Asymptotic Tests of Composite Hypotheses

Peter Reinhard Hansen[1]

Working Paper No. 2003-09

April, 2003

Brown University

Department of Economics

[1]Brown University, Department of Economics, Box B, Brown University, Providence, RI 02912, USA, Phone: (401) 863 9864, Email: Peter_Hansen@brown.edu

Abstract[2]

Test statistics that are suitable for testing composite hypotheses are typically non-pivotal, and conservative bounds are commonly used to test composite hypotheses.

In this paper, we propose a testing procedure for composite hypotheses that incorporates additional sample information. This avoids, as $n \to \infty$, the use of conservative bounds and leads to tests with better power than standard tests. The testing procedure satisfies a novel similarity condition that is relevant for asymptotic tests of composite hypotheses, and we show that this is a necessary condition for a test to be unbiased.

The procedure is particularly useful for simultaneous testing of multiple inequalities, in particular when the number of inequalities is large. This is the situation for the multiple comparisons of forecasting models, and we show that the new testing procedure dominates the 'reality check' of White (2000) and avoids certain pitfalls.

JEL Classification: C12, C30, C52, C53.

Keywords: Composite hypotheses, similarity, unbiased tests, multiple comparisons.

[2]Part of the paper was written while the author visited University of California, Berkeley and much is owed to discussions I had with Michael Jansson. I also thank Sean Campbell, Frank Wolak, and seminar participants at Stanford University and UC Berkeley for valuable comments. All errors are mine. A previous version of this paper carried the title: "A Reality Check for Data Snooping: A Comment on White". This research was supported by the Danish Research Agency Grant No. 54-00-0363.

1 Introduction

Composite Hypothesis are common in econometrics, where budget constraints, presumed convexities, arbitrage conditions, stochastic dominance, etc., may lead to one or more inequalities that characterize a composite null hypothesis. A composite hypothesis does not point to a unique probability measure to be used in the hypothesis testing, and this makes it more challenging to test composite hypotheses than simple hypotheses. The ambiguity in the null distribution is typically solved by applying a conservative bound in hypotheses tests, see, e.g., Perlman (1969) and Robertson, Wright, and Dykstra (1988). This approach is known as the least favorable configuration (LFC).

In this paper, we consider asymptotic tests of composite hypotheses, and the paper makes three contributions. First, we note that the testing problem of composite hypotheses is closely related to the problem of testing hypotheses in the presence of nuisance parameters. As there is additional sample information about the nuisance parameters,[1] we can exploit this information to derive an asymptotically exact test that has better power than the LFC-test. The idea that underlies our results is an asymptotic version of that applied by Dufour (1990), Berger and Boos (1994), and Silvapulle (1996) to various problems, and our asymptotic results yield insight about how the idea should be implemented in finite samples.

Second, we formulate a similarity condition that is relevant for asymptotic tests of composite hypotheses, and we show that the condition is necessary for an asymptotic test to be unbiased in regular problems. We pay special attention to the case where the null hypothesis is characterized by linear inequalities, which is the most common composite testing problem in econometrics. In this context, we show that the LFC is increasingly inferior as the dimension of the testing problem increases. Our result leads to the conclusion that a LFC-based test is inadmissible for testing multiple inequalities.

Third, our results have important implications for the reality check for data snooping (RC) by White (2000). In this framework, the question of interest is whether a benchmark forecast is outperformed by alternative forecasting models, which leads to a composite null hypothesis. We show the advantages of the new testing procedure and the practical relevance of our theoretical results are confirmed by simulation experiments and an empirical application. We also characterize some rather unfortunate properties of the RC that can be avoided by the new procedure. However, a partial pivoting of the RC's test statistic can alleviate some of the RC's problems. Based on these findings it is not advisable to use the RC in its original form.

Composite hypotheses often arise from inequality constraints, and much of the under-

[1] In our framework, it is the parameter of interest that appears as a 'nuisance' parameter, so strictly speaking this is not a nuisance parameter problem. Our problem is not directly related to the problem where the nuisance parameter is only identified under the alternative.

lying theory for hypotheses testing, where either the null or the alternative hypothesis is characterized by linear inequalities, is due to Perlman (1969), see Robertson, Wright, and Dykstra (1988) for a general treatment. Within the framework of the linear regression model, Gouriéroux, Holly, and Monfort (1982) and Wolak (1987, 1989b) derived tests of hypotheses that are given from linear inequalities, and Dufour (1989) derived exact simultaneous tests in this setting. Composite hypotheses testing in non-linear models has been analyzed by Wolak (1989a, 1991).

The main complication in testing composite hypotheses is the lack of pivotal quantities that are suitable for testing. The common solution is to use a quantity, which has a distribution that can be properly bounded over the null hypothesis. This bound is known as the least favorable configuration, because it employs the distribution 'in the null' that is least favorable to the alternative hypothesis. The motivation for using the LFC is that it leads to exact tests, however, the LFC method has drawbacks because it often leads to non-similar tests that are biased and have poor power properties against certain alternatives. It may have been believed that the LFC approach is the only way to construct tests that are exact asymptotically, see e.g., Wolak (1989a, p. 10). Our results show that this is not the case and that the new testing procedure dominates the LFC approach.

In general, the asymptotic similarity condition provides guidance on how to construct unbiased and powerful tests, if such exist. In the context of testing linear inequalities the similarity condition has the same implication for tests as the LFC when testing a single inequality, but differ in dimensions two and higher. Thus the intuition from tests of a single inequality, e.g., $\beta_1 \geq 0$, does not carry over to the situation with two or more inequalities, e.g., $\beta_1 \geq 0$, $\beta_2 \geq 0$. This point was also made by Goldberger (1992).

Simultaneous inference and multiple comparison problems sometimes lead to testing problems of multiple linear inequalities, see, e.g., Gupta and Panchapakesan (1979), Miller (1981), Savin (1984), and Hsu (1996). One such case is when multiple forecasting models are being compared to a benchmark model, which is particularly interesting for certain econometric problems. In this setting White (2000) recently proposed a test, the reality check, which made two valuable contributions to this problem. First, White suggested a bootstrap implementation of the RC. This approach is very useful because it circumvents an explicit estimation of a large covariance matrix, which is infeasible whenever the number of competing forecasts exceeds the sample size. A second contribution is the formulation of the null hypothesis. Rather than testing for equal predictive ability (EPA), as analyzed by Diebold and Mariano (1995) and West (1996), the RC is constructed to test for superior predictive ability (SPA). Indeed, SPA is often more relevant for economic applications than EPA, because the existence of a better forecasting model is typically of more importance than the existence of a worse model. For example, testing for SPA is relevant for forecasters who

want to evaluate whether the forecasting model they currently use is inferior to alternative models. Also, if an economic theory predicts that a particular forecasting model embodies all information about the future, then testing for SPA can be used to falsify the theory.

Testing for superior predictive ability leads to a composite hypothesis and our theoretical results have important implications for the properties of the RC. We find that the RC is sensitive to the inclusion of poor and irrelevant models in the space of competing forecasting models, and the power of the RC is unnecessarily low in most situations. These problems are caused by two aspects, one is that the RC is a LFC-test and the other is that the individual model-statistics, which enter the test statistic, are non-standardized.

We use the following notation: For $x \in \mathbb{R}^p$ we define $x^+ \equiv (\max\{x_1, 0\}, \ldots, \max\{x_p, 0\})'$ and we let $\|x\|_2 = \sqrt{\sum_{i=1}^p x_i^2}$ denote the Euclidian norm of x. The open ball around x with radius $\epsilon > 0$ we denote by $\mathcal{N}_\epsilon(x) = \{y \in \mathbb{R}^p : \|y - x\|_2 < \epsilon\}$. For a constant, $a \in \mathbb{R}$ we let $[a]$ denote its integer part and let $\lim_{u \nearrow a}$ denote the left limit. For a subset, A, of some space, $\mathcal{S} \subset \mathbb{R}^p$, the complement of A is denoted by $\complement A = \{a \in \mathcal{S} : a \notin A\}$. Convergence in probability, distribution, and weak convergence we denote by $\xrightarrow{p}$, $\xrightarrow{d}$, and $\xrightarrow{w}$, respectively.

The remainder of the paper is organized as follows. Section 2 contains the theoretical framework with emphasis on simultaneous testing of multiple inequalities and includes a necessary condition for a test to be unbiased. A simulation study quantifies our theoretical results and reveals substantial gains in power from using the new testing procedure. Section 3 shows the implications that our theoretical results have for the RC of White (2000). The improvements that can be achieved by the new testing procedure are emphasized in an empirical application. Section 4 contains concluding remarks.

2 The Theoretical Framework

We consider a statistical model, $(\Omega, \mathcal{F}, \mathcal{P})$, where Ω is the sample space, $\mathcal{F}$ is an σ-algebra on Ω, and $\mathcal{P} = (P_\theta)_{\theta \in \Theta}$ is a parametric family of probability measures on $(\Omega, \mathcal{F})$. The parameter space, Θ, defines the maintained hypothesis, which is a non-empty subset of $\mathbb{R}^p$, for some integer p.

We consider the hypothesis, $H_0 : \theta \in \Theta_0$, where Θ_0 is a subset of Θ, and we shall be concerned with the case where Θ_0 contains more than a single point, so that H_0 is a composite hypothesis.

2.1 A Simple Illustrative Example

As stated in the introduction, the problem of testing a composite hypothesis is related to that of testing in the presence of nuisance parameters. We illustrate this with a simple example that also serves as an illustration of the idea behind our testing procedure.

Let $X_1, \ldots, X_n$ be independent and identically distributed $N(\mu, \sigma^2)$, where the parameter space for the unknown parameters, $\theta = (\mu, \sigma^2)$, is $\Theta = \mathbb{R} \times [0, \infty)$. Consider the hypothesis

$H_0 : \mu = 0$, in which case σ^2 is a nuisance, and note that H_0 is a composite hypothesis, because it corresponds to $H_0 : \theta \in \Theta_0$, where $\Theta_0 = \{0\} \times [0, \infty)$. We seek to test H_0 at some level, $\alpha \in (0, 1)$, and an obvious test is Gosset's well-known t-test, which has the quality of being similar as the t-statistic is a pivot.

Suppose, for the sake of illustration, that no pivot is available and we instead apply the test statistic, $T_n \equiv n^{1/2}|\bar{X}_n|$, where $\bar{X}_n \equiv n^{-1}\sum_{i=1}^n X_i$ is the sample average. Since $n^{1/2}\bar{X}_n \sim N(0, \sigma^2)$ we see that the distribution of T_n depends on σ^2, so T_n is not a pivot. We now discuss four approaches to handling the non-pivotalness of T_n.[2]

1. The first approach is the LFC, which entails finding a bound for the distribution of T_n. The LFC-test rejects H_0 if $T_n > \sup_{\theta \in \Theta_0} \Phi^{-1}_{0,\sigma^2}(1-\alpha/2)$, where Φ^{-1}_{0,σ^2} is the inverse cdf of the normal variable with mean zero and variance σ^2. Since $\Phi^{-1}_{0,\sigma^2}(1-\alpha/2) \to \infty$ as $\sigma^2 \to \infty$, this problem does not have a solution. However, if the parameter space is given by $\Theta = \mathbb{R} \times [0, \eta^2]$ for some constant $\eta > 0$, we have $\sup_{\theta \in \Theta_0} \Phi^{-1}_{0,\sigma^2}(1-\alpha/2) = \Phi^{-1}_{0,\eta^2}(1-\alpha/2)$, and the LFC-test would reject H_0 if $T_n/\eta > 1.96$, using the level $\alpha = 0.05$.

2. A second approach substitutes a consistent estimator and invokes the asymptotic distribution of T_n. Thus H_0 is rejected if $T_n > \Phi^{-1}_{0,\hat{\sigma}^2}(1-\alpha/2)$, where $\hat{\sigma}^2$ is a consistent estimator for σ^2.

3. A third approach is based on a $(1-\delta)$ confidence interval for the nuisance parameter σ^2, $\mathcal{I}$ say, where for $\delta \in (0, \alpha)$. We can define the test that rejects H_0 if $T_n > \sup_{\sigma^2 \in \mathcal{I}} \Phi^{-1}_{0,\sigma^2}(1-(\alpha-\delta)/2)$, and it is easy to verify that this test has level α.

4. The fourth approach, which illustrates the main result of this paper, is an asymptotic version of the third approach. Rather than holding δ fixed, we let $\delta_n \to 0$ as $n \to \infty$ at an appropriate rate, and use $\sup_{\sigma^2 \in \mathcal{I}_n} \Phi^{-1}_{0,\sigma^2}(1-(\alpha-\delta_n)/2)$ as the critical value for T_n, where $\mathcal{I}_n$ is an $(1-\delta_n)$ confidence interval for σ^2.

Although the test of the first approach has the correct size, (equal to the level α), this approach has obvious drawbacks. The actual rejection probability (Type I error) can be arbitrarily small, and for σ^2 close to zero, this test has very low power compared to the t-test.

The second approach is widely used in econometrics. However, this approach can produce misleading results in a number of cases. Obviously, the approach is not suited for a

[2] None of the four approaches should be viewed as a viable competitor to the t-test in this simple framework. The simple setting is used for illustration only.

problem where the nuisance parameters cannot be consistently estimated, as is the case in the incidental parameter problem. Similarly, the asymptotic approximation can be poor if the nuisance parameter are poorly estimated, which is the case for IV regressions with weak instruments. Better alternatives to these problems include the parameter orthogonalization of Lancaster (2000, 2002) and the tests of Kleibergen (2002) and Moreira (2003). These solutions share the property that they achieve an exact or asymptotic non-dependence of the nuisance parameter, which leads to similar tests (like the t-test in our example). The second approach is also problematic if the asymptotic distribution poorly approximates the finite sample distribution, however, this problem sometimes be avoided through bootstrap methods, see Horowitz and Savin (2000). Hypotheses testing using a heteroskedasticity and autocorrelation consistent (HAC) covariance matrix, in a time-series context, is a case where the asymptotic distribution can be a poor approximation to the finite sample distribution, and alternative tests that avoid the use of HAC covariance matrices can have better finite sample properties, see Kiefer, Vogelsang, and Bunzel (2000) and Jansson (2002).

The second approach, can also be misleading when θ is located on the (relative) boundary of Θ_0, and this is the reason that this approach is not reliable for composite hypothesis testing. For a discussion on how this can affect the reliability of bootstrap methods, see Andrews (2000).

The third approach nests the two previous approaches, as they correspond to $\delta = 0\%$ and $\delta = 100\%$, respectively, although the second approach ignores the δ-term when deriving the critical value of the test statistic. The idea behind the third approach is not new. Dufour (1990) used the idea within the linear regression model with autocorrelated errors, and applied a confidence interval for the autoregressive parameter to make inference about the regression parameters. The idea can also be found in Berger and Boos (1994) who constructed valid p-values using confidence sets for nuisance parameters, and in Cavanagh, Elliott, and Stock (1995) who used a confidence interval for a local-to-unity parameter. See also Dufour and Kiviet (1996, 1998) and Silvapulle (1996). In the context of model discrimination, Loh (1985) applied the idea as an alternative to the test of Cox (1961, 1962). Berger (1996) applied the confidence p-values of Berger and Boos (1994) to test that two binomial coefficients (from different populations) are equal and concluded that this leads to better power properties compared to several standard tests. The idea is closely related to the projection method, where a confidence set for (μ, σ^2) is projected onto the parameter space for μ, see, e.g., Dufour and Taamouti (2001), and sequential testing in instrumental variable regressions by Staiger and Stock (1997) and Stock and Yogo (2002).

It is the fourth approach that is successful in the context of composite hypothesis testing. Asymptotically this approach emulates the second approach, without compromising the size. Like the improved method for dealing with the incidental parameter problem, weak

instruments, and inference without the use of HAC estimators, this approach achieves a form of similarity. Compared to the first approach, this test achieves better power by directing its power towards the 'relevant' alternatives. In this example the relevant alternatives are given by the pairs (μ, σ^2) for which σ^2 in a neighborhood of $\hat{\sigma}^2$, and $\mu \neq 0$. So, the power improvements are achieved in a similar way to that of Andrews (1998) who proposed directed tests to test a simple null against a restricted alternative.

Currently, there is no theory for choosing δ, except that δ must be smaller than the significance level, α, for it to be useful for testing. Dufour and Kiviet (1996, 1998) used $\delta = 0.025$ and $\delta = 0.05$, Berger and Boos (1994) and Berger (1996) use $\delta = 0.001$, and Silvapulle (1996) use $\delta = 0.005$. Our asymptotic results show that the best properties are achieved if δ_n goes to zero at a certain rate, as $n \to \infty$, and a bound for this rate shed light on how δ should be chosen for finite n.

2.2 The Basic Framework

We use the notation, $\mathcal{R}$, to refer to a test of H_0, where $\mathcal{R} \subset \Omega$ is the rejection region that defines the realization that leads to a rejection of H_0.[3] In order to evaluate the probability of the event, $\mathcal{R}$, we assume that $\mathcal{R} \in \mathcal{F}$, and we follow Horowitz (2001) and refer to $P_\theta(\mathcal{R})$ as the *rejection probability*. Naturally, the objective for constructing a test of H_0, is to determine a rejection region, $\mathcal{R}$, for which $P_\theta(\mathcal{R})$ is small for $\theta \in \Theta_0$ and large for $\theta \notin \Theta_0$.

As the reader may recall, a test, $\mathcal{R}$, is said to be *similar* if $P_\theta(\mathcal{R})$ is constant on Θ_0. Similar tests are easily constructed from pivots: e.g., the t-statistic in our example is a pivot and $\mathcal{R} = \{\omega : |t(\omega)| \geq c\}$ defines a similar test for any $c \geq 0$. However, pivots that are suitable for hypothesis testing need not exist, see, e.g., Bahadur and Savage (1956) and Dufour (1997), and when this is the case it is common practice to use the LFC in the hypothesis testing, which was the first approach in our example. This typically leads to a conservative test, in the sense that the rejection probability (Type I error), for some $\theta \in \Theta_0$, is strictly smaller than the level of the test.

Without loss of generality we consider tests that are defined by some test statistic, $T : \omega \mapsto [0, \infty)$, where large values of T favors the alternative hypothesis. So a test will typical have the form: $\mathcal{R} = \{\omega : T(\omega) > c\}$ for some $c \in \mathbb{R}$. In what follows, we let T and T_n, $n = 1, 2, \ldots$ denote test statistics that are measurable mappings from $(\Omega, \mathcal{F})$ into $(\mathbb{R}, \mathcal{B})$, where $\mathcal{B}$ is the Borel σ-algebra under the Euclidian topology. In our asymptotic analysis, $\omega \in \Omega$ can be thought of as a realization of an infinite sequence of random variables and T_n can be thought of as a function of the first n coordinates of ω.

In this paper, we shall be less concerned with the problem of choosing a good test statistic.

[3] To simplify notation, we deviate from the more common notation where a test is represented by a pair, $(\mathcal{A}, \mathcal{R})$, where $\mathcal{A} = \Omega \backslash \mathcal{R}$ is the 'acceptance' region that defines the realizations for which H_0 is not rejected.

Rather, we take T as given and show how standard testing procedures can be improved in the situation where T is non-pivotal. The main complication in composite hypothesis testing is the lack of pivotal test statistics that are suitable for testing. Nevertheless, it will often be desirable to employ a test statistic, T, that is 'close' to being pivotal. As we shall see in our discussion of the RC, a partial pivoting of the RC's test statistic leads to a test with better properties.

2.3 Asymptotic Tests of Composite Hypotheses

We now study asymptotic tests of composite hypotheses. An asymptotic test is characterized by a sequence of rejection region, $\mathcal{R}_n$, $n = 1, 2, \ldots$, and we let the sequence be denoted by, $\{\mathcal{R}_n\}$, and shall refer to $\{\mathcal{R}_n\}$ as an asymptotic test. The asymptotic size is defined by $\alpha \equiv \limsup_{n\to\infty} \sup_{\theta\in\Theta_0} P_\theta(\mathcal{R}_n)$.

The boundary of the null hypothesis is particularly interesting for asymptotic tests of composite hypotheses. See, e.g., Chernoff (1954) who derived the asymptotic properties of likelihood ratio tests when the parameter is on the boundary. The boundary is denoted by $\partial\Theta_0$ and is defined to be the intersection of the closure of Θ_0 and the closure of its compliment, $\complement\Theta_0 \equiv \{\theta \in \Theta : \theta \notin \Theta_0\}$, under the Euclidian topology.

Assumption 1 *(i) For all $\theta_0 \in \Theta$ it holds that $\hat{\theta}_n \xrightarrow{p} \theta_0$ and (ii) uniformly on Θ_0 it holds that $\hat{\theta}_n \xrightarrow{p} \theta_0$ and $T_n \xrightarrow{d} F_{\theta_0}$, where the cdf, F_{θ_0}, is continuous for $\theta_0 \in \partial\Theta_0$.*[4]

Note that F_{θ_0} is only required to be continuous for $\theta_0 \in \partial\Theta_0$ (T_n must be properly normalized in n). The motivation for this is that there will not be any asymptotic evidence against the null, when $\theta_0 \in \Theta_0 \backslash \partial\Theta_0$, since $\hat{\theta}_n \xrightarrow{p} \theta_0$. So in this case F_{θ_0} should be allowed to degenerate (e.g., $T_n \xrightarrow{p} 0$).

For all $\theta \in \Theta_0$ and any $\alpha \in (0, 1)$ we define the half-line $\mathcal{I}_\alpha^\theta = \{a \in \mathbb{R} : \lim_{u \nearrow a} F_\theta(u) \geq 1 - \alpha\}$, which can be interpreted as an asymptotic critical region for the test statistic. This half-line will typically be closed but can be open if F_θ has discontinuities, which is the reason that we choose to work with $\mathcal{I}_{\alpha,n}^\theta$ rather than a critical value, e.g., $\inf_{a\in\mathcal{I}_{\alpha,n}^\theta} a$. For any $D \subset \Theta_0$ we define the half-line $\mathcal{I}_\alpha^D \equiv \bigcap_{\theta\in D} \mathcal{I}_\alpha^\theta$, (with the convention $\mathcal{I}_\alpha^\varnothing \equiv \mathbb{R}$) and we define the 'rejection region',

$$\mathcal{R}_{\alpha,n}^D \equiv \{\omega \in \Omega : T_n(\omega) \in \mathcal{I}_\alpha^D\}.$$

It follows directly that $\mathcal{R}_{\alpha,n}^{D_1} \subset \mathcal{R}_{\alpha,n}^{D_2}$ for $D_2 \subset D_1 \subset \Theta_0$, since $\mathcal{I}_\alpha^{D_1} \subset \mathcal{I}_\alpha^{D_2}$, and we note that the LFC-test (with asymptotic level α) is given by $\{\mathcal{R}_{\alpha,n}^{\Theta_0}\}$.

[4] *The requirements are: For all $\epsilon > 0$ there exists an N_ϵ, such that $|P_\theta(T_n \leq a_\theta) - F_\theta(a_\theta)| \leq \epsilon$ for all $n \geq N_\epsilon$ and for all $\theta \in \Theta_0$, where a_θ is any continuity point of F_θ. Similarly for all $\epsilon, \delta > 0$ there exists an $N_{\epsilon,\delta}$, such that $P_\theta(|\hat{\theta}_n - \theta| > \epsilon) < \delta$ for all $n \geq N_{\epsilon,\delta}$ and for all $\theta \in \Theta_0$. The uniform convergence requires that N_ϵ and $N_{\epsilon,\delta}$ do not depend on θ.*

Lemma 1 *Given Assumption 1 it holds that* $\limsup_{n\to\infty}\sup_{\theta\in\Theta_0} P_\theta(\mathcal{R}^{\Theta_0}_{\alpha,n}) \leq \alpha$ *and* $\lim_{n\to\infty} P_{\theta_0}(\mathcal{R}^{\{\theta_0\}}_{\alpha,n}) = \alpha$ *for* $\theta_0 \in \partial\Theta_0$.

If F_θ is continuous and its inverse is well-defined, we have that $\mathcal{R}^{\Theta_0}_{\alpha,n} \equiv \{\omega \in \Omega : T_n(\omega) \geq c\}$, where $c \equiv \sup_{\theta\in\Theta_0} F^{-1}_\theta(1-\alpha)$ is the critical value. Note that $\mathcal{R}^{\{\theta_0\}}_{\alpha,n}$ is not a (feasible) test, because it depends on the unknown parameter, θ_0.

We now show that a simple modification of the LFC-test leads to a test with better asymptotic properties than the LFC-test.

Lemma 2 *For an arbitrary* $\epsilon > 0$ *we define* $C_\epsilon \equiv \mathcal{N}_\epsilon(\hat{\theta}_n) \cap \Theta_0$, *which is a neighborhood of* $\hat{\theta}_n$ *in* Θ_0. *Given Assumption 1 it holds that:* (i) $\mathcal{R}^{\Theta_0}_{\alpha,n} \subset \mathcal{R}^{C_\epsilon}_{\alpha,n}$ *for all* n *and all* $\theta \in \Theta$; (ii) $\limsup_{n\to\infty}\sup_{\theta\in\Theta_0} P_\theta(\mathcal{R}^{C_\epsilon}_{\alpha,n}) \leq \alpha$, *for all* $\theta \in \Theta_0$.

The lemma shows that a very simple modification of the LFC approach can yield a test that has better (or at least as good) power properties than the LFC-test (i), without compromising the asymptotic size (ii). However, the asymptotic result is not informative about how large C_ϵ should be (define by ϵ). Clearly, the smaller is the volume of C_ϵ the larger is $\mathcal{R}^{C_\epsilon}_{\alpha,n}$ and the more powerful is the test. However, if C_ϵ is chosen too small it may (for a finite n) only contain θ_0 with a small probability, and the size of the test can exceed α to an extent that is unacceptable. Although this problem vanishes as n increases, we need some guidance on how to choose C_ϵ. If $P_{\theta_0}(\complement\mathcal{C}_n)$ is easy to evaluate (or bound from above), then for some $\delta_n \geq \sup_{\theta\in\Theta_0} P_\theta(\complement\mathcal{C}_n)$, one can use the rejection region, $\mathcal{R}^{C_\epsilon}_{\alpha-\delta_n,n}$, as in the finite sample tests, (the third approach in our example).[5]

Assumption 2 *Let* $\{C_n\}$ *be a sequence of subsets of* Θ_0 *that satisfies* (i) $P_\theta(\theta \in C_n) \to 1$ *as* $n \to \infty$ *uniformly in* θ *on* Θ_0. (ii) *For* $\theta_0 \in \Theta_0$ *and* $\epsilon > 0$, *it holds that* $P_{\theta_0}(\{\theta' : \theta' \notin \mathcal{N}_\epsilon(\theta_0)\} \cap C_n \neq \varnothing) \to 0$ *as* $n \to \infty$.

Assumption 2 (i) is crucial for $\{\mathcal{R}^{C_n}_{\alpha,n}\}$ to have correct asymptotic level, whereas (ii) is necessary (but not sufficient) for the test not to be conservative. For $\{\mathcal{R}^{C_n}_{\alpha,n}\}$ to be a feasible test we must specify a sequence $\{C_n\}$ that satisfies Assumption 2, without assuming that θ_0 is known. This will be addressed in the next subsection.

Assumption 3 *For any* $\alpha \in (0,1)$ *the correspondence* $\theta \mapsto \{u : F_\theta(u) \leq 1-\alpha\}$ *is upper semicontinuous on* Θ_0.

Assumption 3 requires a weak form of continuity of F_θ, see Debreu (1959), and the assumption is satisfied in our leading example of simultaneous testing of multiple linear

[5] Naturally, for $\mathcal{R}^{C}_{\alpha-\delta_n,n}$ to be an exact α-level test, knowledge about the finite sample distribution of T_n, beyond the asymptotic distribution, F_{θ_0}, is needed.

inequalities. An alternatively formulation of Assumption 3 is that $\lim_{\epsilon\to 0}\inf_{\theta\in\mathcal{N}_\epsilon(\theta_0)} F_\theta(u) = F_{\theta_0}(u)$ for $\theta_0 \in \partial\Theta_0$, such that a test that is based on a shrinking neighborhood of $\theta_0 \in \partial\Theta_0$ is not conservative. This is not guaranteed to hold under Assumption 2 alone.

We are now ready to formulate our main result.

Theorem 3 *Suppose that Assumption 1 holds and let $\{C_n\}$ satisfy Assumption 2. Then (i) $\mathcal{R}_{\alpha,n}^{\Theta_0} \subset \mathcal{R}_{\alpha,n}^{C_n}$ for all n and all $\theta \in \Theta$; (ii) $\limsup_{n\to\infty}\sup_{\theta\in\Theta_0} P_\theta(\mathcal{R}_{\alpha,n}^{C_n}) \leq \alpha$ for all $\theta \in \Theta_0$; and (iii) Under Assumption 3, $\lim_{n\to\infty} P_\theta(\mathcal{R}_{\alpha,n}^{C_n}) = \alpha$ for all $\theta \in \partial\Theta_0$.*

Theorem 3 (*i*) shows that the shrinking confidence set test, $\mathcal{R}_{\alpha,n}^{C_n}$, is more powerful than (or as powerful as) the LFC-test, and (*ii*) shows that this test has the correct asymptotic size. Finally, (*iii*) states that the test will have an asymptotic rejection probability that exactly equals the level (and size) of the test if $\theta_0 \in \partial\Theta_0$ and Assumption 3 holds. This result holds for for any test statistic, T_n, that satisfies our regularity conditions.

2.4 Regular Testing Problems

We now turn to a framework that is general enough to include most standard problems. The advantage of this framework is that we can use the regularity conditions and make specific suggestions on how $\{C_n\}$ should be constructed.

Assumption 4 *The estimator has the form $\hat{\theta}_n - \theta = M_n(\theta)n^{-1}\sum_{t=1}^n s_t(\theta)$, and uniformly in θ it holds that, $M_n(\theta) \xrightarrow{p} \Phi$ for some full rank matrix Φ and that $\{s_t(\theta)\}$ satisfies*

$$n^{-1/2}\sum_{t=1}^{[nu]} s_t(\theta) \xrightarrow{w} \Sigma_{ss}^{1/2}B(u), \qquad \textit{where} \quad \Sigma_{ss} \equiv \lim_{n\to\infty}\operatorname{var}(n^{-1/2}\sum_{t=1}^{n} s_t(\theta))$$

is positive definite and $B(u)$ is a standard k-dimensional Brownian motion.

The uniform convergence is required in order to control the asymptotic size of tests based on shrinking confidence sets. West (1996) has shown that many common estimators have the form that is required by Assumption 4, including the OLS, ML, IV, and GMM estimators under standard regularity conditions. The asymptotic covariance matrix of $\hat{\theta}_n$ is given by $\Sigma_{\theta\theta} \equiv \Phi\Sigma_{ss}\Phi'$.

Under Assumption 4, by the law of the iterated logarithm,

$$\sup_\lambda \lim_{n\to\infty} \sup_{1\leq t\leq n} \left| \frac{\sum_{\tau=1}^t \lambda' s_\tau(\theta)}{\sqrt{\lambda'\Sigma_{ss}\lambda\, n\, 2\log\log n}} \right| = 1, \tag{1}$$

almost surely, which suggests a bound for the rate at which C_n can shrink to $\{\theta_0\}$.

Assumption 5 *Consider a sequence of sets, $C_n \subset \Theta_0$, $n = 1, 2, \ldots$, and define $d_n \equiv \inf\{d \in \mathbb{R} : C_n \subset \mathcal{N}_d(\hat{\theta}_n)\}$. It holds that (i) $\lim_{n\to\infty} d_n = 0$ almost surely and (ii) $P_\theta(B_{g_n} \subset C_n) \to$*

1 *as* $n \to \infty$ *uniformly in* θ, *where*

$$B_{g_n} \equiv \left\{ y \in \Theta_0 : n(y - \hat{\theta}_n)' \Sigma_{\theta\theta}^{-1} (y - \hat{\theta}_n) \leq g_n \right\}, \tag{2}$$

where $g_n = o(n)$ *and satisfies* $g_n \to \infty$ *as* $n \to \infty$.

The assumption characterizes a class of sequences that shrink to a set will zero volume at a rate that is slow enough to capture B_n, and hence θ_0. The advantage of this assumption is that it gives us some flexibility in our choice of $\{C_n\}$, which can be useful whenever it is difficult to determine B_n, as defined in (2). For example, if $\Sigma_{\theta\theta}$ is unknown and difficult to estimate. Given (1) it is tempting to set $g_n = 2 \log \log(n)$, but a slower increasing sequence will suffice, as (1) yields a stronger result that is required for Assumption 2 to hold (the $\sup_{1 \leq t \leq n}$ in (1) is not needed as only $\hat{\theta}_n$ is required to be close to θ_0).

Lemma 4 *Let* $\{\hat{\theta}_n\}$ *satisfy Assumption 4 and let* $\{C_n\}$ *satisfy Assumption 5. Then Assumption 2 holds.*

It should be observed that Assumption 2 will hold under weaker assumptions than those of Assumptions 4 and 5. It will suffice that $n^{1/2}(\hat{\theta}_n - \theta_0)$ converges uniformly in distribution, but Assumption 4 motivates a particular rate at which C_n could shrink to $\{\theta_0\}$, and prescribed by (1). Further, these results can also be derived in situations where the rate of convergence is different from $\sqrt{n}$ and where the limit distribution is non-Gaussian. Such cases would require a different construction of C_n and will not be explored in this paper.

It is easy to construct a data dependent sequence of sets, C_n, that satisfies Assumption 5, such as the sets in the following example.

Example 1 *Suppose that* $\{\delta_n\}$ *is such that* (i) $\lim_{n \to \infty} \delta_n = 0$; *and* (ii) *there exists an* $N \in \mathbb{N}$, *such that* $\delta_n \geq \sqrt{g_n/n}$ *for all* $n \geq N$. *Then* $C_{1,n} = \{y \in \Theta_0 : (y - \hat{\theta}_n)' \hat{\Sigma}_{\theta\theta}^{-1} (y - \hat{\theta}_n) \leq \delta_n^2\}$, *satisfies Assumption 5 provided that* $\hat{\Sigma}_{\theta\theta} \xrightarrow{p} \Sigma_{\theta\theta}$ *uniformly in* θ; *and so does* $C_{2,n} = \{y \in \Theta_0 : \sup_{1 \leq i \leq p} |y_i - \hat{\theta}_{i,n}| / \hat{\sigma}_i \leq \delta_n\}$, *provided that* $\hat{\sigma}_i^2 \xrightarrow{p} \sigma_i^2$, $i = 1, \ldots, p$, *where* $\hat{\theta}_{i,n}$ *is the* i*th element of* $\hat{\theta}_n$, *and where* σ_i^2 *is the* i*th diagonal element of* $\Sigma_{\theta\theta}$.

Sequences, $\{\delta_n\}$, *that satisfy conditions* (i) *and* (ii) *include* $\{\delta_n = \kappa_0 + \kappa_1 n^{\gamma - 1/2} : \gamma \in (0, 1/2)$ *and* $\kappa_1 > 0\}$, *e.g.,* $\delta_n = n^{-1/4}$, *and* $\{\delta_n = \kappa_0 + \kappa_1 (\log^q(n)/n)^{1/2} : \kappa_1 > 0, q \in \mathbb{N}\}$, *e.g.,* $\delta_n^2 = 2 \log(\log(n))/n$.

2.5 Similarity and Unbiased Tests

Next, we introduce a similarity condition that is relevant for asymptotic tests of composite hypotheses. The equivalent condition for finite sample tests is well known, see, e.g., Cox and Hinkley (1974, p. 150) and Gourieroux and Monfort (1995, chapter 16). The conditions is expressed in terms of the boundary of the null hypothesis, $\partial\Theta_0$.

Definition 1 (Similar on the boundary) *An asymptotic test,* $\{\mathcal{R}_n\}$*, is* asymptotically similar on the boundary of the null hypothesis *if* $\lim_{n\to\infty} P_\theta(\mathcal{R}_n) = \alpha$ *for all* $\theta \in \partial\Theta_0$*, where* α *is the asymptotic size of* $\{\mathcal{R}_n\}$*.*

Definition 2 (Unbiased) *An asymptotic test,* $\{\mathcal{R}_n\}$*, is* asymptotically unbiased *if* $\liminf_{n\to\infty} P_{\theta_n}(\mathcal{R}_n) \geq \alpha$ *for any sequence of alternatives,* $\theta_n \notin \Theta_0$*, where* α *is the asymptotic size of* $\{\mathcal{R}_n\}$*.*

Next, we consider local alternatives (Pittman drifts) that have the following form. For $\theta_0 \in \partial\Theta_0$ and a $y \in \mathbb{R}^p$ that is such that $(\theta_0, \theta_0 + y] \subset \Theta_a = \Theta \backslash \Theta_0$, the local alternative (local to θ_0 in the direction y) is given by $\theta_0 + n^{-1/2}\epsilon y$ for $\epsilon > 0$.

Assumption 6 *For any local alternative,* $\theta_0 + n^{-1/2}\epsilon y$*, the rejection probability* $\rho(\epsilon) \equiv \lim_{n\to\infty} P_{\theta_0 + n^{-1/2}\epsilon y}(\mathcal{R}_n)$ *is continuous in* ϵ*, for* $\epsilon \geq 0$*.*

Assumption 6 typically holds under Assumption 4 for test statistics that are continuous in $\hat{\theta}_n$.

Theorem 5 *Let Assumption 6 hold. A necessary condition for* $\{\mathcal{R}_n\}$ *to be asymptotically unbiased is that* $\{\mathcal{R}_n\}$ *is asymptotically similar on the boundary of* Θ_0*.*

The existence of an unbiased test is not guaranteed, even if a sequence, $\{C_n\}$, that satisfies our assumptions is available. However, the resulting test will dominate the LFC-test as formulated in the following corollary.

Corollary 6 *Let Assumptions 1 and 3 hold and suppose that the LFC-test is non-similar on* $\partial\Theta_0$*. If there exists a sequence,* $\{C_n\}$*, that satisfies Assumption 2 then the LFC-test,* $\mathcal{R}_{\alpha,n}^{\Theta_0}$*, is asymptotically inadmissible.*

Note that the corollary does not claim that $\{\mathcal{R}_{\alpha,n}^{C_n}\}$ is unbiased or admissible, and in fact $\{\mathcal{R}_{\alpha,n}^{C_n}\}$ need not have any of these properties without additional assumptions. In fact, an admissible test in the class of tests, $\{\mathcal{R}_{\alpha,n}^{C_n}\}$, that satisfies our assumptions is unlikely to exist, as a C_n that is constructed using a slower rate of g_n (e.g. $\log(g_n)$) will have unit power against a larger class of local alternatives, without compromising the asymptotic size.

2.6 *p*-VALUES

Given a realization of the test statistic, $T_n(\omega) = \tau$, we define $\mathcal{D}_{n,\tau} \equiv \{\omega' : T_n(\omega') > \tau\}$ and the probability of this event is given by $p_n^{\{\theta_0\}}(\tau) \equiv P_{\theta_0}(\mathcal{D}_{n,\tau})$, which we may refer to as the 'true' p-value. The conventional p-value is defined by $p_n^{\Theta_0}(\tau) \equiv \sup_{\theta\in\Theta_0} P_\theta(\mathcal{D}_{n,\tau})$ and is closely related to the LFC-test, as the rejection region, $\mathcal{R}_n = \{\omega : p_n^{\Theta_0}(\tau) \leq \alpha\}$, defines the LFC-test with asymptotic level α. Our testing procedure also yields a p-value, which is given by $p_n^{C_n}(\tau) \equiv \sup_{\theta\in C_n} P_\theta(\mathcal{D}_{n,\tau})$.

Corollary 7 *Let Assumptions 1 and 2 hold. (i) The p-values, $p_n^{\Theta_0}(\tau)$ and $p_n^{C_n}(\tau)$, are asymptotically valid, and (ii) if, in addition, Assumption 3 holds, then $p_n^{C_n}(\tau) \xrightarrow{p} p_n^{\{\theta_0\}}(\tau)$ for $\theta_0 \in \partial\Theta_0$.*

So the p-value of the new testing procedure possesses a consistency for the true p-value, which is not the case for the conventional p-value in general.

2.7 Simultaneous Testing of Multiple Linear Inequalities

Consider a null hypothesis that is given by linear inequalities, in which case Θ_0 is a convex cone. This problem has been analyzed Perlman (1969) in a general framework, and by Judge and Yancey (1986) and Wolak (1987) in the context of the linear regression models. A related testing problem is that in Gouriéroux, Holly, and Monfort (1982) and Andrews (1998), who considers the case with a simple null hypothesis against a restricted alternative. Goldberger (1992) compares of the tests of Wolak (1987) and Gouriéroux, Holly, and Monfort (1982), and provides valuable insight to the power properties of these tests through graphical illustrations of the tests' rejection regions.

Suppose that $n^{1/2}(\bar{X}_n - \theta_0) \xrightarrow{d} N_m(0, \Sigma)$, and consider the hypothesis $R\theta - r \leq 0$, where R is a full rank $m \times m$ matrix and r is a $m \times 1$ vector. Thus $\Theta_0 = \{\theta : R\theta - r \leq 0\}$ in this case. Wolak (1987) proposed the quadratic test statistic, $T_n = \min_{\theta \in \Theta_0} n(\theta - \bar{X}_n)'\Sigma^{-1}(\theta - \bar{X}_n)$, for testing H_0. It is easy to verify that the point least favorably to the alternative is given by $\theta_{LFC} = R^{-1}r$, which is the unique value of θ for which all inequalities are binding, see Wolak (1987) and Robertson, Wright, and Dykstra (1988, pp. 68–69). So the asymptotic distribution of T_n is bounded by the distribution, $F_{\theta_{LFC}}$, which can be shown to be a mixture of χ^2 distributions, $\sum_{i=0}^{m} \omega_i \chi^2_{(i)}$, where $\sum_{i=0}^{m} \omega_i = 1$, and ω_i, $i = 0, \ldots, m$ are positive constants that depend on R and Σ, see Wolak (1987). It is well-known, that the LFC-test is conservative if any of the inequalities are non-binding ($\theta_0 \neq \theta_{LFC}$), see Wolak (1989b, p. 220). In fact, the discrepancy between F_{θ_0} and $F_{\theta_{LFC}}$ increases with the number of non-binding inequalities, where F_{θ_0} is the asymptotic distribution of T_n and $F_{\theta_{LFC}}$ is that employed by the LFC-test. The reason is that only the binding inequalities matter for the asymptotic distribution, F_{θ_0}. Since Assumption 3 is satisfied in this framework, the LFC-test is inadmissible and will be inferior to the test based on shrinking confidence sets.

One of the conclusions of this paper is that the LFC approach is not necessary in order to construct asymptotically exact tests.[6] In fact, the approach of this paper yields tests that will dominate the corresponding LFC-tests in terms of power. The reason that the LFC approach can be improved (without compromising the asymptotic size) is that there

[6] The literature appear to suggest otherwise, e.g., Wolak (1989a, p. 10) writes: "A least favorable value of $\theta \in \Theta_0$ must be found to construct an asymptotically exact size test of the inequality constraints" [formulated in our notation].

is sufficient information to determine exactly which inequalities that are non-binding, (as $n \to \infty$). A second advantage of the new testing procedure is that it produces unbiased tests, because they will be asymptotically similar on the boundary of Θ_0.

2.8 Simulation Experiment

To quantify the potential gains from the new testing procedure we consider a simple simulation experiment, where we have generated pseudo random numbers, $n^{1/2}(\bar{X}_n - \theta) \sim N_m(0, I)$, for various choices of θ. We consider the null hypothesis, $H_0 : \theta \leq 0$, and cases where the null hypothesis is true are labelled by (*type I error*) whereas cases where $\theta \nleq 0$ are labeled by (*power*). We evaluate the actual rejection probabilities in both cases through simulations.

In the study of type I error, we let the $[\rho m]$ first coordinates of θ be $\theta_i = 0$, (the binding inequalities) and the remaining $m - [\rho m]$ coordinates be $\theta_i = -\frac{1}{4} < 0$ (the non-binding inequalities), where $\rho \in (0, 1]$. So θ is consistent with the hypothesis $H_0 : \theta \leq 0$. The power study is identical to that of the type I error, with the modification that $\theta_1 = \frac{1}{4}$, such that the first inequality is violated and the null hypothesis is false in these simulations.

We compare two tests that are based the test statistic of Wolak (1987), which simplifies to $T_n = n \sum_{i=1}^{m} \left(\bar{X}_{i,n}^{+}\right)^2$ due to the simple (and known) covariance structure. The first test is based on the new testing procedure, which invokes the 'confidence' set $C_n = \{y \in \mathbb{R}^m : y_i \in [\hat{\theta}_{i,n} - c_n, \hat{\theta}_{i,n} + c_n],\ i = 1, \ldots, m\}$, where $\hat{\theta}_n \equiv \bar{X}_n^{+}$ and $c_n = \sqrt{(2 \log\log n)/n}$, for $n = 1, 2, \ldots$. Note that $\{C_n\}$ satisfies Assumption 5. The other test is the LFC-test and the tests are labelled by "log^2" and "LFC", respectively.

Tables 1 and 2 report the results for all possible combinations of $m = 10$, 40, 100 (the number of inequalities); $\rho = 0.1$, 0.2, 0.5, 0.8, 0.9, or 1.0 (the ratio of binding inequalities); and for $n = 40$, 100, 500, and '∞' (the sample size, '∞' corresponds to the results for the largest value of n allowed by the software ($n > 10^{300}$)).[7] Table 1 contains the results at the 5% level and Table 2 reports the results at the 10% level. As can be seen, the distortions from non-binding inequalities can be enormous. When the null is true and less than 20% ($\rho = 0.2$) of the inequalities are binding, the LFC-test has a rejection probability that is close to zero, and this hurts the LFC-test substantially in terms of power. For example for $(\rho, n, m) = (0.1, 200, 40)$, the log^2-test is quite powerful against this alternative, whereas LFC test hardly has any power. The estimated rejection probabilities are 69.2% and 2.1%, respectively, for the tests at the 5% level. An extreme case is $(\rho, n, m) = (0.1, 500, 100)$ where the log^2-test almost has unit power whereas the LFC-test has power close to zero. It can be seen that the distortion of the LFC-test increases with the dimensionality of the

[7] In each simulation, we generate one draw from the multivariate standard Gaussian distribution, which is scaled by $n^{-1/2}$ and recentered about θ, to yield a draw of $\bar{X}_n$.

null hypothesis, m, and as can be expected, the distortion is reduced as ρ get closer to one. However, even for $\rho = 0.9$ we note that the LFC-test is clearly inferior to the $\log^2$-test. For example the estimated rejection probabilities for the 5%-level tests are 31.6% and 18.3%, respectively, for the configuration $(\rho, n, m) = (0.9, 200, 100)$. The case $\rho = 1$ corresponds to the situation where the true parameter is the point least favorably to the alternative. In this situation, one might be concerned about the finite sample properties of the $\log^2$-test. However, as can be seen from the tables the size distortion of the $\log^2$-test is small and vanishes with the sample size, as predicted by the theory. For $m = 10$ the rejection probabilities of the $\log^2$-test never exceeds the intended level by more than 0.5%. Not surprisingly, the $\log^2$-test becomes more liberal as the number of inequalities increases (for the case $\rho = 1$), and it may be desirable to modify c_n to depend on m, e.g., $c_{n,m} = \sqrt{(2\log(\log(mn))/n}$ to reduce the distortion, however, such aspects are beyond the scope of this paper.

3 The Reality Check for Data Snooping

White (2000) proposed a test for superior predictive ability, which amount to testing the hypothesis that a benchmark model is not dominated by a set of alternative models in terms of predictive ability. The framework of White considers $m + 1$ competing models, where the benchmark model is indexed by $k = 0$ and the alternative models are indexed by $k = 1, \ldots, m$. Model k produces a sequence of forecasts, $\hat{Y}_{k,1}, \ldots, \hat{Y}_{k,n}$ of some sequence of random objects, $Y_1, \ldots, Y_n$, and the forecasts are evaluated with an additive loss function $L(Y_t, \hat{Y}_{k,t})$, from which the relative performance variables

$$f_{k,t} = L(Y_t, \hat{Y}_{0,t}) - L(Y_t, \hat{Y}_{k,t}), \qquad k = 1, \ldots, m, \quad t = 1, \ldots, n,$$

are defined. White (2000) makes assumptions that ensure that $n^{1/2}(\bar{f}_n - \mu) \xrightarrow{d} N_m(0, \Omega)$, where $\bar{f}_n = n^{-1}\sum_{t=1}^{n}(f_{1,t}, \ldots, f_{m,t})'$, $\mu = (\mu_1, \ldots, \mu_m)'$, and $\mu_k = E(f_{k,t})$, $k = 1, \ldots, m$, and where Ω is the asymptotic covariance matrix.

A positive μ_k corresponds to model k having a better predictive ability than model 0, so the null hypothesis is given by $H_0 : \mu \leq 0$. An equivalent formulation of the null hypothesis is $\max_{k=1,\ldots,m} \mu_k \leq 0$, which motivates the test statistic, $T_n^{rc} \equiv \max_{k=1,\ldots,m} n^{1/2}\bar{f}_{k,n}$, that is employed by the RC. The asymptotic distribution of T_n^{rc} depends on the nuisance parameters μ and Ω. White (2000) proposes to use the stationary bootstrap to handle the dependence on Ω,[8] whereas a bound is applied to control for the dependence on μ, (approach 1 in the example of Section 2). So, in this respect, the RC is a LFC-test and the bound is given from $\mu = 0$.

[8] Based on theoretical results of Lahiri (1999) there is reason to believe that other bootstrap techniques may preform better in this context. For an overview on bootstrap techniques for dependent time series, see Härdle, Horowitz, and Kreiss (2002).

This implies that the RC asymptotically derives critical values from $\max_{k=1,\ldots,m} Z_k$, where Z is a Gaussian m-dimensional vector with mean zero and variance Ω. It is easily verified that the true asymptotic distribution is given by the distribution of $Z_{\max} \equiv \max_{j=1,\ldots,m_0} Z_j$, where $(Z_1, \ldots, Z_{m_0})' \sim N_{m_0}(0, \Sigma)$, m_0 is the number of models with $\mu_k = 0$, and Σ is the $m_0 \times m_0$ submatrix of Ω that contains the (i,j)'th element of Ω if $\mu_i = \mu_j = 0$. All models are worse than the benchmark when $m_0 = 0$ and in this case $T_n^{rc} \xrightarrow{p} -\infty$, whereas $T_n^{rc} \xrightarrow{p} \infty$ under the alternative ($\max_k \mu_k > 0$). The latter confirms that the asymptotic power of the RC is one.

3.1 Some Unfortunate Properties of the RC

As shown in our theoretical analysis, it is only the binding constraints ($\mu_k = 0$) that matter for the asymptotic distribution when applying the test statistic of Wolak (1987). This is also the case for the test statistic, T_n^{rc}, which implies that the RC is conservative whenever $m_0 < m$. This is highlighted by the following example, where $m = 2$ and $m_0 = 1$.

Example 2 *Consider the case with a benchmark forecast and two alternative forecasts, $m = 2$. Suppose that $f(t)$ is iid $N_2(\mu, \Omega)$, where $\mu = (0, \gamma)'$, $\gamma < 0$ and Ω is a 2×2 diagonal matrix, $\Omega = \text{diag}(1, \omega^2)$. Thus, the first alternative forecast is as good as the benchmark, whereas the other is worse. The exact distribution of $\bar{f}_n$ is given by $n^{1/2}\bar{f}_n \sim N_2\left(n^{1/2}\mu, \Omega\right)$, and since the two relative performance variables, $f_1(t)$ and $f_2(t)$, are independent, the distribution of $T_n^{rc} \equiv n^{1/2} \max_{k=1,2} \bar{f}_{k,n}$ is given by*

$$F_0(x) = P(T_n^{rc} \leq x) = \Phi(x)\Phi(\frac{x - n^{1/2}\gamma}{\omega}),$$

where Φ denotes the cdf of the standard normal.

If γ is small (very negative) and $\omega^2 = |\gamma|$, then $-n^{1/2}\gamma$ is large and $\Phi\left((x - n^{1/2}\gamma)/\omega\right) \approx 1$ for positive values of x. So the critical value will almost entirely be determined from the distribution of $n^{1/2}\bar{f}_{1,n}$, whereas $n^{1/2}\bar{f}_{2,n}$ is almost irrelevant for the distribution of T_n^{rc}.

Asymptotically, the RC derives critical values from $\max_{k=1,2} Z_k$, where $(Z_1, Z_2)' \sim N_2(0, \Omega)$ to derive critical values, and the distribution of $\max_{k=1,2} Z_k$ is given by

$$F_{LFC}(x) = P(Z_{\max} \leq x) = \Phi(x)\Phi(\frac{x}{\omega}).$$

Since ω is large, the upper tail of $F_{LFC}(x)$ is dominated by $\Phi(\frac{x}{\omega})$, which is the distribution of Z_2, thus the critical value will almost entirely be determined from the distribution of Z_2.

The example illustrates the sensitivity of the RC to *irrelevant alternative models*, a role played by $\bar{f}_{2,n}$ in this case. Although the probability that $\bar{f}_{2,n} > \bar{f}_{1,n}$ is negligible, the RC

allows $\bar{f}_{2,n} - \gamma$ to define the critical values. Another implication is that a poor model reduce the power of the RC.

The example also confirms the result of Theorem 5, that the RC is a biased test. In a situation where some forecasts are better than the benchmark whereas other are worse, there will exists local alternatives μ_n, with both positive and negative elements, for which $\lim_{n\to\infty} P_{\mu_n}(\text{RC rejects } H_0) < \alpha$, where $\alpha = \lim_{n\to\infty} P_{\mu=0}(\text{RC rejects } H_0)$.

A situation where a poor model can severely distort the RC is when the (relative) performance is bounded from above, but not necessarily from below. For example if the models are compared using the mean squared prediction error, $L(Y, \hat{Y}) = (Y - \hat{Y})^2$. The variable of interest is here denoted by Y and $\hat{Y}$ represents a prediction of Y. In this case, the expected (relative) performance of model k is given by $\mu_k \equiv E(Y - \hat{Y}_0)^2 - E(Y - \hat{Y}_k)^2$, and the sample equivalent is $\bar{f}_{k,n} \equiv n^{-1} \sum_{t=1}^{n} [(Y(t) - \hat{Y}_0(t))^2 - (Y(t) - \hat{Y}_k(t))^2]$. Given a realization of $(Y(1), \ldots, Y(n))$ and benchmark forecasts, $(\hat{Y}_0(1), \ldots, \hat{Y}_0(n))$, the relative sample performance, $\bar{f}_{k,n}$, takes values in $(-\infty, c]$, where $c = n^{-1} \sum_{t=1}^{n} (Y(t) - \hat{Y}_0(t))^2$. For a test to have any power at all, its critical value (for the test statistic T_n^{rc}) must be less than $n^{1/2}c$. It is therefore reasonable to require that a critical value lies in the interval $(-\infty, n^{1/2}c]$. However, a critical value of the RC can be greater than $n^{1/2}c$, because it is derived from a vector of random variables with the same distribution as $n^{1/2}\left(\bar{f}_n - \mu\right)$. Since $\bar{f}_{k,n} - \mu_k$ can have a substantial amount of its probability mass to the right of c if $\mu_k < 0$, this can result in a critical value that is larger than $n^{1/2}c$. Naturally, this is a small sample phenomenon, because as $n \to \infty$, the distribution of $\bar{f}_{k,n} - \mu_k$ will be concentrated about zero.

We can summarize the unfortunate properties of the RC as follows:

1. The RC is asymptotically biased.

2. The RC is sensitive to the inclusion of poor models that can create an artificial non-rejection of a false null hypothesis. This applies to a situation where it is possible to add forecasting models that are worse than the benchmark model.

3. If forecast are evaluated by a loss function that is bounded from below, then the critical values of the RC can be so large that it requires an unobtainable performance in order to reject the null hypothesis. In this case the RC may have no power, although the problem vanishes as $n \to \infty$.

4. The p-values of the RC are typically inflated.

3.2 Studentization of the Test Statistic

We propose to use a different test statistic than that of the RC. The usual way to combine multiple tests involves a standardization of the individual test statistic or a transformation

to p-values. Several ways of combining (independent) p-values are discussed in Folks (1984), one being the Tippett method, which is based on the smallest p-value, see Tippett (1931). Combining p-values using resampling techniques is discussed in Westfall and Young (1993), and Dufour and Khalaf (2002) analyze a problem where dependent p-values are combined to construct exact tests for contemporaneous correlation in seemingly unrelated regressions.

In the light of the literature, the test statistic, $T_n^{rc} \equiv n^{1/2} \max_{k=1,\ldots,m} \bar{f}_{k,n}$, is non-standard, and we shall take a different approach, which is similar to using the smallest p-value. Specifically we suggest to use tests statistic, $T_n^{sm} \equiv \max_{k=1,\ldots,m} \frac{n^{1/2}\bar{f}_{k,n}}{\hat{\omega}_k}$, where $\hat{\omega}_k^2$ is a consistent estimate of $\text{var}(n^{1/2}\bar{f}_{k,n})$.[9] This test statistic takes supremum over the m standardize statistics, (the t-statistics for relative forecast performance), whereas the test statistic of the RC takes supremum over non-standardized statistics.

It is well known that bootstrapping (asymptotically) pivotal quantities is better than bootstrapping non-pivotal quantities, see, e.g., Babu and Singh (1983) and Singh and Babu (1990). We are considering a situation where a 'good' estimate of Ω is unavailable, so it is not possible to combine the statistics, $\bar{f}_{1,n}, \ldots, \bar{f}_{m,n}$ into a useful statistic that is asymptotically pivotal. Nevertheless, there may benefits from a partial pivoting, and this is what the substitution, T_n^{sm} is place of T_n^{rc}, amounts to. The standardization removes part of the nuisance dependence on Ω, in the sense that the asymptotic distribution of T_n^{sm} depends on (μ, ϱ) where ϱ is the asymptotic correlation matrix of $n^{1/2}\bar{f}_n$. So the asymptotic distribution of T_n^{sm} has fewer nuisance parameters than that of T_n^{rc}, which depends on (μ, Ω).

As will be evident from the empirical application, some of the RC's problems are alleviated by using T_n^{rc} instead of T_n^{rc}.

3.3 An Empirical Illustration

We illustrate the problems of the RC by revisiting the empirical application of White (2000). The question of interest is whether a linear regression models, which is based on technical indicators, is capable of predicting daily returns of the S&P 500 index better than the sample average of historical returns. The comparison of models is made with the mean squared prediction error criterion and we analyze two sample periods. Our first sample is identical to the one analyzed by White (2000), which spans the period from March 29, 1988 through May 31, 1994, and the second is an extended sample, which spans the period, March 29, 1988 through November 15, 2000.

We denote the one-day ahead return of the S&P 500 index by $Y(t)$, $t = -R+1, \ldots, 0, 1, \ldots, n$, where $R = 803$ is the number of observations used for estimation before the first prediction is made. This leaves us with $n = 758$ daily observation for the forecast compar-

[9] The supscript, 'sm', refers to standardized maximum. In our empirical analysis we estimate $\hat{\omega}_k^2$ with the bootstrap, $k = 1, \ldots, m$.

ison. The competing linear models are constructed by taking all possible combinations of 3 out of the 29 technical indicators as regressors, in addition to a constant. This leads to $m = 3,654$ models.[10] The 29 technical indicators are the following: lagged returns ($Z_1(t)$), momentum measures ($Z_2(t), \ldots, Z_{11}(t)$), local trends ($Z_{12}(t), \ldots, Z_{15}(t)$), relative strength indices ($Z_{16}(t), \ldots, Z_{19}(t)$), and moving average oscillators ($Z_{20}(t), \ldots, Z_{29}(t)$). The indicators, $Z_i(t)$, $i = 1, \ldots, 29$, are observable at time $t-1$, see White (2000) for more details on the technical indicators.

The forecasts are given by

$$\hat{Y}_k(t+1) = \hat{\beta}'_{k,t} X_k(t+1), \qquad t = 0, \ldots, n-1, \quad k = 1, \ldots, m,$$

where $X_k(t) = (Z_{i1_k}(t), Z_{i2_k}(t), Z_{i3_k}(t), 1)'$, and $\hat{\beta}_{k,t}$ is the least squares estimator from regressing Y on X_k, using past observations up to time t.

The benchmark model corresponds to a regression model that only includes a constant. This model is nested in any of the competing models and under the null hypothesis it holds that $\beta_k = 0$, $k = 1, \ldots, m$. So in this case the null hypothesis $\mu \leq 0$ is equivalent to the simple hypothesis, $\mu = 0$, and the RC does not suffer from non-binding inequalities. However to illustrate how the RC is affected by a poor model we consider three additional model-sets.

The empirical results are presented in Table 3,[11] where $M^\dagger$ and M^* refer to a 'poor' and a 'good' model, respectively. These were constructed as follows. The forecast errors of the benchmark model are given by $\varepsilon_{0,t} = Y_t - \hat{Y}_{0,t}$, $t = 1, \ldots, n$, and the sequence of 'poor' forecasts is defined by $\hat{Y}_{p,t} = Y_t - (\frac{1}{2} + 2\upsilon_t)\varepsilon_{0,t}$ and the sequence of 'good' forecasts is given $\hat{Y}_{g,t} = Y_t - (0.9 + 0.15\eta_t)\varepsilon_{0,t}$, where $\upsilon_t, \eta_t \sim iid\,\text{uniform}(0,1)$, $t = 1, \ldots, n$. So the forecast errors of $M^\dagger$ are, on average, 150% larger than those of the benchmark, whereas the average sized of M^*'s forecast errors are 97.5% times those of the benchmark. In Table 3, the original set of forecasting model is denoted by $\mathcal{M}_{\text{org}}$ and the set that also includes the poor model is denoted by $\mathcal{M}_{\text{org}} + M^\dagger$. In our analysis of the power properties we consider the original set plus the 'good' forecast, which is denoted by $\mathcal{M}_{\text{org}} + M^*$, and the set that includes both the 'poor' and the 'good' forecast, which is denoted by, $\mathcal{M}_{\text{org}} + M^\dagger + M^*$.

The short sample with the original set of forecasting models, $\mathcal{M}_{\text{org}}$, corresponds to that investigated by White (2000), and we arrive at the same conclusion as White and find no evidence against the null hypothesis in either of the two samples.[12] There is no difference

[10] Some of these models are identical, due to colinearity of the 29 technical indicators.

[11] The analysis was made using Ox, version 3.00, see Doornik (1999). For the sake of comparability we employ the same bootstrap techniques as in White (2000), and p-values are derived using the stationary bootstrap, with a dependence-parameter $q = .5$ and $B = 1,000$ bootstrap resamples. The resamples are used to estimate both $\hat{\omega}_k^2$ and the distribution of the test statistic.

[12] There are unimportant differences between the results in Table 3 and those reported by White (2000). These can be explained by numerical issues, the random number generator used for the bootstrap implementation, and different sources of data.

between the $\log^2$-test and the LFC-test in this case, as can be expected since μ equals zero if the null hypothesis is true. However, the second set of forecasting models, $\mathcal{M}_{\text{org}} + M^{\dagger}$, shows that the p-value of the RC is severely distorted by the inclusion of a single poor model. The p-value of the original RC jumps from 27.3% to 59.4% (48.0% to 74.7% in the extended sample), whereas the $\log^2$-test is unaffected. Since the analysis of the original set of models, led to a non-rejection of the null hypothesis, the inflated p-value of the RC does not affect the conclusion of the test in this case. However, in our analysis of the power we see that the RC is blinded by the inclusion of the poor model. In the set of models that does not include the 'poor' forecast, the null hypothesis is clearly rejected by the RC, as the p-value is 0.0%. So the RC is capable of detecting the 'good' model in this case. When the set of models also includes the poor forecast the p-value jumps to 35.9%, and the RC is no longer capable of detecting the 'good' forecasting model! Even in the extended sample the RC does not get close to rejecting the false null hypothesis. The $\log^2$-test is unaffected by the inclusion of the poor model, which shows the strength of this testing procedure.

The lower half of Table 3, presents the result for the second test statistic, T_n^{sm}, which performs much better than the RC. The standardization of the individual performance statistics reduces the influence of the 'poor' model, and this test leads to the correct conclusion in both cases – regardless of the way the nuisance parameter, μ, is treated ($\log^2$ or LFC). The LFC approach inflates the p-value by 1.0% in the short sample and by 0.2% in the extended sample, when the null hypothesis is (presumed to be) true, and there are no noticeable differences in the results with the sets of forecasting models that address the power of the tests.

Although the partial pivoting of the test statistic led to the correct conclusion in this applications, it does not control the (nuisance) dependence on μ. In general, there will be additional gains from using the methods of shrinking confidence sets, such as the $\log^2$-approach. This is clear from our simulation study, where μ was the only nuisance parameter (θ in the previous notation), and where the test statistic was in a standardized form. Our recommendation for the testing problem considered by White (2000), is to use the (partially) standardized test statistic and the methods of shrinking confidence sets for hypothesis testing.

4 Summary and Concluding Remarks

In this paper, we considered asymptotic tests of composite hypotheses and proposed a testing procedure that avoids the use of conservative bounds as $n \to \infty$. The new testing procedure is superior to standard tests that are based on the least favorable configuration, because it leads to asymptotically unbiased tests that are more powerful than LFC-tests.

The new testing procedure applies to the simultaneous testing of multiple inequalities, and is particularly useful when the number of inequalities is large. Through simulations, we

studied a particular alternative to the LFC-test, the $\log^2$-test, and showed that this leads to a substantial gain in power. In some cases the difference in power was close to 99%.

We introduced an asymptotic similarity condition and showed that it is a necessary condition for a test to be unbiased. In the problem of testing multiple inequalities we showed than it is simple to derive a test that satisfies the similarity condition, and that this test will dominate the corresponding LFC-test.

Testing for superior predictive ability is a test of a composite hypothesis and the new testing procedure will dominate tests that are based on the LFC, such as the reality check of White (2000). In fact, we concluded that the RC does not satisfy a relevant similarity condition and showed that this led to several unfortunate properties. The RC is a biased test and the RC is sensitive to the inclusion of irrelevant alternatives, in the sense that the inclusion of a poor model leads to loss of power. The p-value of the RC is typically inflated and poor models can cause an 'artificial' non-rejection of a false null hypothesis. This adds a high degree of subjectivity to hypotheses testing when the RC is used, because it is possible to 'avoid' a rejection of the null hypothesis by including poor models. One important exception, where the RC need not be affected by these problems, is when the null hypothesis implies that none of the competing models are worse than the benchmark. In this case the null hypothesis reduces to a simple hypothesis, which eliminates the mean parameter, μ, as a nuisance parameter.

The conclusion is that the RC should not be applied to compare forecasting models if there is reason to believe that some of the models could be worse than the benchmark forecast. Given these results, it might be appropriate to revisit the empirical studies that applied the RC. The reason being that a failure to reject the null hypothesis may have been caused by poor forecasting models in the set of competing forecasts. The testing procedure of composite hypotheses, which was introduced in this paper, can greatly improve the power properties, however a partial pivoting of the test statistic is also very advantageous, as can be seen from our empirical application.

The new testing procedure is applicable to several other econometric problem, besides that of comparing forecasting models. When testing inequality constraints, the largest gains in power can be expected when the number of non-binding inequalities is large. The procedure may be particularly useful when testing a null hypothesis that is defined by a continuum of inequalities. This is the case when testing for first or second order stochastic dominance, see McFadden (1989) and Klecan, McFadden, and McFadden (1991), and when testing for a structural change with an unknown change point, Andrews (1993).

Tests of inequality constraints are, perhaps, most frequently used in linear regression models, where the number of constraints is typically a small number. The power improvements in this framework are yet to be seen, and we leave this for further research.

Appendix: Proofs

Proof of Lemma 1. Given the convergence in distribution we have that $P_\theta(T_n \in \mathcal{I}_\alpha^\theta)$ converges to a number less than α for $\theta \in \Theta_0$. The uniform convergence in distribution of T_n (Assumption 1.*ii*), ensures that $\limsup_{n\to\infty} \sup_{\theta\in\Theta_0} P_\theta(T_n \in \mathcal{I}_\alpha^{\Theta_0}) \leq \limsup_{n\to\infty} \sup_{\theta\in\Theta_0} P_\theta(T_n \in \mathcal{I}_\alpha^\theta) \leq \alpha$. Finally, $\lim_{n\to\infty} P_{\theta_0}(\mathcal{R}_{\alpha,n}^{\{\theta_0\}}) = \alpha$ for $\theta_0 \in \partial\Theta_0$ follows by the convergence in distribution and the continuity of F_θ for $\theta \in \partial\Theta_0$. ■

Proof of Lemma 2. (*i*) follows from the fact that $C_\epsilon \subset \Theta_0$, and to prove (*ii*) we define $\mathcal{C}_n \equiv \{\omega \in \Omega : \theta_0 \in \mathcal{N}_\epsilon(\hat{\theta}_n) \cap \Theta_0\}$, such that $(\mathcal{R}_{\alpha,n}^{C_\epsilon} \cap \mathcal{C}_n) \subset \mathcal{R}_{\alpha,n}^{\{\theta_0\}}$. This leads to the inequality $P_{\theta_0}(\mathcal{R}_{\alpha,n}^{C_\epsilon}) = P_{\theta_0}(\mathcal{R}_{\alpha,n}^{C_\epsilon} \cap \mathcal{C}_n) + P_{\theta_0}(\mathcal{R}_{\alpha,n}^{C_\epsilon} \cap \complement\mathcal{C}_n) \leq P_{\theta_0}(\mathcal{R}_{\alpha,n}^{\{\theta_0\}}) + P_{\theta_0}(\complement\mathcal{C}_n)$, and since $\hat{\theta}_n \xrightarrow{p} \theta_0$ (uniformly in θ_0 on Θ_0) we have that $\lim_{n\to\infty} \sup_{\theta\in\Theta_0} P_\theta(\complement\mathcal{C}_n) = 0$, such that $\limsup_{n\to\infty} \sup_{\theta\in\Theta_0} P_\theta(\mathcal{R}_{\alpha,n}^{C_\epsilon}) \leq \limsup_{n\to\infty} \sup_{\theta\in\Theta_0} P_\theta(\mathcal{R}_{\alpha,n}^{\{\theta_0\}}) + \limsup_{n\to\infty} \sup_{\theta\in\Theta_0} P_\theta(\complement\mathcal{C}_n) = \alpha$, where we have used Lemma 1. ■

Proof of Theorem 3. (*i*) follows from Lemma 2. To prove (*ii*) we define $\mathcal{C}_n \equiv \{\omega \in \Omega : \theta_0 \in C_n\}$, such that $(\mathcal{R}_{\alpha,n}^{C_n} \cap \mathcal{C}_n) \subset \mathcal{R}_{\alpha,n}^{\{\theta_0\}}$. This leads to the inequality

$$P_{\theta_0}(\mathcal{R}_{\alpha,n}^{C_n}) = P_{\theta_0}(\mathcal{R}_{\alpha,n}^{C_n} \cap \mathcal{C}_n) + P_{\theta_0}(\mathcal{R}_{\alpha,n}^{C_n} \cap \complement\mathcal{C}_n) \leq P_{\theta_0}(\mathcal{R}_{\alpha,n}^{\{\theta_0\}}) + P_{\theta_0}(\complement\mathcal{C}_n).$$

From Lemma 1 we have that $\lim_{n\to\infty} \sup_{\theta\in\Theta_0} P_\theta(\mathcal{R}_{\alpha,n}^{\{\theta_0\}}) = \alpha$ and by Assumption 2 (*i*) we have that $\lim_{n\to\infty} \sup_{\theta\in\Theta_0} P_\theta(\complement\mathcal{C}_n) = 0$, which proves that $\limsup_{n\to\infty} \sup_{\theta\in\Theta_0} P_{\theta_0}(\mathcal{R}_{\alpha,n}^{C_n}) \leq \alpha$.

To show (*iii*) we use the identity, $P_{\theta_0}(\mathcal{R}_{\alpha,n}^{C_n}) = P_{\theta_0}(\mathcal{R}_{\alpha,n}^{C_n}) - P_{\theta_0}(\mathcal{R}_{\alpha,n}^{\{\theta_0\}}) + P_{\theta_0}(\mathcal{R}_{\alpha,n}^{\{\theta_0\}}) - \alpha + \alpha$, which shows that

$$|P_{\theta_0}(\mathcal{R}_{\alpha,n}^{C_n}) - \alpha| \leq |P_{\theta_0}(\mathcal{R}_{\alpha,n}^{C_n}) - P_{\theta_0}(\mathcal{R}_{\alpha,n}^{\{\theta_0\}})| + |P_{\theta_0}(\mathcal{R}_{\alpha,n}^{\{\theta_0\}}) - \alpha|.$$

The last term equals $|P_{\theta_0}(T_n \geq a) - \alpha|$, where $a \equiv F_{\theta_0}^{-1}(1-\alpha)$, since F_{θ_0} is continuous under Assumption 1, and the convergence in distribution guarantees that this term converges to zero. Since $(\mathcal{R}_{\alpha,n}^{C_n} \cap \mathcal{C}_n) \subset (\mathcal{R}_{\alpha,n}^{\{\theta_0\}} \cap \mathcal{C}_n)$ the other term can be bounded by

$$|P_{\theta_0}(\mathcal{R}_{\alpha,n}^{C_n}) - P_{\theta_0}(\mathcal{R}_{\alpha,n}^{\{\theta_0\}})| \leq P_{\theta_0}(T_n \in [a, b_n)) + P_{\theta_0}(\complement\mathcal{C}_n),$$

where $b_n = \inf\{b : b \in \mathcal{I}_\alpha^{C_n} \cup \mathcal{I}_\alpha^{\theta_0}\}$. Given Assumptions 2 and 3 it follows that $b_n \to a$ and from the continuity of F_{θ_0} it follows that $P_{\theta_0}(T_n \in [a, b_n)) \to 0$ as $n \to \infty$. This completes the proof. ■

Proof of Lemma 4. We have $B_n \subset C_n$ and $\theta_0 \in B_n$ for n sufficiently large. Since the latter is assumed to hold uniformly in θ (Assumption 4) we have shown Assumption 2 (*i*). The relation, $\bigcap_{k=1}^\infty \bigcup_{n=k}^\infty C_n \subset \bigcap_{k=1}^\infty \bigcup_{n=k}^\infty \mathcal{N}_{d_n}(\theta_0) = \bigcap_{k=1}^\infty \mathcal{N}_{d_k}(\theta_0) = \{\theta_0\}$ almost surely,

proves Assumption 2 (ii). ■

Proof of Theorem 5. Suppose that $\mathcal{R}_n$ is non-similar on the boundary, such that there exists $\theta_0 \in \partial\Theta_0$ for which $\lim_{n\to\infty} P_{\theta_0}(\mathcal{R}_n) < \alpha$, where α is the size of the test. Then for some local alternative, $\theta_n = \theta_0 + n^{-1/2}\epsilon y$ where $\epsilon > 0$, it holds that $\rho(\epsilon) \equiv \lim_{n\to\infty} P_{\theta_n}(\mathcal{R}_n) < \alpha$, given the continuity of the normalized rejection probability. This shows that the test is biased. ■

Proof of Corollary 6. From Theorem 3 it follows that $\mathcal{R}^{\Theta_0}_{\alpha,n}$ and $\mathcal{R}^{C_n}_{\alpha,n}$ have the same asymptotic size (α) and that $\mathcal{R}^{C_n}_{\alpha,n}$ is at least as powerful as $\mathcal{R}^{\Theta_0}_{\alpha,n}$. If $\mathcal{R}^{\Theta_0}_{\alpha,n}$ is non-similar on the boundary of the null hypothesis, there will exist local alternatives to some point on the boundary, $\theta_0 \in \partial\Theta_0$, for which $\mathcal{R}^{C_n}_{\alpha,n}$ is more powerful than $\mathcal{R}^{\Theta_0}_{\alpha,n}$, which shows that $\mathcal{R}^{\Theta_0}_{\alpha,n}$ is asymptotically inadmissible.[13] ■

Proof of Corollary 7. From Theorem 3 (ii) we have that $p_n^{\Theta_0}(\tau) \geq p_n^{C_n}(\tau)$ and $\lim_{n\to\infty} p_n^{C_n}(\tau) \geq \lim_{n\to\infty} p_n^{\{\theta_0\}}(\tau)$, which shows that the p-values are valid, and similar to the last result of Theorem 3 (ii), it follows that $p_n^{C_n}(\tau) \xrightarrow{p} p_n^{\{\theta_0\}}(\tau)$ for $\theta_0 \in \partial\Theta_0$. ■

References

Andrews, D. W. K. (1993): "Test for Parameter Instability and Structural Change with Unknown Change Point," *Econometrica*, 61, 821–856.

——— (1998): "Hypothesis Testing with a Restricted Parameter Space," *Jounal 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.

Babu, G. J., and K. Singh (1983): "Inference on Means Using the Bootstrap," *Annals of Statistics*, 11, 999–1003.

Bahadur, R. R., and L. J. Savage (1956): "The Nonexistence of Certain Statistical Procedures in Nonparametric Problems," *Annals of mathematical statistics*, 27, 1115–1122.

Berger, R. L. (1996): "More Powerful Tests from Confidence Interval p Values," *The American Statistician*, 50, 314–318.

Berger, R. L., and D. D. Boos (1994): "P Values Maximized over a Confidence Set for the Nuisance Parameter," *Journal of the American Statistical Association*, 89, 1012–1016.

Cavanagh, C. L., G. Elliott, and J. H. Stock (1995): "Inference in Models with Nearly Nonstationary Regressors," *Econometric Theory*, 11, 1131–1147.

Chernoff, H. (1954): "On the Distribution of the Likelihood Ratio," *Annals of Mathematical Statistics*, 25, 573–578.

Cox, D. R. (1961): "Test of Seperate Families of Hypotheses," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 105–123.

[13] Following the definition of Lehmann (1947), applied to our asymptotic framework.

——— (1962): "Further Results on Tests of Seperate Families of Hypotheses," *Journal of the Royal Statistical Society, Ser. B*, 24, 406–423.

Cox, D. R., and D. V. Hinkley (1974): *Theoretical Statistics*. Chapman and Hall, London.

Debreu, G. (1959): *Theory of Value*. Wiley, New York.

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.

——— (1990): "Exact Tests and Confidence Sets in Linear Regressions with Autocorrelated Errors," *Econometrica*, 58, 475–494.

——— (1997): "Some Impossibility Theorems in Econometrics with Applications to Structural and Dynamic Models," *Econometrica*, 65, 1365–1387.

Dufour, J.-M., and L. Khalaf (2002): "Exact Tests for Contemporaneous Correlation of Disturbances in Seemingly Unrelated Regressions," *Journal of econometrics*, 106, 143–170.

Dufour, J.-M., and J. F. Kiviet (1996): "Exact Tests for Structural Change in First-Order Dynamic Models," *Journal of Econometrics*, 70, 39–68.

——— (1998): "Exact Inference Methods for First-Order Autoregressive Distributed Lag Models," *Econometrica*, 66, 79–104.

Dufour, J.-M., and M. Taamouti (2001): "Projection-Based Statistical Inference in Linear Structural Models with Possibly Weak Instruments," *Discussion Paper, C.R.D.E. University of Montreal*.

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.

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.

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.

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

Härdle, W., J. Horowitz, and J.-P. Kreiss (2002): "Bootstrap Methods for Time Series," *Mimeo*.

Horowitz, J. L. (2001): "The Bootstrap," in *Handbook of Econometrics*, ed. by J. J. Heckman, and E. E. Leamer, vol. 5. North-Holland.

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

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

Jansson, M. (2002): "Autocorrelation Robust Tests with Good Size and Power," *UC Berkeley, Mimeo*.

Judge, G. G., and T. A. Yancey (1986): *Improved Methods of Inference in Econometrics*. North-Holland, Amsterdam.

Kiefer, N. M., T. J. Vogelsang, and H. Bunzel (2000): "Simple Robust Testing of Regression Hypotheses," *Econometrica*, 68, 695–714.

Klecan, L., R. McFadden, and D. L. McFadden (1991): "A Robust Test for Stochastic Dominance," *Working paper, Dept. of Economics, MIT.*

Kleibergen, F. (2002): "Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression," *Econometrica*, 70, 1781–1804.

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

Lancaster, T. (2000): "The Incidental Parameter Problem Since 1948," *Journal of Econometrics*, 95, 391–413.

——— (2002): "Orthogonal Parameters and Panel Data," *Review of Economic Studies*, 69, 647–666.

Lehmann, E. L. (1947): "On Families of Admissible Tests," *Annals of Mathematical Statistics*, 18, 97–104.

Loh, W.-Y. (1985): "A New Method for Testing Seperate Families of Hypotheses," *Journal of the American Statistical Association*, 80, 362–368.

McFadden, D. L. (1989): "Testing for Stochastic Dominance," in *Studies in the Economics of Uncertainty*, ed. by T. Fomby, and T. K. Seo, pp. 113–134, New York. Springer.

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

Moreira, M. J. (2003): "A Conditional Likelihood Ratio Test for Structural Models," *Econometrica (forthcoming)*.

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

Robertson, T., F. T. Wright, and R. L. Dykstra (1988): *Order Restricted Statistical Inference*. John Wiley & Sons, New York.

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.

Silvapulle, M. J. (1996): "A Test in the Presence of Nuisance Parameters," *Journal of the American Statistical Association*, 91, 1690–1693.

Singh, K., and G. Babu (1990): "On Asymptotic Optimality of the Bootstrap," *Scandinavian Journal of Statistics*, 17, 1–9.

Staiger, D., and J. H. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, 557–586.

Stock, J., and M. Yogo (2002): "Testing for Weak Instruments in Linear IV Regression," *NBER Working Paper: T284*.

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.

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: Type I Error and Power Properties ($\alpha = 0.05$)

		Type I Error						*Power*					
ρ	n	$m = 10$		$m = 40$		$m = 100$		$m = 10$		$m = 40$		$m = 100$	
		$\log^2$	LFC	$\log^2$	LFC	$\log^2$	LFC	$\log^2$	LFC	$\log^2$	LFC	$\log^2$	LFC
	40	0.002	0.000	0.000	0.000	0.000	0.000	0.078	0.035	0.001	0.000	0.000	0.000
	100	0.006	0.000	0.000	0.000	0.000	0.000	0.450	0.178	0.070	0.001	0.002	0.000
0.1	200	0.023	0.000	0.009	0.000	0.002	0.000	0.924	0.544	0.692	0.021	0.330	0.000
	500	0.049	0.000	0.048	0.000	0.049	0.000	1.000	0.987	0.999	0.514	0.995	0.008
	'∞'	0.050	0.000	0.050	0.000	0.050	0.000	1.000	1.000	1.000	1.000	1.000	1.000
	40	0.004	0.001	0.000	0.000	0.000	0.000	0.087	0.044	0.002	0.000	0.000	0.000
	100	0.009	0.001	0.001	0.000	0.000	0.000	0.444	0.200	0.087	0.003	0.006	0.000
0.2	200	0.030	0.001	0.015	0.000	0.006	0.000	0.905	0.568	0.640	0.037	0.315	0.000
	500	0.049	0.001	0.050	0.000	0.050	0.000	1.000	0.988	0.997	0.582	0.981	0.024
	'∞'	0.050	0.001	0.050	0.000	0.050	0.000	1.000	1.000	1.000	1.000	1.000	1.000
	40	0.016	0.008	0.003	0.001	0.001	0.000	0.125	0.084	0.018	0.005	0.002	0.000
	100	0.024	0.008	0.010	0.000	0.004	0.000	0.438	0.283	0.141	0.025	0.035	0.001
0.5	200	0.041	0.008	0.031	0.000	0.025	0.000	0.846	0.655	0.556	0.138	0.302	0.006
	500	0.050	0.008	0.052	0.000	0.052	0.000	0.999	0.993	0.986	0.762	0.921	0.173
	'∞'	0.050	0.008	0.050	0.000	0.050	0.000	1.000	1.000	1.000	1.000	1.000	1.000
	40	0.035	0.027	0.024	0.014	0.016	0.006	0.168	0.141	0.067	0.042	0.032	0.014
	100	0.041	0.027	0.033	0.014	0.024	0.006	0.443	0.373	0.204	0.117	0.101	0.033
0.8	200	0.048	0.027	0.045	0.014	0.042	0.006	0.805	0.725	0.514	0.322	0.306	0.101
	500	0.052	0.027	0.052	0.014	0.055	0.006	0.998	0.995	0.963	0.890	0.837	0.543
	'∞'	0.050	0.027	0.050	0.014	0.050	0.006	1.000	1.000	1.000	1.000	1.000	1.000
	40	0.044	0.038	0.036	0.027	0.031	0.020	0.182	0.167	0.092	0.072	0.061	0.039
	100	0.047	0.037	0.042	0.026	0.038	0.020	0.445	0.406	0.224	0.167	0.136	0.077
0.9	200	0.050	0.037	0.049	0.026	0.050	0.020	0.787	0.747	0.505	0.398	0.316	0.183
	500	0.052	0.037	0.053	0.026	0.055	0.020	0.997	0.996	0.954	0.921	0.815	0.667
	'∞'	0.050	0.037	0.050	0.026	0.050	0.020	1.000	1.000	1.000	1.000	1.000	1.000
	40	0.053	0.050	0.056	0.050	0.061	0.050	0.200	0.193	0.125	0.112	0.102	0.087
	100	0.052	0.050	0.054	0.050	0.058	0.050	0.448	0.440	0.242	0.226	0.173	0.154
1.0	200	0.052	0.050	0.054	0.050	0.057	0.050	0.775	0.771	0.496	0.479	0.318	0.299
	500	0.052	0.050	0.053	0.050	0.056	0.050	0.996	0.996	0.945	0.941	0.793	0.781
	'∞'	0.050	0.050	0.050	0.050	0.050	0.050	1.000	1.000	1.000	1.000	1.000	1.000

This table shows the properties of the $\log^2$-test and the LFC-test of $H_0 : \theta_i \leq 0$, $i = 1, \ldots, m$ for various configurations. The proportion of binding inequalities is ρ and the power simulations are based on a violation of the first inequality. The rejection probabilities are estimated from 10,000 simulations.

Table 2: Type I Error and Power Properties ($\alpha = 0.10$)

		Type I Error						*Power*					
ρ	n	$m = 10$		$m = 40$		$m = 100$		$m = 10$		$m = 40$		$m = 100$	
		$\log^2$	LFC	$\log^2$	LFC	$\log^2$	LFC	$\log^2$	LFC	$\log^2$	LFC	$\log^2$	LFC
	40	0.005	0.001	0.000	0.000	0.000	0.000	0.131	0.065	0.003	0.000	0.000	0.000
	100	0.014	0.001	0.001	0.000	0.000	0.000	0.573	0.268	0.125	0.003	0.005	0.000
0.1	200	0.044	0.001	0.020	0.000	0.006	0.000	0.959	0.658	0.788	0.042	0.446	0.000
	500	0.097	0.001	0.099	0.000	0.096	0.000	1.000	0.994	1.000	0.625	0.998	0.019
	'∞'	0.100	0.001	0.100	0.000	0.100	0.000	1.000	1.000	1.000	1.000	1.000	1.000
	40	0.009	0.003	0.000	0.000	0.000	0.000	0.149	0.080	0.007	0.001	0.000	0.000
	100	0.022	0.003	0.004	0.000	0.000	0.000	0.569	0.299	0.150	0.007	0.013	0.000
0.2	200	0.060	0.003	0.035	0.000	0.017	0.000	0.949	0.686	0.747	0.070	0.429	0.000
	500	0.100	0.003	0.101	0.000	0.099	0.000	1.000	0.994	0.999	0.689	0.992	0.045
	'∞'	0.100	0.003	0.100	0.000	0.100	0.000	1.000	1.000	1.000	1.000	1.000	1.000
	40	0.036	0.020	0.010	0.002	0.002	0.000	0.207	0.143	0.041	0.013	0.005	0.001
	100	0.055	0.019	0.025	0.002	0.010	0.000	0.566	0.398	0.231	0.054	0.070	0.002
0.5	200	0.082	0.019	0.065	0.002	0.049	0.000	0.913	0.763	0.680	0.220	0.430	0.014
	500	0.101	0.019	0.103	0.001	0.105	0.000	1.000	0.997	0.993	0.847	0.958	0.268
	'∞'	0.100	0.019	0.100	0.001	0.100	0.000	1.000	1.000	1.000	1.000	1.000	1.000
	40	0.074	0.059	0.053	0.033	0.034	0.016	0.268	0.231	0.127	0.085	0.065	0.033
	100	0.085	0.059	0.068	0.031	0.055	0.014	0.574	0.503	0.314	0.195	0.179	0.066
0.8	200	0.097	0.059	0.092	0.031	0.087	0.014	0.884	0.826	0.646	0.450	0.440	0.171
	500	0.101	0.059	0.104	0.031	0.107	0.014	0.999	0.998	0.983	0.940	0.908	0.666
	'∞'	0.100	0.059	0.100	0.031	0.100	0.014	1.000	1.000	1.000	1.000	1.000	1.000
	40	0.089	0.077	0.078	0.058	0.066	0.042	0.290	0.265	0.167	0.133	0.114	0.077
	100	0.094	0.077	0.087	0.058	0.080	0.041	0.583	0.538	0.341	0.263	0.222	0.140
0.9	200	0.099	0.077	0.099	0.058	0.100	0.041	0.876	0.844	0.640	0.531	0.442	0.288
	500	0.102	0.077	0.104	0.058	0.108	0.041	0.999	0.998	0.980	0.959	0.889	0.781
	'∞'	0.100	0.077	0.100	0.058	0.100	0.041	1.000	1.000	1.000	1.000	1.000	1.000
	40	0.105	0.100	0.111	0.100	0.118	0.100	0.310	0.301	0.209	0.193	0.184	0.157
	100	0.104	0.100	0.108	0.100	0.114	0.100	0.582	0.572	0.364	0.347	0.273	0.254
1.0	200	0.103	0.100	0.106	0.100	0.112	0.100	0.864	0.860	0.622	0.607	0.450	0.429
	500	0.103	0.100	0.105	0.100	0.110	0.100	0.999	0.999	0.974	0.972	0.875	0.866
	'∞'	0.100	0.100	0.100	0.100	0.100	0.100	1.000	1.000	1.000	1.000	1.000	1.000

This table shows the properties of the $\log^2$-test and the LFC-test of $H_0 : \theta_i \leq 0$, $i = 1, \ldots, m$ for various configurations. The proportion of binding inequalities is ρ and the power simulations are based on a violation of the first inequality. The rejection probabilities are estimated from 10,000 simulations.

Table 3: Empirical Results:

		True Null Hypothesis				*False Null Hypothesis*			
		$\mathcal{M}_{\text{org}}$		$\mathcal{M}_{\text{org}} + M^{\dagger}$		$\mathcal{M}_{\text{org}} + M^{*}$		$\mathcal{M}_{\text{org}} + M^{*} + M^{\dagger}$	
Test stat.	n	$\log^2$	LFC	$\log^2$	LFC	$\log^2$	LFC	$\log^2$	LFC
T_n^{rc}	758	0.273	0.273	0.273	0.594	0.000	0.000	0.000	0.359
	2,392	0.480	0.480	0.480	0.747	0.000	0.000	0.000	0.285
T_n^{sm}	758	0.591	0.592	0.591	0.602	0.001	0.001	0.001	0.001
	2,392	0.414	0.420	0.414	0.422	0.000	0.000	0.000	0.000

This table contains the p-values of four tests, applied to eight testing problems. The four tests are the combinations of the test statistics, T_n^{rc} and T_n^{sm}, and the two ways to derive critical values, $\log^2$ and LFC. The eight testing problems are the combination of two samples and four sets of forecasting models, where $\mathcal{M}_{\text{org}}$ is the original set with 3,654 models that were analyzed by White (2000).

The reality check corresponds to the combination with T_n^{rc} and LFC, and it can be seen that the p-value of the RC is severely distorted by the inclusion of a poor (and for the hypothesis irrelevant) model.