A Test for Superior Predictive Ability

Peter Reinhard Hansen*

Stanford University
Department of Economics
579 Serra Mall
Stanford, CA, 94305-6072, USA
Email: peter.hansen@stanford.edu

Abstract

We propose a new test for superior predictive ability. The new test compares favorable to the *reality check for data snooping* (RC), because the former is more powerful and less sensitive to poor and irrelevant alternatives. The improvements are achieved by two modifications of the RC. We employ a studentized test statistic that reduces the influence of erratic forecasts and we invoke a sample dependent null distribution. The advantages of the new test are confirmed by Monte Carlo experiments and in an empirical exercise, where we compare a large number of regression-based forecasts of annual US inflation to a simple random walk forecast. The random walk forecast is found to be inferior to regression-based forecasts and, interestingly, the best sample performance is achieved by models that have a Phillips curve structure.

JEL Classification: C12, C32, C52, C53.

Keywords: Testing for superior predictive ability, forecasting, forecast evaluation, multiple comparison, inequality testing.

*I thank Jinyong Hahn, James D. Hamilton, Søren Johansen, Tony Lancaster, Asger Lunde, Michael McCracken, Barbara Rossi and seminar participants at Princeton, Harvard/MIT, University of Montreal, UBC, New York Fed, Stanford, NBER/NSF Summer Institute, 2001, three anonymous referees and Torben Anderson (editor) for many valuable comments and suggestions. I am also grateful for financial support from the Danish Research Agency, grant no. 24-00-0363.

1 Introduction

To test whether a particular forecasting procedure is outperformed by alternative forecasts is a test of superior predictive ability (SPA). White (2000) developed a framework for comparing multiple forecasting models[1] and proposed a test for SPA that is known as the reality check for data snooping (RC). In his framework m alternative forecasts (where m is a fixed number) are compared to a benchmark forecast, where the predictive abilities are defined by expected loss. The complexity of this inference problem arises from the need to control for the full set of alternatives.

In this paper, we propose a new test for SPA. Our framework is identical to that of White (2000), but we take a different path in our construction of the test. To be specific, we employ a *different test statistic* and we invoke a *sample dependent distribution* under the null hypothesis. Compared to the RC, the new test is more powerful and less sensitive to the inclusion of poor and irrelevant alternatives.

We make three contributions in this paper. First, we provide a theoretical analysis of the testing problem that highlights some important aspects. Our theoretical results reveal that the RC can be manipulated by the inclusion of poor and irrelevant forecasts in the set of alternative forecasts. This problem is alleviated by studentizing the test statistic and by invoking a sample dependent null distribution. The latter is based on a novel procedure that incorporates additional sample information in order to identify the 'relevant' alternatives. Second, we provided a detailed explanation of a bootstrap implementation of our test for SPA, which will make it easy for users to employ these methods in practice. Third, we apply the tests in an empirical analysis of US inflation. Our benchmark is a simple random walk forecast that uses current inflation as the prediction of future inflation. The benchmark is compared to a large number of regression-based forecasts and our empirical results show that the benchmark is significantly outperformed. Interestingly, the strongest evidence is provided by regression models that have a Phillips curve structure.

When testing for SPA, the question of interest is whether any alternative forecast is better than the benchmark forecast, or equivalently, whether the best alternative forecasting model is

[1] The term "model" is here used in a broad sense that includes forecasting rules/methods, which do not necessarily involve a modelling of data.

better than the benchmark. This question can be addressed by testing the null hypothesis that "the benchmark is not inferior to any alternative forecast". This testing problem is relevant for applied econometrics, because several ideas and specifications are often employed before a model is selected. This *mining* may be exacerbated if more than one researcher is searching for a good forecasting model. For a more complete discussion on this issue, see Sullivan, Timmermann, and White (2003) and references therein. Testing for SPA is useful for a forecaster who wants to explore whether a better forecasting model is available, compared to the model currently being used to make predictions. After a search over several alternative models, the relevant question is whether an observed excess performance by an alternative model is significant or not. The test for SPA can also be used to test an economic theory that places restrictions on the predictability of certain variables, such as the *efficient markets hypothesis*, see Sullivan, Timmermann, and White (1999).

Tests for equal predictive ability (EPA), in a general setting, were proposed by Diebold and Mariano (1995) and West (1996), where the framework of the latter can accommodate the situation where forecasts involve estimated parameters. Harvey, Leybourne, and Newbold (1997) suggested a modification of the Diebold-Mariano test that leads to better small sample properties. A test for comparing multiple nested model was given by Harvey and Newbold (2000) and McCracken (2000) derived results for the case with estimated parameters and non-differentiable loss functions, such as the mean absolute deviation loss function. West and McCracken (1998) developed regression-based tests and other extensions were made by Harvey, Leybourne, and Newbold (1998), West (2001), and Clark and McCracken (2001) who considered tests for forecast encompassing, and by Corradi, Swanson, and Olivetti (2001) who compared forecasting models that include cointegrated variables.

Whereas the frameworks of Diebold and Mariano (1995) and West (1996) involve tests for EPA, the testing problem in White's framework is a test for SPA. The distinction is important because the former leads to a simple null hypothesis, whereas the latter leads to a composite hypothesis. One of the main complications in composite hypotheses testing is that (asymptotic) distributions typically depend on nuisance parameters, such that the null distribution is not unique.

The usual way to handle this ambiguity is to use the *least favorable configuration* (LFC), which is sometimes referred to as "the point least favorable to the alternative". Our analysis shows that this approach leads to some rather unfortunate properties when testing for SPA. An edifying example for understanding the advantages of our sample dependent null distribution, is that where a simple Bonferroni bound test is employed. Naturally, our test is quite different from the conservative Bonferroni bound test. If we let $p_{\min}$ denotes the smallest p-value of the m pairwise comparisons (comparing each alternative to the benchmark), then the Bonferroni bound test (at level α) rejects the null hypothesis if $p_{\min} < \alpha/m$. It is now evident that the power of this test can be driven to zero by adding poor and irrelevant alternatives to the comparison, because this increases m (but does not affect $p_{\min}$). Yet, sample information will (at least asymptotically) identify the poor and irrelevant alternative, and this allows us to use a smaller denominator when defining the critical value, e.g. α/m_0 for some $m_0 \leq m$. Our sample dependent null distribution is quite analogous to this improvement of the Bonferroni bound test, although the (presumed) poor alternatives are not discarded entirely from the comparison.

In relation to the existing literature on forecast evaluation and comparison, it is important to acknowledge a limitation of the specific test that we propose in this paper. A comparison of nested models becomes problematic when parameters are estimated recursively, because this situation violates our stationarity assumption. So this situation requires a different bootstrap implementation amongst other things. The advantages of the studentized test statistic and our sample dependent null distribution do not rely on stationarity, and are therefore expected to be useful in a more general context. A related issue concerns optimallity of our test. While the new test dominates the RC we do not claim it to be optimal. The lack of an optimallity result is not surprising, because such results are rare in composite hypothesis testing. It is also worth to observe that leading statisticians continue to quarrel about the suitable criterion (for defining optimallity) in this context, see Perlman and Wu (1999) and the comments on this paper by Roger Berger, David Cox, Michael McDermott and Yining Wang.

This paper is organized as follows. Section 2 introduces the new test for SPA and contains our theoretical results. Section 3 provides the details of the bootstrap implementation. Section

4 contains a simulation-based study of the finite sample properties of the new test for SPA and compares it to those of the RC. Section 5 contains an empirical forecasting exercise of US inflation, and Section 6 gives a summery and some concluding remarks. All proofs are presented in Appendix A.

2 Testing for Superior Predictive Ability

We consider a situation where a decision must be made h periods in advance and let $\{\delta_{k,t-h}, k = 0, 1, \ldots, m\}$ be the (finite) set of possible decision rules. Decisions are evaluated with a real-valued loss function, $L(\xi_t, \delta_{t-h})$, where ξ_t is a random variable that represents the aspects of the decision problem that are unknown at the time the decision is made. An overview of our notation is given in Table 1.

Table 1 about here

This provides a general framework for comparing forecasts and decision rules. Our leading example is the comparison of forecasts, so we shall often refer to $\delta_{k,t-h}$ as the kth forecasting model. The first model, $k = 0$, has a special role and will be referred to as the benchmark.

The decision rule, $\delta_{k,t-h}$, can represent a point forecast, an interval forecast, a density forecasts, or a trading rule for an investor. Next, we give some examples.

Example 1 (Point Forecast) *Let $\delta_{k,t-h}$, $k = 0, 1, \ldots, m$ be different point forecasts of a real random variable ξ_t. The mean squared error loss function, $L(\xi_t, \delta_{k,t-h}) = (\xi_t - \delta_{k,t-h})^2$, is an example of a loss function that could be used to compare the different forecasts.*

Example 2 (Conditional Distribution and Value-at-Risk Forecasts) *Let ξ_t be a conditional density on $\mathbb{R}$, and let $\delta_{k,t-h}$ be a forecast of ξ_t. Then we might evaluate the precision of δ_k by the Kolmogorov-Smirnov statistic, $L(\xi_t, \delta_{k,t-h}) = \sup_{x\in\mathbb{R}} \left|\int_{-\infty}^{x}[\xi_t(y) - \delta_{k,t-h}(y)]dy\right|$, or a Kullback-Leibler measure, $L(\xi_t, \delta_{k,t-h}) = \int_{-\infty}^{\infty} \log[\delta_{k,t-h}(x)/\xi_t(x)]\xi_t(x)dx$.*

Alternatively, $\delta_{k,t-h}$ could be a Value-at-Risk measure (at quantile α), which could be evaluated using $L(\xi_t, \delta_{k,t-h}) = |\int_{-\infty}^{\delta_{k,t-h}} \xi_t(x)dx - \alpha|$.

In Example 2, ξ_t will often be unobserved and this creates additional complication for the empirical evaluation and comparison. When a proxy is substituted for ξ_t it can cause the empirical ranking of alternatives to be inconsistent for the intended (true) ranking, see Hansen and Lunde (2005). Corradi and Swanson (2004) have recently derived a RC-type test for comparing conditional density forecasts, which is closely related to the problem of Example 2. Their test is similar to that of White (2000), so their test might also be improved by the two modifications that we propose in this paper.

Example 3 (Trading Rules) *Let $\delta_{k,t-1}$ be a binary variable that instructs a trader to take either a short or a long position in an asset at time $t-1$. The kth* trading rule *yields the profit $\pi_{k,t} = \delta_{k,t-1} r_t$, where r_t is the return on the asset in period t. A trader who is currently using the rule, δ_0, might be interested to know if an alternative rule has a larger expected profit than δ_0. This can be formulated in our framework by setting $\xi_t = r_t$ and $L(\xi_t, \delta_{k,t-1}) = -\delta_{k,t-1}\xi_t$.*

The benchmark in Example 3 could be $\delta_{0,t} = 1$, which is the rule that is always "long in the market". This was the benchmark used by Sullivan, Timmermann, and White (1999, 2001), who evaluated the significance of technical trading rules and calendar effects in stock returns.

2.1 Hypothesis of Interest

We are interested to know whether any of the models, $k = 1, \ldots, m$, are better than the benchmark in terms of expected loss. So we seek a test of the null hypothesis that *the benchmark is not inferior to any of the alternatives*. The variables that will be key for our analysis are the relative performance variables, which are defined by

$$d_{k,t} \equiv L(\xi_t, \delta_{0,t-h}) - L(\xi_t, \delta_{k,t-h}), \qquad k = 1, \ldots, m.$$

So $d_{k,t}$ denotes the performance of model k relative to the benchmark at time t, and we stack these variables into the vector of relative performances, $\mathbf{d}_t = (d_{1,t}, \ldots, d_{m,t})'$. Provided that $\boldsymbol{\mu} \equiv E(\mathbf{d}_t)$

is well defined, we can now formulate the null hypothesis of interest as

$$H_0 : \boldsymbol{\mu} \leq \mathbf{0}, \tag{1}$$

and our maintained hypothesis is $\boldsymbol{\mu} \in \mathbb{R}^m$.

We work under the assumption that model k is better than the benchmark if and only if $E(d_{k,t}) > 0$. So we focus exclusively on the properties of $\mathbf{d}_t$ and abstract entirely from all aspects that relate to the construction of the δ-variables. So $\mathbf{d}_t$, $t = 1, \ldots, n$, will *de facto* be viewed as our data, and we will therefore state all assumptions in terms $\mathbf{d}_t$. Specifically we make the following assumption.

Assumption 1 *The vector of relative loss variables,* $\{\mathbf{d}_t\}$, *is (strictly) stationary and* α*-mixing of size* $-(2+\delta)(r+\delta)/(r-2)$, *for some* $r > 2$ *and* $\delta > 0$, *where* $E\,|\mathbf{d}_t|^{r+\delta} < \infty$ *and* $\text{var}(d_{k,t}) > 0$ *for all* $k = 1, \ldots, m$.

Assumption 1 is made for two reasons. The first is to ensure that certain population moments, such as $\boldsymbol{\mu}$, are well defined. The second reason is to justify the use of bootstrap techniques that we describe in details in Section 3. Note that Assumption 1 does not require the individual loss variables, $L(\xi_t, \delta_{k,t-h})$, to be stationary. An immediate consequence of Assumption 1 is that a central limit theorem applies, such that

$$n^{1/2}(\bar{\mathbf{d}} - \boldsymbol{\mu}) \xrightarrow{d} N_m(\mathbf{0}, \boldsymbol{\Omega}), \tag{2}$$

where $\bar{\mathbf{d}} \equiv n^{-1} \sum_{t=1}^{n} \mathbf{d}_t$ and $\boldsymbol{\Omega} \equiv \text{avar}(n^{1/2}(\bar{\mathbf{d}} - \boldsymbol{\mu}))$, see, e.g., de Jong (1997).

Diebold and Mariano (1995) and West (1996) provide sufficient conditions that also lead to the asymptotic normality in (2), see also Giacomini and White (2003) who establishes this property for a related testing problem. However, the asymptotic normality does not hold in general. An important exception is that were the benchmark is nested in all alternative models (under the null hypothesis) and the parameters estimated recursively. In this situation the limiting distribution will typically be given as a function of Brownian motions, see, e.g., Clark and McCracken (2001).

When comparing nested models, the null hypothesis simplifies to the simple hypothesis, $\boldsymbol{\mu} = \mathbf{0}$. So in this case it seems more appropriate to apply a test for EPA, such as that of Harvey and Newbold (2000), which can be used to compare multiple nested model.

At this point, all essential aspects of our framework are identical to those in White (2000), and White proceeds by constructing the RC from the test statistic,

$$T_n^{\mathrm{RC}} \equiv \max(n^{1/2}\bar{d}_1, \ldots, n^{1/2}\bar{d}_m),$$

and an asymptotic null distribution that is based on $n^{1/2}\bar{\mathbf{d}} \sim N_m(\mathbf{0}, \hat{\Omega})$, where $\hat{\Omega}$ is a consistent estimator of Ω. Here, it is worth to note that the RC is based on an asymptotic null distribution that assumes $\mu_k = 0$ for all k, even though all negative values of μ_k also conform with the null hypothesis. This aspect is the underlying topic of the subsections 2.3 and 2.4. First we discuss a studentization of the test statistic.

Given the asymptotic normality of $\bar{\mathbf{d}}$, it may seem natural to employ a quadratic-form test statistic to test H_0, such as the likelihood ratio test used in Wolak (1987). However, the situation that we have in mind is one where m is too large to obtain a sensible estimate of all elements of Ω. Instead we consider simpler statistics, such as T_n^{SPA} (defined below) that only requires the diagonal elements of Ω to be estimated. It is not surprising that non-quadratic statistics will be non-pivotal – even asymptotically – because their asymptotic distribution will depend on (some elements of) the covariance matrix, which makes Ω a nuisance parameter. To handle this problem, we follow White (2000) and employ a bootstrap method that implicitly takes care of this nuisance parameter problem. So our motivation for using the bootstrap is not driven by higher-order refinements, but merely to handle this nuisance parameter problem.

We analyze this testing problem in the remainder of this section, and our findings motivate the following two recommendations that spell out the differences between the RC and our new test for SPA.

1. Use the studentized test statistic,

$$T_n^{\text{SPA}} \equiv \max\left[\max_{k=1,\ldots,m} \frac{n^{1/2}\bar{d}_k}{\hat{\omega}_k}, 0\right],$$

where $\hat{\omega}_k^2$ is some consistent estimator of $\omega_k^2 \equiv \text{var}(n^{1/2}\bar{d}_k)$.

2. Invoke a null distribution that is based on $N_m(\hat{\boldsymbol{\mu}}^c, \hat{\Omega})$, where $\hat{\boldsymbol{\mu}}^c$ is a carefully chosen estimator for $\boldsymbol{\mu}$ that conforms with the null hypothesis. Specifically we suggest the estimator

$$\hat{\mu}_k^c = \bar{d}_k 1_{\left\{n^{1/2}\bar{d}_k/\hat{\omega}_k \leq -\sqrt{2\log\log n}\right\}}, \qquad k = 1, \ldots, m,$$

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

The motivations for our choice of $\boldsymbol{\mu}$-estimator will be explained in Section 2.4, but it is important to understand that the use of a consistent estimator of $\boldsymbol{\mu}$ need not produce a valid test.

2.2 Choice of Test Statistic

When the benchmark has the best sample performance ($\bar{\mathbf{d}} \leq \mathbf{0}$) the test statistic is normalized to be zero. In this case there is no evidence against the null hypothesis, and the null should therefore not be rejected. The normalization is convenient for theoretical reasons, because we avoid a divergence problem (to $-\infty$) that would otherwise occur whenever $\boldsymbol{\mu} < \mathbf{0}$.

As we discussed in the introduction, there are few optimallity results in the context of composite hypothesis testing. This is particularly the case for the present problem of testing multiple inequalities. However, some arguments that justify our choice of test statistic, T_n^{SPA} (instead of using T_n^{RC}) are called upon.

While we shall argue that T_n^{SPA} is preferred to T_n^{RC}, it cannot be shown that the former uniformly dominate the latter in terms of power. In fact there are situations where the T_n^{RC} leads to a more powerful test (such as the case where $\omega_j^2 = \omega_k^2$, $\forall\ j, k = 1, \ldots, m$). However, such exceptions are unlikely to be of much empirical relevance, as we discuss below. So we are comfortable recommending the use of T_n^{SPA} in practice and it is worth to point out that a studentization of the

individual statistics is the conventional approach to multiple comparisons, see Miller (1981) and Savin (1984). This studentization is also embedded in the related approach where the individual statistics are converted into m "p-values", and the maximum p-value is used as the test statistic, see Tippett (1931), Folks (1984), Marden (1985), Westfall and Young (1993), and Dufour and Khalaf (2002). In the present context, Romano and Wolf (2003) also adopt the studentized test statistic, see also Lehmann and Romano (2005, chapter 9).

Our main argument for the studentization is that it typically will improve the power. This can be understood from the following simple example.

Example 4 *Consider the case where* $m = 2$ *and suppose that*

$$n^{1/2}(\bar{\mathbf{d}} - \boldsymbol{\mu}) \sim N_2(\mathbf{0}, \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}),$$

where the covariance is zero (*a simplification that is not necessary for our argument*). *Now consider the particular local alternative where* $\mu_2 = 2n^{-1/2} > 0$. *So* $\bar{d}_2$ *is expected to yield a fair amount of evidence against* $H_0 : \boldsymbol{\mu} \leq \mathbf{0}$, *because the t-statistic,* $n^{1/2}\bar{d}_2 \big/ \hat{\omega}_k$, *will be centered about* 2. *It follows that the null distributions* (*using* $\boldsymbol{\mu} = \mathbf{0}$) *are given by* $T_n^{\mathrm{RC}} \sim F_0(x) \equiv \Phi(x/2)\Phi(x)$ *and* $T_n^{\mathrm{SPA}} \overset{a}{\sim} G_0(x) \equiv \Phi(x)\Phi(x)$, *whereas* $T_n^{\mathrm{RC}} \sim F_1(x) \equiv \Phi(x/2)\Phi(x+2)$ *and* $T_n^{\mathrm{SPA}} \overset{a}{\sim} G_1(x) \equiv \Phi(x)\Phi(x+2)$ *under the local alternative. Here we use* $\Phi(\cdot)$ *to denote the standard Gaussian distribution and* $\overset{a}{\sim}$ *refers to "asymptotically distributed as". Figure 1 shows the upper tails of the null distributions,* $1 - F_0(x)$ *and* $1 - G_0(x)$ *(thick lines), and the upper tails of* $1 - F_1(x)$ *and* $1 - G_1(x)$ *(thin lines) that represent the distributions of the test statistics under the local alternative. We use dotted lines for the distributions of* T_n^{RC} *and solid lines for the distributions of* T_n^{SPA}. *The power for a given level of either of the two tests can be read off the figure and we have singled out the powers of the* 5%*-level tests. These reveal that the studentization more than triples the power, from about 15% to about 53%. So the RC is much less likely to detect that the null is false, because the noisy* $\bar{d}_1$ *conceals the evidence against* H_0 *that* $\bar{d}_2$ *provides.*

Figure 1 about here

The previous example highlights the advantages of studentizing the individual statistics, as it avoids a comparison of objects that are measured in different "units of standard deviation", (avoiding a comparison of apples and bananas). There is one exception where the studentization may reduce the power that occurs when the best performing model has the largest variance (i.e. if $\text{var}(\bar{d}_2) \geq \text{var}(\bar{d}_1)$ in the previous example). Since poor performing models also tend to have the most erratic performances, we consider this case to be of little empirical relevance. Also, the loss in power from estimating ω_k^2, $k = 1, \ldots, m$, is quite modest when these are estimated precisely.

In the remainder of this section we derive results that motivate our data dependent choice of null distribution. To make clear that our results are not specific to the two statistics, T_n^{RC} and T_n^{SPA}, we derive our results for a broader class of test statistics. This is also convenient because other statistics (from this class of statistics) may be used in future applied work.

2.3 Theoretical Results for a Class of Test Statistics

We consider a class of test statistics, where each of the statistics satisfies the following conditions.

Assumption 2 *The test statistic has the form* $T_n = \varphi(\mathbf{U}_n, \mathbf{V}_n)$, *where* $\mathbf{U}_n \equiv n^{1/2}\bar{\mathbf{d}}$ *and* $\mathbf{V}_n \xrightarrow{p} \mathbf{v}_o \in \mathbb{R}^q$ *(a constant). The mapping,* $\varphi(\mathbf{u}, \mathbf{v})$, *is continuous in* $\mathbf{u}$ *on* $\mathbb{R}^m$ *and continuous in* $\mathbf{v}$ *in a neighborhood of* $\mathbf{v}_0$. *Further,*

(*a*) $\varphi(\mathbf{u}, \mathbf{v}) \geq 0$ *and* $\varphi(\mathbf{0}, \mathbf{v}) = 0$;

(*b*) $\varphi(\mathbf{u}, \mathbf{v}) = \varphi(\mathbf{u}^+, \mathbf{v})$ *where* $u_k^+ = \max(0, u_k)$, $k = 1, \ldots, m$;

(*c*) $\varphi(\mathbf{u}, \mathbf{v}) \to \infty$, *if* $u_k \to \infty$ *for some* $k = 1, \ldots, m$.

Thus, in addition to the sample average, $\bar{\mathbf{d}}$, the test statistic may depend on the data through $\mathbf{V}_n \equiv v(\mathbf{d}_1, \ldots, \mathbf{d}_n)$, as long as $\mathbf{V}_n$ converges in probability to a constant (or vector of constants). Assumption 2.*a* is a normalization, (if $\bar{\mathbf{d}} = \mathbf{0}$ there is no evidence against H_0); Assumption 2.*b* states that only the positive elements of $\mathbf{u}$ matter for the value of the test statistic; and Assumption

2.*c* requires that the test statistic diverges to infinity as the evidence against the null hypothesis increases (to infinity).

The mapping: $(\boldsymbol{\mu}, \Omega) \mapsto \Omega^o$, given by

$$\Omega^o_{ij} \equiv \Omega_{ij} 1_{\{\mu_i = \mu_j = 0\}}, \qquad i, j = 1, \ldots, m,$$

defines an $m \times m$ covariance matrix, Ω^o, that plays a role in our asymptotic results. So Ω^o is similar to Ω, except that the elements of certain rows and columns have been set to zero. An example of how $\boldsymbol{\mu}$ and Ω translate into Ω^o, is the following:

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ -2 \\ 0 \end{pmatrix}, \qquad \Omega = \begin{pmatrix} \omega_{11}^2 & \omega_{12} & \omega_{13} \\ \omega_{21} & \omega_{22}^2 & \omega_{23} \\ \omega_{31} & \omega_{32} & \omega_{33}^2 \end{pmatrix}, \qquad \Omega^o = \begin{pmatrix} \omega_{11}^2 & 0 & \omega_{13} \\ 0 & 0 & 0 \\ \omega_{31} & 0 & \omega_{33}^2 \end{pmatrix},$$

and Ω^o has at most rank m_o, where m_o is the number of elements in $\boldsymbol{\mu}$ that equals zero.

The following Theorem provides the asymptotic null distribution for all test statistics that satisfies Assumption 2.

Theorem 1 *Suppose Assumptions 1 and 2 hold and let F_o be the cumulative distribution function (cdf) of $\varphi(\mathbf{Z}, \mathbf{v}_0)$, where $\mathbf{Z} \sim N_m(\mathbf{0}, \Omega^o)$. Under the null hypothesis, $\boldsymbol{\mu} \leq \mathbf{0}$, it holds that $\varphi(n^{1/2}\bar{\mathbf{d}}, \mathbf{V}_n) \xrightarrow{d} F_o$, where $\mathbf{v}_o = \operatorname{plim} \mathbf{V}_n$. Under the alternative, $\boldsymbol{\mu} \not\leq \mathbf{0}$, we have that $\varphi(n^{1/2}\bar{\mathbf{d}}, \mathbf{V}_n) \xrightarrow{p} \infty$.*

The test statistic, T_n^{SPA}, satisfies Assumption 2 whereas that of the RC does not. It is nevertheless possible to obtain critical values for T_n^{RC} from Theorem 1. This is done by applying Theorem 1 to the test statistic $T_n^{\mathrm{RC}^+} = \max(T_n^{\mathrm{RC}}, 0)$ that satisfies Assumption 2 and noting that the distributions of $T_n^{\mathrm{RC}^+}$ and T_n^{RC} coincide on the positive axis, which is the relevant support for the critical value. Alternatively, the asymptotic distribution of T_n^{RC} can be obtained directly as we do in the following corollary.

Corollary 2 *Let $m_o \leq m$ be the number of models with $\mu_k = 0$, and define Σ to be the $m_o \times m_o$ submatrix of Ω that contains the (i, j)'th element of Ω if $\mu_i = \mu_j = 0$, and let ζ_Σ denote*

the distribution of $Z_{\max} \equiv \max_{j=1,\ldots,m_o} Z_j^o$, *where* $\mathbf{Z}^o = (Z_1^o, \ldots, Z_{m_o}^o)' \sim N_{m_o}(0, \Sigma)$. *Then* $T_n^{\mathrm{RC}} \xrightarrow{d} \zeta_\Sigma$ *if* $\max_k \mu_k = 0$, *whereas* $T_n^{\mathrm{RC}} \xrightarrow{p} -\infty$ *if* $\mu_k < 0$ *for all* $k = 1, \ldots, m$. *Under the alternative where* $\mu_k > 0$ *for some* k, *it holds that* $T_n^{\mathrm{RC}} \xrightarrow{p} \infty$.

Theorem 1 and Corollary 2 have shown that it is only the binding constraints (those with $\mu_k = 0$) that matter for the asymptotic distribution. Naturally the number of binding constraints can be small relative to the number of inequalities, m, that are being tested. This result is known from the problem of testing linear inequalities in linear (regression) models, see Perlman (1969), Wolak (1987, 1989b), Robertson, Wright, and Dykstra (1988), and Dufour (1989), and see Wolak (1989a, 1991) for tests of nonlinear inequalities. The testing problem is also related to that in Gouriéroux, Holly, and Monfort (1982), King and Smith (1986), and Andrews (1998) where the alternative is constrained by inequalities. See Goldberger (1992) for a nice discussion of the relation between the two testing problems.

An immediate consequence of Corollary 2 is that the RC is easy to manipulated by the inclusion of *irrelevant alternative models*. The p-value can be increased in an artificial way by adding poor forecasts to the set of alternative forecasts (i.e. by increasing m while m_o remains constant). In other words, it is possible to eroded the power of the RC to zero by including poor alternatives in the analysis. Naturally, we would want to avoid such properties, to the extend that this is possible.

Since the test statistics have asymptotic distributions that depend on $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$, these are nuisance parameters. The traditional way to proceed in this case, is to substitute a consistent estimator for $\boldsymbol{\Omega}$ and employ the LFC over the values of $\boldsymbol{\mu}$ that satisfy the null hypothesis. In the present situation, the point least favorable to the alternative is $\boldsymbol{\mu} = \mathbf{0}$, which presumes that all alternatives are as good as the benchmark. In the next subsection we explore an alternative way to handle the nuisance dependence on $\boldsymbol{\mu}$, where we use a data dependent choice for $\boldsymbol{\mu}$, rather than $\boldsymbol{\mu} = \mathbf{0}$ as dictated by the LFC.

Figure 2 about here

Figure 2 illustrates a situation for $m = 2$, where the two-dimensional plane represents the sampling space for $\bar{\mathbf{d}} = (\bar{d}_1, \bar{d}_2)'$. We have plotted a realization of $\bar{\mathbf{d}}$, which is in the neighborhood of its true expected value, $\boldsymbol{\mu} = (\mu_1, \mu_2)'$ and the ellipse around $\boldsymbol{\mu}$ is meant to illustrate the covariance structure of $\bar{\mathbf{d}}$. The shaded area represent the values of $\boldsymbol{\mu}$ that conform with the null hypothesis. Because we have placed $\boldsymbol{\mu}$ outside this shaded area, the situation in Figure 2 is one where the null hypothesis is false. The RC is a LFC-based test, so it derives critical values as if $\boldsymbol{\mu} = \mathbf{0}$ (the origin, $o = (0, 0)'$ of the figure). The critical value, C_{RC}, is illustrated by the dashed line, such that the area above and to the right of the dashed line defines the critical region of the RC. The shape of the critical region follows from the definition of T_n^{RC}. Because $\bar{\mathbf{d}}$ is outside the critical region, the RC fails to reject the (false) null hypothesis in this example.

2.4 The Distribution under the Null Hypothesis

Hansen (2003) proposed an alternative to the LFC approach that leads to more powerful tests of composite hypotheses. The LFC is based on a supremum that is taken over the null hypothesis, whereas the idea in Hansen (2003) is to take the supremum over a smaller (confidence) set that is chosen such that it contains the true parameter with a probability that converges to one. In this paper, we use a closely related procedure that is based directly on the asymptotic distributions of Theorem 1 and Corollary 2.

In the previous subsection, we saw that the poor alternatives are irrelevant for the asymptotic distribution. So a proper test should reduce the influence of these models, while preserving the influence of the models with $\mu_k = 0$. It may be tempting to simply exclude the alternatives with $\bar{d}_k < 0$ from the analysis. However, this approach does not lead to valid inference in general, because the models that are (or appear to be) a little worse than the benchmark, can have a substantial influence on the distribution of the test statistic in finite samples (and even asymptotically if $\mu_k = 0$). So we construct our test in a way that incorporates all models, while it reduces the influence of alternatives that the data suggest are poor.

Our choice of estimator, $\hat{\boldsymbol{\mu}}^c$, is motivated by the law of the iterated logarithm that states that

$$
\begin{aligned}
P\left(\liminf_{n\to\infty} \frac{n^{1/2}(\bar{d}_k - \mu_k)}{\omega_k} = -\sqrt{2\log\log n}\right) &= 1, \quad \text{and} \\
P\left(\limsup_{n\to\infty} \frac{n^{1/2}(\bar{d}_k - \mu_k)}{\omega_k} = +\sqrt{2\log\log n}\right) &= 1.
\end{aligned}
$$

The first equality shows that $\hat{\mu}_k^c$ effectively captures all the elements of $\boldsymbol{\mu}$ that are zero. I.e. $\mu_k = 0 \Rightarrow \hat{\mu}_k^c = 0$ almost surely. Similarly, if $\mu_k < 0$ the second equality states that $\bar{d}_k$ will be very close to μ_k, in fact $n^{1/2}\bar{d}_k$ is smaller than $-n^{1/2-\epsilon}$ for any $\epsilon > 0$ and n sufficiently large. Thus $n^{1/2}\bar{d}_k/\omega_k$ is, in particular, smaller than the threshold rate, $-\sqrt{2\log\log n}$, for n sufficiently large, which shows that $\bar{d}_k$ eventually will stay below the implicit threshold in our definition of $\hat{\mu}_k^c$, such that $\mu_k < 0 \Rightarrow \hat{\mu}_k^c \ll 0$ almost surely. So $\hat{\boldsymbol{\mu}}^c$ meet the necessary asymptotic requirements that we identified in Theorem 1 and Corollary 2.

While the poor alternatives should be discarded asymptotically, this is not true in finite samples as we have discussed earlier. Our estimator, $\hat{\boldsymbol{\mu}}^c$, explicitly accounts for this by keeping all alternatives in the analysis. A poor alternative, $\mu_k < 0$, still has an impact on the critical value whenever $\mu_k/(\omega_k n^{1/2})$ is only moderately negative, say between -1 and 0. This is the reason that *the poor performing alternatives cannot simply be omitted from the analysis*. We emphasize this point because an earlier version of this paper has incorrectly been quoted for "discarding the poor models".

While $\hat{\boldsymbol{\mu}}^c$ leads to a correct separation of good and poor alternatives, there are other threshold rates that also produce valid tests. The rate $\sqrt{2\log\log n}$ is the slowest rate that captures all alternatives with $\mu_k = 0$, whereas the faster rate, $n^{1/2-\epsilon}$, for any $\epsilon > 0$, guarantees that all the poor models are discarded asymptotically. So there is a wide range of rates that can be used to asymptotically discriminate between good and poor alternatives. One example is $\frac{1}{4}n^{1/4}$ that was used in a previous version of this paper. Because different threshold rates will lead to different p-values in finite samples it is convenient to determine an upper and lower bound the p-values that different threshold rates would result in. These are easily obtained by using the "estimators", $\hat{\boldsymbol{\mu}}^l$ and $\hat{\boldsymbol{\mu}}^u$, given by, $\hat{\mu}_k^l \equiv \min(\bar{d}_k, 0)$ and $\hat{\mu}_k^u = 0$, $k = 1, \ldots, m$, where the latter yields the LFC-based test.

It is simple to verify that $\hat{\boldsymbol{\mu}}^l \leq \hat{\boldsymbol{\mu}}^c \leq \hat{\boldsymbol{\mu}}^u$, which, in part, motivates the superscripts, and we have the following result, where F_o is the cdf of $\varphi(\mathbf{Z}, \mathbf{v}_0)$ that we defined in Theorem 1.

Theorem 3 *Let F_n^i be the cdf of $\varphi(n^{1/2}\mathbf{Z}_n^i, \mathbf{V}_n)$, for $i = l$, c, or u, where $n^{1/2}(\mathbf{Z}_n^i - \hat{\boldsymbol{\mu}}^i) \xrightarrow{d} N_m(\mathbf{0}, \boldsymbol{\Omega})$. Suppose that Assumptions 1 and 2 hold, then $F_n^c \to F_o$ as $n \to \infty$, for all continuity points of F_o and $F_n^l(x) \leq F_n^c(x) \leq F_n^u(x)$ for all n and all $x \in \mathbb{R}$.*

Theorem 3 shows that $\hat{\boldsymbol{\mu}}^c$ leads to a consistent estimate of the asymptotic distribution of our test statistic. The theorem also shows that $\hat{\boldsymbol{\mu}}^l$ and $\hat{\boldsymbol{\mu}}^u$ provide an upper and lower bound for the distribution, F_n^c, which can be useful in practice. E.g., a substantial difference between these bounds is indicative of the presence of poor alternatives, in which case the sample dependent null distribution is useful.

Given a value for the test statistic $t = T_n(\mathbf{d}_1, \ldots, \mathbf{d}_n)$ it is natural to define the *true asymptotic p-value* as $p_o(t) \equiv P(T > t)$ where $T \sim F_o$, such that $p_o(t) = 1 - F_o(t)$ whenever F_0 is continuous at t. The empirical p-value is to be deduced from an estimate of F_n^i, $i = l, c, u$, and the following corollary shows that $\hat{\boldsymbol{\mu}}^c$ yields a consistent p-value.

Corollary 4 *Consider the studentized test statistic, $t = T_n^{\mathrm{SPA}}(\mathbf{d}_1, \ldots, \mathbf{d}_n)$. Let the empirical p-value, $\hat{p}_n^c(t)$, be inferred from $\hat{F}_n^c$, where $\hat{F}_n^c(t) - F_n^c(t) = o(1)$ for all t. Then $\hat{p}_n^c(t) \xrightarrow{p} p_o(t)$ for any $t > 0$.*

The two other choices, $\hat{\boldsymbol{\mu}}^l$ and $\hat{\boldsymbol{\mu}}^u$, do not produce consistent p-values in general. It follows directly from Theorem 1 that $\hat{\boldsymbol{\mu}}^u$ will not produce a consistent p-value, unless $\boldsymbol{\mu} = \mathbf{0}$. That the p-value from using $\hat{\boldsymbol{\mu}}^l$ is inconsistent is easily understood by noting that a critical value that is based on $N(\mathbf{0}, \boldsymbol{\Omega})$, will be greater than one that is based on the mixed Gaussian distribution, $N(n^{1/2}\hat{\boldsymbol{\mu}}^l, \boldsymbol{\Omega})$. So a p-value that is based on $\hat{\boldsymbol{\mu}}^l$ is (asymptotically) smaller than the correct p-value, which makes this a liberal test despite of the fact that $\hat{\boldsymbol{\mu}}^l \xrightarrow{p} \boldsymbol{\mu}$ under the null hypothesis. This problem is closely related to the inconsistency of the bootstrap, when a parameter is on the boundary of the parameter space, as analyzed by Andrews (2000). In our situation the inconsistency arises because $\boldsymbol{\mu}$ is on the boundary of the null hypothesis, which leads to a violation of a *similarity on the boundary* condition, see Hansen (2003). See Cox and Hinkley (1974, p. 150)

and Gourieroux and Monfort (1995, chapter 16) for a discussion of the finite-sample version of this similarity condition.

Figure 3 about here

Figure 3 shows how the consistent estimate of the null distribution can improve the power. Recall the situation from Figure 2 where the null hypothesis is false. The data dependent null distribution is defined from a projection of $\bar{\mathbf{d}} = (\bar{d}_1, \bar{d}_2)'$ onto the set of parameter values that conform with the null hypothesis. This yields the point a which represents $\hat{\boldsymbol{\mu}}^l = \hat{\boldsymbol{\mu}}^c$ (assuming that $\bar{d}_2$ is below the relevant $2 \log \log n$-threshold). The critical region of the SPA test (induced by $\bar{\mathbf{d}}$) is the area above and to the right of the dotted line marked by C_{SPA}. Because $\bar{\mathbf{d}}$ is in the critical region, the SPA-test (correctly) rejects the null hypothesis in this case.

3 Bootstrap Implementation of the Test for SPA

In this section we describe a bootstrap implementation of the SPA tests in details. The implementation is based on the stationary bootstrap of Politis and Romano (1994), but it is straight forward to modify the implementation to the block bootstrap of Künsch (1989). While there are arguments that favor the block bootstrap over the stationary bootstrap, see Lahiri (1999), these advantages require the use of an optimal block-length that is hard to determine when m is large relative to n, as will often be the case when testing for SPA.

The stationary bootstrap of Politis and Romano (1994) is based on pseudo time-series of the original data. The pseudo time-series, $\{\mathbf{d}^*_{b,t}\} \equiv \{\mathbf{d}_{\tau_{b,t}}\}$, $b = 1, \ldots, B$, are resamples of $\mathbf{d}_t$, where $\{\tau_{b,1}, \ldots, \tau_{b,n}\}$ is constructed by combining blocks of $\{1, \ldots, n\}$ with random lengths. The leading case is that where the block-length is chosen to be geometrically distributed with parameter $q \in (0, 1]$, but the block-lenght may be randomized differently as discussed in Politis and Romano (1994). Here we follow the conventional setup of the stationary bootstrap. The B resamples[2] can be generated from two random $B \times n$ matrices, $\mathbf{U}$ and $\mathbf{V}$, where the elements, $u_{b,t}$

[2]The number of bootstrap resamples, B, should be chosen to be sufficiently large such that the results are not affected by the actual draws of $\tau_{b,t}$. This can be achieved by increasing B until the results are robust to increments, or one can apply more formal methods, such as the three-step method of Andrews and Buchinsky (2000).

and $v_{b,t}$, are independent and uniformly distributed on (0, 1]. The first element of each resample is defined by $\tau_{b,1} = \lceil n\, u_{b,1} \rceil$, where $\lceil x \rceil$ is the smallest integer that is larger than or equal to x. For $t = 2, \ldots, n$ the elements are given recursively, by

$$\tau_{b,t} = \begin{cases} \lceil n\, u_{b,1} \rceil & \text{if } v_{b,t} < q, \\ 1_{\{\tau_{b,t-1} < n\}} \tau_{b,t-1} + 1 & \text{if } v_{b,t} \geq q. \end{cases}$$

So with probability q, the tth element is chosen uniformly on $\{1, \ldots, n\}$ and with probability $1 - q$, the tth element is chosen to be the integer that follows $\tau_{b,t-1}$, unless $\tau_{b,t-1} = n$ in which case $\tau_{b,t} \equiv 1$. The block bootstrap is very similar to the stationary bootstrap, but instead of using blocks with random length, the block bootstrap combines blocks of equal length.

From the pseudo time-series, we calculate their sample averages, $\bar{\mathbf{d}}^*_b \equiv n^{-1} \sum_{t=1}^{n} \mathbf{d}^*_{b,t}$, $b = 1, \ldots, B$, that can be viewed as (asymptotically) independent draws from the distribution of $\bar{\mathbf{d}}$, under the bootstrap distribution. So this provides an intermediate step to estimate the distribution of our test statistic.

Lemma 5 *Let Assumption 1 hold and suppose that the bootstrap parameter, $q = q_n$, satisfies $q_n \to 0$ and $n\, q_n^2 \to \infty$ as $n \to \infty$. Then*

$$\sup_{z \in \mathbb{R}^m} \left| P^*(n^{1/2}(\bar{\mathbf{d}}^*_b - \bar{\mathbf{d}}) \leq \mathbf{z}) - P(n^{1/2}(\bar{\mathbf{d}} - \boldsymbol{\mu}) \leq \mathbf{z}) \right| \xrightarrow{p} 0,$$

where P^* denotes the bootstrap probability measure.

The theorem shows that the empirical distribution of the pseudo time-series can be used to approximate the distribution of $n^{1/2}(\bar{\mathbf{d}} - \boldsymbol{\mu})$. This result follows directly from Goncalves and de Jong (2003, theorem 2), who derived the result under slightly weaker assumptions than we have stated. (Their assumptions are formulated for near-epoch dependent processes). The test statistic, T_n^{SPA}, requires estimates of ω_k^2, $k = 1, \ldots, m$. An earlier version of this paper was based on the estimator:

$$\hat{\omega}^{*2}_{k,B} \equiv B^{-1} \sum_{b=1}^{B} \left(n^{1/2} \bar{d}^*_{k,b} - n^{1/2} \bar{d}_k \right)^2,$$

where $\bar{d}^*_{k,b} = n^{-1}\sum_{t=1}^{n} d_{k,\tau_{b,t}}$. By the law of large numbers, this estimator is consistent for the bootstrap-population value of the variance, which, in turn, is consistent for the true variance, ω_k^2, see Goncalves and de Jong (2003, theorem 1). However, it is our experience that B needs to be quite large to sufficiently reduce the additional layer of randomness that is introduced by the resampling scheme. So our recommendation is to use the bootstrap-population value directly, which is given by

$$\hat{\omega}_k^2 \equiv \hat{\gamma}_{0,k} + 2\sum_{i=1}^{n-1}\kappa(n,i)\hat{\gamma}_{i,k},$$

where

$$\hat{\gamma}_{i,k} \equiv n^{-1}\sum_{j=1}^{n-i}(d_{k,j} - \bar{d}_k)(d_{k,j+i} - \bar{d}_k), \quad i = 0, 1, \ldots, n-1,$$

are the usual empirical covariances and the kernel weights (under the stationary bootstrap) are given by

$$\kappa(n,i) \equiv \frac{n-i}{n}(1-q)^i + \frac{i}{n}(1-q)^{n-i},$$

see Politis and Romano (1994).

We seek the distribution of the test statistics under the null hypothesis, and we impose the null by recentering the bootstrap variables about $\hat{\boldsymbol{\mu}}^l$, $\hat{\boldsymbol{\mu}}^c$, or $\hat{\boldsymbol{\mu}}^u$. This is done by defining:

$$Z^*_{k,b,t} \equiv d^*_{k,b,t} - g_i(\bar{d}_k), \qquad i = l, c, u, \quad b = 1, \ldots, B, \quad t = 1, \ldots, n,$$

where $g_l(x) = \max(0, x)$, $g_c(x) = x \cdot 1_{\left\{x \geq -\sqrt{(\hat{\omega}_k^2/n)2\log\log n}\right\}}$, and $g_u(x) = x$. It is simple to verify that the expected value of $Z^*_{k,b,t}$ (conditional on $\mathbf{d}_1, \ldots, \mathbf{d}_n$) is given by $\hat{\boldsymbol{\mu}}^l$, $\hat{\boldsymbol{\mu}}^c$, and $\hat{\boldsymbol{\mu}}^u$, respectively.

Corollary 6 *Let Assumption 1 hold and let $\mathbf{Z}^*_{b,t}$ be centered about $\hat{\boldsymbol{\mu}}$, for $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}^l$, $\hat{\boldsymbol{\mu}}^c$, or $\hat{\boldsymbol{\mu}}^u$. Then*

$$\sup_{z\in\mathbb{R}^m}\left|P^*(n^{1/2}(\overline{\mathbf{Z}}^*_b - \hat{\boldsymbol{\mu}}) \leq \mathbf{z}) - P(n^{1/2}(\bar{\mathbf{d}} - \boldsymbol{\mu}) \leq \mathbf{z})\right| \xrightarrow{p} 0,$$

where $\bar{Z}^*_{k,b} = n^{-1}\sum_{t=1}^{n} Z^*_{k,b,t}$, $k = 1, \ldots, m$.

Given our assumptions about the test statistic, Corollary 6 shows that we can approximate the

distribution of our test statistics under the null hypothesis, by the empirical distribution we obtain from the bootstrap resamples $\mathbf{Z}^*_{b,t}$, $t = 1, \ldots, n$. The p-values of the three tests for SPA are now simple to obtain. We calculate $T^{\text{SPA}*}_{b,n} = \max\{0, \max_{k=1,\ldots,m}[n^{1/2}\bar{Z}^*_{k,b}/\hat{\omega}_k]\}$ for $b = 1, \ldots, B$ and the bootstrap p-value is given by

$$\hat{p}_{\text{SPA}} \equiv \sum_{b=1}^{B} \frac{1_{\{T^{\text{SPA}*}_{b,n} > T^{\text{SPA}}_n\}}}{B},$$

where the null hypothesis should be rejected for small p-values. Thus we obtain three p-values, one for each of the estimators, $\hat{\boldsymbol{\mu}}^l$, $\hat{\boldsymbol{\mu}}^c$, and $\hat{\boldsymbol{\mu}}^u$. The p-values that are based on the test statistic, T^{RC}_n, can be derived similarly.

Note that we are using the same estimate of ω^2_k to calculate T^{SPA}_n and $T^{\text{SPA}*}_{b,n}$, $b = 1, \ldots, B$. A nice robustness of the SPA-test is that it is valid even if $\hat{\omega}^2_k$ is inconsistent for ω^2_k. This is easy to understand by recalling that $\hat{\omega}^2_k = 1$ for all k leads to the RC, (and 1 is generally inconsistent for ω^2_k). While this robustness is convenient, it is desirable that $\hat{\omega}^2_k$ be close to ω^2_k, such that the individual statistics, $n^{1/2}\bar{d}_k/\hat{\omega}_k$, are close to having the same scale, due to the power issues we discussed in Section 2.

4 Size and Power Comparison by Monte Carlo Simulations

The two test statistics, T^{RC}_n and T^{SPA}_n, and the three null distributions (about $\hat{\boldsymbol{\mu}}^l$, $\hat{\boldsymbol{\mu}}^c$, and $\hat{\boldsymbol{\mu}}^u$) result in six different tests. In this section, we study the size and and power properties of these tests in a simple Monte Carlo experiment.

We generate $L_{k,t} \sim \text{iid}\, N(\lambda_k/\sqrt{n}, \sigma^2_k)$, for $k = 0, 1, \ldots, m$ and $t = 1, \ldots, n$, where the benchmark model has $\lambda_0 = 0$. So positive values ($\lambda_k > 0$) correspond to alternatives that are worse than the benchmark whereas negative values ($\lambda_k < 0$) correspond to alternatives that are better than the benchmark.

In our experiment we have $\lambda_1 \leq 0$ and $\lambda_k \geq 0$ for $k = 2, \ldots, m$, such that the first alternative ($k = 1$) defines whether the rejection probability corresponds to a Type I error ($\lambda_1 = 0$) or a power ($\lambda_1 < 0$). The performance of the "poor" models, are such that their mean-values are

spread evenly between zero and $\lambda_m = \Lambda_0$ (the worst model). So the vector of λ_ks are given by

$$\boldsymbol{\lambda} \equiv \begin{pmatrix} \lambda_0 \\ \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \vdots \\ \lambda_{m-1} \\ \lambda_m \end{pmatrix} = \begin{pmatrix} 0 \\ \Lambda_1 \\ \frac{1}{m-1}\Lambda_0 \\ \frac{2}{m-1}\Lambda_0 \\ \vdots \\ \frac{m-2}{m-1}\Lambda_0 \\ \Lambda_0 \end{pmatrix}.$$

In our experiments we use $\Lambda_0 = 0, 1, 2, 5, 10$ to control the extent to which the inequalities are binding ($\Lambda_0 = 0$ corresponds to the case where all inequalities are binding). The first alternative model has $\Lambda_1 = 0, -1, -2, -3, -4, -5$. So $\lambda_1 = \Lambda_1$ defines the local alternative that is being analyzed (unless $\Lambda_1 = 0$ which conforms with the null hypothesis). To make the experiment more realistic, we tie the variance, σ_k^2, to the "quality" of the model. Specifically we set

$$\sigma_k^2 = \tfrac{1}{2}\exp(\arctan(\lambda_k)),$$

such that a good model has a smaller variance than poor model. Note that this implies that

$$\sqrt{n}\bar{d}_k \sim N(\mu_k, \omega_k^2), \qquad \text{where } \mu_k = \lambda_k/\sqrt{n} \quad \text{and} \quad \omega_k^2 \simeq 1 + \tfrac{1}{2}\lambda_k + \tfrac{1}{4}\lambda_k^2 - \tfrac{1}{12}\lambda_k^3,$$

where the expression for ω_k^2 now follows from $\text{var}(d_{k,t}) = \text{var}(L_{0,t} - L_{k,t}) = \frac{1}{2} + \text{var}(L_{k,t})$ and the Taylor expansion (about zero)

$$\tfrac{1}{2}\exp(\arctan(x)) = \tfrac{1}{2}[1 + x + \tfrac{1}{2}x^2 - \tfrac{1}{6}x^3 - \tfrac{7}{24}x^4 + O\left(x^5\right)].$$

4.1 Simulation Results

We consider first the case with $m = 100$, where we generate result for the sample sizes, $n = 200$ and $n = 1000$. Next, we consider the case with $m = 1000$ using the sample size $n = 200$. The

rejection frequencies we report are based on 10,000 independent samples, where we used $q = 1$ in accordance with the lack of time-dependence in $\mathbf{d}_t$, $t = 1, \ldots, n$.[3] The rejection frequencies of the tests at levels 5% and 10% are reported in Tables 2, 3, and 4. Numbers in italic font are used when the null hypothesis is true, ($\Lambda_1 = 0$). So these frequencies correspond to Type I errors. Numbers in standard font represent powers for the various local alternatives ($\Lambda_1 < 0$).

Tables 2 and 3 about here

Table 2 contains the results for the case where $m = 100$ and $n = 200$. In the situation where all 100 inequalities are binding, ($\Lambda_0 = \Lambda_1 = 0$), we see that the rejection probabilities are close to the nominal levels for all the tests. The SPA_c-test has an over-rejection by 1%, this over-rejection appears to be a small sample problem, because it disappears when the sample size is increased to $n = 1000$, see Table 3. The fact that the liberal null distribution does not lead to a larger over-rejection is interesting. This finding may be due to the positive correlation across alternatives, $\text{cov}(d_{i,t}, d_{j,t}) = \text{var}(L_{0,t}) > 0$, which creates a positive correlation between the test statistic and $\hat{\boldsymbol{\mu}}^l$. So the critical value will tend to be (too) small when the test statistic is small and this correlation will reduce the over-rejection of the tests that are based on $\hat{\boldsymbol{\mu}}^l$. This suggest that our test may be improved additionally, if there is a reliable way to incorporate information about the off-diagonal elements of $\boldsymbol{\Omega}$. We do not pursue this aspect in this paper.

Panel A corresponds to the case where $\boldsymbol{\mu} = \mathbf{0}$, and is therefore the best possible situation for LFC-based tests. So this is the (unique) situation where the LFC-based tests apply the correct asymptotic distribution, and it is therefore not surprising that the tests that are based on $\hat{\boldsymbol{\mu}}^u = \mathbf{0}$ do well. Fortunately, our new test, SPA_c, also performs well in this case. When we turn to the configurations where $\Lambda_0 > 0$ we immediately see the advantages of using the sample dependent null distribution. A somewhat extreme situation is observed in Table 2 Panel E for $(\Lambda_0, \Lambda_1) = (10, -3)$, where the RC almost never rejects the null hypothesis, while the new SPA_c-test has a power that is close to 84%.

Table 4 about here

[3] All simulations were made using Ox 3.30, see Doornik (1999).

Table 4 is quite interesting because this is a situation where $m = 1000$ exceeds the sample size $n = 200$, such that it would be impossible to estimate $\mathbf{\Omega}$ without imposing a restrictive structure on its coefficients. So using standard first-order asymptotics is not a viable alternative to the bootstrap implementation in this situation. Since the bootstrap invokes an implicit estimate of $\mathbf{\Omega}$ one might worry about its properties in this situation, where an explicit estimated is unavailable. Nevertheless, the bootstrap does surprisingly well and we only notice a slight over-rejection when all inequalities are binding, ($\Lambda_0 = \Lambda_1 = 0$). The power properties are quite good, despite the fact that 1000 alternative are being compared to the benchmark.

Figure 4 about here

The power curves for the tests that employ $\hat{\boldsymbol{\mu}}^c$ and $\hat{\boldsymbol{\mu}}^u$ are shown in Figure 4, for the case where $m = 100$, $n = 200$ and $\Lambda_0 = 20$. The power curves are based on tests that aim at a 5% significance level, and their rejection frequencies are plotted against a range of local alternatives. These rejection frequencies have not been adjusted for their under-rejection at $\Lambda_1 = 0$. This is a fair comparison, because it would not be possible to make such an adjustment in practice without exceeding the intended level of the test for other configurations – particularly the case where $\Lambda_0 = \Lambda_1 = 0$. See Horowitz and Savin (2000) for a criticism of reporting 'size'-adjusted powers. From the power curves in Figure 4, it is clear that the RC is dominated by the three other tests. There is a substantial increase in power from using the consistent distribution, and a similar improvement is achieved by using the standardized test statistic, T_n^{SPA}. For example, the local alternative $\Lambda_1 = -4$ is rejected by the RC in about 5.5%. Using either the data dependent null distribution (RC_c) or the studentization (SPA_u) improves the power to about 73.6% and 96.4%, respectively. Invoking both modifications (as advocated in this paper) improves the power to 99.7% in this simulation experiment. So both modifications are very useful and the combination of the two substantially improves the power.

Comparing the sample sizes that would result in the same power is an effective way to convey the relative efficiency of the tests. For the configuration that was used in Figure 4, we see that the four tests have 50% power at the local alternatives, $\mu_1/\sqrt{n} \simeq 2.13$, 2.60, 3.63, and 5.28, respectively. Thus we would need a sample size that is $(2.60/2.13)^2 = 1.49$ times larger to

regain the power that is lost by using the LFC instead of the sample dependent null distribution. In other words, using the LFC is equivalent to tossing away about 33% of the data. Similarly, dropping the studentization is equivalent to tossing away about 65% of the data, and dropping both modifications, (i.e. using the RC instead of SPA_c), is equivalent to tossing away about 84% of the data in this simulation design.

5 Forecasting US Inflation using Linear Regression Models

In an attempt to forecast annual US inflation, we estimate a large number of linear regression models that are used to construct competing forecasts. The annual US inflation is defined by $Y_t \equiv \log[P_t/P_{t-4}]$, where P_t is the GDP price deflator for the tth quarter. Inflation and most of the variables are not observed instantaneously. For this reason, we let the set of potential regressors consists of variables that are lagged five quarters or more. This leaves time (one quarter) for most of our regressors to be observed at the beginning of the 12-months period that we attempt to forecast the inflation for.

The linear regression models include one, two, or three regressors out of the pool of 27 regressors, $X_{1,t}, \ldots, X_{27,t}$, which leads to a total of 3303 regression models. Descriptions and definitions of the regressors are given in Table 5.

Table 5 about here

The sequence of forecasts that is produced by the kth regression model is given by

$$\hat{Y}_{k,\tau+5} \equiv \hat{\boldsymbol{\beta}}'_{(k),\tau}\mathbf{X}_{(k),\tau}, \qquad \tau = 0, \ldots, n-1,$$

where $\mathbf{X}_{(k),\tau}$ contains the regressors included in model k and $\hat{\boldsymbol{\beta}}'_{(k),\tau}$ is the least squares estimator based on the 32 most recent observations (a rolling window). Thus, $\hat{\boldsymbol{\beta}}_{(k),\tau} \equiv (\mathbf{X}'_{k,\tau}\mathbf{X}_{k,\tau})^{-1}\mathbf{X}'_{k,\tau}\mathbf{Y}_\tau$, where the rows of $\mathbf{X}_{k,\tau}$ are given by $\mathbf{X}'_{(k),t-5}$, $t = \tau - 32 + 1, \ldots, \tau$, and similarly the elements of $\mathbf{Y}_\tau$ are given by Y_τ, $t = \tau - 32 + 1, \ldots, \tau$. Using a rolling-window estimation scheme ensures that stationarity of $(\mathbf{X}_t, Y_t)$ is carried over to $L(Y_{t+h}, \hat{\boldsymbol{\beta}}'_{(k),t}\mathbf{X}_{(k),t})$, whereby an obvious violation

of Assumption 1 is avoided. For example, it is difficult to reconcile Assumption 1 with the case where $\boldsymbol{\beta}_{(k)}$ is estimated recursively (using an expanding window of observation).

The first forecast of annual inflation is made at time 1959:Q4 (predicting inflation for the year 1960:Q1–1961:Q1). So the evaluation period includes $n = 160$ quarters.

$$t = \underbrace{1952\text{:Q1}, \ldots, 1959\text{:Q4}}_{\text{initial estimation period}}, \quad \underbrace{1961\text{:Q1}, \ldots, 2000\text{:Q4}}_{\text{evaluation period}}.$$

The models are evaluated using a mean absolute error criterion (MAE) given by, $L(Y_t, \hat{Y}_{k,t}) = |Y_t - \hat{Y}_{k,t}|$, and the best performing models have a Phillips curve structure. In fact, the best forecasts are produced by regressors that measure (changes in): inflation, interest rates, employment, and GDP, and the very best sample performance is achieved by the three regressors, $X_{1,t}$, $X_{8,t}$, and $X_{13,t}$, that represent *annual inflation*, *employment relative to the previous year's employment*, and the *change in GDP*, respectively, see Table 5. We also include the average forecast (average across all regression-based forecasts), because this simple combination of forecasts is often found to dominate the individual forecasts, see for example Stock and Watson (1999). In addition to the average forecast, the 27 regressors leads 3303 regression-based forecasts when we consider all possible subset regressions with one, two, or three regressors. So we are to compare $m = 3304$ forecasts to the random walk benchmark and we refer to this set of competing forecasts as the Large Universe.

Table 6 about here

Panel A of Table 6 contains the output produced by the tests for SPA for the Large Universe. Since the SPA_c p-value is .832 there is no statistical evidence that any of the regression-based forecasts (including the average of them) is better than the random walk forecast. Note the discrepancy between the p-values that are based on $\hat{\boldsymbol{\mu}}^l$ and $\hat{\boldsymbol{\mu}}^u$. This difference suggests that some of the alternatives are poor forecasts and a closer inspection of the Large Universe verifies that several models have a performance that is substantially worse than the benchmark.

The ability to construct better forecasts using models with additional regressors is made difficult by the need to estimate additional parameters. In a forecasting exercise there is a trade-off

between estimating a parameter and imposing it to have a particular value (typically zero, which is implicitly imposed on the coefficient of an omitted regressor). Imposing a particular value will (most likely) introduce a 'bias', but if the bias is small it may be less severe for out-of-sample predictions than the prediction error that is introduced by estimation error, see, e.g., Clements and Hendry (1998). Exploiting this bias-variance trade-off is particularly useful whenever the estimator is based on a moderate number of observations, as is the case in our application. For this reason we also consider a Small Universe of regression-based alternatives that all include lagged inflation, $X_{1,t}$, as a regressor with a coefficient that is set to unity. The remaining parameters are estimated by ridge regression that shrinks these parameters towards zero.

So the regression models have the form

$$Y_{\tau+5} - Y_\tau \equiv \boldsymbol{\beta}'_{(k)}\mathbf{X}_{(k),\tau} + \varepsilon_{(k),t}, \qquad \tau = 0, \ldots, n-1,$$

where $\mathbf{X}_{(k),\tau}$ is a vector that includes either one or two regressors. As before we use a rolling window scheme (32 quarters), but the estimator for $\boldsymbol{\beta}_{(k)}$ is now given by $\hat{\boldsymbol{\beta}}_{(k),\tau} \equiv (\mathbf{X}'_{k,\tau}\mathbf{X}_{k,\tau} + \lambda\mathbf{I})^{-1}\mathbf{X}'_{k,\tau}\tilde{\mathbf{Y}}_\tau$, where we use $\lambda = 0.1$ as the shrinkage parameter and the elements of $\tilde{\mathbf{Y}}_\tau$ are given by $Y_t - Y_{t-5}$ for $t = \tau - 32 + 1, \ldots, \tau$.

This results in 351 regression-based forecasts plus the average-forecast, such that the total number of alternative forecasts in the Small Universe is $m = 352$. The most accurate forecast in the Small Universe is produced by the regression model with the regressors, $X_{8,t}$ and $X_{9,t}$, that are two measures of (relative) employment. The most significant excess performance (over the benchmark) is produced by the regressions, $X_{6,t}$ and $X_{10,t}$, that represent changes in employment and inventories, respectively. So our findings support a conclusion reached by Stock and Watson (1999), that forecasts that are based on Phillips curve specifications are useful for forecasting inflation.

The empirical results for this universe are presented in Panel B of Table 6. The SPA_c p-value for this universe is 0.048, which suggests that the benchmark is outperformed by the regression-based forecasts. For each of the test statistics we note that the p-values are quite similar. This agreement is not surprising because the worsts forecast is only slightly worse than the benchmark,

such that $\hat{\boldsymbol{\mu}}^l$ and $\hat{\boldsymbol{\mu}}^u$ are similar. The difference in p-values across the two test statistics is fairly modest but do suggest some variation in the variances, ω_k^2, $k = 1, \ldots, 352$.

Reporting the results in Panel B is not fully consistent with the spirit of this paper. The reason is that the results in Panel B do not control for the 3304 forecasting models that were compared to the benchmark in the initial analysis of the Large Universe. So the significant p-values in Panel B are subject to criticizes for data mining. A way to address this issue seriously, is to perform the tests for SPA over the union of the Large Universe and the Small Universe. We refer to this set of forecasts as the Full Universe and the results for this set of alternatives are presented in Panel C. Adding the large number of insignificant alternatives to the comparison reduces the significance, although the excess performance continue to be borderline significant with a SPA_c p-value of 10%. Note that the RC's p-value increases from 10.6% to 96.3% by "adding" the Large Universe to the Small Universe. This jump in the p-value is most likely due to the RC's sensitivity to poor and irrelevant alternatives. The SPA_c-test's p-value increases from 4.8% to 10%. While this increment is more moderate, it reveals that the new test is not entirely immune to the inclusion of (a large number of) poor forecasts. This reminds us that excessive data mining can be costly in terms of the conclusions that can be drawn from the analysis, because it may prevent the researcher from concluding that a particular finding is significant. Given the scarcity of macroeconomic data, it can therefore be useful to confine the set of alternatives to those that are motivated by theoretical considerations, instead of a blind search over a large number of alternatives.

6 Summary and Concluding Remarks

We have analyzed the problem of comparing multiple forecasts to a given benchmark through tests for superior predictive ability. We have shown that the power can be improved (often substantially) by using a studentized test statistic and incorporating additional sample information by means of a data dependent null distribution. The latter serves to identify the irrelevant alternatives and reduce their influence on the test for SPA.

The power improvements were quantified in simulation experiments and an empirical forecasting exercise of US inflation. These also highlighted that the RC is sensitive to poor and

irrelevant alternatives. Two researchers are therefore more likely to arrive at the same conclusion when they use the SPA_c-test, than is the case when they use the RC – even if they do not fully agree on the set of forecasts that is relevant for the analysis.

Interesting we found that the best (and most significant) predictions of US inflation were produced by regression-based forecasts that had a Phillips curve structure. In our Full Universe of alternatives we found that the (random walk) benchmark forecast is outperformed by the regression-based forecasts if a moderate significance level is employed. While the SPA_c-test yields a p-value of 10% the RC yields a p-value of about 96%, such that the two tests arrive at opposite conclusions (weak evidence against H_0 *versus* no evidence). This occurs because the poor alternatives conceal the evidence against the null hypothesis when the RC is used. This phenomenon is also found in Hansen and Lunde (2004) who compared a large number of volatility models, using the theoretical results of the present paper.

While there are several advantages of our new test, there are important issues that need to be addressed in future research. In the present paper, we have proposed two modifications and adopted these in a stationary framework. This framework does not permit the comparison of nested models if the recursive scheme is used to estimate the parameters. So it would be interesting to construct a suitable test for comparing a large number of nested models and analyze our two modifications in this framework.

Despite its many pitfalls, data mining is a constructive device for the discovery of true phenomena and has become a popular tool in many applied areas, such as genetics, e-commerce, and financial services. However it is necessary to account for the full data exploration, before a legitimate statement about significance can be made. Increasing the number of comparisons makes it more difficult to establish significance (other things being equal). This aspect is particularly problematic for economic applications where data are often scarce. In this context it is particularly useful to confine the exploration to alternatives that are motivated by theoretical considerations. Our empirical application provides a good illustration of this issue. Within the Small Universe we found fairly compelling evidence against the null hypothesis, and *ex post* is it easy to produce arguments that motivate the use of shrinkage methods, which led to the Small Universe

of regression-based forecasts. However, because the Large Universe was explored in the initial analysis, we cannot exclude the possibility that the largest t-statistic would have been found in this universe. The weaker evidence against the null hypothesis that is found in the Full Universe is the price we have to pay for the (perhaps unnecessary) data exploration that preceded our analysis of the Small Universe.

A Appendix of Proofs

Proof of Theorem 1. We define the vectors, $\mathbf{W}_n, \mathbf{Z}_n \in \mathbb{R}^m$, whose elements are given by $W_{n,k} = n^{1/2}\bar{d}_k 1_{\{\mu_k<0\}}$ and $Z_{n,k} = n^{1/2}\bar{d}_k 1_{\{\mu_k=0\}}$, $k = 1, \ldots, m$. Thus $\mathbf{U}_n = \mathbf{W}_n + \mathbf{Z}_n$ under the null hypothesis. The mappings (coordinate selectors) that transform $\mathbf{U}_n$ into $\mathbf{W}_n$ and $\mathbf{Z}_n$ are continuous, so that $(\mathbf{W}_n - n^{1/2}\boldsymbol{\mu}) \xrightarrow{d} N(\mathbf{0}, \Omega - \Omega^o)$ and $\mathbf{Z}_n \xrightarrow{d} N(\mathbf{0}, \Omega^o)$, by the continuous mapping theorem. This implies that

$$\varphi(\mathbf{U}_n, \mathbf{V}_n) = \varphi(\mathbf{W}_n + \mathbf{Z}_n, \mathbf{V}_n) = \varphi(\mathbf{Z}_n, \mathbf{V}_n) + o_p(1) \xrightarrow{d} \varphi(\mathbf{Z}, \mathbf{v}_0),$$

where the second equality uses Assumption 2.*b* and the fact that the elements of $\mathbf{W}_n$ are either zero ($\mu_k = 0$) or diverges to minus infinity in probability ($\mu_k < 0$). Under the alternative hypothesis there will be an element of $n^{1/2}\bar{\mathbf{d}}$ that diverges to infinity. So the last result of the theorem follows by Assumption 2.*c*. ■

Proof of Corollary 2. The results follow from $n^{1/2}(\bar{\mathbf{d}} - \boldsymbol{\mu}) \xrightarrow{d} N(\mathbf{0}, \Omega)$ and the continuous mapping theorem, applied to the mapping $\bar{\mathbf{d}} \mapsto \bar{d}_{\max}$. ■

Proof of Theorem 3. Without loss of generality, suppose that $\mu_k = 0$ for $k = 1, \ldots, m_o$ and that $\mu_k < 0$ for $k = m_o + 1, \ldots, m$. Given our definition of $\hat{\boldsymbol{\mu}}^c$ it holds that $P(\hat{\mu}^c_1 = \cdots = \hat{\mu}^c_{m_o} = 0, \hat{\mu}^c_{m_o+1} < \epsilon, \ldots, \hat{\mu}^c_m < \epsilon)$ almost surely as $n \to 0$, for some $\epsilon < 0$. So for n sufficiently large, the last $m - m_o$ elements of $\mathbf{Z}^i_n$ are bounded below zero in probability, which shows that $\hat{\boldsymbol{\mu}}^c$ leads to the correct limiting distribution and $F^c_n \to F_o$. That $F^l_n(x) \leq F^c_n(x) \leq F^u_n(x)$ follows from $\hat{\boldsymbol{\mu}}^l \leq \hat{\boldsymbol{\mu}}^c \leq \hat{\boldsymbol{\mu}}^u$. ■

Proof of Corollary 4. The test statistic T^{SPA}_n leads to a continuous asymptotic distribution, $F_o(t)$,

for all $t > 0$. Since $\hat{F}_n^c(t) - F_o(t) = [\hat{F}_n^c(t) - F_n^c(t)] + [F_n^c(t) - F_o(t)]$, the result now follows, because the first term is $o(1)$ by assumption and the second term is $o(1)$ by Theorem 3. ■

Proof of Lemma 5. Follows from Goncalves and de Jong (2003, theorem 2). ■

Proof of Corollary 6. Since $Z_{k,b,t}^* - \hat{\mu}_k = (d_{k,b,t}^* - g_i(\bar{d}_k)) - (\bar{d}_k - g_i(\bar{d}_k)) = d_{k,b,t}^* - \bar{d}_k$ for all $k = 1, \ldots, m$, the Corollary follows trivially from Lemma **5**. ■

References

ANDREWS, D. W. K. (1998): "Hypothesis Testing with a Restricted Parameter Space," *Journal of Econometrics*, 84, 155–199.

——— (2000): "Inconsistency of the Bootstrap When a Parameter is on the Boundary of the Parameter Space," *Econometrica*, 68, 399–405.

ANDREWS, D. W. K., AND M. BUCHINSKY (2000): "A Three-Steep Method for Choosing the Number of Bootstrap Repetitions," *Econometrica*, 68, 23–52.

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

CLEMENTS, M. P., AND D. F. HENDRY (1998): *Forecasting Economic Time Series*. Cambridge University Press, Cambridge.

CORRADI, V., AND N. R. SWANSON (2004): "Predictive Density and Conditional Confidence Interval Accuracy Tests," Mimeo: http://econweb.rutgers.edu/nswanson/.

CORRADI, V., N. R. SWANSON, AND C. OLIVETTI (2001): "Predictive Ability with Cointegrated Variables," *Journal of Econometrics*, 104, 315–358.

COX, D. R., AND D. V. HINKLEY (1974): *Theoretical Statistics*. Chapman and Hall, London.

DE JONG, R. (1997): "Central Limit Theorems for Dependent Heterogeneous Random Variables," *Econometric Theory*, 13, 353–367.

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

DOORNIK, J. A. (1999): *Ox: An Object-Oriented Matrix Language*. Timberlake Consulants Press, London, 3rd edn.

DUFOUR, J.-M. (1989): "Nonlinear Hypotheses, Inequality Restrictions, and Non-Nested Hypotheses: Exact Simultaneous Test in Linear Regressions," *Econometrica*, 57, 335–355.

DUFOUR, J.-M., AND L. KHALAF (2002): "Exact Tests for Contemporaneous Correlation of Disturbances in Seemingly Unrelated Regressions," *Journal of Econometrics*, 106, 143–170.

FOLKS, L. (1984): "Combination of Independent Tests," in *Handbook of Statistics 4: Nonparametric Methods*, ed. by P. R. Krishnaiah, and P. K. Sen, pp. 113–121. North-Holland, New York.

GIACOMINI, R., AND H. WHITE (2003): "Tests of Conditional Predictive Ability," Boston College Working Paper 572.

GOLDBERGER, A. S. (1992): "One-Sided and Inequality Tests for a Pair of Means," in *Contributions to Consumer Demand and Econometrics*, ed. by R. Bewley, and T. V. Hoa, pp. 140–162. St. Martin's Press, New York.

GONCALVES, S., AND R. DE JONG (2003): "Consistency of the Stationary Bootstrap under Weak Moment Conditions," *Economics Letters*, 81, 273–278.

GOURIÉROUX, C., A. HOLLY, AND A. MONFORT (1982): "Likelihood Ratio Test, Wald Test, and Kuhn-Tucker Test in Linear Models with Inequality Constraints on the Regression Parameters," *Econometrica*, 50, 63–80.

GOURIEROUX, C., AND A. MONFORT (1995): *Statistics and Econometric Models*. Cambridge University Press, Cambridge.

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

HANSEN, P. R., AND A. LUNDE (2004): "A Forecast Comparison of Volatility Models: Does Anything Beat a GARCH(1,1)?," *Journal of Applied Econometrics*, Forthcoming.

——— (2005): "Consistent Ranking of Volatility Models," *Journal of Econometrics*, Forthcoming.

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

HARVEY, D. I., S. J. LEYBOURNE, AND P. NEWBOLD (1997): "Testing the Equality of Prediction Mean Squared Errors," *International Journal of Forecasting*, 13, 281–291.

——— (1998): "Tests for Forecast Encompassing," *Journal of Business and Economic Statistics*, 16, 254–259.

HOROWITZ, J. L., AND N. E. SAVIN (2000): "Empirically Relevant Critical Values for Hypothesis Tests: A Bootstrap Approach," *Journal of Econometrics*, 95, 375–389.

KING, M. L., AND M. D. SMITH (1986): "Joint One-Sided Tests of Linear Regression Coefficients," *Journal of Econometrics*, 32, 367–387.

KÜNSCH, H. R. (1989): "The Jackknife and the Bootstrap for General Stationary Observations," *Annals of Statistics*, 17, 1217–1241.

LAHIRI, S. N. (1999): "Theoretical Comparisons of Block Bootstrap Methods," *Annals of Statistics*, 27, 386–404.

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

MARDEN, J. I. (1985): "Combining Independent One-Sided Noncentral t or Normal Mean Tests," *Annals og Statistics*, 13, 1535–1553.

MCCRACKEN, M. W. (2000): "Robust Out-of-Sample Inference," *Journal of Econometrics*, 99, 195–223.

MILLER, R. G. (1981): *Simultaneous Statistical Inference*. Springer-Verlag, New York, 2nd edn.

PERLMAN, M. D. (1969): "One-Sided Testing Problems in Multivariate Analysis," *The Annals of Mathematical Statistics*, 40, 549–567.

PERLMAN, M. D., AND L. WU (1999): "The Emperor's New Tests," *Statistical Science*, 14, 355–381.

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

ROBERTSON, T., F. T. WRIGHT, AND R. L. DYKSTRA (1988): *Order Restricted Statistical Inference*. John Wiley & Sons, New York.

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

SAVIN, N. E. (1984): "Multiple Hypothesis Testing," in *Handbook of Econometrics*, ed. by K. J. Arrow, and M. D. Intriligator, vol. 2, pp. 827–879. North-Holland, Amsterdam.

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

SULLIVAN, R., A. TIMMERMANN, AND H. WHITE (1999): "Data-Snooping, Technical Trading Rules, and the Bootstrap.," *Journal of Finance*, 54, 1647–1692.

——— (2001): "Dangers of Data-Driven Inference: The Case of Calendar Effects in Stock Returns," *Journal of Econometrics*, 105, 249–286.

——— (2003): "Forecast Evaluation with Shared Data Sets," *International Journal of Forecasting*, 19, 217–227.

TIPPETT, L. H. C. (1931): *The Methods of Statistics*. Williams and Norgate, London.

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

——— (2001): "Tests for Forecast Encompassing When Forecasts Depend on Estimated Regression Parameters," *Journal of Business and Economic Statistics*, 19, 29–33.

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

WESTFALL, P. H., AND S. S. YOUNG (1993): *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustments*. Wiley, New York.

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

WOLAK, F. A. (1987): "An Exact Test for Multiple Inequality and Equality Constrants in the Linear Regression Model," *Journal of the American Statistical Association*, 82, 782–793.

——— (1989a): "Local and Global Testing of Linear and Nonlinear Inequality Constraints in Nonlinear Econometric Models," *Econometric Theory*, 5, 1–35.

——— (1989b): "Testing Inequality Constraints in Linear Econometric Models," *Journal of Econometrics*, 41, 205–235.

——— (1991): "The Local Nature of Hypothesis Tests Involving Inequality Constraints in Nonlinear Models," *Econometrica*, 59, 981–995.

Table 1: Overview of Notation and Definitions

$t = 1, \ldots, n$	Sample period for the model comparison.
$k = 0, 1, \ldots, m$	Model index ($k = 0$ is the benchmark).
ξ_t	Object (variable) of interest.
$\delta_{k,t-h}$	The kth decision rule (e.g. h-step-ahead forecast of ξ_t).
$L_{k,t} \equiv L(\xi_t, \delta_{k,t-h})$	Observed loss of the kth decision rule/forecast.
$d_{k,t} \equiv L_{0,t} - L_{k,t}$	Performance of model k relative to the benchmark.
$\bar{d}_k \equiv n^{-1} \sum_{t=1}^{n} d_{k,t}$	Average relative performance of model k.
$\mathbf{d}_t \equiv (d_{1,t}, \ldots, d_{m,t})'$	Vector of relative performances at time t.
$\bar{\mathbf{d}} \equiv n^{-1} \sum_{t=1}^{n} \mathbf{d}_t$	Vector of average relative performance.
$\mu_k \equiv E(d_{k,t})$	Expected excess performance of model k.
$\mu \equiv (\mu_1, \ldots, \mu_m)'$	Vector of expected excess performances.
$\Omega \equiv \text{avar}(n^{1/2}\bar{\mathbf{d}})$	Asymptotic $m \times m$ covariance matrix.

Table 2: Rejection Frequencies under the Null and Alternative ($m = 100$ and $n = 200$)

	Level: $\alpha = 0.05$						Level: $\alpha = 0.10$					
Λ_1	RC_l	RC_c	RC_u	SPA_l	SPA_c	SPA_u	RC_l	RC_c	RC_u	SPA_l	SPA_c	SPA_u
Panel A: $\Lambda_0 = 0$												
0	*0.055*	*0.053*	*0.053*	*0.062*	*0.060*	*0.060*	*0.108*	*0.101*	*0.101*	*0.116*	*0.110*	*0.109*
-1	0.057	0.054	0.054	0.077	0.074	0.074	0.112	0.105	0.105	0.136	0.129	0.129
-2	0.121	0.111	0.111	0.310	0.280	0.280	0.219	0.197	0.197	0.436	0.389	0.388
-3	0.550	0.471	0.470	0.848	0.764	0.761	0.727	0.620	0.618	0.921	0.845	0.841
-4	0.968	0.888	0.882	0.997	0.979	0.976	0.993	0.947	0.941	1.000	0.990	0.987
-5	1.000	0.996	0.992	1.000	1.000	1.000	1.000	0.999	0.998	1.000	1.000	1.000
Panel B: $\Lambda_0 = 1$												
0	*0.013*	*0.010*	*0.010*	*0.026*	*0.022*	*0.022*	*0.035*	*0.025*	*0.025*	*0.055*	*0.044*	*0.044*
-1	0.013	0.010	0.010	0.047	0.041	0.040	0.036	0.027	0.027	0.087	0.072	0.071
-2	0.036	0.028	0.028	0.312	0.252	0.250	0.084	0.060	0.060	0.436	0.345	0.342
-3	0.301	0.201	0.197	0.862	0.744	0.733	0.516	0.334	0.327	0.928	0.829	0.814
-4	0.896	0.677	0.658	0.998	0.977	0.971	0.971	0.816	0.793	1.000	0.989	0.984
-5	1.000	0.968	0.952	1.000	1.000	0.999	1.000	0.991	0.980	1.000	1.000	1.000
Panel C: $\Lambda_0 = 2$												
0	*0.004*	*0.002*	*0.002*	*0.018*	*0.012*	*0.012*	*0.013*	*0.007*	*0.006*	*0.039*	*0.026*	*0.026*
-1	0.004	0.002	0.002	0.044	0.032	0.032	0.014	0.007	0.006	0.080	0.058	0.056
-2	0.013	0.007	0.006	0.336	0.244	0.238	0.041	0.020	0.019	0.464	0.336	0.324
-3	0.195	0.077	0.073	0.881	0.745	0.721	0.401	0.167	0.152	0.941	0.827	0.799
-4	0.842	0.460	0.414	0.999	0.978	0.968	0.957	0.659	0.598	1.000	0.989	0.982
-5	0.999	0.911	0.855	1.000	1.000	0.999	1.000	0.971	0.934	1.000	1.000	1.000
Panel D: $\Lambda_0 = 5$												
0	*0.002*	*0.000*	*0.000*	*0.014*	*0.007*	*0.005*	*0.008*	*0.001*	*0.000*	*0.032*	*0.013*	*0.011*
-1	0.002	0.000	0.000	0.056	0.031	0.025	0.009	0.001	0.000	0.101	0.054	0.044
-2	0.012	0.001	0.001	0.433	0.273	0.227	0.047	0.005	0.003	0.573	0.370	0.306
-3	0.262	0.032	0.017	0.929	0.787	0.710	0.533	0.088	0.045	0.968	0.860	0.784
-4	0.913	0.336	0.167	1.000	0.986	0.966	0.983	0.581	0.312	1.000	0.995	0.979
-5	1.000	0.894	0.620	1.000	1.000	0.999	1.000	0.974	0.786	1.000	1.000	1.000
Panel E: $\Lambda_0 = 10$												
0	*0.003*	*0.000*	*0.000*	*0.016*	*0.007*	*0.002*	*0.011*	*0.001*	*0.000*	*0.036*	*0.015*	*0.006*
-1	0.004	0.000	0.000	0.080	0.043	0.022	0.014	0.001	0.000	0.149	0.073	0.039
-2	0.037	0.002	0.000	0.532	0.340	0.221	0.128	0.011	0.001	0.675	0.455	0.298
-3	0.487	0.064	0.006	0.953	0.843	0.703	0.768	0.181	0.021	0.980	0.907	0.779
-4	0.973	0.526	0.091	1.000	0.992	0.964	0.997	0.772	0.196	1.000	0.998	0.979
-5	1.000	0.963	0.462	1.000	1.000	0.999	1.000	0.993	0.662	1.000	1.000	1.000

Estimated rejection frequencies for the six tests for SPA under the null hypothesis ($\Lambda_1 = 0$) and local alternatives ($\Lambda_1 < 0$). Thus the rejection frequencies in italic font correspond to Type I errors and those in normal font are correspond to local powers. The reality check of White (2000) is denoted by RC_u and the test advocated by this paper is denoted by SPA_c.

Table 3: Rejection Frequencies under the Null and Alterantive ($m = 100$ and $n = 1000$)

	Level: $\alpha = 0.05$						Level: $\alpha = 0.10$					
Λ_1	RC_l	RC_c	RC_u	SPA_l	SPA_c	SPA_u	RC_l	RC_c	RC_u	SPA_l	SPA_c	SPA_u
Panel A: $\Lambda_0 = 0$												
0	*0.051*	*0.048*	*0.048*	*0.051*	*0.048*	*0.048*	*0.104*	*0.098*	*0.098*	*0.107*	*0.100*	*0.100*
-1	0.054	0.051	0.051	0.068	0.064	0.064	0.110	0.103	0.103	0.131	0.122	0.122
-2	0.125	0.116	0.116	0.309	0.282	0.282	0.223	0.202	0.202	0.435	0.391	0.390
-3	0.556	0.480	0.479	0.843	0.762	0.760	0.729	0.624	0.622	0.918	0.842	0.840
-4	0.970	0.889	0.886	0.998	0.980	0.977	0.995	0.945	0.941	1.000	0.992	0.990
-5	1.000	0.996	0.994	1.000	1.000	1.000	1.000	0.999	0.997	1.000	1.000	1.000
Panel B: $\Lambda_0 = 1$												
0	*0.011*	*0.009*	*0.009*	*0.020*	*0.017*	*0.017*	*0.031*	*0.024*	*0.023*	*0.050*	*0.040*	*0.039*
-1	0.011	0.009	0.009	0.043	0.036	0.035	0.033	0.025	0.025	0.086	0.069	0.069
-2	0.034	0.026	0.026	0.312	0.252	0.250	0.084	0.059	0.059	0.436	0.346	0.342
-3	0.316	0.205	0.203	0.859	0.740	0.732	0.520	0.338	0.331	0.927	0.822	0.814
-4	0.900	0.682	0.666	0.999	0.978	0.972	0.973	0.816	0.797	1.000	0.990	0.985
-5	1.000	0.968	0.955	1.000	1.000	0.999	1.000	0.991	0.982	1.000	1.000	1.000
Panel B: $\Lambda_0 = 2$												
0	*0.003*	*0.001*	*0.001*	*0.014*	*0.009*	*0.009*	*0.012*	*0.004*	*0.004*	*0.034*	*0.022*	*0.021*
-1	0.003	0.002	0.002	0.042	0.029	0.028	0.013	0.004	0.004	0.079	0.055	0.054
-2	0.014	0.006	0.006	0.338	0.242	0.236	0.042	0.018	0.017	0.465	0.330	0.322
-3	0.202	0.082	0.077	0.881	0.737	0.720	0.411	0.169	0.159	0.941	0.820	0.798
-4	0.844	0.461	0.428	0.999	0.979	0.969	0.959	0.652	0.602	1.000	0.991	0.983
-5	1.000	0.906	0.861	1.000	1.000	0.999	1.000	0.969	0.936	1.000	1.000	1.000
Panel B: $\Lambda_0 = 5$												
0	*0.002*	*0.000*	*0.000*	*0.012*	*0.005*	*0.004*	*0.006*	*0.000*	*0.000*	*0.029*	*0.011*	*0.008*
-1	0.002	0.000	0.000	0.057	0.028	0.024	0.007	0.001	0.000	0.103	0.051	0.042
-2	0.014	0.001	0.000	0.435	0.267	0.225	0.047	0.004	0.002	0.572	0.364	0.306
-3	0.270	0.029	0.017	0.930	0.777	0.708	0.540	0.084	0.044	0.968	0.851	0.784
-4	0.917	0.328	0.175	0.999	0.987	0.966	0.987	0.554	0.320	1.000	0.995	0.981
-5	1.000	0.877	0.632	1.000	1.000	0.999	1.000	0.966	0.791	1.000	1.000	1.000
Panel B: $\Lambda_0 = 10$												
0	*0.003*	*0.000*	*0.000*	*0.013*	*0.005*	*0.003*	*0.010*	*0.001*	*0.000*	*0.033*	*0.012*	*0.005*
-1	0.003	0.000	0.000	0.083	0.042	0.022	0.013	0.001	0.000	0.145	0.070	0.039
-2	0.039	0.002	0.000	0.534	0.335	0.220	0.128	0.010	0.000	0.672	0.444	0.299
-3	0.498	0.060	0.006	0.954	0.835	0.703	0.762	0.165	0.020	0.980	0.900	0.778
-4	0.974	0.496	0.095	0.999	0.994	0.965	0.997	0.737	0.203	1.000	0.998	0.980
-5	1.000	0.953	0.480	1.000	1.000	0.999	1.000	0.993	0.669	1.000	1.000	1.000

Estimated rejection frequencies for the six tests for SPA under the null hypothesis ($\Lambda_1 = 0$) and local alternatives ($\Lambda_1 < 0$). Thus the rejection frequencies in italic font correspond to Type I errors and those in normal font are correspond to local powers. The reality check of White (2000) is denoted by RC_u and the test advocated by this paper is denoted by SPA_c.

Table 4: Rejection Frequencies under the Null and Alterantive ($m = 1000$ and $n = 200$)

	Level: $\alpha = 0.05$						Level: $\alpha = 0.10$					
Λ_1	RC_l	RC_c	RC_u	SPA_l	SPA_c	SPA_u	RC_l	RC_c	RC_u	SPA_l	SPA_c	SPA_u
Panel A: $\Lambda_0 = 0$												
0	*0.049*	*0.047*	*0.047*	*0.064*	*0.062*	*0.062*	*0.106*	*0.100*	*0.100*	*0.125*	*0.119*	*0.119*
-1	0.049	0.047	0.047	0.066	0.064	0.064	0.106	0.101	0.100	0.128	0.122	0.122
-2	0.061	0.058	0.058	0.173	0.164	0.164	0.128	0.121	0.121	0.269	0.252	0.252
-3	0.288	0.262	0.262	0.658	0.598	0.596	0.434	0.388	0.388	0.770	0.699	0.697
-4	0.815	0.720	0.719	0.980	0.937	0.933	0.917	0.828	0.824	0.994	0.967	0.963
-5	0.998	0.971	0.967	1.000	0.999	0.998	1.000	0.991	0.988	1.000	1.000	1.000
Panel B: $\Lambda_0 = 1$												
0	*0.009*	*0.007*	*0.007*	*0.025*	*0.022*	*0.022*	*0.022*	*0.017*	*0.017*	*0.054*	*0.045*	*0.045*
-1	0.009	0.007	0.007	0.029	0.025	0.025	0.022	0.017	0.017	0.059	0.050	0.050
-2	0.010	0.008	0.008	0.150	0.127	0.127	0.026	0.020	0.020	0.229	0.192	0.191
-3	0.066	0.049	0.049	0.652	0.555	0.548	0.150	0.103	0.102	0.759	0.652	0.643
-4	0.502	0.345	0.339	0.980	0.924	0.916	0.701	0.500	0.488	0.993	0.956	0.947
-5	0.965	0.813	0.794	1.000	0.998	0.997	0.994	0.907	0.886	1.000	1.000	0.999
Panel C: $\Lambda_0 = 2$												
0	*0.001*	*0.000*	*0.000*	*0.015*	*0.011*	*0.011*	*0.005*	*0.002*	*0.002*	*0.035*	*0.026*	*0.025*
-1	0.001	0.000	0.000	0.020	0.015	0.015	0.005	0.002	0.002	0.043	0.032	0.032
-2	0.002	0.000	0.000	0.155	0.115	0.113	0.006	0.003	0.003	0.233	0.172	0.167
-3	0.016	0.007	0.007	0.669	0.544	0.525	0.054	0.022	0.022	0.779	0.636	0.616
-4	0.291	0.125	0.117	0.985	0.923	0.906	0.516	0.243	0.224	0.994	0.954	0.940
-5	0.901	0.576	0.529	1.000	0.999	0.996	0.980	0.744	0.683	1.000	1.000	0.998
Panel D: $\Lambda_0 = 5$												
0	*0.000*	*0.000*	*0.000*	*0.011*	*0.005*	*0.004*	*0.002*	*0.000*	*0.000*	*0.029*	*0.012*	*0.009*
-1	0.000	0.000	0.000	0.019	0.010	0.008	0.002	0.000	0.000	0.044	0.020	0.016
-2	0.000	0.000	0.000	0.199	0.122	0.101	0.002	0.000	0.000	0.291	0.180	0.148
-3	0.011	0.000	0.000	0.748	0.570	0.505	0.045	0.004	0.002	0.843	0.664	0.589
-4	0.303	0.036	0.017	0.993	0.939	0.897	0.575	0.098	0.050	0.998	0.967	0.930
-5	0.936	0.387	0.207	1.000	0.999	0.996	0.992	0.605	0.356	1.000	1.000	0.998
Panel E: $\Lambda_0 = 10$												
0	*0.001*	*0.000*	*0.000*	*0.012*	*0.004*	*0.003*	*0.002*	*0.000*	*0.000*	*0.029*	*0.011*	*0.004*
-1	0.001	0.000	0.000	0.025	0.012	0.007	0.002	0.000	0.000	0.054	0.024	0.011
-2	0.001	0.000	0.000	0.259	0.156	0.097	0.004	0.000	0.000	0.366	0.226	0.141
-3	0.031	0.001	0.000	0.815	0.633	0.495	0.109	0.006	0.000	0.891	0.726	0.579
-4	0.508	0.064	0.005	0.996	0.958	0.892	0.765	0.175	0.018	0.999	0.981	0.926
-5	0.983	0.531	0.099	1.000	1.000	0.995	0.998	0.753	0.210	1.000	1.000	0.998

Estimated rejection frequencies for the six tests for SPA under the null hypothesis ($\Lambda_1 = 0$) and local alternatives ($\Lambda_1 < 0$). Thus the rejection frequencies in italic font correspond to Type I errors and those in normal font are correspond to local powers. The reality check of White (2000) is denoted by RC_u and the test advocated by this paper is denoted by SPA_c.

Table 5: Definitions of Variables

Panel A: Description of Variables

Y_t	Annual inflation
$X_{1,t}$, $X_{2,t}$	Annual inflation (lags of Y_t)
$X_{3,t}$, $X_{4,t}$	Quarterly inflation
$X_{5,t}$	Quarterly inflation relative to previous year's inflation
$X_{6,t}$, $X_{7,t}$	Changes in employment in manufacturing sector
$X_{8,t}$	Quarterly employment relative to average of previous year
$X_{9,t}$	Quarterly employment relative to average of previous two years
$X_{10,t}$, $X_{11,t}$	Quarterly changes in real inventory
$X_{12,t}$, $X_{13,t}$	Quarterly changes in quarterly GDP
$X_{14,t}$	Interest paid on 3-month T-bill
$X_{15,t}$, $X_{16,t}$	Changes in 3-month T-bill
$X_{17,t}$, $X_{18,t}$	Changes in 3-month T-bill relative to level of T-bill
$X_{19,t}$, $X_{20,t}$	Changes in prices of fuel and energy
$X_{21,t}$, $X_{22,t}$	Changes in prices of food
$X_{23,t} - X_{26,t}$	Quarterly dummies: first, second, third, and fourth quarter
$X_{27,t}$	Constant

Panel B: Definitions of Variables

$$
\begin{aligned}
Y_t &= \log(\text{GDPCTPI}_t) - \log(\text{GDPCTPI}_{t-4}),\ X_{1,t} = Y_{t-5},\ X_{2,t} = Y_{t-8} \\
X_{3,t} &= 4[\log(\text{GDPCTPI}_t) - \log(\text{GDPCTPI}_{t-1}],\ X_{4,t} = X_{3,t-1} \\
X_{5,t} &= \log(1 + X_{3,t}) - \log(1 + X_{1,t-1}) \\
X_{6,t} &= \log(\text{MANEMP}_t) - \log(\text{MANEMP}_{t-1}),\ X_{7,t} = X_{6,t-1} \\
X_{8,t} &= \log(\text{MANEMP}_t) - \log(\tfrac{1}{4}\textstyle\sum_{i=1}^{4} \text{MANEMP}_{t-i}) \\
X_{9,t} &= \log(\text{MANEMP}_t) - \log(\tfrac{1}{8}\textstyle\sum_{i=1}^{8} \text{MANEMP}_{t-i}) \\
X_{10,t} &= \log(\text{CBI}_t) - \log(\text{GDP}_t),\ X_{11,t} = X_{10,t-1} \\
X_{12,t} &= \log(\text{GDP}_t) - \log(\text{GDP}_{t-1}),\ X_{13,t} = X_{12,t-1} \\
X_{14,t} &= \text{TB3MS}_t,\ X_{15,t} = \Delta X_{14,t},\ X_{16,t} = X_{15,t-1},\ X_{17,t} = \Delta X_{14,t}/X_{14,t},\ X_{18,t} = X_{17,t-1} \\
X_{19,t} &= \log(\text{PPIENG}_t) - \log(\text{PPIENG}_{t-1}),\ X_{20,t} = X_{19,t-1} \\
X_{21,t} &= \log(\text{PPIFCF}_t) - \log(\text{PPIFCF}_{t-1}),\ X_{22,t} = X_{21,t-1} \\
X_{23,t} &= 1,\ X_{24,t} = X_{23,t-1},\ X_{25,t} = X_{23,t-2},\ X_{26,t} = X_{23,t-3},\ X_{27,t} = 1
\end{aligned}
$$

Raw Data: GDPCTPI = Gross Domestic Product: Chain-type Price Index; CBI = Change in Private Inventories; GDP = Gross Domestic Product; TB3MS = 3-Month Treasury Bill Rate, Secondary Market*; PPIENG = Producer Price Index: Fuels & Related Products & Power**; PPIFCF = Producer Price Index: Finished Consumer Foods**; MANEMP = Employees on Nonfarm Payrolls:Manufacturing.

* Quarterly data are defined to be the average of the monthly observations over the quarter.

** Quarterly data are defined to be the last monthly observation of the quarter.

Table 6: Tests for Superior Predictive Ability

Panel A: Results for the Large Universe of Forecasts

				Loss	t-statistic	"p-value"
Evaluated by MAE			Benchmark:	0.0098	–	–
$m = 3304$ (number of models)			Best Performing:	0.0084	1.2363	0.120
$n = 160$ (sample size)			Most Significant:	0.0085	1.2628	0.112
$B = 10,000$ (resamples)			Median:	0.0141	−2.7694	–
$q = 0.25$ (dependence)			Worst:	0.0416	−7.8939	–
	RC_l	RC_c	RC_u	SPA_l	SPA_c	SPA_u
SPA p-values:	0.503	0.781	0.978	0.571	**0.741**	0.903

Panel B: Results for the Small Universe of Forecasts

				Loss	t-statistic	"p-value"
Evaluated by MAE			Benchmark:	0.0098	–	–
$m = 352$ (number of models)			Best Performing:	0.0082	2.7547	0.006
$n = 160$ (sample size)			Most Significant:	0.0096	2.9399	0.004
$B = 10,000$ (resamples)			Median:	0.0097	0.0657	–
$q = 0.25$ (dependence)			Worst:	0.0107	−1.3272	–
	RC_l	RC_c	RC_u	SPA_l	SPA_c	SPA_u
SPA p-values:	0.071	0.106	0.106	0.045	**0.048**	0.048

Panel C: Results for the Full Universe of Forecasts

				Loss	t-statistic	"p-value"
Evaluated by MAE			Benchmark:	0.0098	–	–
$m = 3656$ (number of models)			Best Performing:	0.0082	2.7547	0.006
$n = 160$ (sample size)			Most Significant:	0.0096	2.9399	0.004
$B = 10,000$ (resamples)			Median:	0.0135	−1.9398	–
$q = 0.25$ (dependence)			Worst:	0.0416	−7.8939	–
	RC_l	RC_c	RC_u	SPA_l	SPA_c	SPA_u
SPA p-values:	0.395	0.691	0.963	0.078	**0.100**	0.135

The table reports SPA p-values for three sets of regression-based forecasts that are compared to a random walk forecast. The p-value of the new test, SPA_c, is in bold font.

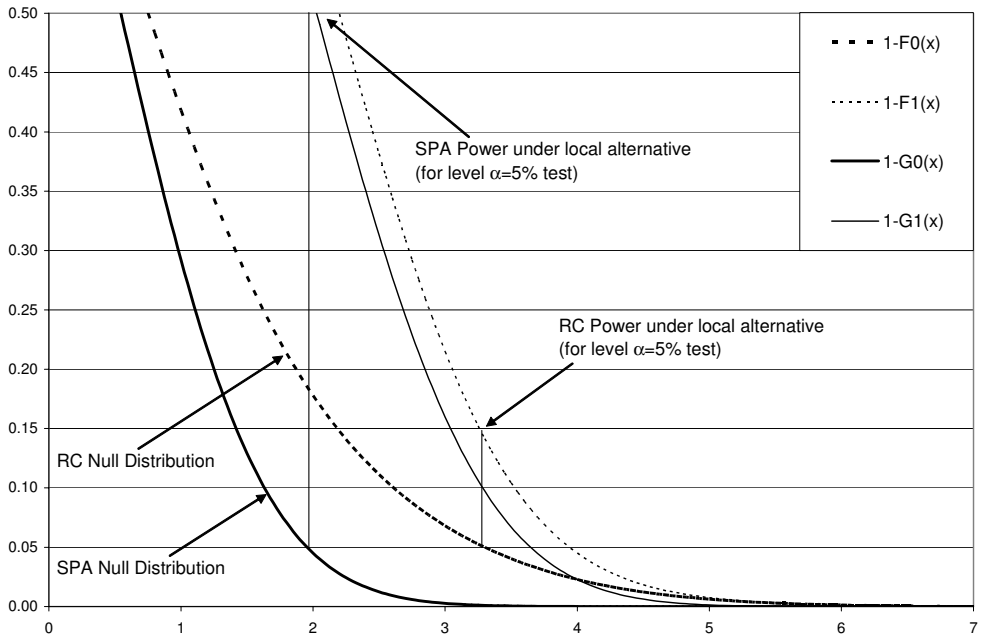

Figure 1: This Figure shows (one minus) the cdfs for the test statistics T^{RC} and T^{SPA} under both the null hypothesis, $\mu_1 = \mu_2 = 0$, and the local alternative where $\mu_2 = 2/\sqrt{n} > 0$. The studentization improves the power from about 15% to about 53%.

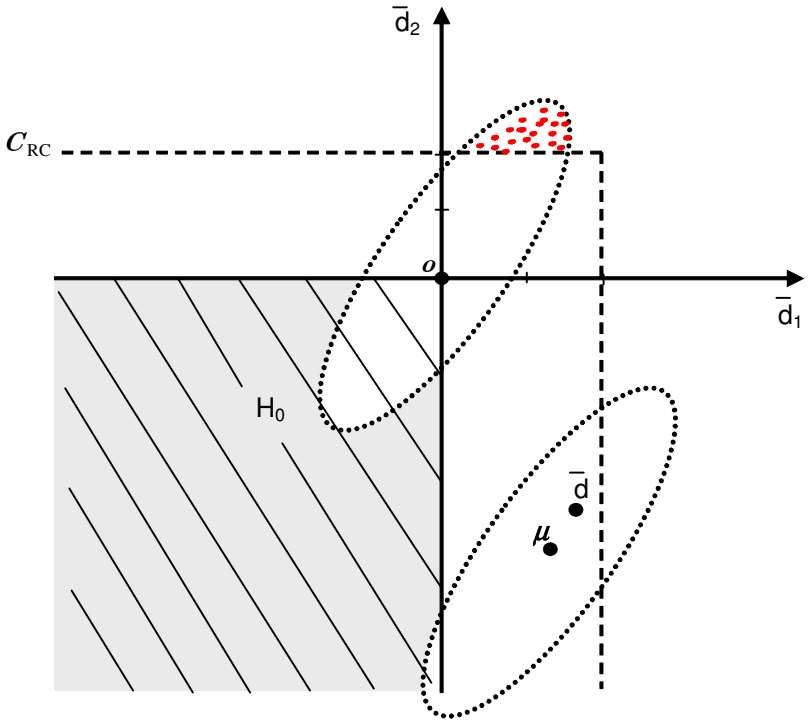

Figure 2: A situation where the RC fails to reject a false null hypothesis. The true parameter value is $\boldsymbol{\mu} = (\mu_1, \mu_2)'$, the sample estimate is $\bar{d} = (\bar{d}_1, \bar{d}_2)'$, and C_{RC} illustrates the critical value derived from a distribution that tacitly assumes $\boldsymbol{\mu} = (0, 0)'$.

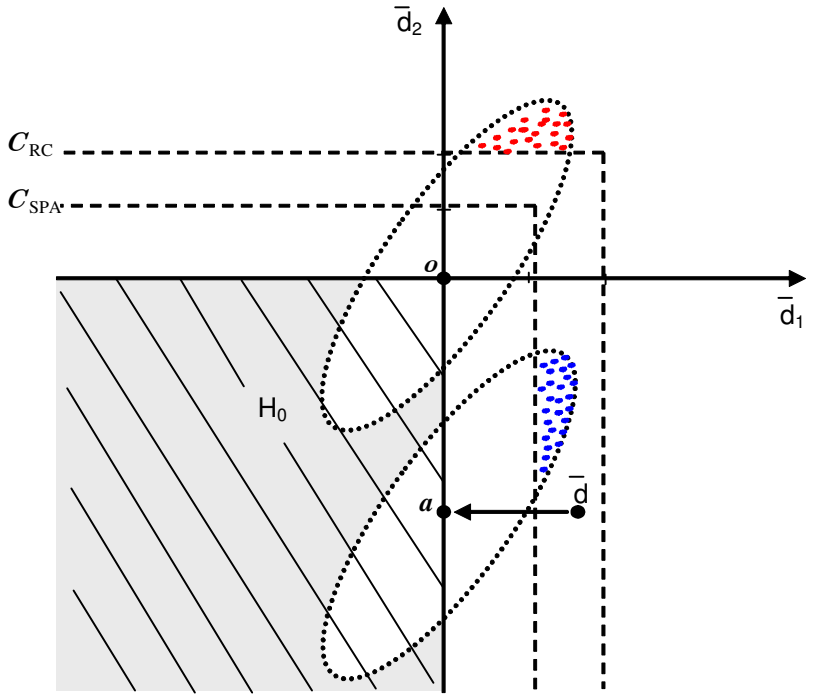

Figure 3: The figure shows how the power is improved by using the sample dependent null distribution. This distribution is centered about $\hat{\mu}^c = a$, which leads to the critical value C_{SPA}. In contrast, the RC fails to reject the LFC null distribution leads to the larger critical value C_{RC}.

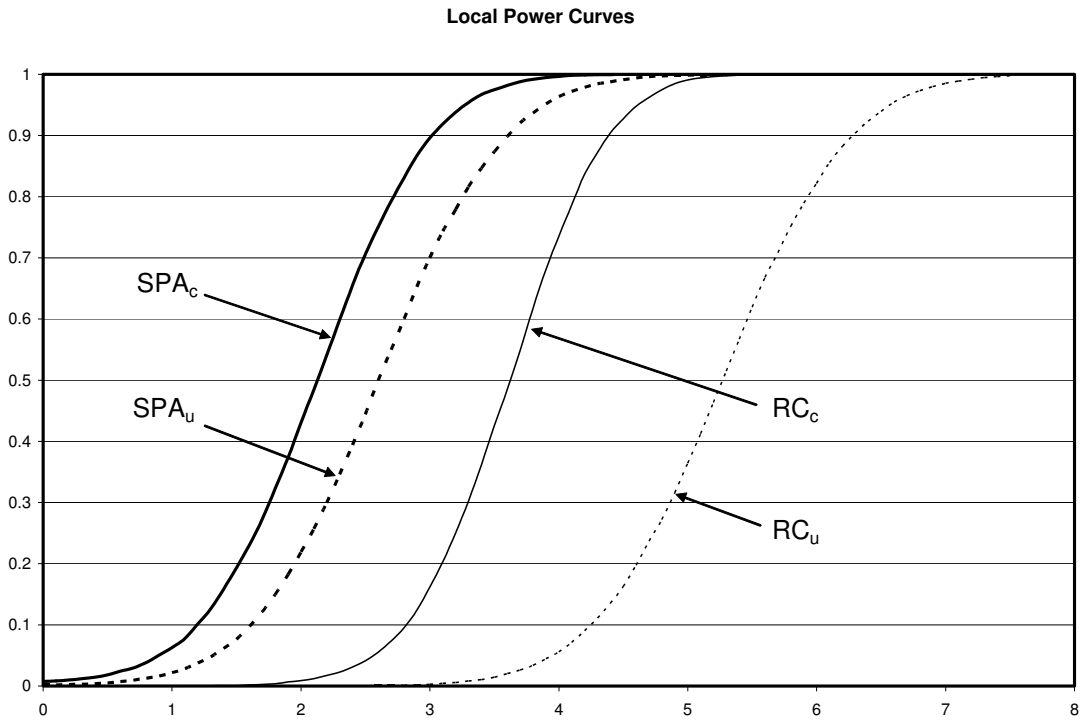

Figure 4: Local power curves of the four tests, SPA$_c$, SPA$_u$, RC$_c$, and RC$_u$, for the simulation experiment where $m = 100$, $\Lambda_0 = 20$, and $\mu_1/\sqrt{n}$ $(= -\Lambda_1)$ ranges from 0 to 8 (the x-axis). The power curves quantify the power improvements from the two modifications of the Reality Check. Both the studentization and the data dependent null distribution lead to substantial gains in power for this configuration.