widehat{\mathcal{M}}^*$ includes the best model(s) with a certain probability in the same sense that a confidence interval covers a population parameter.

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

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

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 'models' loosely. It can refer to econometric models, competing forecasts, or alternatives that need not involve any modelling of data, such as trading rules. So the MCS procedure is not specific to comparisons of models. For example, one could construct a MCS for a set of different 'treatments' by comparing sample estimates of the corresponding treatment effects, or a MCS for trading rules with the best Sharpe ratio.

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

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

1.2 Bootstrap Implementation and Simulation Results

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

We study the properties of our bootstrap implementation of the MCS procedure through simulation experiments. The results are very encouraging 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.

1.3   Empirical Analysis of Inflation Forecasts and Taylor Rules

We apply the MCS to two empirical problems. First, the MCS is used to study the inflation forecasting problem. The choice of an inflation forecasting model is an especially important issue for central banks, treasuries, and private sector agents. The fifty plus year tradition of the Phillips curve suggests it remains an effective vehicle for the task of inflation forecasting. 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 (1997b), 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 simple no-change forecast in out-of-sample point prediction.

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

Our second empirical example shows that the MCS approach is a useful tool for in-sample evaluation of regression models. This example applies the MCS to choosing from a set of competing (nominal) interest rate rule regressions on a quarterly U.S. sample that runs from 1979 through 2006. These regressions fall into the class of interest rate rules promoted by Taylor (1993). His Taylor rule forms the basis of a class of monetary policy rules that gauge the success of monetary policy at keeping inflation low and the real economy close to trend. The MCS does not reveal which Taylor rule regressions best describe the actual U.S. monetary policy, nor does it identify the best policy rule. Rather the MCS selects the Taylor rule regressions that have the best empirical fit of the U.S. federal funds rate in this sample period, where the 'best fit' is defined by different likelihood criteria.

The MCS procedure begins with 25 regression models. We include a pure first-order autoregression, AR(1), of the federal funds rate in the initial MCS. The remaining 24 models are Taylor rule regressions that contain different combinations of lagged inflation, lags of various definitions of real economic activity (i.e., the output gap, the unemployment rate gap, or real marginal cost), and in some cases the lagged federal funds rate.

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

1.4 Outline of Paper

The paper is organized as follows. We present the theoretical framework of the MCS in Section 2. Section 3 outlines practical bootstrap methods to implement the MCS. Multiple model comparison methods related to the MCS are discussed in Section 4. Section 5 reports the results of simulation experiments. The MCS is applied to two empirical examples in Section 6. Section 7 concludes.

2 General Theory for Model Confidence Set

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

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

Define the relative performance variables

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

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

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

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

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

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

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

2.1 The MCS Algorithm and Its Properties

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

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

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

Assumption 1 *For any $\mathcal{M} \subset \mathcal{M}^0$ we assume 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) $\lim_{n\to\infty} P(e_{\mathcal{M}} \in \mathcal{M}^*|H_{A,\mathcal{M}}) = 0$.*

The conditions that Assumption 1 states for $\delta_{\mathcal{M}}$ are standard requirements for hypothesis tests. (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 1, it holds that* (i) $\liminf_{n \to \infty} P(\mathcal{M}^* \subset \widehat{\mathcal{M}}^*_{1-\alpha}) \geq 1 - \alpha$, *and* (ii) $\lim_{n \to \infty} P(i \in \widehat{\mathcal{M}}^*_{1-\alpha}) = 0$ *for all* $i \notin \mathcal{M}^*$.

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

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. Since lack of power is tied to the lack of information in the data, the MCS should be large when there is insufficient information to distinguish good and bad models.

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

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

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

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

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

2.2 Coherency between Test and Elimination Rule

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

Theorem 3 *Suppose that* $P(\delta_{\mathcal{M}} = 1, e_{\mathcal{M}} \in \mathcal{M}^*) \leq \alpha$ *then we have*

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

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

The property that $P(\delta_{\mathcal{M}} = 1, e_{\mathcal{M}} \in \mathcal{M}^*) \leq \alpha$ holds under both the null hypothesis and the alternative hypothesis is key for the result in Theorem 3. For a test with the correct size, we have $P(\delta_{\mathcal{M}} = 1|H_{0,\mathcal{M}}) \leq \alpha$, which implies $P(\delta_{\mathcal{M}} = 1, e_{\mathcal{M}} \in \mathcal{M}^*|H_{0,\mathcal{M}}) \leq \alpha$. The additional condition, $P(\delta_{\mathcal{M}} = 1, e_{\mathcal{M}} \in \mathcal{M}^*|H_{A,\mathcal{M}}) \leq \alpha$, ensures that a rejection, $\delta_{\mathcal{M}} = 1$, can be taken as significant evidence that $e_{\mathcal{M}}$ is not in $\mathcal{M}^*$.

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

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

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

The coherency in conjunction with an asymptotic control of the Type I error, $\limsup_{n \to \infty} P_0(\delta_{\mathcal{M}} = 1) \leq \alpha$, translate into an asymptotic version of the assumption we made in Theorem 3. Coherency places restrictions on the combinations of tests and elimination rules we can employ. These restrictions go beyond those imposed by the asymptotic conditions we formulated in Assumption 1. In fact, coherency serves to

curb the reliance on asymptotic properties, in order to avoid perverse outcomes in finite samples that could result from absurd combinations of test and elimination rule. Coherency prevents us from adopting the most powerful test of the hypothesis $H_{0,\mathcal{M}}$ in some situations. The reason is that tests do not necessarily identify a single element as the cause for the rejection. A good analogy is found in the standard regression model, where a F-test may reject the joint hypothesis that all regression coefficients are zero, even though all t-statistics are insignificant.[1]

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

2.3   MCS p-Values

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

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

The advantage of this definition of MCS p-values will be evident from Theorem 4 that is stated below. Since $\mathcal{M}_{m_0}$ consists of a single model, the null hypothesis, $H_{0,\mathcal{M}_{m_0}}$, simply states that the last surviving model is as good as itself, making the convention, $P_{H_{0,\mathcal{M}_{m_0}}} \equiv 1$, logical.

Table 1 illustrates how MCS p-values are computed and how they relate to p-values of the individual tests, $P_{H_{0,\mathcal{M}_i}}$, $i = 1, \ldots, m_0$. The MCS p-values are convenient because they make it easy to determine whether a particular object is in $\widehat{\mathcal{M}}^*_{1-\alpha}$ or not, for any α. Thus, the MCS p-values are an effective way of conveying the information in the data.

Theorem 4 *Let the elements of $\mathcal{M}^0$ be indexed by $i = 1, \ldots, m_0$. The MCS p-value, $\hat{p}_i$, is such that $i \in \widehat{\mathcal{M}}^*_{1-\alpha}$ if and only if $\hat{p}_i \geq \alpha$, for any $i \in \mathcal{M}^0$.*

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

---

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

Table 1: Computation of MCS p-values

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

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

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

3 Bootstrap Implementation

3.1 Equivalence Tests and Elimination Rules

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

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

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

3.1.1 Quadratic-Form Test

Let $\mathcal{M}$ be some subset of $\mathcal{M}^0$ and let m be the number of models in $\mathcal{M} = \{i_1, \ldots, i_m\}$. 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 $\iota'_\perp \iota = 0$ (a vector of zeros). The $m-1$ dimensional vector $X_t \equiv \iota'_\perp L_t$ can be viewed as $m-1$ contrasts because each element of X_t is a linear combination of $d_{ij,t}$, $i, j \in \mathcal{M}$, which has mean zero under the null hypothesis.

Lemma 5 *Given Assumption 2, let $X_t \equiv \iota'_\perp L_t$ and define $\theta \equiv \mathrm{E}(X_t)$. The null hypothesis $H_{0,\mathcal{M}}$ is equivalent to $\theta = 0$ and it holds that $n^{1/2}(\bar{X} - \theta) \overset{d}{\to} N(0, \Sigma)$, where $\bar{X} \equiv n^{-1} \sum_{t=1}^{n} X_t$ and $\Sigma \equiv \lim_{n \to \infty} \mathrm{var}(n^{1/2}\bar{X})$.*

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

Lemma 5 shows that $H_{0,\mathcal{M}}$ can be tested using traditional quadratic-form statistics. An example is $T_Q \equiv n\bar{X}'\hat{\Sigma}^\#\bar{X}$, where $\hat{\Sigma}$ is some consistent estimator of Σ and $\hat{\Sigma}^\#$ denotes the Moore-Penrose inverse of $\hat{\Sigma}$.[2] The rank $q \equiv \mathrm{rank}(\hat{\Sigma})$ represents the effective number of *contrasts* (the number of linearly independent comparisons) under $H_{0,\mathcal{M}}$. Since $\hat{\Sigma} \overset{p}{\to} \Sigma$ (by assumption) it follows that $T_Q \overset{d}{\to} \chi^2_{(q)}$ where $\chi^2_{(q)}$ denotes the χ^2-distribution with q degrees of freedom. Under the alternative hypothesis, T_Q diverge to infinity with probability one. So the test $\delta_\mathcal{M}$ will meet the requirements of Assumption 1 when constructed from T_Q. Although the matrix $\iota_\perp$ is not fully identified by the requirements: $\iota'_\perp \iota = 0$ and $\det(\iota'_\perp \iota_\perp) \neq 0$ (but the sub-space spanned by the columns of $\iota_\perp$ is), there is no problem because the statistic T_Q is invariant to the choice for $\iota_\perp$.

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

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

---

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

3.1.2 Tests Constructed from t-Statistics

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

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

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

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

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

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

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

which are available to test the hypothesis $H_{0,\mathcal{M}}$.[3] The asymptotic distributions of these test statistics are non-standard because they depend on nuisance parameters (under both the null and the alternative). However, the nuisance parameters pose few obstacles as the relevant distributions can be estimated with bootstrap methods that implicitly deal with the nuisance parameter problem. This feature of the bootstrap has previously been used in this context by Kilian (1999), White (2000b), Hansen (2003b, 2005), and Clark and McCracken (2005).

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

---

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

this model is such that $t_{e_{R,\mathcal{M}}j} = T_{R,\mathcal{M}}$, for some $j \in \mathcal{M}$. These combinations of test and elimination rule will satisfy the required coherency.

Proposition 6 *Let $\delta_{\max,\mathcal{M}}$ and $\delta_{R,\mathcal{M}}$ denote the tests based on the statistics $T_{\max,\mathcal{M}}$ and $T_{R,\mathcal{M}}$, respectively. Then $(\delta_{\max,\mathcal{M}}, e_{\max,\mathcal{M}})$ and $(\delta_{R,\mathcal{M}}, e_{R,\mathcal{M}})$ satisfy the coherency of Definition 3.*

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

$$P(T > c, e_{\mathcal{M}} = i \text{ for some } i \in \mathcal{M}^*) = P(T > c, T = T_i \text{ for some } i \in \mathcal{M}^*)$$

$$= P(\max_{i\in\mathcal{M}\cap\mathcal{M}^*} T_i > c, T_i \geq T_j \text{ for all } j \in \mathcal{M}) \leq P(\max_{i\in\mathcal{M}\cap\mathcal{M}^*} T_i > c)$$

$$\leq P_0(\max_{i\in\mathcal{M}\cap\mathcal{M}^*} T_i > c) \leq P_0(\max_{i\in\mathcal{M}} T_i > c) = P_0(T > c).$$

This completes the proof. ■

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

Lemma 7 *Suppose that Assumption 2 holds and define $\bar{Z} = (\bar{d}_{1.}, \ldots, \bar{d}_{m.})'$. Then*

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

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

Proof. From the identity $\bar{d}_{i.} = \bar{L}_i - \bar{L}. = \bar{L}_i - m^{-1}\sum_{j\in\mathcal{M}}\bar{L}_j = m^{-1}\sum_{j\in\mathcal{M}}(\bar{L}_i - \bar{L}_j) = m^{-1}\sum_{j\in\mathcal{M}}\bar{d}_{ij}$, we see that the elements of $\bar{Z}$ are linear transformations of $\bar{X}$ from Lemma 5. Thus for some $(m-1) \times m$ matrix G we have $\bar{Z} = G'\bar{X}$, and the result now follows, where $\psi = G'\theta$ and $\Omega = G'\Sigma G$. (The $m \times m$ covariance matrix, Ω, has reduced rank, as $\mathrm{rank}(\Omega) \leq m-1$.) ■

In the following, we let ϱ denote the $m\times m$ correlation matrix that is implied by the covariance matrix, Ω, of Lemma 7. Further, given the vector of random variables $\xi \sim N_m(0, \varrho)$, we let F_ϱ denote the distribution of $\max_i \xi_i$.

Theorem 8 *Let Assumption 2 hold and suppose that $\hat{\omega}_i^2 \equiv \widehat{\mathrm{var}}(n^{1/2}\bar{d}_{i.}) = n\,\widehat{\mathrm{var}}(\bar{d}_{i.}) \xrightarrow{p} \omega_i^2$, where ω_i^2, $i = 1, \ldots, m$ are the diagonal elements of Ω. Under $H_{0,\mathcal{M}}$ we have $T_{\max,\mathcal{M}} \xrightarrow{d} F_\varrho$ and under the alternative hypothesis, $H_{A,\mathcal{M}}$, we have that $T_{\max,\mathcal{M}} \to \infty$ in probability. Moreover, under the alternative hypothesis we have $T_{\max,\mathcal{M}} = t_{j.}$ where $j = e_{\max,\mathcal{M}} \notin \mathcal{M}^*$, for n sufficiently large.*

Proof. Let $D \equiv \mathrm{diag}(\omega_1^2, \ldots, \omega_m^2)$ and $\hat{D} \equiv \mathrm{diag}(\hat{\omega}_1^2, \ldots, \hat{\omega}_m^2)$. From Lemma 7 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 $t_{i.} = \bar{d}_{i.}/\sqrt{\widehat{\mathrm{var}}(\bar{d}_{i.})} =$

$n^{1/2}\bar{d}_{i\cdot}/\hat{\omega}_i = \xi_{i,n}\frac{\omega_i}{\hat{\omega}_i}$ it now follows that $T_{\max,\mathcal{M}} = \max_i t_{i\cdot}^2 = \max_i(\hat{D}^{-1/2}n^{1/2}\bar{Z})_i \xrightarrow{d} F_\varrho$. Under the alternative hypothesis we have $\bar{d}_{j\cdot} \xrightarrow{p} \mu_{j\cdot} > 0$ for any $j \notin \mathcal{M}^*$, so that both $t_{j\cdot}$ and $T_{\max,\mathcal{M}}$ diverge to infinity at rate $n^{1/2}$ in probability. Moreover, it follows that $e_{\max,\mathcal{M}} \notin \mathcal{M}^*$, for n sufficiently large. $\blacksquare$

Theorem 8 shows that the asymptotic distribution of $T_{\max,\mathcal{M}}$ depends on the correlation matrix, ϱ. Nonetheless, as discussed earlier, bootstrap methods can be employed to deal with this nuisance parameter problem. Thus, we construct a test of $H_{0,\mathcal{M}}$ by comparing the test statistic $T_{\max,\mathcal{M}}$ to an estimate of the 95%-quantile, say, of its limit distribution under the null hypothesis. Although the quantile may depend on ϱ, our bootstrap implementation leads to an asymptotically valid test because the bootstrap consistently estimates the desired quantile. A detailed description of our bootstrap implementation is available in a separate appendix, Hansen, Lunde, and Nason (2009).

Theorem 8 formulates results for the situation where the MCS is constructed with $T_{\max,\mathcal{M}}$ and $e_{\max,\mathcal{M}} = \arg\max_i t_{i\cdot}$. Similar results hold for the MCS that is constructed from $T_{R,\mathcal{M}}$ and $e_{R,\mathcal{M}}$. The arguments are almost identical to those used for Theorem 8.

3.2 MCS for Regression Models

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

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

3.2.1 Framework and Assumptions

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

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

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

function is, apart from a constant, given by

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

3.2.2 MCS by Kullback-Leibler Divergence

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

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

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

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

which (as always) permits the existence of more than one best model.[4] The extension to other criteria, such as the AIC and the BIC, is straight forward. For instance, the set of best models in terms of the AIC is given by $\mathcal{M}_{\mathrm{AIC}}^* = \left\{ j : E[Q(\mathcal{Z}, \theta_{0j}) + 2k_j] = \min_i E[Q(\mathcal{Z}, \theta_{0i}) + 2k_i] \right\}$, where k_j is the degrees of freedom in the j-th model.

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

$$Q(\mathcal{Z}, \hat{\theta}_j) - Q(\mathcal{Z}, \theta_{0j}) = n \left\{ \left(\log \sigma_{0j}^2 - \log \hat{\sigma}_j^2 \right) + \left(n^{-1} \sum_{t=1}^{n} \varepsilon_{j,t}^2 / \sigma_{0j}^2 - 1 \right) \right\},$$

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

In the event that the j-th model is correctly specified, it is well known that the limit distribution of $Q(\mathcal{Z}, \hat{\theta}_j) - Q(\mathcal{Z}, \theta_{0j})$ is $\chi_{(k_j)}^2$, where the degrees of freedom, k_j, is given by the dimension of $\theta_{0j} = (\beta_{0j}', \sigma_{0j}^2)'$. In the present multi-model setup, it is unlikely that all models are correctly specified. More generally, the limit distribution of the QLR statistic has the form, $\sum_{i=1}^{k_j} \lambda_{i,j} Z_{i,j}^2$, where $\lambda_{1,j}, \ldots, \lambda_{k_j,j}$ are

---

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

the eigenvalues of $\mathcal{I}_j^{-1}\mathcal{J}_j$ and $Z_{1,j}, \ldots, Z_{k_j,j} \sim \text{iid} N(0, 1)$. The information matrices $\mathcal{I}_j$ and $\mathcal{J}_j$ are those associated with the j-th model, $\mathcal{I}_j = \text{diag}\left(\sigma_{0j}^{-2}\text{E}(X_{j,t}X_{j,t}'), \frac{1}{2}\sigma_{0j}^{-4}\right)$ and

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

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

$$k_j^\star = \lambda_{1,j} + \cdots + \lambda_{k_j,j} = \text{tr}\{\mathcal{I}_j^{-1}\mathcal{J}_j\}$$

$$= \text{tr}\left\{[\text{E}(X_{j,t}X_{j,t}')]^{-1}\sigma_{0j}^{-2}n^{-1}\sum_{s,t=1}^n \text{E}(X_{j,s}\varepsilon_{j,s}X_{j,t}'\varepsilon_{j,t})\right\} + n^{-1}\frac{1}{2}\sum_{s,t=1}^n \text{E}(\frac{\varepsilon_{j,s}^2\varepsilon_{j,t}^2}{\sigma_{0j}^4} - 1).$$

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

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

3.2.3 The MCS Procedure

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

The literature has proposed several bootstrap estimators of the effective degrees of freedom, $k_j^\star = \text{E}[Q(\mathcal{Z}, \theta_{0j}) - Q(\mathcal{Z}, \hat{\theta}_j)]$, see e.g. Efron (1983, 1986) and Cavanaugh and Shumway (1997). These and additional estimators are analyzed and compared in Shibata (1997). We adopt the estimator for $k_j^\star$ that is labelled B_3 in Shibata (1997). In the regression context this estimator takes the form

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

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

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

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

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

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

$$(Q(\mathcal{Z}, \hat{\theta}_1) + k_1^\star - \mathrm{E}[Q(\mathcal{Z}, \theta_{01})], \ldots, Q(\mathcal{Z}, \hat{\theta}_{m_0}) + k_{m_0}^\star - \mathrm{E}[Q(\mathcal{Z}, \theta_{0m_0})]),$$

is estimated by the empirical distribution of

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

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

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

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

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

The empirical distribution of $T_{b,R,\mathcal{M}}^*$ yields an estimate of the distribution of $T_{R,\mathcal{M}}$, as $n, B \to \infty$. The p-value for the hypothesis test with which we are concerned is computed by

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

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

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

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

Under the null hypothesis we recenter the bootstrap statistics about zero and this offsets the location shift $c_i - c_j$.

When two models are nested, the null hypothesis used with KLIC, $\mathrm{E}[Q(\mathcal{Z}, \theta_{0i})] = \mathrm{E}[Q(\mathcal{Z}, \theta_{0j})]$, has the strong implication that $Q(\mathcal{Z}, \theta_{0i}) = Q(\mathcal{Z}, \theta_{0j})$ a.e. (almost everywhere), and this causes the limit distribution of the quasi likelihood ratio statistic, $Q(\mathcal{Z}, \hat{\theta}_i) - Q(\mathcal{Z}, \hat{\theta}_j)$, to differ for nested or non-nested comparisons, see Vuong (1989). This property of nested comparisons can be imposed on the bootstrap resamples, by replacing $Q(\mathcal{Z}, \hat{\theta}_j)$ with $Q(\mathcal{Z}^*, \hat{\theta}_j)$, because the latter is the bootstrap variant of $Q(\mathcal{Z}, \theta_{0j})$. The MCS procedure can be adapted, so that different bootstrap schemes are used for nested and non-nested comparisons. Imposing the stronger null hypothesis, $Q(\mathcal{Z}, \theta_{0i}) = Q(\mathcal{Z}, \theta_{0j})$ a.e. may improve the power of the procedure. However, such a bootstrap implementation is only applicable to KLIC because the underlying null hypotheses of other criteria, such as AIC* and BIC*, do not imply $Q(\mathcal{Z}, \theta_{0i}) = Q(\mathcal{Z}, \theta_{0j})$ for nested models. Therefore, this paper does not pursue an adaptive bootstrap implementation in this paper.

4 Relation to Existing Multiple Comparisons Methods

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

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

Our MCS employs a sequential testing procedure that mimics step-down procedures for multiple hypothesis testing, see e.g. Dudoit, Shaffer, and Boldrick (2003), Lehmann and Romano (2005, chapter 9) or Romano, Shaikh, and Wolf (2008). Our definition of MCS p-values implies the monotonicity, $\hat{p}_{e\mathcal{M}_1} \leq \hat{p}_{e\mathcal{M}_2} \leq \cdots \leq \hat{p}_{e\mathcal{M}_{m_0}}$, that is key for the result of Theorem 4. This monotonicity is also a feature of the so-called *step-down Holm adjusted p-values*.

4.1 Relation to Tests for Superior Predictive Ability

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*. In the context of forecast comparisons, this is the problem that arises when testing for *superior predictive ability* (SPA), see White (2000b), Hansen (2005), and Romano and Wolf (2005).

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

The MCS procedure has the advantage that it can be employed for model selection, whereas a SPA-test is ill-suited for this problem. A rejection of the SPA-test only identifies one or more models as significantly better than the benchmark.[5] Thus, the SPA-test offers little guidance about which models reside in $\mathcal{M}^*$. We are also faced with a similar problem in the event that the null hypothesis is not rejected by the SPA-test. In this case the benchmark may be the best model, but this label may also be applied to other models. This issue can be resolved if all models serve as the benchmark in a series of comparisons. The result is a sequence of SPA-tests that define the MCS to be the set of 'benchmark' models that are found not to be significantly inferior to the alternatives. However, the level of individual SPA-tests need to be adjusted for the number of tests that are computed to control the FWE. For example, if the level in each of the SPA-tests is α/m, the Bonferroni bound states that the resulting set of 'surviving' benchmarks is a MCS with coverage $(1 - \alpha)$. Nonetheless, there is a substantial loss of power associated with the small level applied to the individual tests. The loss of power highlights a major pitfall of sequential SPA-tests.

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

---

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

4.2   Related Sequential Testing Procedures for Model Selection

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

Several papers have studied the problem of selecting the best forecasting model from a set of competing models. For example, Engle and Brown (1985) 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 (2006) 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 is the true model, nor is the null hypothesis defined by a single model. Instead, all models are treated equally in the comparison and only evaluated on out-of-sample predictive ability.

4.3   Aspects of Parameter Uncertainty and Forecasting

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

4.4   Bayesian Interpretation

The MCS procedure is based on frequentist principles, but resembles some aspects of Bayesian model selection techniques. By specifying a prior over the models in $\mathcal{M}^0$, a Bayesian procedure would produce a posterior distribution for each model, conditional on the actual data. This approach to MCS construction

includes those models with the largest posteriors 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 random data-dependent set of models that is the MCS. It therefore is meaningful to state that the best model can be found in the MCS with a certain probability. The MCS also places moderate computational demands on the researcher, unlike the synthetic data creation methods on which Bayesian Markov chain-Monte Carlo methods rely.

5 Simulation Results

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

5.1 Simulation Experiment I

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

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

where ρ controls the degree of correlation between alternatives.

Design I.A (Symmetric distributed loss): Define the (vector of) loss variables to be

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

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

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

Table 2: Simulation Design I.A

	$m = 10$				$m = 40$				$m = 100$			

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
$\lambda = 0$	0.879	0.889	0.880	0.880	0.865	0.866	0.859	0.857	0.846	0.844	0.854	0.844
$\lambda = 5$	0.989	0.988	0.991	1.000	0.978	0.978	0.974	0.984	0.972	0.972	0.971	0.976
$\lambda = 10$	0.994	0.997	0.999	1.000	0.979	0.982	0.984	0.993	0.970	0.972	0.976	0.980
$\lambda = 20$	0.998	1.000	1.000	1.000	0.987	0.981	0.992	1.000	0.976	0.977	0.984	0.991
$\lambda = 40$	1.000	1.000	1.000	1.000	0.993	0.995	0.998	1.000	0.983	0.984	0.990	0.998

Average number of elements in $\widehat{\mathcal{M}}^*_{90\%}$ (power)

$\lambda = 0$	9.590	9.628	9.624	9.606	38.41	38.52	38.62	38.57	96.23	95.92	96.45	96.20
$\lambda = 5$	6.501	4.715	3.251	1.549	25.00	18.64	13.32	6.397	58.59	43.18	32.10	14.99
$\lambda = 10$	3.369	2.400	1.732	1.027	13.54	9.807	7.156	3.280	31.84	22.82	16.88	7.915
$\lambda = 20$	1.704	1.308	1.062	1.000	7.074	5.034	3.636	1.678	16.92	12.39	8.818	4.079
$\lambda = 40$	1.072	1.005	1.000	1.000	3.576	2.608	1.842	1.052	8.784	6.376	4.539	2.096

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
$\lambda = 0$	0.898	0.894	0.900	0.892	0.898	0.888	0.891	0.892	0.898	0.897	0.890	0.889
$\lambda = 5$	0.984	0.990	0.995	1.000	0.968	0.973	0.977	0.984	0.967	0.968	0.964	0.969
$\lambda = 10$	0.990	0.999	1.000	1.000	0.980	0.981	0.982	0.995	0.970	0.967	0.981	0.985
$\lambda = 20$	0.999	1.000	1.000	1.000	0.987	0.989	0.989	1.000	0.978	0.971	0.979	0.992
$\lambda = 40$	1.000	1.000	1.000	1.000	0.996	0.996	1.000	1.000	0.978	0.982	0.990	0.999

Average number of elements in $\widehat{\mathcal{M}}^*_{90\%}$ (power)

$\lambda = 0$	9.644	9.637	9.650	9.624	38.67	38.64	38.75	38.73	97.19	97.21	96.83	96.69
$\lambda = 5$	6.171	4.597	3.276	1.587	24.00	17.52	13.04	6.168	56.10	40.14	29.79	14.43
$\lambda = 10$	3.255	2.354	1.730	1.037	12.90	9.290	6.884	3.218	29.92	21.91	16.30	7.562
$\lambda = 20$	1.746	1.288	1.073	1.000	6.896	4.974	3.621	1.695	15.86	11.55	8.526	3.978
$\lambda = 40$	1.088	1.009	1.000	1.000	3.586	2.596	1.850	1.064	8.425	6.245	4.434	2.072

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
$\lambda = 0$	0.928	0.922	0.928	0.936	0.946	0.938	0.938	0.935	0.940	0.945	0.948	0.938
$\lambda = 5$	0.990	0.997	0.997	1.000	0.979	0.979	0.986	0.991	0.962	0.967	0.971	0.978
$\lambda = 10$	0.998	1.000	1.000	1.000	0.980	0.982	0.993	0.998	0.973	0.972	0.974	0.991
$\lambda = 20$	1.000	1.000	1.000	1.000	0.990	0.991	0.996	1.000	0.977	0.981	0.990	0.996
$\lambda = 40$	1.000	1.000	1.000	1.000	0.998	0.999	1.000	1.000	0.991	0.994	0.994	0.999

Average number of elements in $\widehat{\mathcal{M}}^*_{90\%}$ (power)

$\lambda = 0$	9.738	9.759	9.771	9.796	39.39	39.33	39.33	39.27	98.33	98.51	98.63	98.43
$\lambda = 5$	4.546	3.578	2.587	1.414	16.64	12.62	9.395	4.670	38.12	27.64	20.60	10.63
$\lambda = 10$	2.622	2.058	1.540	1.102	9.580	6.896	5.121	2.605	20.70	14.88	11.53	5.822
$\lambda = 20$	1.574	1.284	1.146	1.016	5.066	3.791	2.784	1.546	11.24	8.438	6.255	3.132
$\lambda = 40$	1.134	1.055	1.022	1.003	2.888	2.149	1.613	1.115	6.284	4.664	3.480	1.796

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.

We report two statistics from our simulation experiment based on $\alpha = 10\%$. One is the frequency at which $\widehat{\mathcal{M}}^*_{90\%}$ contains $\mathcal{M}^*$ 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 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 the μ_{ij}s become more disperse (e.g., as λ increases). Note also that a larger correlation makes it easier to separate inferior models from superior model. This is not surprising because $\text{var}(d_{ij,t}) = \text{var}(L_{it}) + \text{var}(L_{jt}) - 2\text{cov}(L_{it}, L_{jt}) = 2(1 - \rho)$, which is decreasing in ρ. Thus, a larger correlation (holding the individual variances fixed) is associated with more information that allows the MCS to separate good from bad models. Finally, the effects of heteroskedasticity are relatively small, but heteroskedasticity does appear to add power to the MCS procedure. The average number of models in $\widehat{\mathcal{M}}^*_{90\%}$ tends to fall as φ increases.

Corollary 2 has a consistency result that applies when $\lambda > 0$. The implication is that only one model enters $\mathcal{M}^*$ under this restriction. Table 2 shows that $\mathcal{M}^*$ often contains only one model given $\lambda > 0$. The MCS matches this theoretical prediction in Table 2 because $\widehat{\mathcal{M}}^*_{90\%} = \mathcal{M}^*$ in a large number of simulations. This equality holds especially when λ and ρ are large. These are also the simulation experiments that yield size and power statistics equal (or nearly equal) to one. With size close to one or equal to one, observe that $\mathcal{M}^* \subset \widehat{\mathcal{M}}^*_{90\%}$ (in all the synthetic samples). On the other hand, $\widehat{\mathcal{M}}^*_{90\%}$ is reduced to a single model (in all the synthetic samples) when power is close to one or equal to one.

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

The simulation results are presented in Figure 1. The left panels display the frequency at which $\widehat{\mathcal{M}}^*_{90\%}$ contains $\mathcal{M}^*$ ('size') at various sample sizes. The right panels present the average number of models in $\widehat{\mathcal{M}}^*_{90\%}$ ('power'). The two upper panels contain the results for the case where $\mathcal{M}^*$ is a single model. The upper-left panel indicates that the best model is almost always contained in the MCS. This agrees with Corollary 2, that states that $\widehat{\mathcal{M}}^*_{1-\alpha} \xrightarrow{p} \mathcal{M}^*$ as $n \to \infty$, whenever $\mathcal{M}^*$ consists of a single model. The upper-right panel illustrates the 'power' of the procedure based on $T_{\max,\mathcal{M}} = \max_{i \in \mathcal{M}} t_{i\cdot}$. We note that it takes about 800 observations to weed out the nine inferior models in this design. The MCS procedure is barely affected by the correlation parameter, ρ, but we note that a larger ρ results in a small loss in 'power'. In the lower-left panel we see that the frequency at which $\mathcal{M}^*$ is contained in $\widehat{\mathcal{M}}^*_{90\%}$ is reasonably close to 90% except for the very short sample sizes. From the middle-right and lower-right panel we see that it takes about 500 observations to remove all the poor models. We also note that the correlation parameter has a larger effect on the 'power' when $\mathcal{M}^*$ contains more than a one model.
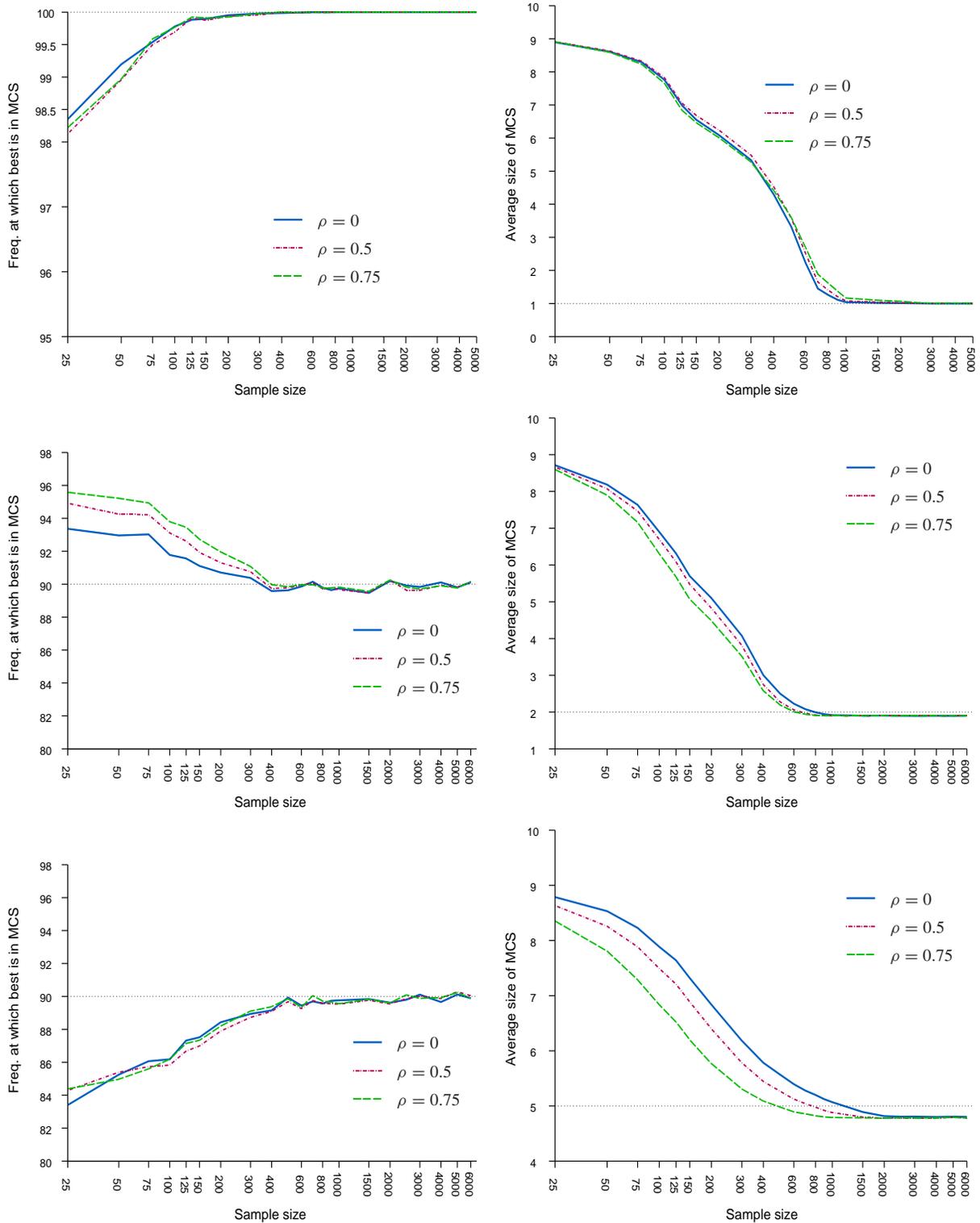
Figure 1: Simulation design I.B with 10 alternatives and 1, 2, or 5 elements in $\mathcal{M}^*$. The left panels report the frequency at which $\mathcal{M}^*$ is contained in $\widehat{\mathcal{M}}^*_{90\%}$ (size properties) and the right panels report the average number of models in $\widehat{\mathcal{M}}^*_{90\%}$ (power properties).

The middle-right and lower-right panels illustrate another aspect of the MCS procedure. For large sample sizes we note that the average number of models in $\widehat{\mathcal{M}}^*_{90\%}$ falls below the number of models in $\mathcal{M}^*$. The explanation is simple. After all poor models have been eliminated, as occurs with probability approaching one as $n \to \infty$, there is a positive probability that $H_{0,\mathcal{M}^*}$ is rejected, which causes the MCS procedure to eliminate a good model. Thus, the inference we draw from the simulation results are quite encouraging for the $T_{\max,\mathcal{M}}$ test.

5.2 Simulation Experiment II: Regression Models

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

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

where ρ measures the degree of dependence between the regressors. We define the dependent variable by $Y_t = \mu + \beta X_{1,t} + \sqrt{1 - \beta^2}\varepsilon_t$, where $\varepsilon_t \sim \text{iid } N(0, 1)$. In addition to the six variables in X_t, we include a constant, $X_{0,t} = 1$, in all regression models. The set of regressions being estimated is given by the twelve regression models that are listed in each of the panels in Table 3.

We report simulation results based on 10,000 repetitions, using a design with an $R^2 = 50\%$ (i.e. $\beta^2 = 0.5$) and either $\rho = 0.3$ or $\rho = 0.9$.[6] For the number of bootstrap resamples we use $B = 1,000$. Since $X_{0,t} = 1$ is included in all regression models, the relevant MCS statistics are invariant to the actual value for μ, so we set $\mu = 0$ in our simulations.

The definition of $\mathcal{M}^*$ will depend on the criterion. With KLIC the set of best models, $\mathcal{M}^*$, is given by the set of regression models that includes X_1. The reason is that KLIC does not favor parsimonious models, unlike the AIC* and BIC*. With these two criteria, $\mathcal{M}^*$, is defined to be the most parsimonious regression model that includes X_1. The models in $\mathcal{M}^*$ are identified by the shaded regions in Table 3.

Our simulation results are reported in Table 3. The average value of $Q(\mathcal{Z}_j, \hat{\theta}_j)$ is given in the first pair of columns, followed by the average estimate of the effective degrees of freedom, $\hat{k}^\star$. The Gaussian setup is such that all models are correctly specified. So the effective degrees of freedom is simply the number of free parameters, which is the number of regressors plus one for σ_j^2. Table 3 shows that the average value of $\hat{k}_j^\star$, is very close to the number of free parameters in the j-th regression model. The last three pairs of columns report the frequency that each of the models are in $\widehat{\mathcal{M}}^*_{90\%}$. We want large numbers inside the shaded region and small numbers outside the shaded region. The results are intuitive. As the sample size increases, from 50 to 100 and then to 500, the MCS procedure becomes better at eliminating the models that do not reside in $\mathcal{M}^*$. With a sample size of $n = 500$, the consistent criterion, BIC*, has reduced the MCS to the single

---

[6]Simulation results for $\beta^2 = 0.1$ and 0.9 are available in a separate appendix, see Hansen, Lunde, and Nason (2009)

Table 3: Simulation Experiment II

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

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

best model in the majority of simulations. This is not true for the AIC* criterion. Although it tends to settle on more parsimonious models than the KLIC, the AIC* has a penalty that makes it possible for an overparameterized model to have the best AIC*. The bootstrap testing procedure is conservative when the comparisons involve nested models under KLIC, see our discussion in the last paragraph of Section 3.2. This explains that both Type I and Type II errors are close to zero when $n = 500$, an ideal outcome that is not guaranteed when $\mathcal{M}^*_{\text{klic}}$ includes non-nested models.[7]

6 Empirical Applications

6.1 US Inflation Forecasts: Stock & Watson (JME, 1999) Revisited

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

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

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

where u_t is the unemployment rate, $\mathbf{L}$ is the lag polynomial operator, and e_{t+h} is the long-horizon inflation forecast innovation. Note that the natural rate hypothesis is not imposed on the Phillips curve (4) and that inflation as a regressor is in its first difference. Stock and Watson also forecast inflation with (4) where the unemployment rate u_t is replaced with different macro variables.[8]

The entire sample runs from 1959:M1 to 1997:M9. Following Stock and Watson, we study the properties of their forecasting models on the pre- and post-1984 subsamples of 1970:M1-1983:M12 and 1984:M1-1996:M9.[9] The former subsample contains the great inflation of the 1970s and the rapid disinflation of the early 1980s. Inflation does not exhibit this volatile behavior in the post-1984 subsample. We follow Stock and Watson in order to replicate their inflation forecasts. However, our MCS bootstrap implementation, which is described in Section 3, relies on an assumption that $d_{ij,t}$ is stationary. This is not plausible when the parameters are estimated with a recursive estimation scheme, as was used in Stock and Watson (1999). We

---

[7]In an unreported simulation study where $\mathcal{M}^*_{\text{klic}}$ was designed to include non-nested models, we found the frequency by which $\mathcal{M}^*_{\text{klic}} \subset \widehat{\mathcal{M}}^*_{90\%}$ converge to 90%.

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

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

avoid this problem by following Giacomini and White (2006) and present empirical results that are based on parameters estimated over a rolling window with a fixed number of observations.[10] Regressions are estimated on data that begins no earlier than 1960:M2, although lagged regressors impinge on observations back to 1959:M1.

Table 4: MCS for simple regression-based inflation forecasts.

| Variable | Trans | PUNEW | | | | GMDC | | | |
| | | 1970-1983 | | 1984-1996 | | 1970-1983 | | 1984-1996 | |
		RMSE	p_{MCS}	RMSE	p_{MCS}	RMSE	p_{MCS}	RMSE	p_{MCS}
No change (month)		3.290	.005	2.140	.099	2.208	.083	1.751	.000
No change (year)	-	2.798	.015	1.207	1.00**	2.100	.204*	0.888	1.00**
uniar	-	2.802	.011	1.330	.851**	2.026	.278**	1.070	.547**
'Gaps' specifications									
dtip	DT	2.597	.110*	1.475	.785**	2.103	.177*	1.050	.554**
dtgmpyq	DT	2.751	.046	1.691	.467**	2.090	.296**	1.125	.468**
dtmsmtq	DT	2.202	.896**	1.704	.671**	1.806	.575**	1.046	.554**
dtlpnag	DT	2.591	.140*	1.433	.825**	2.132	.137*	1.026	.554**
ipxmca	LV	2.609	.122*	1.318	.851**	2.040	.468**	1.034	.554**
hsbp	LN	2.114	1.00**	1.582	.754**	1.967	.550**	1.034	.554**
lhmu25	LV	2.968	.015	1.439	.754**	2.231	.122*	1.040	.554**
First difference specifications									
ip	DLN	2.344	.485**	1.393	.851**	1.946	.509**	1.058	.554**
gmpyq	DLN	2.306	.881**	1.524	.624**	1.709	1.00**	1.158	.254**
msmtq	DLN	2.158	.896**	1.391	.851**	1.857	.575**	1.066	.554**
lpnag	DLN	2.408	.606**	1.341	.851**	1.940	.509**	1.027	.554**
dipxmca	DLV	2.379	.261**	1.353	.851**	1.903	.575**	1.041	.554**
dhsbp	DLN	2.850	.008	1.456	.804**	2.076	.137*	1.070	.547**
dlhmu25	DLV	2.383	.314**	1.440	.754**	2.035	.190*	1.065	.547**
dlhur	DLV	2.296	.763**	1.429	.804**	1.904	.516**	1.067	.547**
Phillips curve									
LHUR		2.637	.071	1.388	.851**	2.076	.184*	1.162	.480**

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

We compute the MCS across all of the Stock and Watson inflation forecasting models. This includes the Phillips curve model (4), the inflation forecasting equation that runs through all of the macro variables considered by Stock and Watson, a univariate autoregressive model, and two no-change forecasts. The first no-change forecast is the past month's inflation rate and the second no-change forecast uses the past year's inflation rate as its forecast. The former matches the no-change forecast in Stock and Watson (1999) and

---

[10]The corresponding empirical results that are based on parameters that are estimated with the recursive scheme, as was used in Stock and Watson (1999), are available in a separate appendix, see Hansen, Lunde, and Nason (2009). Although our assumption do not justify the recursive estimation scheme, it produces pseudo-MCS results that are very similar to those obtained under the rolling window estimation scheme.

the latter matches the no-change forecast in Atkeson and Ohanian (2001). Stock and Watson also present results for forecast combinations and forecasts based on principal component indicator variables.[11]

Tables 4-5 report (the level of) the root mean square error (RMSE) and MCS p-values for each of the inflation forecasting models. The first column of Table 4 also lists the transformation of the macro variable employed by the forecasting equation.

Our Table 4 matches the results reported in Stock and Watson (1999, table 2). The initial model space $\mathcal{M}^0$ is filled with a total of 19 models. The results for the two no-change forecasts and the AR(p) are the first three rows of Table 4. The RMSEs and the p-values for the Phillips curve forecasting model (4) appear in the bottom row of our Table 4. The rest of the rows of Table 4 are the 'gap' and 'first difference' specifications of Stock and Watson's aggregate activity variables that appear in place of u_t in inflation forecasting equation (4). The 'gap' variables are computed with a one-sided Hodrick and Prescott (1997) filter, see Stock and Watson (1999, p. 301) for details.[12]

A glance at Table 4 reveals that the MCS of subsamples 1970:M1-1983:M12 and 1984:M1-1996:M9 are strikingly different for both *PUNEW* and *GMDC*. The MCS of the pre-1984 subsample places nine forecasting models in *PUNEW*-$\widehat{\mathcal{M}}^*_{75\%}$ and 11 models in *GMDC*-$\widehat{\mathcal{M}}^*_{75\%}$. For the post-1984 subsample, all but one model ends up in $\widehat{\mathcal{M}}^*_{75\%}$ for both *PUNEW* and *GMDC*. The only model that is consistently kicked out of these MCSs is the 'monthly' no-change forecast, which uses last month's inflation rate as its forecast.

Another intriguing feature of Table 4 is the inflation forecasting models that reside in the MCS when faced with the 1970:M1-1983:M12 subsample. The nine models that are in *PUNEW*-$\widehat{\mathcal{M}}^*_{75\%}$ are driven by macro variables related either to real economic activity (e.g., manufacturing and trade, capacity utilization, and building permits) or to the labor market. The labor market variables are *lpnag* (employees on nonagricultural payrolls) and *dlhur* (first difference of the unemployment rate, all workers 16 years and older). Thus, there is labor market information that is important for predicting inflation during the pre-1984 subsample. This result is consistent with traditional Keynesian measures of aggregate demand.

Table 4 also shows that there are two levels and six first difference specifications of the forecasting equation that consistently appear in $\widehat{\mathcal{M}}^*_{75\%}$ using the 1970:M1-1983:M12 subsample. On this subsample, only *msmtq* (total real manufacturing and trade) is consistently embraced by *PUNEW*- and *GMDC*-$\widehat{\mathcal{M}}^*_{75\%}$ whether in levels or first differences. In summary, we interpret these variables as signals about the anticipated path either of real aggregate demand or real aggregate supply that help to predict inflation out-of-sample in the pre-1984 subsample.

There are several more inferences to draw from Table 4. These concern the two types of no change forecasts whose predictive accuracy is strikingly different. The no-change (month) forecast fails to appear in $\widehat{\mathcal{M}}^*_{75\%}$ either on the pre-1984 or on the post-1984 subsamples, whereas the no-change (year) forecast

---

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

[12]The MCS p-values are computed using a block size of $l = 12$ in the bootstrap implementation. The MCS p-values are qualitatively similar when computed with $l = 6$, and $l = 9$. These are reported in a separate appendix, see Hansen, Lunde, and Nason (2009).

finds its way into $\widehat{\mathcal{M}}^*_{75\%}$ for the post-1984 subsample, but not the 1970:M1-1983:M12 subsample. These results are especially of interest because the no-change (year) forecast yields the best inflation forecasts on the 1984:M1-1996:M9 subsample for both *PUNEW* and *GMDC*. These empirical results for the no-change inflation forecasts are interesting because they reconcile the results of Stock and Watson (1999) with those of Atkeson and Ohanian (2001). Stock and Watson (1999, p. 327) find that: "The conventionally speci-fied Phillips curve, based on the unemployment rate, was found to perform reasonably well. Its forecasts are better than univariate forecasting models (both autoregressions and random walk models)". In contrast, Atkeson and Ohanian (2001, p. 10) conclude that: "economists have not produced a version of the Phillips curve that makes more accurate inflation forecasts than those from a naive model that presumes inflation over the next four quarters will be equal to inflation over the last four quarters." The source of the disagree-ment is that Stock and Watson and Atkeson and Ohanian study different no-change inflation forecasts. The no-change forecast Stock and Watson (1999) deploy is last month's inflation rate, whereas the no-change forecasts in Atkeson and Ohanian (2001) is the past year's inflation rate.

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

Table 5 generates MCSs using factor models and forecast combination methods that replicates the set of forecasts in Stock and Watson (1999, table 4). They combine a large set of inflation forecasts from an array of 168 models using sample means, sample medians, and ridge estimation to produce forecast weighting schemes. The other forecasting approach depends on principal components of the 168 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.

The factor models and forecast combination methods produce inflation forecasts that are in general better than those in Table 4. The forecasts constructed from 'All indicators' and 'Real activity indicators' in Panels A and B do particularly well across the board. Interestingly, the best forecast during the 1970:M1-1983:M12 subsample is the one-factor 'All indicators' model, while the second best is the one-factor 'Real activity indicators' model. Most of the forecasts constructed from the 'Money' variables do not find their way into the MCSs.

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

Table 5: MCS results for shrinkage-type inflation forecasts.

| Variable | PUNEW | | | | GMDC | | | |
| | 1970-1983 | | 1984-1996 | | 1970-1983 | | 1984-1996 | |
	RMSE	p_{MCS}	RMSE	p_{MCS}	RMSE	p_{MCS}	RMSE	p_{MCS}
No change (month)	3.290	.012	2.140	.000	2.208	.009	1.751	.000
No change (year)	2.798	.039	1.207	1.00**	2.100	.132*	0.888	1.00**
Univariate	2.802	.019	1.330	.834**	2.026	.067	1.070	.540**
Panel A. All indicators								
Mul. factors	2.367	.385**	1.407	.101*	2.105	.218*	1.013	.721**
1 factor	2.106	1.00**	1.351	.258**	1.746	1.00**	1.038	.721**
Comb. mean	2.423	.158*	1.269	.939**	1.880	.664**	1.030	.721**
Comb. median	2.585	.051	1.294	.939**	1.939	.445**	1.055	.598**
Comb. ridge reg.	2.121	.985**	1.318	.939**	1.918	.664**	1.013	.721**
Panel B. Real activity indicators								
Mul. factors	2.245	.737**	1.416	.035	1.959	.445**	0.990	.721**
1 factor	2.115	.985**	1.347	.478**	1.774	.718**	1.041	.721**
Comb. mean	2.284	.737**	1.263	.939**	1.827	.718**	1.012	.721**
Comb. median	2.329	.633**	1.284	.939**	1.854	.664**	1.038	.704**
Comb. ridge reg.	2.160	.737**	1.326	.923**	1.888	.664**	1.013	.721**
Panel C. Interest rates								
Mul. factors	2.828	.037	1.512	.008	2.215	.013	1.294	.017
1 factor	2.776	.054	1.463	.006	2.111	.012	1.102	.243*
Comb. mean	2.474	.151*	1.349	.174*	1.935	.445**	1.060	.598**
Comb. median	2.567	.127*	1.377	.049	1.974	.402**	1.066	.592**
Comb. ridge reg.	2.436	.266**	1.372	.101*	1.962	.313**	1.052	.598**
Panel D. Money								
Mul. factors	2.801	.027	1.340	.740**	2.028	.029	1.075	.084
1 factor	2.805	.022	1.352	.258**	2.027	.045	1.104	.038
Comb. mean	2.742	.037	1.390	.035	2.033	.017	1.088	.023
Comb. median	2.752	.037	1.340	.514**	2.032	.013	1.077	.140*
Comb. ridge reg.	2.721	.037	1.446	.010	2.013	.192*	1.088	.017
Phillips curve								
LHUR	2.637	.054	1.388	.035	2.076	.045	1.162	.598**

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

6.2 Likelihood-Based Comparison of Taylor-Rule Models

Monetary policy is often evaluated with the Taylor (1993) rule. A Taylor rule summarizes the objectives and constraints that defines monetary policy by mapping (implicitly) from this decision problem to the path of the short-term nominal interest rate. A canonical monetary policy loss function penalizes the decision maker for volatility in inflation against its target and output volatility around its trend. The mapping generates a Taylor rule that has the interest rate respond to inflation and output deviations from trend. Thus, Taylor rules measure ex post the success monetary policy has had at meeting the goals of keeping inflation close

to target and output at trend. Taylor (1999), Clarida, Galí, and Gertler (2000), and Orphanides (2003) are leading examples of using Taylor rules to evaluate actual monetary policy, while McCallum (1999) provides an introduction for consumers of monetary policy rules.

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

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

where R_t denotes the short-term nominal interest rate, π_t is inflation, y_t equals deviations of output from trend (*i.e.*, the output gap), and the error term, v_t, is assumed to be a martingale difference process. The Taylor principle is satisfied if $\sum_{j=1}^{p_\pi} \gamma_{\pi,j}$ exceeds one because a one percent rise in the sum of p_π lags of inflation indicates that R_t should rise by more than 100 basis points. The monetary policy response to real side fluctuations is given by $\sum_{j=1}^{p_y} \gamma_{y,j}$ on the p_y lags of the output gap. The intercept γ_0 is the equilibrium steady state real rate plus the target inflation rate (weighted by $1 - \sum_{j=1}^{p_\pi} \gamma_{\pi,j}$). The Taylor rule regression (5) includes lagged interest, R_{t-1}, which may be interpreted as interest rate smoothing by the central bank. Alternatively, the lagged interest rate could be interpreted as a proxy for other determinants of the interest rate that are not captured by the regression (5). Note also that the Taylor rule regression (5) avoids issues that arise in the estimation of simultaneous equation systems because contemporaneous inflation, π_t, and the output gap, y_t, are not regressors, only lags of these variables are. In this case, structural interpretations have to be applied to the Taylor rule regression (5) with care.

The Taylor rule regression (5) is estimated by ordinary least squares on a U.S. sample that runs from 1979:Q1 to 2006:Q4. Table 6 provides details about the data used to estimate the Taylor rule regression.[13] The (effective) federal funds rate defines the Taylor rule policy rate, R_t. The growth rate of the implicit GDP deflator is our measure of inflation, π_t. The cyclical component of the Hodrick and Prescott (1997) filter is applied to real GDP to obtain estimates of the output gap, y_t. We also employ two real activity variables to fill out the model space and to act as alternatives to the output gap. These real activity variables are the Baxter and King (1999) filtered unemployment rate gap, ur_t, and the Nason and Smith (2008) measure of real unit labor costs, $rulc_t$. We compute the Baxter-King ur_t using the maximum likelihood-Kalman filter methods of Harvey and Trimbur (2003).

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

---

[13]We have generated results on a shorter post-1984 sample. Omitting the volatile 1979-1983 period from the analysis does not substantially change our results, beyond the loss of information that one would expect with a shorter sample. These results available in a separate appendix found in Hansen, Lunde, and Nason (2009).

Table 6: Taylor Rule Regression Data Set.

	Observable	Construction
Dependent Variable		
R_t: Interest rate	Effective Fed Funds Rate (EFFR), $R_{fed\,funds,t}$	Temporally aggregate daily return (annual rate) to quarterly, $R_t = 100 \times \ln[1 + \frac{R_{fed\,funds,t}}{100}]$
Independent Variables		
π_t: Inflation,	Implicit GDP Deflator, P_t, Seasonally Adjusted (SA)	$\pi_t = 400 \times \ln[P_t/P_{t-1}]$
y_t: Output gap	$\ln Q_t -$ trend Q_t, *i.e.* Transitory Component of Output. Where Q_t is Real GDP in Billions of Chained 2000 \$, SA at Annual Rates.	Apply Hodrick-Prescott filter to $\ln Q_t$
ur_t: Unemployment rate gap	$UR_t -$ trend UR_t, *i.e.* transitory component of UR_t. Where UR_t is the is the civilian unemployment rate, SA.	Temporally aggregate monthly to quarterly frequency to get UR_t. Apply Baxter-King Filter to UR_t
$rulc_t$: Real unit labor costs	The cointegrating residual of Nominal $ULC_t(= LS_t - LS_t)$ and $\ln P_t$. LS_t is Labor Share, i.e. log of compensation per hour in the non-farm business sector. LP_t is Labor Productivity, i.e. log of output per hour of all persons non-farm business sector	$rulc_t = LS_t - LP_t$ $-\hat{a}_0 - \hat{a}_1 t - \hat{a}_2 \ln P_t$

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

We present results of applying the MCS and likelihood-based criteria to the choice of the best Taylor rule regression (5) and AR(1) regressions in Tables 7 and 8. Table 7 reports $Q(\mathcal{Z}_j, \hat{\theta}_j)$ (the log-likelihood function multiplied by minus two); the bootstrap estimate of the effective degrees of freedom, $\hat{k}^\star$; and the realizations of the three empirical criteria, KLIC, AIC*, and BIC*. The numbers surrounded by parentheses in columns headed by KLIC, AIC*, and BIC*, are the MCS p-values, and an asterisk identifies the specifications that enter $\widehat{\mathcal{M}}^*_{90\%}$. Table 8 lists estimates of the regressions models that are in $\widehat{\mathcal{M}}^*_{90\%}$ along with their corresponding t-statistics in parentheses. The t-statistics are based on robust standard errors following Newey and West (1987).

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

We report the maximized log-likelihood function (multiplied by minus two), the effective degress of freedom, and the three criteria, KLIC, AIC* and BIC*, along with the corresponding MCS p-values. The regression models in $\widehat{\mathcal{M}}^*_{90\%}$ and $\widehat{\mathcal{M}}^*_{75\%}$ are identified by one and two asterisks, respectively. See the text and Table 6 for variable mnemonics and definitions.

Table 7 shows that the MCS procedure selects 10 to 13 of the 25 possible regressions depending on the information criteria. The lagged nominal rate R_{t-1} is the one regressor common to the regressions that enter $\widehat{\mathcal{M}}^*_{90\%}$ for the KLIC, AIC*, and BIC*. Besides the AR(1), $\widehat{\mathcal{M}}^*_{90\%}$ consists of the six Taylor rule specifications that nest the AR(1). Under the KLIC and AIC*, the Taylor rule regressions include all one or two lag combinations of π_t, y_t, ur_t, and $rulc_t$. The BIC produces a smaller $\widehat{\mathcal{M}}^*_{90\%}$ because it ejects the two lag Taylor rule specifications that exclude lagged π_t. Thus, the Taylor rule regression-MCS example finds the BIC tends to settle on more parsimonious models. This is to be expected, given its larger penalty on model complexity.

The AR(1) falls into $\widehat{\mathcal{M}}^*_{90\%}$ under the KLIC, AIC*, and BIC*. Although the first line of Table 7 shows that the AR(1) has the largest $Q(\mathcal{Z}_j, \hat{\theta}_j)$ of the regressions covered by $\widehat{\mathcal{M}}^*_{90\%}$, the MCS recruits the AR(1) because it has a relatively small estimate of the effective degrees of freedom, $\hat{k}^\star$. It is important to keep in mind that estimates of the effective degrees of freedom are larger than the number of free parameters in each of the models. This reflects the fact that the Gaussian model is misspecified. For example, the conventional AIC penalty (that doubles the number of free parameters) is misleading in the context of misspecified models, see Takeuchi (1976), Sin and White (1996), and Hong and Preston (2008).

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

Table 8 provides information about the regressions in $\widehat{\mathcal{M}}^*_{90\%}$-KLIC. The shaded area identifies the models in $\widehat{\mathcal{M}}^*_{75\%}$-BIC*. First, note that the estimated Taylor rules always satisfy the Taylor principle (*i.e.*, $\hat{\gamma}_{\pi,1} > 1$ or $\hat{\gamma}_{\pi,1} + \hat{\gamma}_{\pi,2} > 1$). The coefficients associated with real activity variables have insignificant t-statistics in most cases. Only the first lag of the output gap produces a positive coefficient with a t-ratio above two in the first Taylor rule regression listed in Table 8. Moreover, the statistically insignificant coefficients for the unemployment rate gap and real unit labor costs variables often have counter intuitive signs. Finally, the estimates of ρ are between 0.83 and 0.87 in the Taylor rule regressions that include a lag of π_t, which suggests interest rate smoothing.[14]

The fact that the MCS cannot settle on a single specification is not a surprising result. Monetary policy-makers almost surely rely on a more complex information set than can be summarized by a simple model. Furthermore, any real activity variable is an imperfect measure of the underlying state of the economy, and

---

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

there are important and unresolved issues regarding the measurement of 'gap' and marginal cost variables, which translate into uncertainty about the proper definitions of the real activity variables.

Table 8: Regression models in $\widehat{\mathcal{M}}^*_{90\%}$-KLIC.

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

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

7   Summary and Concluding Remarks

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

It is important to understand the principle of the MCS procedure in applications. The MCS is constructed such that inference about the 'best' follows the conventional meaning of the word 'significance'. Although the MCS will contain only the best model(s) asymptotically, it may contain several poor models in finite samples. A key feature of the MCS procedure is that a model is discarded only if it is found to be significantly inferior to another model. Models remain in the MCS until proven inferior, which has the implication that not all models in the MCS may be judged good models.[15]

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

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

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

---

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

funds rate. Further we find that the first-order autoregression also enters the MCS. Thus, the U.S. data lack the information to resolve precisely which Taylor rule specification best describes the data.

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

References

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

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

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

BAXTER, M., AND R. G. KING (1999): "Measuring business cycles: Approximate bandpass filters for economic time series," *Review of Economics and Statistics*, 81, 575–593.

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

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

CAVANAUGH, J. E., AND R. H. SHUMWAY (1997): "A bootstrap variant of AIC for state-space model selection," *Statistica Sinica*, 7, 473–496.

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

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

CLARIDA, R., J. GALÍ, AND M. GERTLER (2000): "Monetary policy rules and macroeconomic stability: Evidence and some theory," *Quarterly Journal of Economics*, 115, 147–180.

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

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

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

DOORNIK, J. A. (2006): *Ox: An Object-Orientated Matrix Programming Langguage*. Timberlake Consultants Ltd., London, 5 edn.

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

EFRON, B. (1983): "Estimating the error rate of a preduction rule: improvement on cross-validation," *Journal of the American Statistical Association*, 78, 316–331.

——— (1986): "How biased is the apparent error rate of a prediction rule?," *Journal of the American Statistical Association*, 81, 461–470.

ENGLE, R. F., AND S. J. BROWN (1985): "Model Selection for Forecasting," *Journal of Computation in Statistics*.

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

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

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. (2003a): "Asymptotic Tests of Composite Hypotheses," http://www.stanford.edu/people/peter.hansen.

———— (2003b): "Regression Analysis with Many Specifications: A Bootstrap Method to Robust Inference," Mimeo http://www.stanford.edu/people/peter.hansen.

———— (2005): "A Test for Superior Predictive Ability," *Journal of Business and Economic Statistics*, 23, 365–380.

HANSEN, P. R., A. LUNDE, AND J. M. NASON (2009): "Supplements for Model Confidence Sets," Web appendix, Aarhus School of Business, Aarhus University.

HARVEY, A. C., AND T. M. TRIMBUR (2003): "General model-based filters for extracting cycles and trend in economic time series.," *Review of Economics and Statistics*, 85, 244–255.

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.

HONG, H., AND B. PRESTON (2008): "Bayesian Averaging, Prediction and Nonnested Model Selection," NBER Working Paper No. W14284.

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 (2006): "On the Selection of Forecasting Models," *Journal of Econometrics*, 130, 273–306.

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

KILIAN, 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.

LEHMANN, E. L., AND J. P. ROMANO (2005): *Testing Statistical Hypotheses*. John Wiley and Sons, New York, 3nd edn.

MCCALLUM, B. T. (1999): "Issues in the design of monetary policy rules," in *Handbook of Macroeconomics*, ed. J. B. Taylor, and M. Woodford, vol. IV, pp. 1483–1530. New York: North-Holland Elsevier.

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

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

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

ORPHANIDES, A. (2003): "Historical monetary policy analysis and the Taylor rule," *JME*, 50, 983–1022.

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

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

ROMANO, J. P., A. M. SHAIKH, AND M. WOLF (2008): "Formalized data snooping based on generalized error rates," *Econometric Theory*, 24, 404–447.

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

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

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

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 the Phillips Curve with Gun and Camara," in *Inflation, Trade, and Taxes*, ed. 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 (1997a): "How Precise are Estimates of the Natural Rate of Unemployment?," in *Reducing Inflation: Motivation and Strategy*, ed. C. Romer, and D. Romer, pp. 195–242. Chicago: University of Chicago Press.

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

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

——— (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.

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

TAYLOR, J. B. (1993): "Discretion versus policy rules in practice," *Carnegie-Rochester Conference Series on Public Policy*, 39, 195–214.

——— (1999): "A historical analysis of monetary policy rules," in *Monetary Policy Rules*, ed. J. B. Taylor, pp. 319–341. Chicago: University of Chicago Press.

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

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. (1994): *Estimation, Inference and Specification Analysis*. Cambridge University Press, Cambridge.

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

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

Appendix: Model Confidence Sets

Bootstrap Procedure, Inflation forecasting, Regression simulations and Taylor Rules

Peter R. Hansen[a], **Asger Lunde**[b], and **James M. Nason**[c]

[a] *Stanford University, Department of Economics, 579 Serra Mall,*
Stanford, CA 94305-6072, USA & CREATES

[b] *Aarhus School of Business, University of Aarhus, Department of Marketing and Statistics,*
Haslegaardsvej 10, 8210 Aarhus V, Denmark & CREATES

[c] *Federal Reserve Bank of Atlanta, Research Department, 1000 Peachtree St., NE,*
Atlanta, GA 30309, USA

May 5, 2009

Contents

List of Tables

1. Bootstrap Procedure

This section describes the bootstrap implementation of the MCS procedure used in the forecasting application.

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 $b = 1, \ldots, B$:

      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.

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

   (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 $\zeta^*_{b,i} \equiv \bar{L}^*_{b,i} - \bar{L}_i$, 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

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

Alternatively one can define $\widehat{\text{var}}(\cdot)$ to be its analytical value given the employed bootstrap scheme. Now define $t_{i\cdot} \equiv \bar{d}_{i\cdot}/\sqrt{\widehat{\text{var}}(\bar{d}_{i\cdot})}$ and calculate the test statistic $T_{\max} = \max_i t_{i\cdot}$.

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

$$T^*_{b,\max} = \max_i t^*_{b,i\cdot}, \quad \text{for} \quad b = 1, \dots, B,$$

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 $H_{0,\mathcal{M}}$ is given by

$$P_{H_{0,\mathcal{M}}} \equiv \frac{1}{B} \sum_{b=1}^{B} 1_{\left\{ T_{\max} > T^*_{b,\max} \right\}},$$

where $1_{\{\cdot\}}$ is the indicator function.

(d) If $P_{H_{0,\mathcal{M}}} < \alpha$, where α is the level of the test, then $H_{0,\mathcal{M}}$ is rejected and $e_{\mathcal{M}} \equiv \arg\max_i t_{i\cdot}$ is eliminated from $\mathcal{M}$.

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

1.1. *Justification of bootstrap implementation*

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

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

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

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

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

Under the null hypothesis, the distribution of $T_{\max}$ is approximated by

$$\max_i \left(\hat{D}^{-1/2} n^{1/2} (\bar{Z}^*_b - \bar{Z}) \right)_i = \max_i \left(\text{diag}(\widehat{\text{var}}(\bar{d}_{1\cdot}), \dots, \widehat{\text{var}}(\bar{d}_{1\cdot}))^{-1/2} (\bar{Z}^*_b - \bar{Z}) \right)_i$$

$$= \max_i \frac{\bar{d}^*_{b,i\cdot} - \bar{d}_{i\cdot}}{\sqrt{\widehat{\text{var}}(\bar{d}_{i\cdot})}} = \max_i \frac{\zeta^*_{b,i} - \zeta^*_{b,\cdot}}{\sqrt{\widehat{\text{var}}(\bar{d}_{i\cdot})}} = \max_i t^*_{b,i\cdot}$$

$$= T^*_{b,\max}.$$

2. Inflation Forecasting

Here we investigate the sensitivity of our MCS for Stock and Watson (JME,1999) to the choice of estimation scheme and equivalence test.

2.1. *Sensitivity Analysis of MCS's to estimation scheme and the choice of test for EPA*

The Tables corresponding to Table 2 in Stock and Watson (JME,1999) are as follows.

Table A.1 use an expanding recursive estimation scheme.

Table A.2 use a rolling estimation scheme.

The Tables corresponding to Table 4 in Stock and Watson (JME,1999) are as follows.

Table A.3 use a expanding recursive estimation scheme.

Table A.4 use an rolling estimation scheme.

These tables also display the root MSE of each model.

3. Regression Simulation

Here we present simulation results for $\beta^2 = 0.1, 0.5, 0.9$ for the simulation experiment in section 5.2.

Tables A.5-A.7 report the fraction that each of the specifications is in the MCS for the KLIC, AIC* and BIC*.

Tables A.9-A.10 report the average MCS p-value for the KLIC, AIC* and BIC*.

4. Taylor Rules

The Tables are as follows.

Table A.11 gives MCS results when the models are estimated on a sample period covering 1979Q1 to

2006Q4.

Table A.12 gives MCS results when the sample period only span 1984Q1 to 2006Q4 .

References

POLITIS, D. N., AND J. P. ROMANO (1994): "The Stationary Bootstrap," *Journal of the American Statistical Association*, 89, 1303–1313. 1

Table A.1: MCS p-values for Stock and Watson JME (1999, table 2) (recursive scheme).

Variable	Trans	PUNEW: 1970-1983				PUNEW: 1984-1996				GMDC: 1970-1983				GMDC: 1984-1996			
		RMSE	p_6	p_9	p_{12}	RMSE	p_6	p_9	p_{12}	RMSE	p_6	p_9	p_{12}	RMSE	p_6	p_9	p_{12}
No change (month)		3.290	.003	.005	.010	2.140	.007	.009	.007	2.208	.023	.045	.057	1.751	.003	.001	.001
No change (year)	-	2.798	.017	.029	.034	1.207	1.00**	1.00**	1.00**	2.100	.040	.078	.089	0.888	1.00**	1.00**	1.00**
uniar	-	2.675	.014	.024	.029	1.360	.900**	.899**	.897**	1.941	.078	.112*	.121*	1.082	.311**	.320**	.305**
'Gaps' specifications																	
dtip	DT	2.519	.045	.061	.060	1.310	.900**	.921**	.936**	1.913	.094	.119*	.121*	1.043	.404**	.417**	.408**
dtgmpyq	DT	2.644	.011	.019	.024	1.446	.548**	.585**	.552**	2.067	.038	.074	.088	1.103	.254**	.253**	.235*
dtmsmtq	DT	2.341	.158*	.166*	.158*	1.280	.900**	.921**	.936**	1.844	.109*	.152*	.139*	1.007	.444**	.468**	.472**
dtlpnag	DT	2.482	.051	.065	.060	1.323	.900**	.921**	.936**	2.024	.065	.108*	.116*	1.012	.444**	.468**	.472**
ipxmca	LV	2.373	.107*	.119*	.118*	1.264	.900**	.921**	.936**	1.887	.098	.142*	.139*	1.026	.444**	.468**	.472**
hsbp	LN	2.205	.765**	.769**	.766**	1.392	.873**	.884**	.884**	1.829	.109*	.152*	.139*	0.993	.444**	.468**	.472**
lhmu25	LV	2.433	.054	.065	.060	1.401	.866**	.871**	.860**	1.937	.071	.112*	.114*	1.055	.410**	.424**	.447**
First difference specifications																	
ip	DLN	2.384	.090	.065	.060	1.429	.704**	.687**	.668**	1.819	.109*	.152*	.139*	1.115	.252**	.248*	.228*
gmpyq	DLN	2.233	.622**	.576**	.506**	1.532	.405**	.450**	.417**	1.565	1.00**	1.00**	1.00**	1.149	.260**	.259**	.240*
msmtq	DLN	2.169	1.00**	1.00**	1.00**	1.353	.900**	.921**	.936**	1.778	.109*	.152*	.139*	1.062	.410**	.424**	.447**
lpnag	DLN	2.308	.176*	.175*	.165*	1.317	.900**	.921**	.936**	1.809	.109*	.152*	.139*	1.009	.444**	.468**	.472**
dipxmca	DLV	2.355	.107*	.119*	.118*	1.456	.643**	.684**	.668**	1.839	.109*	.144*	.139*	1.128	.229*	.225*	.203*
dhsbp	DLN	2.701	.012	.019	.024	1.405	.866**	.874**	.860**	1.969	.059	.100	.105*	1.077	.358**	.367**	.354**
dlhmu25	DLV	2.352	.107*	.119*	.118*	1.474	.223*	.244*	.224*	1.878	.098	.137*	.124*	1.103	.241*	.241*	.222*
dlhur	DLV	2.321	.176*	.175*	.165*	1.451	.432**	.462**	.436**	1.843	.109*	.144*	.139*	1.088	.300**	.313**	.298**
Phillips curve																	
LHUR		2.387	.051	.065	.060	1.371	.866**	.874**	.860**	1.939	.078	.119*	.121*	1.050	.410**	.424**	.418**

RMSEs and MCS p-values for the different forecasts. For each p-value the bootstrap block length is identified by the subscript. The forecasts in $\widehat{\mathcal{M}}^*_{90\%}$ and $\widehat{\mathcal{M}}^*_{75\%}$ are identified by one and two asterisks, respectively.

Table A.2: MCS p-values for Stock and Watson JME (1999, table 2) (rolling scheme).

Variable	Trans	PUNEW: 1970-1983				PUNEW: 1984-1996				GMDC: 1970-1983				GMDC: 1984-1996			
		RMSE	p_6	p_9	p_{12}	RMSE	p_6	p_9	p_{12}	RMSE	p_6	p_9	p_{12}	RMSE	p_6	p_9	p_{12}
No change (month)		3.290	.001	.002	.005	2.140	.165*	.132*	.099	2.208	.073	.087	.083	1.751	.002	.001	.000
No change (year)	-	2.798	.008	.012	.015	1.207	1.00**	1.00**	1.00**	2.100	.147*	.202*	.204*	0.888	1.00**	1.00**	1.00**
uniar	-	2.802	.002	.004	.011	1.330	.853**	.861**	.851**	2.026	.245*	.291**	.278**	1.070	.524**	.550**	.547**
'Gaps' specifications																	
dtip	DT	2.597	.045	.059	.110*	1.475	.793**	.791**	.785**	2.103	.120*	.177*	.177*	1.050	.524**	.552**	.554**
dtgmpyq	DT	2.751	.006	.009	.046	1.691	.414**	.470**	.467**	2.090	.263**	.305**	.296**	1.125	.448**	.470**	.468**
dtmsmtq	DT	2.202	.881**	.891**	.896**	1.704	.578**	.635**	.671**	1.806	.555**	.572**	.575**	1.046	.524**	.552**	.554**
dtlpnag	DT	2.591	.103*	.123*	.140*	1.433	.817**	.834**	.825**	2.132	.120*	.154*	.137*	1.026	.524**	.552**	.554**
ipxmca	LV	2.609	.090	.108*	.122*	1.318	.853**	.861**	.851**	2.040	.434**	.475**	.468**	1.034	.524**	.552**	.554**
hsbp	LN	2.114	1.00**	1.00**	1.00**	1.582	.718**	.762**	.754**	1.967	.527**	.549**	.550**	1.034	.524**	.552**	.554**
lhmu25	LV	2.968	.006	.009	.015	1.439	.793**	.777**	.754**	2.231	.058	.129*	.122*	1.040	.524**	.552**	.554**
First difference specifications																	
ip	DLN	2.344	.408**	.456**	.485**	1.393	.853**	.861**	.851**	1.946	.498**	.532**	.509**	1.058	.524**	.552**	.554**
gmpyq	DLN	2.306	.853**	.872**	.881**	1.524	.718**	.737**	.624**	1.709	1.00**	1.00**	1.00**	1.158	.303**	.301**	.254**
msmtq	DLN	2.158	.881**	.891**	.896**	1.391	.853**	.861**	.851**	1.857	.555**	.572**	.575**	1.066	.524**	.552**	.554**
lpnag	DLN	2.408	.554**	.583**	.606**	1.341	.853**	.861**	.851**	1.940	.527**	.545**	.509**	1.027	.524**	.552**	.554**
dipxmca	DLV	2.379	.199*	.231*	.261**	1.353	.853**	.861**	.851**	1.903	.555**	.572**	.575**	1.041	.524**	.552**	.554**
dhsbp	DLN	2.850	.003	.005	.008	1.456	.798**	.814**	.804**	2.076	.130*	.154*	.137*	1.070	.524**	.550**	.547**
dlhmu25	DLV	2.383	.253*	.285*	.314**	1.440	.793**	.777**	.754**	2.035	.228*	.191*	.190*	1.065	.524**	.550**	.547**
dlhur	DLV	2.296	.728**	.747**	.763**	1.429	.798**	.814**	.804**	1.904	.555**	.545**	.516**	1.067	.524**	.550**	.547**
Phillips curve																	
LHUR		2.637	.050	.064	.071	1.388	.853**	.861**	.851**	2.076	.144*	.186*	.184*	1.162	.463**	.483**	.480**

RMSEs and MCS p-values for the different forecasts. For each p-value the bootstrap block length is identified by the subscript. The forecasts in $\widehat{\mathcal{M}}^*_{90\%}$ and $\widehat{\mathcal{M}}^*_{75\%}$ are identified by one and two asterisks, respectively.

Table A.3: MCS p-values for Stock and Watson JME (1999, table 4) (recursive scheme).

Variable	PUNEW: 1970-1983				PUNEW: 1984-1996				GMDC: 1970-1983				GMDC: 1984-1996			
	RMSE	p_6	p_9	p_{12}	RMSE	p_6	p_9	p_{12}	RMSE	p_6	p_9	p_{12}	RMSE	p_6	p_9	p_{12}
No change (month)	3.290	.009	.015	.012	2.140	.000	.000	.000	2.208	.000	.002	.005	1.751	.000	.000	.000
No change (year)	2.798	.016	.025	.035	1.207	1.00**	1.00**	1.00**	2.100	.003	.008	.017	0.888	1.00**	1.00**	1.00**
Univariate	2.675	.016	.025	.035	1.360	.827**	.842**	.835**	1.941	.039	.080	.109*	1.082	.212*	.199*	.175*
Panel A. All indicators																
Mul. factors	2.158	.353**	.394**	.395**	1.291	.958**	.969**	.971**	1.894	.146*	.176*	.214*	0.964	.714**	.698**	.659**
1 factor	2.069	.781**	.772**	.718**	1.274	.958**	.969**	.971**	1.692	1.00**	1.00**	1.00**	1.002	.714**	.698**	.659**
Comb. mean	2.439	.022	.030	.037	1.289	.958**	.969**	.971**	1.853	.189*	.196*	.218*	1.036	.577**	.606**	.564**
Comb. median	2.550	.022	.025	.037	1.316	.944**	.952**	.955**	1.895	.129*	.149*	.177*	1.063	.340**	.342**	.325**
Comb. ridge reg.	2.209	.123*	.129*	.109*	1.280	.958**	.969**	.971**	1.842	.200*	.210*	.230*	1.019	.577**	.549**	.545**
Panel B. Real activity indicators																
Mul. factors	2.019	1.00**	1.00**	1.00**	1.357	.871**	.888**	.889**	1.792	.234*	.260**	.288**	0.946	.714**	.698**	.659**
1 factor	2.079	.781**	.772**	.718**	1.281	.958**	.969**	.971**	1.753	.256**	.273**	.292**	1.017	.577**	.606**	.564**
Comb. mean	2.346	.033	.042	.052	1.284	.958**	.969**	.971**	1.807	.221*	.224*	.246*	1.020	.577**	.558**	.545**
Comb. median	2.381	.026	.036	.052	1.299	.958**	.951**	.954**	1.831	.207*	.210*	.230*	1.036	.375**	.342**	.325**
Comb. ridge reg.	2.192	.191*	.205*	.192*	1.298	.958**	.969**	.971**	1.773	.256**	.260**	.288**	1.022	.577**	.499**	.485**
Panel C. Interest rates																
Mul. factors	2.585	.022	.030	.037	1.495	.088	.052	.026	1.976	.090	.111*	.141*	1.173	.176*	.177*	.155*
1 factor	2.524	.030	.042	.052	1.495	.015	.008	.003	2.038	.002	.006	.014	1.077	.306**	.302**	.284**
Comb. mean	2.424	.033	.042	.052	1.341	.903**	.913**	.915**	1.900	.129*	.149*	.169*	1.079	.199*	.187*	.166*
Comb. median	2.513	.022	.030	.037	1.336	.923**	.930**	.934**	1.912	.103*	.126*	.156*	1.078	.278**	.218*	.194*
Comb. ridge reg.	2.432	.038	.044	.052	1.368	.628**	.583**	.494**	1.943	.010	.015	.025	1.123	.163*	.153*	.134*
Panel D. Money																
Mul. factors	2.679	.016	.025	.031	1.360	.724**	.704**	.645**	1.933	.090	.102*	.124*	1.080	.228*	.239*	.215*
1 factor	2.679	.016	.025	.035	1.360	.805**	.819**	.810**	1.933	.090	.104*	.130*	1.080	.252**	.267**	.246*
Comb. mean	2.664	.016	.025	.035	1.350	.850**	.867**	.865**	1.964	.004	.017	.028	1.066	.449**	.499**	.485**
Comb. median	2.670	.016	.025	.035	1.348	.883**	.893**	.891**	1.954	.016	.037	.054	1.070	.378**	.388**	.370**
Comb. ridge reg.	2.638	.016	.025	.035	1.385	.363**	.296**	.214*	1.934	.090	.111*	.141*	1.121	.187*	.164*	.144*
Phillips curve																
LHUR	2.387	.026	.036	.045	1.371	.604**	.552**	.470**	1.939	.108*	.134*	.173*	1.050	.532**	.469**	.450**

RMSEs and MCS p-values for the different forecasts. For each p-value the bootstrap block length is identified by the subscript. The forecasts in $\widehat{\mathcal{M}}^*_{90\%}$ and $\widehat{\mathcal{M}}^*_{75\%}$ are identified by one and two asterisks, respectively.

Variable	PUNEW: 1970-1983				PUNEW: 1984-1996				GMDC: 1970-1983				GMDC: 1984-1996			
	RMSE	p_6	p_9	p_{12}	RMSE	p_6	p_9	p_{12}	RMSE	p_6	p_9	p_{12}	RMSE	p_6	p_9	p_{12}
No change (month)	3.290	.008	.013	.012	2.140	.000	.000	.000	2.208	.001	.004	.009	1.751	.000	.000	.000
No change (year)	2.798	.020	.036	.039	1.207	1.00**	1.00**	1.00**	2.100	.118*	.162*	.132*	0.888	1.00**	1.00**	1.00**
Univariate	2.802	.014	.018	.019	1.330	.804**	.829**	.834**	2.026	.029	.048	.067	1.070	.477**	.534**	.540**
Panel A. All indicators																
Mul. factors	2.367	.378**	.390**	.385**	1.407	.085	.125*	.101*	2.105	.068	.104*	.218*	1.013	.667**	.712**	.721**
1 factor	2.106	1.00**	1.00**	1.00**	1.351	.177*	.235*	.258**	1.746	1.00**	1.00**	1.00**	1.038	.667**	.712**	.721**
Comb. mean	2.423	.197*	.202*	.158*	1.269	.909**	.928**	.939**	1.880	.493**	.532**	.664**	1.030	.667**	.712**	.721**
Comb. median	2.585	.050	.053	.051	1.294	.909**	.928**	.939**	1.939	.409**	.448**	.445**	1.055	.548**	.596**	.598**
Comb. ridge reg.	2.121	.982**	.985**	.985**	1.318	.909**	.928**	.939**	1.918	.434**	.456**	.664**	1.013	.667**	.712**	.721**
Panel B. Real activity indicators																
Mul. factors	2.245	.732**	.698**	.737**	1.416	.021	.035	.035	1.959	.434**	.456**	.445**	0.990	.667**	.712**	.721**
1 factor	2.115	.982**	.985**	.985**	1.347	.406**	.464**	.478**	1.774	.685**	.715**	.718**	1.041	.667**	.712**	.721**
Comb. mean	2.284	.658**	.698**	.737**	1.263	.909**	.928**	.939**	1.827	.671**	.694**	.718**	1.012	.667**	.712**	.721**
Comb. median	2.329	.526**	.580**	.633**	1.284	.909**	.928**	.939**	1.854	.539**	.585**	.664**	1.038	.654**	.695**	.704**
Comb. ridge reg.	2.160	.783**	.729**	.737**	1.326	.892**	.913**	.923**	1.888	.532**	.532**	.664**	1.013	.667**	.712**	.721**
Panel C. Interest rates																
Mul. factors	2.828	.039	.028	.037	1.512	.005	.008	.008	2.215	.002	.005	.013	1.294	.008	.016	.017
1 factor	2.776	.060	.061	.054	1.463	.003	.004	.006	2.111	.002	.005	.012	1.102	.044	.128*	.243*
Comb. mean	2.474	.266**	.255**	.151*	1.349	.125*	.168*	.174*	1.935	.398**	.439**	.445**	1.060	.477**	.596**	.598**
Comb. median	2.567	.059	.061	.127*	1.377	.065	.049	.049	1.974	.321**	.376**	.402**	1.066	.478**	.596**	.592**
Comb. ridge reg.	2.436	.310**	.307**	.266**	1.372	.021	.103*	.101*	1.962	.226*	.281**	.313**	1.052	.524**	.596**	.598**
Panel D. Money																
Mul. factors	2.801	.016	.022	.027	1.340	.477**	.721**	.740**	2.028	.008	.017	.029	1.075	.150*	.072	.084
1 factor	2.805	.017	.028	.022	1.352	.210*	.252**	.258**	2.027	.016	.031	.045	1.104	.009	.019	.038
Comb. mean	2.742	.017	.028	.037	1.390	.021	.035	.035	2.033	.004	.009	.017	1.088	.021	.039	.023
Comb. median	2.752	.017	.028	.037	1.340	.724**	.512**	.514**	2.032	.002	.005	.013	1.077	.273**	.341**	.140*
Comb. ridge reg.	2.721	.018	.028	.037	1.446	.006	.010	.010	2.013	.068	.104*	.192*	1.088	.008	.016	.017
Phillips curve																
LHUR	2.637	.060	.061	.054	1.388	.021	.035	.035	2.076	.016	.031	.045	1.162	.477**	.596**	.598**

RMSEs and MCS p-values for the different forecasts. For each p-value the bootstrap block length is identified by the subscript. The forecasts in $\widehat{\mathcal{M}}^*_{90\%}$ and $\widehat{\mathcal{M}}^*_{75\%}$ are identified by one and two asterisks, respectively.

Table A.5: Simulation Experiment II: $\beta^2 = 0.1$, fraction in MCS

	$Q(\mathcal{Z}_j, \hat{\theta}_j)$		$\hat{k}^*$		KLIC		AIC* (TIC)		BIC*	
Panel A: n = 50										
$\rho =$	0.3	0.9	0.3	0.9	0.3	0.9	0.3	0.9	0.3	0.9
X_0	47.8	48.1	2.00	1.99	0.974	0.956	0.985	0.980	0.988	0.989
X_0, X_1	41.4	41.8	3.03	3.03	1.000	0.999	1.000	0.999	0.979	0.960
$X_0, \ldots, X_2$	40.3	40.8	4.09	4.09	1.000	0.999	0.998	0.999	0.928	0.922
$X_0, \ldots, X_3$	39.3	39.7	5.19	5.19	1.000	1.000	0.997	0.998	0.840	0.832
$X_0, \ldots, X_4$	38.2	38.5	6.33	6.33	1.000	1.000	0.994	0.996	0.676	0.653
$X_0, \ldots, X_5$	37.0	37.4	7.51	7.51	1.000	1.000	0.987	0.989	0.482	0.435
$X_0, \ldots, X_6$	35.8	36.2	8.75	8.74	1.000	1.000	0.970	0.965	0.312	0.250
X_0, X_2	46.3	42.8	3.03	3.03	0.979	0.997	0.984	0.999	0.944	0.959
X_0, X_2, X_3	45.0	41.5	4.09	4.09	0.982	0.998	0.978	0.999	0.812	0.914
$X_0, X_2, \ldots, X_4$	43.7	40.3	5.19	5.18	0.983	0.999	0.966	0.998	0.612	0.808
$X_0, X_2, \ldots, X_5$	42.5	39.2	6.33	6.32	0.983	0.999	0.946	0.995	0.422	0.609
$X_0, X_2, \ldots, X_6$	41.3	38.0	7.51	7.51	0.983	0.999	0.906	0.987	0.271	0.396
Panel B: n = 100										
$\rho =$	0.3	0.9	0.3	0.9	0.3	0.9	0.3	0.9	0.3	0.9
X_0	97.9	98.2	2.00	1.99	0.859	0.806	0.903	0.886	0.930	0.943
X_0, X_1	86.2	86.7	3.01	3.01	1.000	0.999	1.000	1.000	0.996	0.991
$X_0, \ldots, X_2$	85.2	85.6	4.03	4.03	1.000	1.000	0.994	1.000	0.926	0.973
$X_0, \ldots, X_3$	84.2	84.6	5.07	5.07	1.000	1.000	0.990	1.000	0.864	0.900
$X_0, \ldots, X_4$	83.1	83.5	6.12	6.12	1.000	1.000	0.987	1.000	0.677	0.684
$X_0, \ldots, X_5$	82.0	82.4	7.19	7.19	1.000	1.000	0.980	0.999	0.447	0.412
$X_0, \ldots, X_6$	80.9	81.4	8.28	8.27	1.000	1.000	0.972	0.993	0.268	0.220
X_0, X_2	95.9	88.7	3.01	3.00	0.881	0.994	0.900	0.998	0.847	0.988
X_0, X_2, X_3	94.4	87.2	4.03	4.03	0.894	0.997	0.884	0.999	0.628	0.957
$X_0, X_2, \ldots, X_4$	93.1	86.0	5.07	5.07	0.901	0.998	0.865	0.998	0.406	0.837
$X_0, X_2, \ldots, X_5$	91.8	84.9	6.12	6.12	0.904	0.998	0.826	0.997	0.242	0.591
$X_0, X_2, \ldots, X_6$	90.5	83.7	7.19	7.18	0.906	0.998	0.770	0.991	0.140	0.343
Panel C: n = 500										
$\rho =$	0.3	0.9	0.3	0.9	0.3	0.9	0.3	0.9	0.3	0.9
X_0	498	498	2.00	2.00	0.002	0.001	0.005	0.003	0.015	0.018
X_0, X_1	444	444	3.00	3.00	1.000	1.000	1.000	1.000	1.000	1.000
$X_0, \ldots, X_2$	443	443	4.01	4.01	1.000	1.000	0.957	0.999	0.215	0.779
$X_0, \ldots, X_3$	442	442	5.01	5.01	1.000	1.000	0.937	0.998	0.109	0.554
$X_0, \ldots, X_4$	441	441	6.02	6.02	1.000	1.000	0.903	0.996	0.054	0.208
$X_0, \ldots, X_5$	440	440	7.03	7.03	1.000	1.000	0.859	0.992	0.031	0.085
$X_0, \ldots, X_6$	439	439	8.04	8.04	1.000	1.000	0.796	0.986	0.014	0.035
X_0, X_2	493	454	3.00	3.00	0.007	0.849	0.011	0.886	0.013	0.717
X_0, X_2, X_3	489	451	4.00	4.00	0.011	0.915	0.013	0.916	0.008	0.478
$X_0, X_2, \ldots, X_4$	486	449	5.01	5.01	0.016	0.932	0.015	0.909	0.003	0.241
$X_0, X_2, \ldots, X_5$	484	448	6.02	6.02	0.019	0.939	0.016	0.886	0.002	0.100
$X_0, X_2, \ldots, X_6$	483	447	7.03	7.02	0.022	0.943	0.016	0.843	0.001	0.042

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

Table A.6: Simulation Experiment II: $\beta^2 = 0.5$, fraction in MCS

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

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

Table A.7: Simulation Experiment II: $\beta^2 = 0.9$, fraction in MCS

	$Q(\mathcal{Z}_j, \hat{\theta}_j)$		$\hat{k}^*$		KLIC		AIC* (TIC)		BIC*	
Panel A: $n = 50$										
$\rho =$	0.3	0.9	0.3	0.9	0.3	0.9	0.3	0.9	0.3	0.9
X_0	47.9	48.1	2.00	1.99	0.000	0.000	0.000	0.000	0.000	0.000
X_0, X_1	-68.4	-68.3	3.03	3.03	0.999	0.999	1.000	1.000	1.000	1.000
$X_0, \ldots, X_2$	-69.5	-69.3	4.09	4.09	0.999	0.999	0.959	0.962	0.521	0.521
$X_0, \ldots, X_3$	-70.6	-70.4	5.19	5.19	0.999	0.999	0.939	0.941	0.405	0.406
$X_0, \ldots, X_4$	-71.7	-71.5	6.33	6.32	1.000	0.999	0.909	0.908	0.283	0.289
$X_0, \ldots, X_5$	-72.8	-72.7	7.51	7.51	1.000	1.000	0.858	0.864	0.190	0.202
$X_0, \ldots, X_6$	-74.0	-73.9	8.75	8.75	1.000	1.000	0.786	0.797	0.119	0.135
X_0, X_2	42.6	-18.5	3.03	3.02	0.000	0.005	0.000	0.007	0.000	0.009
X_0, X_2, X_3	39.2	-27.2	4.09	4.08	0.000	0.021	0.000	0.023	0.000	0.016
$X_0, X_2, \ldots, X_4$	36.5	-31.4	5.19	5.18	0.000	0.036	0.000	0.032	0.000	0.018
$X_0, X_2, \ldots, X_5$	34.3	-34.2	6.33	6.32	0.000	0.048	0.000	0.036	0.000	0.014
$X_0, X_2, \ldots, X_6$	32.3	-36.4	7.51	7.50	0.000	0.056	0.000	0.038	0.000	0.010
Panel B: $n = 100$										
$\rho =$	0.3	0.9	0.3	0.9	0.3	0.9	0.3	0.9	0.3	0.9
X_0	98.0	98.0	1.99	1.99	0.000	0.000	0.000	0.000	0.000	0.000
X_0, X_1	-134	-133	3.01	3.00	0.999	0.999	1.000	1.000	1.000	1.000
$X_0, \ldots, X_2$	-135	-134	4.03	4.02	1.000	0.999	0.958	0.958	0.400	0.400
$X_0, \ldots, X_3$	-136	-135	5.07	5.06	1.000	1.000	0.937	0.937	0.277	0.275
$X_0, \ldots, X_4$	-137	-137	6.12	6.11	1.000	1.000	0.903	0.904	0.176	0.166
$X_0, \ldots, X_5$	-138	-138	7.19	7.18	1.000	1.000	0.855	0.859	0.103	0.100
$X_0, \ldots, X_6$	-139	-139	8.28	8.27	1.000	1.000	0.796	0.796	0.057	0.053
X_0, X_2	88.5	-33.9	3.00	3.00	0.000	0.000	0.000	0.000	0.000	0.000
X_0, X_2, X_3	82.7	-50.2	4.02	4.03	0.000	0.000	0.000	0.000	0.000	0.000
$X_0, X_2, \ldots, X_4$	78.5	-57.4	5.06	5.06	0.000	0.000	0.000	0.000	0.000	0.000
$X_0, X_2, \ldots, X_5$	75.2	-61.8	6.11	6.12	0.000	0.000	0.000	0.000	0.000	0.000
$X_0, X_2, \ldots, X_6$	72.4	-64.8	7.18	7.19	0.000	0.000	0.000	0.000	0.000	0.000
Panel C: $n = 500$										
$\rho =$	0.3	0.9	0.3	0.9	0.3	0.9	0.3	0.9	0.3	0.9
X_0	499	499	2.00	2.00	0.000	0.000	0.000	0.000	0.000	0.000
X_0, X_1	-654	-654	3.00	3.00	0.999	0.999	1.000	1.000	1.000	1.000
$X_0, \ldots, X_2$	-656	-655	4.01	4.00	1.000	1.000	0.957	0.956	0.206	0.202
$X_0, \ldots, X_3$	-657	-656	5.01	5.01	1.000	1.000	0.937	0.936	0.097	0.093
$X_0, \ldots, X_4$	-658	-657	6.02	6.02	1.000	1.000	0.902	0.901	0.040	0.037
$X_0, \ldots, X_5$	-659	-658	7.03	7.03	1.000	1.000	0.858	0.852	0.019	0.015
$X_0, \ldots, X_6$	-660	-659	8.04	8.04	1.000	1.000	0.796	0.789	0.006	0.006
X_0, X_2	455	-156	3.00	3.00	0.000	0.000	0.000	0.000	0.000	0.000
X_0, X_2, X_3	430	-233	4.00	4.00	0.000	0.000	0.000	0.000	0.000	0.000
$X_0, X_2, \ldots, X_4$	413	-264	5.01	5.01	0.000	0.000	0.000	0.000	0.000	0.000
$X_0, X_2, \ldots, X_5$	401	-282	6.01	6.02	0.000	0.000	0.000	0.000	0.000	0.000
$X_0, X_2, \ldots, X_6$	392	-293	7.02	7.02	0.000	0.000	0.000	0.000	0.000	0.000

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

Table A.8: Simulation Experiment II: $\beta^2 = 0.1$, average MCS p-value

	$Q(\mathcal{Z}_j, \hat{\theta}_j)$		$\hat{k}^*$		KLIC		AIC* (TIC)		BIC*	
Panel A: $n = 50$										
$\rho =$	0.3	0.9	0.3	0.9	0.3	0.9	0.3	0.9	0.3	0.9
X_0	47.8	48.1	2.00	1.99	0.682	0.659	0.729	0.739	0.744	0.768
X_0, X_1	41.4	41.8	3.03	3.03	0.927	0.918	0.938	0.908	0.841	0.794
$X_0, \ldots, X_2$	40.3	40.8	4.09	4.09	0.905	0.921	0.784	0.820	0.504	0.515
$X_0, \ldots, X_3$	39.3	39.7	5.19	5.19	0.905	0.925	0.738	0.765	0.366	0.359
$X_0, \ldots, X_4$	38.2	38.5	6.33	6.33	0.907	0.928	0.690	0.705	0.258	0.242
$X_0, \ldots, X_5$	37.0	37.4	7.51	7.51	0.905	0.927	0.635	0.635	0.176	0.158
$X_0, \ldots, X_6$	35.8	36.2	8.75	8.74	0.903	0.919	0.575	0.560	0.117	0.099
X_0, X_2	46.3	42.8	3.03	3.03	0.702	0.880	0.676	0.848	0.474	0.696
X_0, X_2, X_3	45.0	41.5	4.09	4.09	0.711	0.893	0.631	0.791	0.334	0.486
$X_0, X_2, \ldots, X_4$	43.7	40.3	5.19	5.18	0.716	0.900	0.579	0.738	0.234	0.339
$X_0, X_2, \ldots, X_5$	42.5	39.2	6.33	6.32	0.717	0.901	0.522	0.675	0.161	0.228
$X_0, X_2, \ldots, X_6$	41.3	38.0	7.51	7.51	0.714	0.897	0.465	0.606	0.108	0.150
Panel B: $n = 100$										
$\rho =$	0.3	0.9	0.3	0.9	0.3	0.9	0.3	0.9	0.3	0.9
X_0	97.9	98.2	2.00	1.99	0.421	0.390	0.485	0.490	0.521	0.552
X_0, X_1	86.2	86.7	3.01	3.01	0.919	0.920	0.964	0.936	0.950	0.893
$X_0, \ldots, X_2$	85.2	85.6	4.03	4.03	0.883	0.916	0.732	0.814	0.450	0.490
$X_0, \ldots, X_3$	84.2	84.6	5.07	5.07	0.883	0.921	0.689	0.768	0.324	0.333
$X_0, \ldots, X_4$	83.1	83.5	6.12	6.12	0.882	0.925	0.650	0.720	0.224	0.215
$X_0, \ldots, X_5$	82.0	82.4	7.19	7.19	0.880	0.925	0.609	0.663	0.148	0.134
$X_0, \ldots, X_6$	80.9	81.4	8.28	8.27	0.879	0.921	0.566	0.603	0.094	0.080
X_0, X_2	95.9	88.7	3.01	3.00	0.451	0.824	0.471	0.802	0.345	0.674
X_0, X_2, X_3	94.4	87.2	4.03	4.03	0.467	0.849	0.444	0.756	0.227	0.447
$X_0, X_2, \ldots, X_4$	93.1	86.0	5.07	5.07	0.476	0.858	0.412	0.712	0.149	0.298
$X_0, X_2, \ldots, X_5$	91.8	84.9	6.12	6.12	0.483	0.861	0.378	0.660	0.095	0.190
$X_0, X_2, \ldots, X_6$	90.5	83.7	7.19	7.18	0.488	0.859	0.343	0.605	0.058	0.117
Panel C: $n = 500$										
$\rho =$	0.3	0.9	0.3	0.9	0.3	0.9	0.3	0.9	0.3	0.9
X_0	498	498	2.00	2.00	0.006	0.005	0.008	0.006	0.013	0.014
X_0, X_1	444	444	3.00	3.00	0.909	0.925	0.962	0.968	0.997	0.984
$X_0, \ldots, X_2$	443	443	4.01	4.01	0.868	0.883	0.646	0.706	0.114	0.264
$X_0, \ldots, X_3$	442	442	5.01	5.01	0.866	0.883	0.588	0.670	0.053	0.149
$X_0, \ldots, X_4$	441	441	6.02	6.02	0.866	0.882	0.535	0.635	0.031	0.080
$X_0, \ldots, X_5$	440	440	7.03	7.03	0.862	0.880	0.480	0.598	0.021	0.045
$X_0, \ldots, X_6$	439	439	8.04	8.04	0.858	0.880	0.427	0.559	0.015	0.026
X_0, X_2	493	454	3.00	3.00	0.009	0.410	0.011	0.427	0.011	0.284
X_0, X_2, X_3	489	451	4.00	4.00	0.011	0.490	0.012	0.455	0.008	0.168
$X_0, X_2, \ldots, X_4$	486	449	5.01	5.01	0.013	0.521	0.012	0.441	0.005	0.088
$X_0, X_2, \ldots, X_5$	484	448	6.02	6.02	0.014	0.537	0.013	0.417	0.003	0.046
$X_0, X_2, \ldots, X_6$	483	447	7.03	7.02	0.016	0.546	0.013	0.388	0.002	0.026

The average value of the maximized log-likelihood function multiplied by minus two is reported in the first two columns. The next pair of columns has the average estimate of the degrees of freedom. The last three pairs of columns report the average MCS p-value for each of the three criteria, KLIC, AIC* and BIC*.

Table A.9: Simulation Experiment II: $\beta^2 = 0.5$, average MCS p-value

	$Q(\mathcal{Z}_j, \hat{\theta}_j)$		$\hat{k}^*$		KLIC		AIC* (TIC)		BIC*	
Panel A: n = 50										
$\rho =$	0.3	0.9	0.3	0.9	0.3	0.9	0.3	0.9	0.3	0.9
X_0	48.1	48.1	1.99	2.00	0.027	0.021	0.035	0.031	0.043	0.046
X_0, X_1	12.4	12.4	3.02	3.02	0.897	0.918	0.955	0.958	0.983	0.971
$X_0, \ldots, X_2$	11.3	11.3	4.08	4.08	0.858	0.885	0.644	0.725	0.293	0.438
$X_0, \ldots, X_3$	10.2	10.2	5.18	5.18	0.857	0.888	0.584	0.690	0.204	0.345
$X_0, \ldots, X_4$	9.09	9.04	6.32	6.32	0.856	0.889	0.529	0.650	0.148	0.258
$X_0, \ldots, X_5$	7.95	7.88	7.50	7.50	0.852	0.888	0.470	0.605	0.108	0.189
$X_0, \ldots, X_6$	6.77	6.69	8.73	8.74	0.848	0.887	0.417	0.554	0.079	0.133
X_0, X_2	44.7	21.0	3.02	3.02	0.035	0.491	0.040	0.494	0.038	0.383
X_0, X_2, X_3	42.3	18.1	4.08	4.08	0.041	0.566	0.041	0.511	0.032	0.314
$X_0, X_2, \ldots, X_4$	40.4	16.3	5.18	5.18	0.045	0.594	0.040	0.491	0.024	0.241
$X_0, X_2, \ldots, X_5$	38.8	14.8	6.32	6.32	0.048	0.608	0.038	0.457	0.018	0.177
$X_0, X_2, \ldots, X_6$	37.2	13.4	7.50	7.51	0.051	0.615	0.037	0.418	0.013	0.125
Panel B: n = 100										
$\rho =$	0.3	0.9	0.3	0.9	0.3	0.9	0.3	0.9	0.3	0.9
X_0	98.0	98.1	1.99	1.99	0.001	0.000	0.001	0.001	0.001	0.001
X_0, X_1	27.6	27.8	3.00	3.00	0.898	0.914	0.957	0.964	0.990	0.990
$X_0, \ldots, X_2$	26.6	26.7	4.03	4.03	0.862	0.873	0.652	0.667	0.213	0.290
$X_0, \ldots, X_3$	25.5	25.7	5.07	5.06	0.862	0.873	0.592	0.618	0.130	0.229
$X_0, \ldots, X_4$	24.4	24.6	6.12	6.12	0.863	0.870	0.539	0.574	0.081	0.174
$X_0, \ldots, X_5$	23.4	23.6	7.19	7.18	0.860	0.868	0.482	0.535	0.051	0.126
$X_0, \ldots, X_6$	22.3	22.5	8.28	8.27	0.858	0.865	0.429	0.496	0.031	0.086
X_0, X_2	92.4	45.1	3.00	3.01	0.001	0.207	0.001	0.230	0.001	0.183
X_0, X_2, X_3	88.8	40.4	4.03	4.03	0.001	0.286	0.001	0.276	0.001	0.164
$X_0, X_2, \ldots, X_4$	86.1	38.1	5.07	5.07	0.002	0.319	0.001	0.280	0.001	0.124
$X_0, X_2, \ldots, X_5$	83.9	36.3	6.12	6.12	0.002	0.338	0.002	0.271	0.001	0.087
$X_0, X_2, \ldots, X_6$	82.0	34.8	7.19	7.19	0.002	0.351	0.002	0.256	0.000	0.056
Panel C: n = 500										
$\rho =$	0.3	0.9	0.3	0.9	0.3	0.9	0.3	0.9	0.3	0.9
X_0	498	498	2.00	2.00	0.000	0.000	0.000	0.000	0.000	0.000
X_0, X_1	151	151	3.00	3.00	0.903	0.905	0.960	0.962	0.997	0.997
$X_0, \ldots, X_2$	150	150	4.00	4.00	0.864	0.864	0.644	0.648	0.112	0.110
$X_0, \ldots, X_3$	149	149	5.01	5.01	0.864	0.865	0.587	0.592	0.050	0.050
$X_0, \ldots, X_4$	148	148	6.02	6.01	0.864	0.864	0.533	0.536	0.023	0.022
$X_0, \ldots, X_5$	147	147	7.03	7.02	0.860	0.859	0.480	0.482	0.011	0.011
$X_0, \ldots, X_6$	145	146	8.04	8.03	0.858	0.855	0.430	0.432	0.005	0.006
X_0, X_2	474	238	3.00	3.00	0.000	0.000	0.000	0.000	0.000	0.001
X_0, X_2, X_3	460	219	4.00	4.00	0.000	0.002	0.000	0.002	0.000	0.002
$X_0, X_2, \ldots, X_4$	451	211	5.01	5.01	0.000	0.004	0.000	0.004	0.000	0.002
$X_0, X_2, \ldots, X_5$	444	206	6.02	6.01	0.000	0.005	0.000	0.005	0.000	0.001
$X_0, X_2, \ldots, X_6$	439	203	7.03	7.02	0.000	0.006	0.000	0.005	0.000	0.001

The average value of the maximized log-likelihood function multiplied by minus two is reported in the first two columns. The next pair of columns has the average estimate of the degrees of freedom. The last three pairs of columns report the average MCS *p*-value for each of the three criteria, KLIC, AIC* and BIC*.

Table A.10: Simulation Experiment II: $\beta^2 = 0.9$, average MCS p-value

	$Q(\mathcal{Z}_j, \hat{\theta}_j)$		$\hat{k}^*$		KLIC		AIC* (TIC)		BIC*	
Panel A: $n = 50$										
$\rho =$	0.3	0.9	0.3	0.9	0.3	0.9	0.3	0.9	0.3	0.9
X_0	47.9	48.1	2.00	1.99	0.000	0.000	0.000	0.000	0.000	0.000
X_0, X_1	-68.4	-68.3	3.03	3.03	0.902	0.899	0.957	0.955	0.984	0.983
$X_0, \ldots, X_2$	-69.5	-69.3	4.09	4.09	0.862	0.861	0.644	0.647	0.279	0.285
$X_0, \ldots, X_3$	-70.6	-70.4	5.19	5.19	0.860	0.860	0.582	0.584	0.188	0.191
$X_0, \ldots, X_4$	-71.7	-71.5	6.33	6.32	0.860	0.859	0.524	0.526	0.127	0.131
$X_0, \ldots, X_5$	-72.8	-72.7	7.51	7.51	0.855	0.856	0.465	0.469	0.085	0.090
$X_0, \ldots, X_6$	-74.0	-73.9	8.75	8.75	0.851	0.850	0.411	0.413	0.057	0.061
X_0, X_2	42.6	-18.5	3.03	3.02	0.000	0.005	0.000	0.006	0.000	0.007
X_0, X_2, X_3	39.2	-27.2	4.09	4.08	0.000	0.013	0.000	0.013	0.000	0.010
$X_0, X_2, \ldots, X_4$	36.5	-31.4	5.19	5.18	0.000	0.018	0.000	0.016	0.000	0.010
$X_0, X_2, \ldots, X_5$	34.3	-34.2	6.33	6.32	0.000	0.021	0.000	0.017	0.000	0.009
$X_0, X_2, \ldots, X_6$	32.3	-36.4	7.51	7.50	0.000	0.024	0.000	0.018	0.000	0.007
Panel B: $n = 100$										
$\rho =$	0.3	0.9	0.3	0.9	0.3	0.9	0.3	0.9	0.3	0.9
X_0	98.0	98.0	1.99	1.99	0.000	0.000	0.000	0.000	0.000	0.000
X_0, X_1	-134	-133	3.01	3.00	0.902	0.900	0.960	0.958	0.991	0.990
$X_0, \ldots, X_2$	-135	-134	4.03	4.02	0.861	0.863	0.644	0.654	0.211	0.214
$X_0, \ldots, X_3$	-136	-135	5.07	5.06	0.861	0.862	0.586	0.592	0.127	0.128
$X_0, \ldots, X_4$	-137	-137	6.12	6.11	0.861	0.862	0.531	0.536	0.079	0.079
$X_0, \ldots, X_5$	-138	-138	7.19	7.18	0.857	0.858	0.477	0.480	0.048	0.047
$X_0, \ldots, X_6$	-139	-139	8.28	8.27	0.855	0.854	0.426	0.428	0.028	0.028
X_0, X_2	88.5	-33.9	3.00	3.00	0.000	0.000	0.000	0.000	0.000	0.000
X_0, X_2, X_3	82.7	-50.2	4.02	4.03	0.000	0.000	0.000	0.000	0.000	0.000
$X_0, X_2, \ldots, X_4$	78.5	-57.4	5.06	5.06	0.000	0.000	0.000	0.000	0.000	0.000
$X_0, X_2, \ldots, X_5$	75.2	-61.8	6.11	6.12	0.000	0.000	0.000	0.000	0.000	0.000
$X_0, X_2, \ldots, X_6$	72.4	-64.8	7.18	7.19	0.000	0.001	0.000	0.000	0.000	0.000
Panel C: $n = 500$										
$\rho =$	0.3	0.9	0.3	0.9	0.3	0.9	0.3	0.9	0.3	0.9
X_0	499	499	2.00	2.00	0.000	0.000	0.000	0.000	0.000	0.000
X_0, X_1	-654	-654	3.00	3.00	0.908	0.910	0.962	0.963	0.997	0.997
$X_0, \ldots, X_2$	-656	-655	4.01	4.00	0.868	0.868	0.646	0.646	0.111	0.110
$X_0, \ldots, X_3$	-657	-656	5.01	5.01	0.866	0.866	0.588	0.587	0.047	0.047
$X_0, \ldots, X_4$	-658	-657	6.02	6.02	0.866	0.864	0.535	0.532	0.022	0.021
$X_0, \ldots, X_5$	-659	-658	7.03	7.03	0.862	0.860	0.480	0.478	0.011	0.010
$X_0, \ldots, X_6$	-660	-659	8.04	8.04	0.858	0.854	0.427	0.426	0.005	0.005
X_0, X_2	455	-156	3.00	3.00	0.000	0.000	0.000	0.000	0.000	0.000
X_0, X_2, X_3	430	-233	4.00	4.00	0.000	0.000	0.000	0.000	0.000	0.000
$X_0, X_2, \ldots, X_4$	413	-264	5.01	5.01	0.000	0.000	0.000	0.000	0.000	0.000
$X_0, X_2, \ldots, X_5$	401	-282	6.01	6.02	0.000	0.000	0.000	0.000	0.000	0.000
$X_0, X_2, \ldots, X_6$	392	-293	7.02	7.02	0.000	0.000	0.000	0.000	0.000	0.000

The average value of the maximized log-likelihood function multiplied by minus two is reported in the first two columns. The next pair of columns has the average estimate of the degrees of freedom. The last three pairs of columns report the average MCS *p*-value for each of the three criteria, KLIC, AIC* and BIC*.

Table A.11: MCS for Taylor Rules: 1979:Q1 to 2006:Q4

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

We report the maximized log-likelihood function (multiplied by minus two), the effective degress of freedom, and the three criteria, KLIC, AIC* and BIC*, along with the corresponding MCS p-values. The regression models in $\widehat{\mathcal{M}}^*_{90\%}$ and $\widehat{\mathcal{M}}^*_{75\%}$ are identified by one and two asterisks, respectively. See the text and Table 6 for variable mnemonics and definitions.

Model Specification			$Q(\mathcal{Z}_j, \hat{\theta}_j)$	$\hat{k}^\star$	KLIC	AIC*	BIC*
R_{t-1}			-38.90	6.97	-31.93 (0.56)**	-24.96 (0.93)**	-7.39 (1.00)**
	π_{t-1}	y_{t-1}	208.37	11.57	219.93 (0.00)	231.50 (0.00)	260.68 (0.00)
	$\pi_{t-j,\ j=1,2}$	$y_{t-j,\ j=1,2}$	190.30	13.76	204.06 (0.00)	217.83 (0.00)	252.54 (0.00)
	π_{t-1}	ur_{t-1}	227.79	11.91	239.70 (0.00)	251.61 (0.00)	281.63 (0.00)
	$\pi_{t-j,\ j=1,2}$	$ur_{t-j,\ j=1,2}$	220.78	14.10	234.88 (0.00)	248.97 (0.00)	284.52 (0.00)
	π_{t-1}	$rulc_{t-1}$	228.07	10.34	238.41 (0.00)	248.75 (0.00)	274.83 (0.00)
	$\pi_{t-j,\ j=1,2}$	$rulc_{t-j,\ j=1,2}$	213.53	12.06	225.59 (0.00)	237.65 (0.00)	268.07 (0.00)
	y_{t-1}	ur_{t-1}	226.30	12.82	239.12 (0.00)	251.93 (0.00)	284.25 (0.00)
	$y_{t-j,\ j=1,2}$	$ur_{t-j,\ j=1,2}$	216.60	16.04	232.65 (0.00)	248.69 (0.00)	289.15 (0.00)
	y_{t-1}	$rulc_{t-1}$	225.63	12.38	238.01 (0.00)	250.39 (0.00)	281.62 (0.00)
	$y_{t-j,\ j=1,2}$	$rulc_{t-j,\ j=1,2}$	216.68	14.39	231.07 (0.00)	245.46 (0.00)	281.76 (0.00)
	ur_{t-1}	$rulc_{t-1}$	238.39	12.46	250.85 (0.00)	263.31 (0.00)	294.74 (0.00)
	$ur_{t-j,\ j=1,2}$	$rulc_{t-j,\ j=1,2}$	233.41	14.90	248.31 (0.00)	263.21 (0.00)	300.78 (0.00)
R_{t-1}	π_{t-1}	y_{t-1}	-66.21	14.74	-51.47 (0.78)**	-36.73 (0.96)**	0.44 (0.92)**
R_{t-1}	$\pi_{t-j,\ j=1,2}$	$y_{t-j,\ j=1,2}$	-70.85	16.93	-53.92 (1.00)**	-37.00 (1.00)**	5.69 (0.86)**
R_{t-1}	π_{t-1}	ur_{t-1}	-45.39	9.63	-35.76 (0.64)**	-26.13 (0.93)**	-1.84 (0.92)**
R_{t-1}	$\pi_{t-j,\ j=1,2}$	$ur_{t-j,\ j=1,2}$	-45.55	13.23	-32.31 (0.56)**	-19.08 (0.74)**	14.30 (0.57)**
R_{t-1}	π_{t-1}	$rulc_{t-1}$	-51.84	11.31	-40.53 (0.72)**	-29.22 (0.93)**	-0.71 (0.92)**
R_{t-1}	$\pi_{t-j,\ j=1,2}$	$rulc_{t-j,\ j=1,2}$	-53.68	12.74	-40.94 (0.72)**	-28.21 (0.93)**	3.91 (0.88)**
R_{t-1}	y_{t-1}	ur_{t-1}	-58.05	13.50	-44.55 (0.72)**	-31.05 (0.93)**	2.99 (0.89)**
R_{t-1}	$y_{t-j,\ j=1,2}$	$ur_{t-j,\ j=1,2}$	-62.18	16.65	-45.53 (0.72)**	-28.88 (0.93)**	13.12 (0.60)**
R_{t-1}	y_{t-1}	$rulc_{t-1}$	-58.72	14.20	-44.52 (0.72)**	-30.32 (0.93)**	5.50 (0.86)**
R_{t-1}	$y_{t-j,\ j=1,2}$	$rulc_{t-j,\ j=1,2}$	-64.74	15.81	-48.94 (0.78)**	-33.13 (0.93)**	6.74 (0.86)**
R_{t-1}	ur_{t-1}	$rulc_{t-1}$	-50.00	12.00	-37.99 (0.64)**	-25.99 (0.93)**	4.28 (0.87)**
R_{t-1}	$ur_{t-j,\ j=1,2}$	$rulc_{t-j,\ j=1,2}$	-50.96	15.73	-35.22 (0.62)**	-19.49 (0.74)**	20.19 (0.35)**

We report the maximized log-likelihood function (multiplied by minus two), the effective degress of freedom, and the three criteria, KLIC, AIC* and BIC*, along with the corresponding MCS p-values. The regression models in $\widehat{\mathcal{M}}^*_{90\%}$ and $\widehat{\mathcal{M}}^*_{75\%}$ are identified by one and two asterisks, respectively. See the text and Table 6 for variable mnemonics and definitions.