Working Papers

02/2012

Testing the predictive ability of two classes of models

Ignacio Arbués, Cristina Casaseca, Ramiro Ledo and Silvia Rama

First draft: April 2012

This draft: April 2012

Testing the predictive ability of two classes of models

Abstract

We propose tests for the null that the best model of a class produces as good forecasts as the best model of another one. Forecasts are evaluated using a loss function. Thus, causality can be tested if only the models in one class use a certain input. This is applied to the unemployment/inflation and industrial orders/production relationships. We find causality for the USA, but neither for France nor Spain.

Keywords

Evaluating forecasts, Loss function, Model selection, Causality, Bootstrap, Monte Carlo.

Authors and Affiliations

Ignacio Arbués, Cristina Casaseca, Ramiro Ledo and Silvia Rama

D. G. of Methodology, Quality and Information and Communications Technology

National Statistics Institute

Testing the predictive ability of two classes of models

Ignacio Arbués[1,*], Cristina Casaseca[2], Ramiro Ledo[3] and Silvia Rama[4]

[1,2,3,4]D. G. de Metodología, Calidad y
Tecnologías de la Información y las Comunicaciones
Instituto Nacional de Estadística

[2]E-mail: cristina.casaseca.polo@ine.es, [3]E-mail: ramiro.ledo.arias@ine.es,
[4]E-mail: silvia.rama.garcia@ine.es

June 22, 2012

Abstract

We propose several tests to compare two classes of forecasting models. The null hypothesis is that the best model of class A produces at least as good forecasts as the best one of class B. The quality of the forecasts is measured by the mean of a certain loss function. This generalizes the Reality Check and Superior Predictive Ability tests that deal with the case of one class of models compared to a unique benchmark model.

*Corresponding author. Address: Instituto Nacional de Estadística. Castellana, 183, 28071, Madrid, Spain. E-mail: iarbues@ine.es. Telephone: +34 915834641. Fax: +34 915839499.

In particular, these tests can be used to causality. We consider the class A of the models that do not use X to predict Y and B the class that do use it. Then, rejecting the null indicates causality. We apply this to the relationships between unemployment and inflation and between industrial orders and production. We see that the generalized SPA indicates causality in both cases for the data from the USA, but neither for French nor Spanish data.

Keywords: Evaluating forecasts, Loss function, Model selection, Causality, Bootstrap, Monte Carlo.

1 Introduction

The aim of this paper is to provide a multi-model test to compare the predictive ability of two families of forecasting models. For this, we build on the Reality Check (White 2000) and the modifications performed by Hansen in his Superior Predictive Ability test (Hansen 2005). In the framework of these tests, a family of forecasting models are compared to a benchmark, where the predictive abilities are measured by the expected loss. We generalize these tests to the case of comparing the predictive ability of two families of forecast models. In our generalized framework we test the null hypothesis that the best model of the first class is as good as the best of the second one.

One useful application of this multi-model superior predictive ability test arises in the context of testing causality. Consider a class I of models that do not use a certain input X to forecast the variable of interest Y. Then, we can test the null that in the class J of the models that do include X as input, there is no model that performs better that the best of I.

Let us now review some of the literature. First of all, when we want to compare models from different classes, there is the possibility of picking one model from each class and then make a one-on-one comparison. For this comparisons, there are many tests available. Among the best-known ones it is the test of equal predictive ability (EPA) of Diebold and Mariano (1995), that has been generalized in Giacomini and White (2006) by considering conditional expectations of forecasts to the information set at a given point in time. The null in all these tests is that the expected loss function of the forecasts obtained with two models are equal. In the same framework, West (1996) develops an asymptotic procedure for EPA but taking into account the effects of uncertainty associated with estimated model parameters.

When the competing models are nested, one can use the encompassing tests by Clark and McCracken (2001) and Clark and West (2007). Here the null is that the forecasts of the less parsimonious model do not contain relevant information further than that is included in the most parsimonious one.

The Granger-Causality tests are similar to the encompassing tests, but they do not use the forecasts. Instead, the parameters of the bivariate (or, in general multivariate) model, are tested for a certain null hypothesis, which implies that there is a nested univariate model that would encompass the larger one. Nevertheless, this relationship between causality and encompassing test implies no similarity whatsoever in the results, since Granger-Causality tests reject the null more often.

All the tests considered above assume that there are two pre-determined models. Consequently they depend on the method used to select them. This selection can be made by standard identification methods, for instance, by using information criteria. Under this procedure, the results would only be

valid if the identification of the models is very reliable. This has prompted us to try a completely different approach, in which identification is not required, but just the specification of the two classes of models.

We have drawn from the idea of the Reality Check (White 2000; hereafter referred as White). This test is proposed for the usual scenario in which a large family of models are compared against a benchmark. In a forecasting exercise, it is likely that even if no model beats the benchmark in population terms, for a finite sample, some of them do. Consequently, his goal is to avoid the danger of data snooping (to mistake apparently good results generated by luck for true good results) by measuring the significance of the difference in performance between the best alternative model and the benchmark. In his paper he provides a method to test the null that the best model in a specification search has no predictive superiority over a given benchmark. The alternative is that at least the best model is superior to the benchmark.

Thus, whereas the above-mentioned tests of EPA consider the null as a simple hypothesis, the Reality Check formulates it as a composite one. This can be tackled controlling the significance level of each individual test using Bonferroni inequality, but this approach is not practically useful when the number of unilateral hypothesis, i.e., the number of models to compare with the benchmark, is large.

This leads White to devise a joint test naturally based on the minimum of the differences between the expected loss functions of the benchmark and the models to compare. Given that the null is a collection of composite hypothesis, it is necessary to decide what particular null is to be used for drawing the critical values. White chooses the Least Favorable Configuration (LFC), that is, the simple null that makes more probable to reject. If we ensure that the size of the test is correct for the LFC then, *a fortiori*, it is so

for any other null. The asymptotic p-value is then approximated by either Monte Carlo or Bootstrap methods.

Hansen (2005), hereafter referred as Hansen, points out that the RC is too conservative because the distribution under the null is obtained assuming the LFC. Thus, when we include poor models to compare, the test may suffer from a far from negligible lose of power due to the unreal increase of the empirical p-value. The purpose of the superior predictive ability (SPA) test proposed by Hansen is improving the power of RC test by re-centering the null distribution around a sample-dependent mean and so avoiding the use of the LFC.

Recently, Clark and McCracken (2011) have generalized Hansen's test to the case when the benchmark model is nested in all the alternatives. We also intend to generalize the RC and SPA tests, but in a different direction. Our goal is a generalization to the case in which we want to compare two classes of forecasting models with more than one model. The structure of the article is as follows. In section 2 we describe the RC and SPA tests and introduce their generalizations, GRC and GSPA to the framework of two families of forecasting models. The simulation results for these tests are reported in section 3. Section 4 presents some examples of application of the tests to analyze causality relationships. We conclude with some remarks.

2 Theory

In this section we will briefly describe the RC (White) and SPA (Hansen) tests together with their generalization to the case of two classes of models.

2.1 The framework in RC and SPA tests

The problem considered by White and Hansen is, given a family of m forecasting models, say indexed by $j = 1, \ldots, m$, whether there is one of them whose forecasts beat a benchmark model. The accuracy of the forecasts will be measured by a certain loss function $L(\cdot, \cdot)$, so if x_t is the value of the variable of interest at time t and $\{\hat{x}_{jt}\}_j$ are its forecasts according to the different models, then the preference criteria of a model j over j' is given by the relationship $\mathbb{E}L(x_t, \hat{x}_{jt}) \leq \mathbb{E}L(x_t, \hat{x}_{j't})$. We will abbreviate the notation by setting $L_{jt} = L(x_t, \hat{x}_{jt})$. If λ_j is the expectation of the loss function L_{jt} of model j (assuming stationarity), the null of the test is $H_0 : \lambda_0 \leq \min_{j=1,\ldots,m} \lambda_j$, where λ_0 is the expected loss function of the benchmark model. The alternative is that for at least one j, $\lambda_0 > \lambda_j$.

In terms of the performance relative to the benchmark we can formulate the null as $H_0 : \max_{j=1,\ldots,m} \mu_j \leq 0$, where $d_{jt} = L_{0t} - L_{jt}$ and $\mu_j = \mathbb{E}[d_{jt}] = \lambda_0 - \lambda_j$.

Let us assume that we have a sample $\{L_{jt}\}_{t=1,\ldots,T}$. Then, we can estimate λ_j with $\bar{L}_j = T^{-1} \sum_{t=1}^{T} L_{jt}$. Alternatively, we can work with the differences and get $\bar{d}_j = T^{-1} \sum_{t=1}^{T} d_{jt}$. Let us also introduce the notation $d_t = (d_{1t}, \ldots, d_{mt})'$, $\mu = \mathbb{E}[d_t]$ and $\bar{d} = (\bar{d}_1, \ldots, \bar{d}_m)'$. Under certain assumptions (originally, from West, 1996), it can be proved that $T^{1/2}(\bar{d} - \mu)$ converges in distribution to a multivariate normal. Thus, we write $T^{1/2}(\bar{d} - \mu) \xrightarrow{d} N(0, \Omega)$, where $\Omega = \lim_{T\to\infty} \mathrm{var}(T^{1/2}(\bar{d} - \mu))$.

2.2 White's Reality Check

White's RC test is based on the statistic $T^{\mathrm{RC}} = T^{1/2} \max_{j=1,\ldots,m} \bar{d}_j$. Since the null is a composite hypothesis, a choice must be done about which of

the different simple nulls is used to obtain the critical values. White uses the LFC to obtain the critical region of the RC test. That is, $\lambda_0 = \lambda_j$, i.e., $\mu = 0$.

Under the LFC, $T^{1/2}\bar{d} \xrightarrow{d} N(0, \Omega)$. Therefore, the distribution that we have to use to obtain critical values is the distribution of the maximum of a vector of zero-mean correlated normals, for which there is no known closed form. White proposes to approximate it by two ways: the first is Monte Carlo simulation, but the computational complexity of this method increases with m^2, so White suggests a bootstrap procedure applicable to dependent processes, more precisely, the stationary bootstrap of Politis and Romano (1994), hereafter referred as PR, whose complexity is instead linear in m. Then, the null is rejected when the statistic is above the approximate critical values obtained either with the Monte Carlo or with the Bootstrap procedures.

2.3 Generalized Reality Check

We want now to translate the RC ideas to the framework of the comparison of two classes of models. Let us assume them indexed by $i \in I = \{1, \ldots, n\}$ and $j \in J = \{n+1, \ldots, n+m\}$ respectively. Our goal is to test if the best model of the first family is as good as the best of the second one. Now the null is $H_0 : \min_{i \in I} \lambda_i \leq \min_{j \in J} \lambda_j$.

NOTE: hereafter, when we write $\min_i$ and $\min_j$, we mean $\min_{i \in I}$ and $\min_{j \in J}$ unless explicitly indicated otherwise.

The statistic of the RC test can be generalized to this case as

$$T^{\mathrm{GRC}} = T^{1/2}\big(\min_i \bar{L}_i - \min_j \bar{L}_j\big). \tag{1}$$

Now, we stack $\{\lambda_i\}_{i \in I}$ and $\{\lambda_j\}_{j \in J}$ into a vector λ and we build its

sample counterpart $\bar{L}$. Under the assumptions of the previous section, as in the RC, it holds that $T^{1/2}(\bar{L}-\lambda)$ converges in distribution to a multivariate normal $N(0,\Xi)$.

The adaptation of the RC to the new framework has been straightforward up to this point, but now arises the question what is the LFC for this test? The answer is that there is not in fact a LFC in the null, but we can approach asymptotically to the supremum of the rejection probability in H_0.

Proposition 1. *Let* $\Theta_0 = \{\theta \in \mathbb{R}_+^{n+m} : \min_i \theta_i \leq \min_j \theta_j\}$ *and for any* $\theta \in \Theta_0$, $z_\theta := \min_i(\theta_i + \eta_i) - \min_j(\theta_j + \eta_j)$, *where* η *is a vector of random variables. On the other hand, for any* $i \in I$, $x_i := \eta_i - \min_j \eta_j$. *Then for any* $a \in \mathbb{R}$,

$$\sup_{\theta\in\Theta_0} P[z_\theta \geq a] = \max_i P[x_i \geq a]. \tag{2}$$

This proposition means that we can obtain the critical values of the GRC test at significance level α by following these two steps: (i) for each $i \in I$, compute the critical value $\zeta_{\alpha,i}$ as in White's Reality Check when the benchmark is the ith model; (ii) obtain the critical value of the causality test as $\zeta_\alpha = \max_{i\in I}\{\zeta_{\alpha,i}\}$.

We have thus reduced the computation of the critical values of the GRC to computing the critical values of n RC tests. Note, however, that this is not equivalent to perform n RC tests and reject when any of them rejects, because what we compare to the ith critical value $\zeta_{\alpha,i}$ is not the statistic of the ith RC test, but the statistic of the GRC, T^{GRC}.

Then, we can obtain the distribution under the null by Monte Carlo using the following proposition.

Proposition 2. *Under assumptions A and B in White,*

(i) *If* $\min_i \lambda_i > \min_j \lambda_j$, *then* $T^{GRC} \xrightarrow{p} +\infty$.

(ii) $T^{1/2}(\bar{L}_i - \min_j \bar{L}_j) \xrightarrow{d} Z_i - \min_j Z_j$, *where* $(Z_i, Z_{n+1}, \ldots, Z_{n+m})' \sim N(0, \Xi^i)$ *and* $\Xi^i = (\Xi_{\alpha,\beta})_{\alpha,\beta=i,n+1,\ldots,n+m}$.

The proof is a straightforward consequence of the asymptotic normality and the continuous mapping theorem. See, for example, (Billingsley 1968).

It is also possible to use the PR stationary bootstrap. The validity of the bootstrap estimates is guaranteed by theorem 2.3 and corollary 2.4 in White, which, for the sake of brevity, we do not reproduce here.

2.4 Hansen's Superior Predictive Ability test

The null in Hansen's SPA test is the same as in White, but now, the test is constructed in a different way, by employing a studentized test statistic and invoking a sample dependent distribution under the null. The main advantage achieved with these modifications is to reduce the sensitivity to the inclusion of poor forecasting models and so improving the power property of the test.

The statistic of the test is

$$T^{\text{SPA}} = \max\left\{ \max_{j=1,\ldots,m} T^{1/2} \frac{\bar{L}_0 - \bar{L}_j}{\hat{\omega}_j}, 0 \right\} \tag{3}$$

where $\hat{\omega}_j^2$ is a consistent estimator of the variance of $T^{1/2}(\bar{L}_0 - \bar{L}_j)$. Thus, when the benchmark has the best sample performance ($\bar{L}_0 - \bar{L}_j \leq 0$) the test statistic is normalized to be zero and there is no evidence to reject the null.

The idea of this test is to avoid the LFC approach to the null distribution by recentering it around a data dependent choice for μ rather than $\mu = 0$ as in LFC. In Hansen's Theorem 1 and Corollary 2 he shows that the

asymptotic distribution of the statistic depends only on the models with $\mu_j = 0$ but not on the models with a positive mean. It could seem natural to exclude directly the models with $\bar{d}_j < 0$, but this does not lead to valid inference results. This is related with the inconsistency of the bootstrap for constrained estimators when the parameter is in the boundary of the feasible region (Andrews 2000; Hansen 2003). That is why several threshold rates to separate the good and poor alternatives are proposed. Accordingly, Hansen defines $\hat{\mu}^c$ as the vector with the jth element $\hat{\mu}_j^c = \bar{d}_j 1_{\{T^{1/2}\bar{d}_j/\hat{\omega}_j \geq \sqrt{2\log\log T}\}}$, where $1_{\{.\}}$ denotes the indicator function. It can be seen that $\hat{\mu}_i^c$ converges in probability to μ_j. Recentering the bootstraped distribution around $\hat{\mu}^c$ yields a better approximation of the asymptotic distribution of the statistic under the null: $T^{\text{SPA}} \xrightarrow{d} \max\{N(0, \Omega^0)\}$, where Ω^0 is a matrix obtained from Ω with the jth row and column annihilated when $\mu_j < 0$.

An important condition that has to be met when models are estimated recursively is that the benchmark cannot be nested in all the alternatives, because in that case, the asymptotic normality of the averaged differences is compromised.

2.5 Generalized Superior Predictive Ability test

In this section, we generalize the method of Hansen to our framework. One could, in principle, think of two ways to generalize the T^{SPA} to our case. Either

$$T^{\text{GSPA}} = \max\left\{ \max_j \min_i T^{1/2} \frac{\bar{L}_i - \bar{L}_j}{\hat{\omega}_{ij}}, 0 \right\} \tag{4}$$

or

$$T^{\text{GSPA}} = \max\left\{ \min_i \max_j T^{1/2} \frac{\bar{L}_i - \bar{L}_j}{\hat{\omega}_{ij}}, 0 \right\}, \tag{5}$$

where $\hat{\omega}_{ij}^2$ is a consistent estimator (for simplicity, we assume almost sure consistency) of the variance of $T^{1/2}(\bar{L}_i - \bar{L}_j)$. In both cases, it is clear that $T^{\text{GSPA}} = T^{\text{SPA}}$ when $n = 1$. However, the form of (5) equals the minimum of the statistics of the n SPA tests where each of the n models in the first class are taken in turn as benchmarks. Since this makes easier the interpretation of the GSPA test, will deal hereafter only with the second version. Experimentation with (4) remains for future work.

Let us denote by $d_{ijt} = L_{it} - L_{jt}$ the relative loss functions. We stack them in a vector as $d_t = \text{vec}(D_t)$ with $D_t = \{d_{ijt}\}_{i=1,j=1}^{n,m}$. Their means are accordingly stacked in a vector $\mu = \text{vec}(M)$, where $M = \{\mu_{ij}\}_{i=1,j=1}^{n,m}$ and $\mu_{ij} = \mathbb{E}(L_i - L_j)$. Their sample means are in $\bar{d} = T^{-1}\sum_t d_t$.

In order to ensure the existence of μ_{ij} and justify the use of bootstrap techniques and the consistency of the covariance matrix estimator, we need to suppose the stationarity of d_t and a mixing condition. The following assumption is a straightforward generalization of assumption 1 in Hansen.

Assumption 1. *d_t is strictly stationary and satisfies an $\alpha-$mixing condition of size $-(2+\delta)(r+\delta)/(r-2)$, with $r > 2$ and $\delta > 0$. Besides this, $\mathbb{E}\|d_{ijt}\|^{r+\delta} < +\infty$ and $\mathbb{V}[d_{ijt}] > 0$ for all $i = 1, \ldots, n, j = n+1, \ldots, n+m$.*

If Ξ is the $(n+m)\times(n+m)$ asymptotic covariance matrix of $T^{1/2}(\bar{L}-\lambda)$, then we denote by Ξ^0 the matrix such that $\Xi_{ij}^0 = \Xi_{ij}1_{\{\lambda_i=\lambda_j=\lambda^*\}}$, where $\lambda^* = \min\{\lambda_\ell : \ell = 1, \ldots, n+m\}$. We can write the statistic of the test as $T^{\text{GSPA}} = \varphi(T^{1/2}\bar{L}, \hat{\omega})$, where $\varphi : (u, w) \in \mathbb{R}^{n+m} \times \mathbb{R}^{nm} \mapsto \varphi(u, w) = \max\{\min_i \max_j (u_i - u_j)/w_{ij}, 0\}$. The asymptotic distribution of T^{GSPA} is given in the following proposition.

Proposition 3. *Suppose assumption 1 holds and let F_0 be the cumulative distribution function of $\varphi(Z, w_0)$, where $Z \sim N_{n+m}(0, \Xi^0)$. Then, under the*

null hypothesis,

(i) If $\min_i \lambda_i = \min_j \lambda_j$*, then* $\varphi(T^{1/2}\bar{L}, \hat{\omega}) \xrightarrow{d} F_0$*, if* $\hat{\omega} \xrightarrow{p} w_0$*.*

(ii) If $\min_i \lambda_i < \min_j \lambda_j$*, then* $\varphi(T^{1/2}\bar{L}, \hat{\omega}) = 0$ *almost surely for large* T*.*

Under the alternative $\min_i \lambda_i > \min_j \lambda_j$*,* $\varphi(T^{1/2}\bar{L}, \hat{\omega}) \xrightarrow{a.s.} \infty$*.*

As in the SPA, we will obtain the critical values (or p-values) by the Bootstrap, recentering the averages of the loss functions so that they satisfy a sample-dependent null. This null is required to preserve the asymptotic validity of the test and reduce the influence of the poor alternatives. Let us define first $\Delta_{ij} = T^{1/2}(\bar{L}_i - \bar{L}_j)/\hat{\omega}_{ij}$. Then, the sample-dependent null will be

$$\hat{\lambda}_i^c = \begin{cases} \bar{L}_i - \bar{L}_{\hat{i}^*} & \text{if} \quad \Delta_{i,\hat{i}^*} \geq g(T) \\ 0 & \text{if} \quad \Delta_{i,\hat{i}^*} < g(T) \end{cases}, \tag{6}$$

where $\hat{i}^* = \arg\min_i \bar{L}_i$ and $g(T)$ is a function such that $\limsup_T g(T)T^{-1/2} = 0$ and $\liminf_T g(T)/\sqrt{2\log\log T} > 1$. For $j > n$, $\hat{\lambda}_j^c$ is defined accordingly.

Then, the bootstrap is used to obtain the critical values as as follows. We take the sample of differences $\{d_t\}_t$, where $1 \leq t \leq T$. By an adequate resampling (for example, the PR stationary bootstrap) we obtain the bootstrapped difference averages $\bar{d}^{*,k}$, with $k = 1, \ldots, M$. Then, we re-center the values around $\hat{\mu}^c = \text{vec}(\hat{M}^c)$, where the matrix $\hat{M}^c$ contains the elements $\hat{\mu}_{ij}^c = \hat{\lambda}_i^c - \hat{\lambda}_j^c$.

The re-centered values are $\bar{d}^{*,c,k} = \bar{d}^{*,k} - \bar{d} + \hat{\mu}^c$. With them, we obtain M bootstrapped values of $T^{GSPA,*,k} = \psi(T^{1/2}\bar{d}^{*,c,k}, \hat{\omega}^{*,k})$.

In order to establish the validity of the boostrapped critical values we need to prove that $T^{GSPA,*,k}$ converges in distribution to F_0. In other words, we see that using $\hat{\mu}^c$, we have a theoretical null distribution that converges to

the true one, when the latter satisfies the null. This is proved in proposition 4, but before, we have to define the statistic of the test in terms of the differences instead of the loss functions as $T^{\text{GSPA}} = \psi(T^{1/2}\bar{d}, \hat{\omega})$, where $\psi : (v, w) \in \mathbb{R}^{nm} \times \mathbb{R}^{nm} \mapsto \psi(v, w) = \max\{\min_i \max_j v_{ij}/w_{ij}, 0\}$. Then, the asymptotic distribution F_0 is given by $\psi(V, w_0)$, where $V \sim N_{nm}(0, \Omega^0)$ and $\Omega^0_{i+mj,i'+mj'} = \Xi^0_{ii'} + \Xi^0_{jj'} - \Xi^0_{ji'} - \Xi^0_{ij'}$.

Proposition 4. *Let F_0 be as in proposition 3 and let F^c_T be the cumulative distribution function of $\psi(T^{1/2}\bar{d}^{*,c,k}, \hat{\omega}^{*,k})$. If $\rho(T^{1/2}(\bar{d}^{*,k} - \bar{d}), T^{1/2}(\bar{d} - \mu)) \to 0$, where ρ is some metric metrizing the convergence in distribution, then $F^c_T \longrightarrow F_0$ as $T \longrightarrow \infty$, for all continuity points of F_0.*

The condition $\rho(T^{1/2}(\bar{d}^{*,k} - \bar{d}), T^{1/2}(\bar{d} - \mu)) \to 0$ can be checked, for example, by applying theorem 2.3 in White or lemma 1 in Hansen.

3 Simulations

In this section we want to check by Monte Carlo simulations that the empirical size of the tests is correct and to compare the power under some alternatives. For this purpose we simulate values of the vector of loss functions $L_t \sim N(\lambda, \Xi)$, where $t = 1, \ldots, T$.

We will use the same covariance matrix in all cases. We wanted to mimic to some extent the behavior of real data, as observed in the examples of next section, in which the covariance matrix has only one large eigenvalue and the $n + m - 1$ remaining ones are much smaller. We get this by setting $\Xi = \alpha I_{n+m} + \beta \mathbf{1}_{n+m}\mathbf{1}'_{n+m}$, where $\mathbf{1}_{n+m}$ is a $(n + m) \times 1$ vector of ones ($\alpha = 1$, $\beta = 2$). We fix the level of significance in $\alpha = 0.10$ (the results for $\alpha = 0.05$ seemed to be similar).

We have obtained results for the GSPA with two choices of the function $g(T)$, namely $g_2(T) = \sqrt{2 \log \log T}$ (the used by Hansen) and $g_3(T) = \sqrt{3 \log \log T}$. With g_2 it is not almost sure that the good models are captured, because $\limsup_T x_T = 1$ does not preclude that there are infinite x_T such that $x_T > 1$. We distinguish the two variants of GSPA with a subscript.

The different scenarios are defined thus by the value of the vector of means λ. We consider five different cases to show the size and power properties of our tests. In all the cases we set the number of simulation replications $M = 1,000$ and a set of sample sizes given by $T = 50, 100, 200, 400, 800, 5,000, 20,000$. In the stationary bootstrap we set $S = 1,000$ the number of resamplings.

Case 1: as our GRC and GSPA test have been conceived as a generalization of RC and SPA tests, the first result presented in table 1 is the restriction of the formers to the univariate case, in order to check that the rejection probabilities under the null are close to the nominal levels in both tests. So we consider $n = 1$ and $m = 30$ models to compare, with means $\lambda_1 = 0$ and $\lambda_j = 0$ for $j = n + 1, \ldots, n + m$. In this scenario, where we are comparing a family of models to a benchmark model and all of them have the same mean, the GRC test, based in the LFC approach, seems to perform reasonably well for all the sample sizes. The rejection frequency of GSPA_2 remains somewhat high even with very long samples (with lengths in the order of 10^5, this effect disappears).

Case 2: in table 2 we present the rejection frequencies for the GRC and GSPA tests when comparing two families of $n = m = 30$ models with all means set to 0. While in this situation, the GRC does not work well (it never rejects the null), the GSPA seems to reach rejection frequencies closer to the nominal level.

Table 1: Rejection Frequencies for case 1.

T	50	100	200	400	800	5,000	20,000
GRC	0.104	0.082	0.086	0.088	0.080	0.108	0.092
$GSPA_2$	0.180	0.140	0.142	0.124	0.152	0.140	0.134
$GSPA_3$	0.126	0.118	0.132	0.120	0.138	0.104	0.096

Note: rejection frequencies for $n = 1$, $m = 30$, $\lambda_1 = 0$; and $\lambda_j = 0$ for $j = n + 1, \ldots, n + m$.

Table 2: Rejection Frequencies for case 2.

T	50	100	200	400	800	5,000	20,000
GRC	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$GSPA_2$	0.150	0.136	0.152	0.084	0.164	0.106	0.158
$GSPA_3$	0.114	0.128	0.120	0.116	0.118	0.128	0.118

Note: rejection frequencies for $n = 30$, $m = 30$ and $\lambda_i = 0$ for $i = 1, \ldots, n$, $\lambda_j = 0$ for $j = n + 1, \ldots, n + m$.

Case 3: now we try to simulate a situation as similar as the LFC as possible, by increasing the means of the $n - 1$ last models of the first class. In this case, the GRC performs better than the GSPA since its asymptotic distribution under the null is based in the LFC approach. These results are contained in table 3.

Case 4: table 4 contains the rejection frequencies under the alternative when the poor models in both families have means equal to zero. It shows that the GSPA presents important power improvements over the GRC. We can observe this extreme situation clearly when $T = 400$ where the GRC test almost never rejects the null (6.4%) while the GSPA test has a power over 86%. These results, together with results of Table 5, show how the advan-

Table 3: Rejection Frequencies for case 3.

T	50	100	200	400	800	5,000	20,000
GRC	0.064	0.076	0.068	0.090	0.102	0.100	0.078
$GSPA_2$	0.164	0.186	0.122	0.146	0.156	0.140	0.122
$GSPA_3$	0.152	0.124	0.106	0.100	0.144	0.122	0.120

Note: rejection frequencies for $m = 30$, $n = 30$ and $\lambda_1 = 0$ $\lambda_i = 2$ for $i = 2, \ldots n$; $\lambda_j = 0$ for $j = n+1, \ldots, n+m$.

tages of the SPA test proposed by Hansen over the RC test when irrelevant alternatives are included among the models to compare, are conserved in our generalized tests.

Table 4: Rejection Frequencies for case 4.

T	50	100	200	400	800	5,000	20,000
GRC	0.000	0.000	0.008	0.064	0.470	1.000	1.000
$GSPA_2$	0.256	0.290	0.550	0.862	0.996	1.000	1.000
$GSPA_3$	0.194	0.290	0.534	0.828	0.993	1.000	1.000

Note: rejection frequencies under the alternative for $n = 30$, $m = 30$ and $\lambda_i = 0$ for $i = 1, \ldots, n$ and $\lambda_{n+1} = -0.1$, $\lambda_j = 0$ for $j = n+2, \ldots, n+m$.

Case 5: in table 5 we have raised the means of the poor models of the first family to reproduce a situation more favorable to the GRC test. Although the power properties of the GRC improve over case 4, the GSPA has still a better performance, especially obvious with small sample sizes.

Table 5: Rejection Frequencies for case 5.

T	50	100	200	400	800	5,000	20,000
GRC	0.110	0.166	0.292	0.582	0.918	1.000	1.000
$GSPA_2$	0.246	0.290	0.444	0.774	0.984	1.000	1.000
$GSPA_3$	0.200	0.238	0.436	0.746	0.990	1.000	1.000

Note: rejection frequencies under the alternative for $n = 30$, $m = 30$ and $\lambda_1 = 0, \lambda_i = 2$ for $i = 2, \ldots, n$ and $\lambda_{n+1} = -0.1, \lambda_j = 0$ for $j = n + 2, \ldots, n + m$.

4 Real data results

In this section, we apply the generalized RC and SPA tests to several different forecasting experiments. The aim of these experiments is to assess whether some bivariate models outperform the univariate ones to forecast certain time series at one step distance. In addition, in one case, we employ also a model with a third series to check the possible improvement in the predictive power over the two-series models. This is related to the idea of Granger causality (Granger 1969)

We analyze for three countries (Spain, France and the USA), two sets of variables: (a) the first set comprises the Industrial Production Index (IPI) as variable of interest and the Industrial New Orders (NO) as an input to forecast the first one; (b) in the second one, we forecast the Consumer Price Index (CPI) (in the case of USA, the CPI without food and energy) using the Unemployment Rate (UNEM) as input. For USA, we also use as input the CPI of Energy (CPIE). The details of the data are in table 6.

Thus, we have six combination plus the additional trivariate model for the USA, that is, seven forecasting exercises. The details of the time series

Table 6: Origin of the data.

Serie	National survey	Dates	Freq.	Scope	Source
NO					
	Spain:IEP	2002:1-2011:6	M	B,C (CNAE09)	INE
	France:NOI	2000:1-2011:6	M	B-E (NACE)	Eurostat
	USA:M3	1992:10-2010:10	M	31-33 (NAICS)	Census Bureau
IPI					
	Spain:IPI	2002:1-2011:1	M	B,C,D (CNAE09)	INE
	France:IPI	2000:1-2011:6	M	B,C,D (NACE)	Eurostat
	USA:IPI	1992:10-2010:10	M	31-33,212 (NAICS)	Federal Reserve
UNEM					
	Spain:EPA	1977:Q1-2010:Q4	Q	ILO recommendations (16 to over years)	INE
	France:LFS	1990:Q1-2011:Q1	Q	ILO recommendations (15 to 74 years)	Eurostat
	USA:CPS	1957:1-2010:12	M	LFS definition (16 to over years)	BLS
CPI					
	Spain:IPC	1977:Q1-2010:Q4	Q	Households, all items	INE
	France:CPI	1990:Q1-2011:Q1	Q	Households, all items	INSEE
	USA:CPI	1957:1-2010:12	M	Urban. All items, Less food and energy	BLS
CPIE					
	USA:CPI	1957:1-2010:12	M	Energy items	BLS

Note: in the fourth column, M=monthly, Q=quarterly;The NO of USA are given in U.S. $, the remaining NO, IPI and CPI series are indexes and the UNEM series are rates.

can be found in table1. No series has been seasonally adjusted because this would render the experiment unrealistic. Instead, seasonality has been taken into account in the forecasting models. From the total length of the series, we have used the last third for out-of-sample forecasting. The model

parameters are estimated using a rolling window of constant length.

The series are submitted to some transformations before fitting the models: (a) $x_{0t} = \nabla\nabla_{12} \log \text{IPI}$, $x_{1t} = \nabla\nabla_{12} \log \text{NO}$; (b) $x_{0t} = \nabla\nabla \log \text{CPIC}$, $x_{1t} = \nabla \text{UNEM}$, where ∇ and ∇_{12} are the regular difference and (monthly) seasonal difference operators. We considered that the seasonality of the CPIC was not so significant as in the case of the industrial production.

The univariate forecasts for x_{0t} will be obtained by autoregressive models seasonal or nonseasonal and with or without intercept. Thus, for monthly series, $\hat{x}_{0t} = \sum_k \beta_{0,k} x_{0,t-k}$ or $\hat{x}_{0t} = \sum_k \beta_{0,k} x_{0,t-k} + \gamma$, where the lags included are either $\{k = r + 12s : r \leq p, s \leq q\}$, with $p = 0, \ldots, 6, q = 0, 1$ or just $\{k = 0, \ldots, 6\}$ depending on whether the model is seasonal or not.

The bivariate forecasts are $\hat{x}_{0t} = \sum_k \beta_{0,k} x_{0,t-k} + \sum_\ell \beta_{1,\ell} x_{1,t-\ell}$ or $\hat{x}_{0t} = \sum_k \beta_{0,k} x_{0,t-k} + \sum_\ell \beta_{1,\ell} x_{1,t-\ell} + \gamma$. The lags of x_{0t} and x_{1t} are chosen independently among the same set that the univariate models.

Finally we will define the class of three-variable models,

$$\begin{aligned} \hat{x}_{0t} &= \sum_k \beta_{0,k} x_{0,t-k} + \sum_\ell \beta_{1,\ell} x_{1,t-\ell} + \sum_m \beta_{2,m} x_{2,t-m} \\ \hat{x}_{0t} &= \sum_k \beta_{0,k} x_{0,t-k} + \sum_\ell \beta_{1,\ell} x_{1,t-\ell} + \sum_m \beta_{2,m} x_{2,t-m} + \gamma. \end{aligned}$$

Depending on the different frequencies of the series and the presence of seasonality (or the lack thereof), we obtain different sets of models for each experiment. Seasonal models are used for IPI/NO but not for CPI/UNEM. We summarize this in table 7.

Once we have obtained the forecasts, we calculate the forecasting errors $\hat{\varepsilon}_t = x_{0t} - \hat{x}_{0t}$ and apply two loss functions: absolute $L_t = |\hat{\varepsilon}_t|$ and quadratic $L_t = \hat{\varepsilon}_t^2$.

We obtain different forecasting errors for each of the models considered.

Table 7: Length of series and number of models.

Series	Nationality	Length	Univariate Models	Multivariate Models
IPI/NO				
	Spain	109	28	392
	France	138	28	392
	USA	224	28	392
CPI/UNEM				
	Spain	136	14	98
	France	85	14	98
	USA	648	14	98
CPI/UNEM/CPIE				
	USA	648	98	686

Following the notation of previous sections, we can index them as $\hat{\varepsilon}_{it}$ with $i \in I$ and $\hat{\varepsilon}_{jt}$ with $j \in I$ and accordingly for the loss functions and their means. Then we will test the null hypothesis $H_0 : \min_i \lambda_i \leq \min_j \lambda_j$.

In the six two-series forecasting exercises, I includes the indexes of the univariate models and J the bivariate ones. In the one three-series exercise, I includes the indexes of the bivariate models and J the trivariate ones. Since we consider two loss functions and seven forecasting exercises, we have fourteen different nulls.

We will try four different tests: (i) GRC with p-values obtained with Monte Carlo; (ii) GRC with p-values obtained with Bootstrap; (iii) GSPA_2 and (iv) GSPA_3. For the GSPA, we perform the bootstrap on the Δ_{ij} instead of the differences $\bar{d}_{ij}$ to reproduce better the asymptotic distribution.

We have used 1,000 simulations in the Monte Carlo and 1,000 resamplings in the Bootstrap. We present the p-values in tables 8,9,10 and 11. The results confirm the fact that the GSPA test, by getting the p-values

with the sample-dependent null is more powerful than the GRC. On the other hand, even the GSPA does not find evidence of better performance of the multivariate models in the cases of France and Spain, but only in the case of USA. Also, no gain is found in using the Energy CPI against the bivariate models that use the core CPI and Unemployment.

Table 8: P-values of GRC with Monte Carlo.

Series	Nationality	Absolute Loss	Quadratic Loss
		p-value	p-value
IPI/NO			
	Spain	0.986	1.00
	France	0.943	0.979
	USA	0.585	0.498
CPI/UNEM			
	Spain	1.000	1.000
	France	0.968	0.974
	USA	0.555	0.591
CPI/UNEM/CPIE			
	USA	1.000	0.998

We will show the multivariate models for the cases when GSPA test finds evidence of better performance than the univariate ones. For the case of the USA UNEM and CPI series, we get the model $x_{0t} = -0.510x_{0,t-1} - 0.549x_{0,t-2} - 0.562x_{0,t-3} - 0.494x_{0,t-4} - 0.236x_{0,t-5} - 0.122x_{0,t-6} + 0.016x_{1,t-1} + 0.011x_{1,t-2} - 0.218x_{1,t-3} - 0.014x_{1,t-4} - 0.276x_{1,t-5} - 0.469x_{1,t-6}$, that is the best no matter which loss function is chosen. The coefficients of the lags of UNEM are negative or very small, confirming the findings of previous research (Stock and Watson 1999) and (Clark and McCracken 2001).

For IPI and NO from USA, we get the model $x_{0t} = -0.514x_{0,t-12} + 0.059x_{1,t-1} + 0.070x_{1,t-2} + 0.046x_{1,t-3} + 0.026x_{1,t-4} + 0.039x_{0,t-5} + 0.050x_{0,t-6}$.

In this case, all the coefficients of the lags of NO are positive, which is coherent with the meaning of the variables. An increase in the new orders, produces an increase in production. Since different branches of activity have

Table 9: P-values of GRC with Bootstrap.

Series	Nationality	Absolute Loss	Quadratic Loss
		p-value	p-value
IPI/NO			
	Spain	1.000	1.000
	France	0.998	1.000
	USA	0.574	0.486
CPI/UNEM			
	Spain	0.998	1.000
	France	0.974	0.966
	USA	0.543	0.604
CPI/UNEM/CPIE			
	USA	1.000	0.999

Table 10: P-values of $GSPA_2$ (with Bootstrap).

Series	Nationality	Absolute Loss	Quadratic Loss
		p-value	p-value
IPI/NO			
	Spain	0.568	0.600
	France	0.284	0.272
	USA	0.046	0.007
CPI/UNEM			
	Spain	0.666	0.582
	France	0.617	0.712
	USA	0.021	0.009
CPI/UNEM/CPIE			
	USA	0.888	0.824

different manufacturing times, the effect is spread along the different lags.

5 Conclusions

We have generalized the Reality Check and the Superior Predictive Ability tests to the case when two classes of models of arbitrary size are to be compared. Simulation suggests that the loss of power due to the Least Favorable Configuration makes the GRC test less useful than the GSPA.

These tests can be applied to test for causality without identifying a specific model. We test the null that no model using a certain input beats the best model among the class that do not use that input. With this approach, we see that the industrial new orders and the unemployment rate have predictive power for the industrial production and the inflation respectively, using data from the USA. On the other hand, we fail to detect predictive power for the series of France and Spain.

Table 11: P-values of $GSPA_3$ (with Bootstrap).

Series	Nationality	Absolute Loss	Quadratic Loss
		p-value	p-value
IPI/NO			
	Spain	0.554	0.654
	France	0.392	0.236
	USA	0.050	0.006
CPI/UNEM			
	Spain	0.728	0.670
	France	0.668	0.702
	USA	0.014	0.012
CPI/UNEM/CPIE			
	USA	0.861	0.862

References

[1] Andrews, D. W. K. (2000), "Inconsistency of the Bootstrap when a Parameter is on the Boundary of the Parameter Space," *Econometrica*, 68, 399-405.

[2] Billingsley, P. (1968), *Probability and Measure*. John Wiley and Sons, New York.

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

[4] – (2005), "Evaluating Direct Multistep Forecasts," *Econometric Reviews* 24, 369-404 .

[5] – (2011), "Reality Checks and Comparisons of Nested Predictive Models," *Journal of Business and Economics Statistics* (ahead of print) doi:10.1198/jbes.2011.10278 .

[6] Clark T. E. and West K. D. (2007), "Approximately Normal Tests for Equal Predictive Accuracy in Nested Models," *Journal of Econometrics* 138, 291-311.

[7] Diebold F. and Mariano R. (1995), "Comparing Predictive Accuracy," *Journal of Business and Economics Statistics* 13, 252-263.

[8] Giacomini R. and White H. (2006), "Tests of Conditional Predictive Ability," *Econometrica* 74(6), 1545-1578.

[9] Granger C. W. J. (1969), "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods," *Econometrica* 37, 424-438.

[10] Hannan E. J. and Deistler M. (1988), "The Statistical Theory of Linear Systems," John Wiley and Sons, New York.

[11] Hansen, P. R. (2003), "Asymptotic Tests of Composite Hypotheses," unpublished report.

[12] Hansen, P. R. (2005), "A Test for Superior Predictive Ability," *Journal of Business and Economic Statistics* 23, 365-380.

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

[14] Stock, J.H., Watson, M.W. (1999), "Forecasting Inflation," *Journal of Monetary Economics* 44, 293–335.

[15] West, K. D. (1996), "Asymptotic Inference about Predictive Ability," *Econometrica* 64, 1067-1084.

[16] White, H. (2000), "A Reality Check for Data Snooping," *Econometrica* 68, 1098–1126.

A Appendix: proofs

Proof of proposition 1. We will write, for ease of notation, the power function as $\pi(\theta) = P[z_\theta \geq a]$. Let us see three preliminary results. First, if the inequalities $v_i \geq u_i$ and $v_j \leq u_j$ hold for all $i \in I$ and $j \in J$, then

$$\pi(v) \geq \pi(u). \tag{A.1}$$

On the other hand, for any constant $c \in \mathbb{R}$, we get

$$\pi(u_1 + c, \ldots, u_{n+m} + c) = \pi(u_1, \ldots, u_{n+m}). \tag{A.2}$$

The third preliminary result is the following. Let $\xi_\ell = \theta_\ell + \eta_\ell$. For any $k \in I$, since $\min_i \xi_i - \min_j \xi_j \leq \xi_k - \min_j \xi_j$,

$$\pi(u) \leq P\big[\xi_k - \min_j \xi_j \geq a\big]. \tag{A.3}$$

Now from (A.1) and (A.2), we obtain, $\pi(u) = \pi(u_1 - u_{(1)}, \ldots, u_{n+m} - u_{(1)}) \leq \pi(u_1 - u_{(1)}, \ldots, u_n - u_{(1)}, 0, \ldots, 0)$, where $u_{(1)} = \min_{i \in I} u_i$. Now, if this minimum is attained at k, then using (A.3) we get $\pi(u_1 - u_{(1)}, \ldots, u_n - u_{(1)}, 0, \ldots, 0) \leq P[x_k \geq a] \leq \max_{i=1,\ldots,n} P[x_i \geq a]$. Thus, we have proved that the right hand side of (2) is an upper bound to $\{P[z_\theta \geq a]\}_{\theta \in \Theta_0}$.

On the other hand, if we define for any $u \in \mathbb{R}$, $\theta_u = (u_1, \ldots, u_n, 0, \ldots, 0)$, where $u_k = 0$ and $\forall i \neq k, u_i = u$ then $\theta_u \in \Theta_0$ and $\pi(\theta_u) = P[\min\{\eta_k\} \cap \{\eta_i + u : i \neq k\} - \min_j \eta_j \geq a]$, where $\mathbb{E}\eta_\ell = 0$. If χ_u is the indicator function of $\min\{\eta_k\} \cap \{\eta_i + u : i \neq k\} - \min_j \eta_j \geq a$, we can prove that χ_u converges pointwise to the indicator function of $\eta_k - \min_j \eta_j \geq a$ when $u \to \infty$. Then, by the Dominated Convergence Theorem,

$$\lim_{u \to \infty} \pi(\theta_u) = \lim_{u \to \infty} \int \chi_u dP = P[x_k \geq a]. \tag{A.4}$$

Consequently, $\max_i P[x_i \leq a]$ is in fact a supremum. □

Proof of proposition 3. Let us prove (i). We will see that almost surely, for large T,

$$T^{\mathrm{GSPA}} = \max\Big\{ \min_{i \in I^*} \max_{j \in J^*} \Delta_{ij}, 0 \Big\} \tag{A.5}$$

where $I^* = \{i : 1 \leq i \leq n, \lambda_i = \lambda_0\}$, $J^* = \{j : n+1 \leq j \leq n+m, \lambda_j = \lambda_0\}$ and $\lambda_0 = \min\{\lambda_\ell : \ell = 1, \ldots, n+m\}$.

For this, we will prove that the minimax cannot be attained at any pair $(i, j) \notin I^* \times J^*$. Let us assume first that $i \notin I^*$. Then, for any j,

$$\Delta_{ij} = T^{1/2} \frac{\lambda_i - \lambda_j}{\omega_{ij}} + O(\sqrt{2 \log\log T}) \tag{A.6}$$

consequently, $\max_j \Delta_{ij} \geq T^{1/2}\delta + O(\sqrt{2\log\log T})$, with $\delta > 0$, whereas for $i_0 \in I^*$, $\max_j \Delta_{i_0 j} = O(\sqrt{2\log\log T})$. Thus, almost surely for large T, the minimum is never attained at i.

Now, let $j \notin J^*$. From (A.6), we get $\Delta_{ij_0} - \Delta_{ij} = T^{1/2}(\lambda_j - \lambda_{j_0}) + O(\sqrt{2\log\log T}) \overset{a.s}{\to} +\infty$, so for large T, the maximum cannot be attained at j. Hence we get (i), since (A.5) implies $T^{\mathrm{GSPA}} = \varphi(T^{1/2}\bar{L}^0, \hat{\omega})$, with $\bar{L}^0_\ell = \bar{L}_\ell$ for $\ell \in I^* \cap J^*$ and $\bar{L}^0_\ell = 0$ for $\ell \notin I^* \cap J^*$.

Now, if $\min_i \lambda_i < \min_j \lambda_j$, then $\min_i \max_j \Delta_{ij} \leq \max_j \Delta_{i_0 j}$, with $i_0 \in I^*$, but $\Delta_{i_0 j} \overset{a.s}{\to} -\infty$, so for large T, $\min_i \max_j \Delta_{ij} < 0$ and then $T^{GSPA} = 0$, so (ii) is also proved.

The case of the alternative hypothesis can be proved easily by the same arguments as (ii). □

Proof of proposition 4. First of all, we will prove that with probability one, for large T, the function $g(T)$ in (6) serves to distinguish the models in I^* –the good ones– from the others.

Let us assume that $i \notin I^*$. By the Law of the Iterated logarithm, we have that for any $\epsilon > 0$, there is some T_0 such that w.p.1, if $T \geq T_0$, then $\bar{L}_i \geq \lambda_i - \sqrt{\frac{2\log\log T}{T}}(1+\epsilon)$. On the other hand, for any $i^* \in I^*$, there is T_{i^*} such that w.p.1, if $T \geq T_{i^*}$, $\bar{L}_{i^*} \leq \lambda_{i^*} + \sqrt{\frac{2\log\log T}{T}}(1+\epsilon)$. Then, if $T_1 = \max\{T_0, \max_{i^* \in I^*} T_{i^*}\}$, for $T \geq T_1$, $\bar{L}_i > \bar{L}_{i^*}$ for any $i^* \in I^*$. Consequently, w.p.1, for large T, $\hat{i}^* \in I^*$.

Now, if $i \notin I^*$, then w.p.1, for large T, $\Delta_{i,i^*} \geq T^{1/2}(\lambda_i - \lambda_{i^*}) - O(\sqrt{2\log\log T})$ for all $i^* \in I^*$ and then, $\Delta_{i,\hat{i}^*} \geq g(T)$. On the other hand, for $i, i^* \in I^*$, w.p.1, for large T, $\Delta_{i,\hat{i}^*} < g(T)$. Consequently, w.p.1, for large T, for $i \in I^*$, the bootstrapped loss functions are re-centered around 0 whereas for $i \notin I^*$ are re-centered around the distance from the best model to the ith one.

As a consequence of above, we get that the boostrapped differences satisfy

$$T^{1/2}\bar{d}_{ij}^{*,c,k} = \begin{cases} T^{1/2}(\bar{L}_i^{*,k} - \bar{L}_j^{*,k}) + T^{1/2}(\bar{L}_{\hat{j}^*} - \bar{L}_{\hat{i}^*}) & i \notin I^*, j \notin J^* \\ T^{1/2}(\bar{L}_i^{*,k} - \bar{L}_j^{*,k}) + T^{1/2}(\bar{L}_{\hat{j}^*} - \bar{L}_i) & i \in I^*, j \notin J^* \\ T^{1/2}(\bar{L}_i^{*,k} - \bar{L}_j^{*,k}) + T^{1/2}(\bar{L}_j - \bar{L}_{\hat{i}^*}) & i \notin I^*, j \in J^* \\ T^{1/2}(\bar{L}_i^{*,k} - \bar{L}_j^{*,k}) + T^{1/2}(\bar{L}_j - \bar{L}_i) & i \in I^*, j \in J^*. \end{cases} \tag{A.7}$$

We can see that the second case diverges to $-\infty$ and the third to $+\infty$. This implies that asymptotically, the minimax is attained, as in proposition 3, at points in $I^* \times J^*$, but then, $T^{1/2}(\bar{L}_i^{*,k} - \bar{L}_j^{*,k}) + T^{1/2}(\bar{L}_j - \bar{L}_i) = T^{1/2}(\bar{d}_{ij}^{*,k} - \bar{d}_{ij})$ and we know that with probability one, this converges jointly in distribution to the asymptotic distribution of $T^{1/2}\bar{d}$. □