# Multiple forecast model evaluation[*]

Valentina Corradi[†] Walter Distaso[‡]

University of Warwick Imperial College, London

February 2010

[*] Prepared for the Oxford Handbook of Economic Forecasting, Oxford University Press. Distaso Gratefully acknowledges financial support from the ESRC, grant RES062230790.

[†] Department of Economics, University of Warwick, Coventry CV4 7AL, UK, email: v.corradi@warwick.ac.uk .

[‡] Imperial College Business School, Imperial College, London, South Kensington campus, London, SW7 2AZ, UK, email: w.distaso@imperial.ac.uk .

# 1 Introduction

In many situations arising in Economics it is of interest to compare the forecasting performance of a competing model to that of a reference or benchmark model. The latter could either be suggested by economic theory or simply be the reigning champion resulting from past competitions.

This is typically done by performing a test based on comparing the (out-of-sample) loss functions associated with the benchmark and the competing model. Under the null hypothesis that the two models have the same forecasting ability, the weighted distance between the two loss function is small, and asymptotically the test statistic converges to a well behaved, albeit possibly non-standard distribution (see, e.g., Diebold and Mariano, 1995, West, 1996, 2006, Clark and McCracken, 2001, Giacomini and White, 2006, Clark and West, 2007, McCracken, 2007). Under the alternative hypothesis, the competing model outperforms the benchmark, hence the test statistic diverges, giving rise to a consistent test.

As the number of competing models gets large though, this procedure may run into problems. This is because the null hypothesis is a composite hypothesis, formed by the intersection of several individual hypothesis. When we estimate and compare a large number of models using the same data set, there is a problem of data mining or data snooping. The problem of data snooping is that a model appears to be superior to the benchmark because of luck, not because of its intrinsic merit.

Indeed, if using the same data set, we compare the benchmark model against a large enough set of alternative models, eventually we would find some models outperforming the benchmark, even if none has superior predictive ability. This happens because of a sequential testing bias problem. If we test each of the hypothesis composing the null separately, at a nominal level $\alpha$, then the overall size increases whenever we test a new hypothesis, and eventually reaches one.

The data snooping problem arises because a composite hypothesis is treated as a sequence of individual independent hypotheses. For each single test there is a probability ($\alpha$) that the null hypothesis will be rejected even when it is true.

In the statistical literature, the event of rejecting the null hypothesis when it is true is often referred to as a false discovery. In the present context, we define a model for which the null hypothesis has been rejected despite being true as a "lucky" model.

Suppose that we are testing one hypothesis with a level of confidence equal to $\alpha$. $\alpha$ gives us the probability of rejecting the null hypothesis when it is true. It gives us, in other words, the probability of a false discovery. Suppose, instead, that we are testing $K$ hypotheses, based on data coming from the same sample. If the $K$ test statistics are independent,

$$\Pr(\text{at least 1 false discovery}) = 1 - (1-\alpha)^K = \alpha_K.$$

For example, when $K = 40$, $\alpha = 0.05$, $\alpha_K = 0.87$. Therefore, for a moderate value of $K$, the probability of having at least one false rejection is much higher than $\alpha$, and quickly tends towards 1.

The literature has proposed different, multiple testing procedures to solve this problem. The first solution is to use Bonferroni's one-step procedure, which consists of testing each single null hypothesis independently from the outcome of the others, fixing a level of confidence equal to $\alpha/K$. In this way, the multiple level of significance (also called the family wise error rate, FWER) is equal to $\alpha$. In the example given above, this would mean testing each single hypothesis at $\alpha = 0.00125$, a very conservative level.[1]

Another possibility, proposed by Holm (1979), consists of sorting the $p$-values $p_j$ obtained

[1] Although it is important to distinguish between exact and asymptotic control of the overall level of significance, this is beyond the scope of the Chapter. We refer the interested reader to Romano, Shaikh and Wolf (2008b).

for testing the hypothesis $\mathrm{H}_{0,j}$, $p_{(1)} \leq \ldots \leq p_{(K)}$ and labeling the corresponding hypotheses accordingly, $\mathrm{H}_{0,(1)}, \ldots, \mathrm{H}_{0,(K)}$. Then, $\mathrm{H}_{0,(k)}$ is rejected at level $\alpha$ if $p_{(j)} \leq \alpha/(K-j+1)$, for $j = 1, \ldots, k$. This method is called a *stepdown* method, because it starts with the lowest $p$-value. It controls for the FWER and improves on the Bonferroni's approach, but still produces conservative inference.

White (2000) introduces a formal approach, named reality check (RC), for testing the hypothesis that the best competing model does not outperform the benchmark. By jointly considering all competing models, this approach controls the FWER, and circumvents the data snooping problem. In fact, the reality check procedure ensures that the probability of rejecting the null when is false is smaller than or equal to $\alpha$.

Power improvements are obtainable at the cost of relaxing the controlled type I error. Suppose that one is concerned with Pr(at least $l$ false discoveries), with $l > 1$. Then, methods controlling for $l$-FWER have been developed, based on modified Bonferroni and Holm procedures (for a review of these modifications, see Romano, Shaikh and Wolf, 2008b).

A viable alternative is the control of the less conservative false discovery rate (FDR), defined as the expected value of the ratio between the false discoveries and the total number of rejections. The first successful attempt to control for FDR has been proposed by Benjamini and Hochberg (1995, 2000). Their approach has been subsequently refined in a series of papers (see Storey, 2002, 2003, Storey and Tibshirani, 2003, Storey, Taylor and Siegmund, 2004, Romano, Shaikh and Wolf, 2008a).

Methods for controlling the type I error have already found their way into financial and economic applications. Barras, Scaillet and Wermers (2005), McCracken and Sapp (2005), Bajgrowicz and Scaillet (2008) and Avramov, Barras and Kosowski (2009) control for FDR to evaluate, respectively, the performance of US mutual funds, competing structural models in predicting exchange rates, technical trading strategies and the forecasting power of economic variables for hedge fund returns. Sullivan, Timmermann and White (1999, 2001), Awartani and Corradi (2005) and Hansen and Lunde (2005), Wolf and Wunderly (2009) control for FWER in order to evaluate, respectively, different calendar or trading rules against the benchmark rule of holding cash, the forecasting performance of different GARCH models against the GARCH(1,1), and to construct a fund of truly outperforming hedge funds.

There may be situations where we do not have a benchmark model, but simply want to eliminate poor models and keep all models sharing the same predictive accuracy. This is accomplished by the Model Confidence Set (MCS) approach of Hansen, Lunde and Nason (2009). We suggest an alternative to the MCS approach which ensures that both the probability of eliminating a relevant model or failing to eliminate an irrelevant one approach zero.

The focus of this chapter is on recent development in the forecasting literature on how to simultaneously control both the overall error rate and the contribution of irrelevant models. In this sense, it begins where West's (2006) chapter ends. As a novel contribution, we derive a general class of superior predictive ability tests, which controls for FWER *and* the contribution of irrelevant models. This is accomplished by applying the same methodology currently used to construct confidence intervals for the validity of moment conditions defined by multiple weak inequalities (see, e.g., Chernozhukov, Hong and Tamer, 2007, Andrews and Jia, 2008, Bugni, 2008, Rosen, 2008, Andrews and Guggenberger, 2009, Andrews and Soares, 2010 and Romano and Shaikh, 2010).

The chapter is organized as follows. Section 2 defines the setup. Section 3 reviews the approaches that control for the conservative FWER. Section 4 considers a general class of tests characterized by multiple joint inequalities. Section 5 presents results allowing for control of the less conservative FDR. Finally, Section 6 considers the MCS approach and offers a simple alternative, which reduces the influence of irrelevant models in the initial set.

# 2 Setup

In this section we introduce the notation and the setup we shall use in the sequel. We consider a collection of $K+1$ models, where model 0 is treated as the benchmark or reference model and models $1,\ldots,K$ compose the set of competing models. Formally,

$$y_t = g_k\left(\boldsymbol{X}_{k,t}, \boldsymbol{\beta}_k\right) + u_{k,t},$$

where, in general, $\boldsymbol{X}_{k,t}$ contains lags of $y_t$ and of some other variables used for prediction. Models $i$ and $j$ are non-nested if $g_i \neq g_j$ and/or neither $\boldsymbol{X}_{i,t} \subseteq \boldsymbol{X}_{j,t}$ nor $\boldsymbol{X}_{i,t} \supseteq \boldsymbol{X}_{j,t}$, otherwise one is nested in the other. As $\boldsymbol{\beta}_k$ is unknown, we don't observe the prediction error $u_{k,t}$.

Following the common practice in out-of-sample prediction, we split the total sample made of $T$ observations in two segments $R$, $P$ with $R+P=T$. We use the first $R$ observations to estimate a candidate model, say model $k$, and construct the first $\tau-$step ahead prediction error. Then, we use $R+1$ observations to re-estimate the model and compute the second $\tau-$step ahead prediction error, and so on, until we have a sequence of $(P-\tau+1)$ $\tau-$step ahead prediction errors.[2] We define the estimated parameter vector at each step as

$$\widehat{\boldsymbol{\beta}}_{k,t} = \arg\max_{\boldsymbol{\beta}_k} \left\{ \frac{1}{t} \sum_{j=1}^{t} q_{k,j}\left(\boldsymbol{X}_{k,t}, \boldsymbol{\beta}_k\right) \right\} \text{ for } t \geq R,$$

where $q_{k,j}$ can be thought of as the quasi-likelihood function implied by model $k$. Analogously, we can define

$$\boldsymbol{\beta}_k^{\dagger} = \arg\max_{\boldsymbol{\beta}_k} \left\{ \frac{1}{t} \sum_{j=1}^{t} \mathrm{E}\left(q_{k,j}\left(\boldsymbol{X}_{k,t}, \boldsymbol{\beta}_k\right)\right) \right\} \text{ for } t \geq R.$$

[2] Here we use a recursive estimation scheme, where data up to time $t \geq R$ are used. West and McCracken (1998) also consider a rolling estimation scheme, in which a rolling windows of $R$ observations is used for estimation.

This setup allows a formal analysis of the effect of parameter estimation error on tests for predictive ability. For clarity of exposition, here we only consider the simple case of pairwise model comparisons (Diebold and Mariano, 1995). The hypotheses of interest can be stated as

$$\mathrm{H}_0 : \mathrm{E}\left(f\left(u_{0,t}\right) - f\left(u_{k,t}\right)\right) = 0$$

vs

$$\mathrm{H}_A : \mathrm{E}\left(f\left(u_{0,t}\right) - f\left(u_{k,t}\right)\right) \neq 0,$$

where $f$ denotes a generic loss function.[3] Hence, the null hypothesis implies equal predictive ability of models 0 (the benchmark) and $k$. The test statistic for the hypotheses above is given by

$$\sqrt{P}\widehat{m}_{k,P} = \frac{\sqrt{P}}{P-\tau+1}\sum_{t=R+\tau}^{T}\widehat{m}_{k,t} = \frac{\sqrt{P}}{P-\tau+1}\sum_{t=R+\tau}^{T}\left(f\left(\widehat{u}_{0,t}\right) - f\left(\widehat{u}_{k,t}\right)\right),$$

where the prediction error associated with model $k$ is defined as $\widehat{u}_{k,t+\tau} = y_{t+\tau} - g_k\left(\boldsymbol{X}_{k,t}, \widehat{\boldsymbol{\beta}}_{k,t}\right)$. Assuming that $f$ is twice differentiable in a neighborhood of $\boldsymbol{\beta}_0^{\dagger}$ and $\boldsymbol{\beta}_k^{\dagger}$, it follows that

$$\begin{aligned}\sqrt{P}\left(\widehat{m}_{k,P} - m_{k,P}\right) &= \mathrm{E}\left(\nabla_{\beta_0} f\left(u_{0,t}\right)\right)\frac{\sqrt{P}}{P-\tau+1}\sum_{t=R+\tau}^{T}\left(\widehat{\boldsymbol{\beta}}_{0,t} - \boldsymbol{\beta}_0^{\dagger}\right) \\ &\quad -\mathrm{E}\left(\nabla_{\beta_k} f\left(u_{k,t}\right)\right)\frac{\sqrt{P}}{P-\tau+1}\sum_{t=R+\tau}^{T}\left(\widehat{\boldsymbol{\beta}}_{k,t} - \boldsymbol{\beta}_k^{\dagger}\right) + o_p(1), \qquad (1)\end{aligned}$$

where, similarly to before, the unfeasible statistic is defined as

$$\sqrt{P}m_{k,P} = \frac{\sqrt{P}}{P-\tau+1}\sum_{t=R+\tau}^{T}m_{k,t} = \frac{\sqrt{P}}{P-\tau+1}\sum_{t=R+\tau}^{T}\left(f\left(u_{0,t}\right) - f\left(u_{k,t}\right)\right).$$

The limiting distribution of the right hand side of (1) has been derived by West (1996). Here it

[3] Here, we only consider two-sided alternatives. The same methodologies can also be applied to one-sided alternatives, which are a natural choice when comparing nested models.

suffices to notice that, as for all $t \geq R$, $\sup_{t \geq R} \left| \widehat{\beta}_{k,t} - \beta_k^{\dagger} \right| = O_p \left( R^{-1/2} \right)$ the contribution of parameter estimation error is negligible if either $P/R \rightarrow 0$, as $P, R \rightarrow \infty$, or/and $\mathrm{E} \left( \nabla_{\beta_0} f \left( u_{0,t} \right) \right) = \mathrm{E} \left( \nabla_{\beta_0} f \left( u_{k,t} \right) \right) = 0$. Otherwise, parameter estimation error matters and affects the asymptotic variance of $\sqrt{P} \widehat{m}_{k,P}$. Therefore, it has to be taken into account for inference purposes, when constructing variance estimators and/or bootstrap critical values (see, e.g., West, 1996, West and McCracken, 1998, Corradi and Swanson, 2006a, 2006b, 2007).

Note that $P/R \rightarrow 0$ is satisfied when the number of observations used for estimation grows at a faster rate than the number of observations used for out of sample prediction. On the other hand, because of the first order conditions, $\mathrm{E} \left( \nabla_{\beta_0} f \left( u_{0,t} \right) \right) = \mathrm{E} \left( \nabla_{\beta_0} f \left( u_{k,t} \right) \right) = 0$ when the same loss function is used for estimation and out of sample prediction, i.e. $q_0 = q_k = f$. A typical example is when we estimate parameter by Ordinary Least Squares (using a gaussian quasi-likelihood function), and choose a quadratic loss function for out of sample forecast evaluation.

Giacomini and White (2006) suggest an alternative approach, in which parameters are estimated using a fixed rolling window of observations, thus preserving the effect of estimation uncertainty on the relative models performance. To keep notation simpler, in the following Sections we assume that parameter estimation error is asymptotically negligible.

# 3 Methods for controlling the FWER

## 3.1 Reality Check

As explained in the Introduction, in Economics a frequently used approach controlling the FWER is the Reality Check of White (2000). Formally, the hypotheses of interest are:

$$\mathrm{H}_0 : \max_{k=1,\ldots,K} m_k \leq 0 \tag{2}$$

vs

$$\mathrm{H}_A : \max_{k=1,\ldots,K} m_k > 0.$$

Note that the composite hypothesis in (2) is the intersection of the $K$ null hypotheses $\mathrm{H}_{0,k} : m_k \leq 0$. Under the alternative, there is at least one competing model outperforming the benchmark. We have the following result.

**Proposition 1 (from Proposition 2.2 in White, 2000)** *Let* $\boldsymbol{m}_P = (m_{1,P}, \ldots, m_{K,P})^\top$, $\widehat{\boldsymbol{m}}_P = (\widehat{m}_{1,P}, \ldots, \widehat{m}_{K,P})^\top$, *and assume that, as* $P \to \infty$, $\sqrt{P}\left(\widehat{\boldsymbol{m}}_P - \boldsymbol{m}_P\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}\right)$, *with* $\boldsymbol{V}$ *positive semi-definite. Then,*

$$\max_{k=1,\ldots,K} \left\{ \sqrt{P}\left(\widehat{m}_{k,P} - m_k\right) \right\} \xrightarrow{d} \max_{k=1,\ldots,K} Z_k, \tag{3}$$

*where* $\boldsymbol{Z} = (\ Z_1, \ldots,\ Z_k)^\top$ *is distributed as* $\mathrm{N}\left(\mathbf{0}, \boldsymbol{V}\right)$ *and* $\boldsymbol{V}$ *has typical element*

$$v_{j,k} = \lim_{P\to\infty} \mathrm{E}\left( \frac{1}{\sqrt{P}} \sum_{t=R+\tau}^{T} \left(\widehat{m}_{j,t} - m_j\right) \frac{1}{\sqrt{P}} \sum_{t=R+\tau}^{T} \left(\widehat{m}_{k,t} - m_k\right) \right).$$

Notice that, because $\boldsymbol{V}$ has to be positive semi-definite, at least one competitor has to be non-nested within and non-nesting the benchmark model.[4]

Proposition 1 establishes the limiting distribution of the test statistic using the least favorable distribution to the alternative. This happens when $m_k = 0$ for all $k$, which means that all models share the same predictive accuracy. White defines as RC any methodology able to deliver asymptotically valid $p$-values for the limiting distribution in (3).

Because the maximum of a Gaussian process is not a Gaussian process, the construction of $p$-values for the limiting distribution in (3) is not straightforward. White proposes two alternatives: (i) a simulation-based approach and (ii) a bootstrap-based approach. The first approach starts from a consistent estimator of $\boldsymbol{V}$, say $\widehat{\boldsymbol{V}}$. Then, for each simulation $s = 1, \ldots, S$, we construct

$$\widehat{\boldsymbol{d}}_P^{(s)} = \begin{pmatrix} \widehat{d}_{1,P}^{(s)} \\ \vdots \\ \widehat{d}_{K,P}^{(s)} \end{pmatrix} = \begin{pmatrix} \widehat{v}_{1,1} & \cdots & \widehat{v}_{1,K} \\ \vdots & \ddots & \vdots \\ \widehat{v}_{K,1} & \cdots & \widehat{v}_{K,K} \end{pmatrix}^{1/2} \begin{pmatrix} \eta_1^{(s)} \\ \vdots \\ \eta_K^{(s)} \end{pmatrix},$$

where $\left(\eta_1^{(s)}, \ldots, \eta_K^{(s)}\right)^\top$ is drawn from a $\mathrm{N}(\mathbf{0}, \mathbf{I}_K)$. Then, we compute $\max_{k=1,\ldots,K} \left|\widehat{\boldsymbol{d}}_P^{(s)}\right|$, and the $(1-\alpha)-$percentile of its empirical distribution, say $c_{\alpha,P,S}$.

The simulation based approach requires the estimation of $\boldsymbol{V}$. If $K$ is large, and forecasting errors exhibit a high degree of time dependence, estimators of the long-run variance become imprecise and ill-conditioned, making inference unreliable, especially in small samples. This problem can be overcome using bootstrap critical values.

White (2000) outlines the construction of bootstrap critical values when the contribution of parameter estimation error to the asymptotic covariance matrix vanishes. In this case, we

[4] To the best of our knowledge, there are no testing procedures for predictive evaluation of multiple nested models. On the other hand, there are several tests for pairwise comparison of nested models, see e.g. Clark and McCracken (2001), McCracken (2007), Clark and West (2007).

resample blocks of $\widehat{m}_{k,t}$ and, for each bootstrap replication $b = 1, \ldots, B$, calculate

$$\widehat{m}^{*(b)}_{k,P} = \frac{1}{\sqrt{P}} \sum_{t=R+\tau}^{T} \widehat{m}^{*(b)}_{k,t}.$$

Finally, we compute the bootstrap statistic $\max_{k=1,\ldots,K} \left| \widehat{m}^{*(b)}_{k,P} - \widehat{m}_{k,P} \right|$ and the $(1-\alpha)$-percentile of its empirical distribution, say $c_{\alpha,P,B}$. White has shown that, as $P, S \to \infty$, $c_{\alpha,P,S}$ and, as $P, B \to \infty$, $c_{\alpha,P,B}$ converge to the $(1-\alpha)-$percentile of the limiting distribution on the right hand side of (3). The rule is to reject $\mathrm{H}_0$ if $\max_{k=1,\ldots,K} \sqrt{P}\widehat{m}_{k,P}$ is larger than $c_{\alpha,P,S}$ or $c_{\alpha,P,B}$ (depending on the chosen method to calculate critical values), and do not reject otherwise.

Because the right hand side of (3) represents the limiting distribution for the least favorable case under the null, the rules above ensure that the probability of false rejection is at most $\alpha$. In particular, if all competitors are as good as the benchmark, $m_k = 0$ for all $k$, then the probability of falsely rejecting the null is asymptotically $\alpha$. However, if some model is worse than the benchmark, i.e. $m_k < 0$ for some $k$, then the probability of falsely rejecting the null is smaller than $\alpha$.

Suppose the we consider an additional model, which is worse than both the benchmark and the best competing model, i.e. $m_{K+1} < 0$ and $m_{K+1} < \max_{k=1,\ldots,K} m_k$. The inclusion of this additional model will not change the asymptotic behavior of the statistic, because trivially $\max_{k=1,\ldots,K+1} \sqrt{P}\widehat{m}_{k,P} = \max_{k=1,\ldots,K} \sqrt{P}\widehat{m}_{k,P}$. However, the percentiles of the limiting distribution of $\max_{k=1,\ldots,K+1} \left\{ \sqrt{P} \left( \widehat{m}_{k,P} - m_k \right) \right\}$ will be larger than those of the limiting distribution of $\max_{k=1,\ldots,K} \left\{ \sqrt{P} \left( \widehat{m}_{k,P} - m_k \right) \right\}$.

Hence, the probability of rejecting the null decreases as a direct consequence of the introduction of a poor model. Stretching this argument, RC $p$-values may become larger and larger because of the introduction of more and more irrelevant models. Indeed, the power of the test can be pushed to zero through the inclusion of a large number of poor models. In this sense, RC may become quite conservative.

## 3.2 Hansen's SPA Test

Hansen (2005) suggests a variant of RC, the Superior Predictive Ability test, which is much less sensitive to the inclusion of poor models and thus less conservative. The SPA statistic is given by

$$T_P = \max\left\{0, \max_{k=1,\ldots,K} \frac{\widehat{m}_{k,P}}{\sqrt{\widehat{v}_{k,k}}}\right\}, \tag{4}$$

where $\widehat{v}_{k,k}$ is a consistent estimator of $\lim_{P\to\infty}\text{var}\left(\sqrt{P}\widehat{m}_{k,P}\right)$. Hansen considers the case when parameter estimation error vanishes as $P \to \infty$. Notice that the SPA statistic requires that, for all $k = 1,\ldots,K$, $\widehat{v}_{k,k} \xrightarrow{p} v_{k,k} > 0$, which in turn requires that all competing models are neither nested within nor nesting the benchmark. This is in contrast with RC, which only requires that $v_{k,k} > 0$ for at least one $k$. On the other hand, the SPA approach allows for the case when some competing models are nested within or nesting other competing models. In particular, the matrix $\boldsymbol{V}$ does not need to be positive definite, but the elements of its main diagonal must be strictly positive.

Inspection of (4) reveals that only nonnegative moment conditions contribute to $T_P$ and to its limiting distribution, and that the case of $\widehat{m}_{k,P} < 0$ for all $k$ is considered sufficient evidence in favor of the null. In order to construct a bootstrap analog, Hansen uses the law of the iterated logarithm, according to which

$$\Pr\left(\limsup_{P\to\infty} \sqrt{P}\left(\frac{m_{k,P} - m_k}{\sqrt{v_{k,k}}}\right) = \sqrt{2\ln\ln P}\right) = 1,$$

$$\Pr\left(\liminf_{P\to\infty} \sqrt{P}\left(\frac{m_{k,P} - m_k}{\sqrt{v_{k,k}}}\right) = -\sqrt{2\ln\ln P}\right) = 1.$$

Because for all $k$, $\widehat{v}_{k,k} - v_{k,k} = o_p(1)$ and $\widehat{m}_{k,P} - m_{k,P} = o_p(1)$, it follows that

$$\lim_{P\to\infty} \Pr\left(\sqrt{P}\left(\frac{\widehat{m}_{k,P} - m_k}{\sqrt{\widehat{v}_{k,k}}}\right) = \sqrt{2\ln\ln P}\right) = 1,$$

$$\lim_{P\to\infty} \Pr\left(\sqrt{P}\left(\frac{\widehat{m}_{k,P} - m_k}{\sqrt{\widehat{v}_{k,k}}}\right) = -\sqrt{2\ln\ln P}\right) = 1,$$

and so one can discard the contribution of $\widehat{m}_{k,P}$ when $\widehat{m}_{k,P} \leq -\sqrt{\widehat{v}_{k,k}}\sqrt{2\ln\ln P/P}$. Hence, the bootstrap counterpart to $T_P$ is given by

$$T_P^{*(b)} = \max\left\{0, \max_{k=1,\ldots,K}\left\{\frac{\sqrt{P}\left(\widehat{m}_{k,P}^{*(b)} - \widehat{m}_{k,P}1_{\left\{\widehat{m}_{k,P} > -\widehat{v}_{k,k}\sqrt{2\ln\ln P/P}\right\}}\right)}{\sqrt{\widehat{v}_{k,k}}}\right\}\right\}. \tag{5}$$

Finally, $p$-values for the SPA statistic are given by $1/B\sum_{b=1}^{B} 1_{\left\{T_P^{*(b)} > T_P\right\}}$. The logic underlying the construction of the SPA $p$-values is the following. When $\widehat{m}_{k,P} < -\sqrt{\widehat{v}_{k,k}}\sqrt{2\ln\ln P/P}$, implying that $\widehat{m}_{k,P}$ does not contribute to $T_P$, the corresponding bootstrap moment condition is not recentered. Therefore, also $\widehat{m}_{k,P}^{*(b)}$, which is negative, does not contribute to the bootstrap limiting distribution. The fact that “very” negative moment conditions do not contribute to the $p$-values ensures that SPA $p$-values are less conservative than RC $p$-values. Nevertheless, it cannot be established that the SPA test is uniformly more powerful than the RC test.

## 3.3 Stepwise Multiple testing

A further improvement on the RC procedure is provided by the stepwise multiple testing approach (stepM) of Romano and Wolf (2005). The idea behind stepM is that, by adopting a multi-step approach, power improvements are obtained with respect to the single step Reality Check. Like Holm’s, stepM is a *stepdown* method. The algorithm for implementing the stepM methodology is given below:

1. Relabel the models in descending order of the test statistics $\widehat{m}_{k,P}$: model $r_1$ corresponds to the largest test statistic and model $r_K$ to the smallest.

2. Let the index $j$ and the number $R_j$ denote, respectively, the step of the procedure and the total number of rejections at the end of step $j$. Set $j = 1$ and $R_0 = 0$.

3. For $R_{j-1} + 1 \leq k \leq K$, if $0 \notin [\widehat{m}_{r_k,P} \pm \widehat{c}_j]$, reject the null hypothesis $\mathrm{H}_{0,r_k}$. A data dependent algorithm to calculate the critical values $\widehat{c}_j$ is explained below.

4. (a) If no (further) null hypotheses are rejected, stop.

   (b) Otherwise, let $j = j + 1$ and return to step 3.

The algorithm described above highlights the power improvements achievable by iterating over $j$, which are similar in spirit to those enjoyed by the Holm's methodology over the single step Bonferroni's method. Romano and Wolf (2005) also provide a consistent, bootstrap based, procedure to calculate the critical values, which works as follows:

1. Generate $B$ bootstrap data matrices.

2. From each bootstrapped data matrix, compute the individual test statistics $\widehat{\boldsymbol{m}}_P^{*(b)}$.

3. (a) For $b = 1, \ldots, B$, compute $\max_j^{*(b)} = \max_{R_{j-1}+1 \leq k \leq K} \left\{ \widehat{m}_{r_k,P}^{*(b)} - \widehat{m}_{r_k,P} \right\}$.

   (b) Compute $\widehat{c}_j$ as the $(1-\alpha/2)$-empirical quantile of the $B$ values $\max_j^{*(b)}$, $b = 1, \ldots, B$.

This method is readily generalized to control for $l$-FWER.

# 4 A Class of Tests for Superior Predictive Ability

In this Section we introduce a novel approach to testing for superior predictive ability. In particular, we outline how the methodology currently used to construct confidence intervals for the validity of moment conditions defined by weak inequalities can be applied to the construction of tests for superior predictive ability. The null hypothesis $\mathrm{H}_0 : \max_{k=1,\dots,K} m_k \leq 0$ of Section 3 can be rewritten as

$$\mathrm{H}_0 : m_k \leq 0 \text{ for } k = 1, \dots, K \tag{6}$$

and the corresponding alternative $\mathrm{H}_A : \max_{k=1,\dots,K} m_k > 0$ can be reformulated as follows

$$\mathrm{H}_A : m_k > 0 \text{ for at least one } k \in \{1, \dots, K\}.$$

Confidence sets for the null in (6) have recently been investigated by many researchers (see, e.g., Chernozhukov, Hong and Tamer, 2007, Andrews and Jia, 2008, Bugni, 2008, Rosen, 2008, Andrews and Guggenberger, 2009 and Andrews and Soares, 2010). The literature has suggested two different approaches for testing the null in (6).

The first uses a Quasi Likelihood Ratio (QLR) statistic (Andrews and Jia, 2008, Rosen, 2008 and Andrews and Soares, 2010), defined as

$$S_P = \inf_{\boldsymbol{\varrho} \in \mathbb{R}^K_-} \left( \sqrt{P} \widehat{\boldsymbol{m}}_P - \boldsymbol{\varrho} \right)^\top \widehat{\boldsymbol{V}}_P^{-1} \left( \sqrt{P} \widehat{\boldsymbol{m}}_P - \boldsymbol{\varrho} \right). \tag{7}$$

The second statistic (Chernozhukov, Hong and Tamer, 2007 and Bugni, 2008) is defined as

$$S_P^+ = \sum_{k=1}^{K} \left( \sqrt{P} \widehat{m}_{k,P}^+ / \sqrt{\widehat{v}_{k,k}} \right)^2, \text{ where } x^+ = \begin{cases} x, & \text{if } x > 0 \\ 0, & \text{if } x \leq 0 \end{cases}. \tag{8}$$

Notice that the construction of the QLR statistic requires $\widehat{\boldsymbol{V}}_P$ to be invertible, which is a stronger condition than $\widehat{v}_{k,k} > 0$ for all $k$. When $\widehat{m}_{k,P} \leq 0$ for all $k$, both $S_P$ and $S_P^+$ are equal to zero almost surely. In fact, only those models yielding $\widehat{m}_{k,P} > 0$ contribute to the statistic.

Moreover, under the null, $\widehat{m}_{k,P}$ has to tend to zero as $P \to \infty$, to ensure that $\sqrt{P}\widehat{m}_{k,P}$ is bounded in probability and that both $S_P$, $S_P^+$ have well definite limiting distributions. Under the alternative, as $P \to \infty$, at least one $\widehat{m}_{k,P}$ does not tend to zero and both statistics diverge. Both statistics introduced above are developed in the spirit of those for testing the null of a parameter being on the boundary (see, e.g., Andrews, 1999 and Beg, Silvapulle and Silvapulle, 2001). In fact, the element of the null least favorable to the alternative, i.e. $\widehat{m}_{k,P} = 0$ for each $k$ is treated as a boundary, and only positive values of the statistics contribute to the limiting distribution. Notice that the statistics (7) and (8) are a function of $\widehat{\boldsymbol{m}}_P$ and of $\widehat{\boldsymbol{V}}_P$, and therefore can be written as

$$Z_P = S\left(\widehat{\boldsymbol{m}}_P, \widehat{\boldsymbol{V}}_P\right). \tag{9}$$

Andrews and Guggenberger (2009) provide a set of sufficient conditions on $S$ (Assumptions 1-4), ensuring that, under the null, as $P \to \infty$,

$$S\left(\widehat{\boldsymbol{m}}_P, \widehat{\boldsymbol{V}}_P\right) \xrightarrow{d} S\left(\boldsymbol{\Omega}^{1/2}\boldsymbol{Z} + \boldsymbol{h}, \boldsymbol{\Omega}\right),$$

where $\boldsymbol{Z} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_K)$, $\boldsymbol{\Omega} = \boldsymbol{D}^{-1/2}\boldsymbol{V}\boldsymbol{D}^{-1/2}$, $\boldsymbol{D} = \operatorname{diag}(\boldsymbol{V})$, $\boldsymbol{V} = \operatorname{p\,lim}_{P\to\infty} \widehat{\boldsymbol{V}}_P$ and $\boldsymbol{h} = (h_1, \ldots, h_K)^\top$ is a vector measuring the slackness of the moment conditions, i.e. $h_k = \lim_{P\to\infty} \sqrt{P}\mathrm{E}\left(m_{k,P}/\sqrt{v_{k,k}}\right)$. In particular, $S_P$ and $S_P^+$ satisfy the Assumptions on $S$ of Andrews and Guggenberger (2009), and, under $\mathrm{H}_0$, as $P \to \infty$,

$$S_P \xrightarrow{d} \inf_{\boldsymbol{\varrho} \in \mathbb{R}_-^K} \left(\boldsymbol{\Omega}^{1/2}\boldsymbol{Z} + \boldsymbol{h} - \boldsymbol{\varrho}\right)^\top \boldsymbol{\Omega}^{-1} \left(\boldsymbol{\Omega}^{1/2}\boldsymbol{Z} + \boldsymbol{h} - \boldsymbol{\varrho}\right)$$

and

$$S_P^+ \xrightarrow{d} \sum_{i=1}^{K} \left( \left( \sum_{j=1}^{K} \omega_{i,j}\, Z_j + h_i \right)^+ \right)^2,$$

where $\omega_{i,j}$ is the $[i,j]-$th element of $\boldsymbol{\Omega}$.

The main problem in the computation of the critical values or confidence sets, is that the vector $\boldsymbol{h}$ cannot be consistently estimated. Intuitively, except for the least favorable case under the null, $\mathrm{E}\left(m_{k,P}/\sqrt{v_{k,k}}\right) < 0$ and so $\lim_{P\to\infty}\sqrt{P}\mathrm{E}\left(m_{k,P}/\sqrt{v_{k,k}}\right)$ tends to minus infinity, and cannot be consistently estimated. Andrews and Soares (2010) suggest different, asymptotically valid, rules for approximating the vector $\boldsymbol{h}$. These include for example,

(i) $h_j = 0$ if $\xi_j \geq -1$, and $h_j = -\infty$ if $\xi_j < -1$, where $\xi_j = \sqrt{P/2\ln\ln P}\left(\widehat{m}_{j,P}/\sqrt{\widehat{v}_{j,j}}\right)$

(ii) $h_j = 0$ if $\xi_j \geq -1$, and $h_j = \xi_j$ if $\xi_j < -1$.

The intuition behind rule (i) is that if $\widehat{m}_{j,P} \leq -\sqrt{\widehat{v}_{j,j}}\sqrt{2\ln\ln P/P}$, then the $j-$moment condition is too slack, i.e. model $j$ is too poor to contribute to the limiting distribution.

In what follows, without loss of generality, we follow rule (i) and suppose that we select the first $\widetilde{K}$ competing models, with $\widetilde{K} \leq K$. Critical values can be computed either via a simulation-based approach or via a bootstrap-based approach. Bootstrap critical values for $S_P$ can be obtained by computing the percentiles of the empirical distribution of

$$\begin{aligned} & S_P^* \\ = & \inf_{\boldsymbol{\varrho}\in\mathbb{R}_-^{\widetilde{K}}}\left(\sqrt{P}\widehat{\boldsymbol{D}}^{*^{-1/2}}_{P,(\widetilde{K})}\left(\widehat{\boldsymbol{m}}^*_{P,(\widetilde{K})} - \widehat{\boldsymbol{m}}_{P,(\widetilde{K})}\right) - \boldsymbol{\varrho}\right)^{\top}\widehat{\boldsymbol{\Omega}}^{*^{-1}}_{P,(\widetilde{K})}\left(\sqrt{P}\widehat{\boldsymbol{D}}^{*^{-1/2}}_{P,(\widetilde{K})}\left(\widehat{\boldsymbol{m}}^*_{P,(\widetilde{K})} - \widehat{\boldsymbol{m}}_{P,(\widetilde{K})}\right) - \boldsymbol{\varrho}\right), \end{aligned}$$

where the subscript $(\widetilde{K})$ indicates that we are considering only the selected $\widetilde{K}$ moment conditions, and the starred quantities are bootstrap analogues of sample counterparts.

Alternatively, $\left(\widehat{\boldsymbol{m}}^*_{P,(\widetilde{K})} - \widehat{\boldsymbol{m}}_{P,(\widetilde{K})}\right)$, $\widehat{\boldsymbol{D}}^{*^{-1/2}}_{P,(\widetilde{K})}$ and $\widehat{\boldsymbol{\Omega}}^{*^{-1}}_{P,(\widetilde{K})}$ can be replaced, respectively, by draws from a $\widehat{\boldsymbol{\Omega}}^{-1}_{P,(\widetilde{K})}\mathrm{N}(\boldsymbol{0},\boldsymbol{I}_{\widetilde{K}})$, by $\widehat{\boldsymbol{D}}^{-1/2}_{P,(\widetilde{K})}$ and $\widehat{\boldsymbol{\Omega}}^{-1}_{P,(\widetilde{K})}$. Bootstrap critical values for $S_P^+$ can be obtained from the percentiles of the empirical distribution of

$$S_P^{*+} = \sum_{k=1}^{K} \left( \sqrt{P} \left( \frac{\widehat{m}_{k,P}^{*} - \widehat{m}_{k,P}}{\sqrt{\widehat{v}_{k,k}^{*}}} \right)^{+} 1_{\left\{ \widehat{m}_{k,P} \geq -\sqrt{\widehat{v}_{k,k}} \sqrt{2 \ln \ln P / P} \right\}} \right)^{2}.$$

Let $c_{P,\alpha}^{*}$ and $c_{P,\alpha}^{*+}$ be the $(1-\alpha)-$percentile of empirical distribution of $S_P^{*}$ and $S_P^{*+}$, respectively. From Theorem 1 in Andrews and Soares (2010), we have that, under $\mathrm{H}_0$,

$$\lim_{P\to\infty} \Pr\left(S_P \leq c_{P,\alpha}^{*}\right) \geq 1-\alpha \text{ and } \lim_{P\to\infty} \Pr\left(S_P^{+} \leq c_{P,\alpha}^{*+}\right) \geq 1-\alpha.$$

Thus, the asymptotic size is at most $\alpha$.

By comparing $S_P$ against $c_{P,\alpha}^{*}$ and $S_P^{+}$ against $c_{P,\alpha}^{*+}$ we have a test for $\mathrm{H}_0$ versus $\mathrm{H}_A$, where FWER is controlled for, and the probability is false discovery is at most $\alpha$. Furthermore, the selection rule for eliminating slack moment conditions (i.e. poor forecasting models) limits the risk of driving the power to zero by including weak models.

In principle, any function $S$ in (9) satisfying Assumptions 1-4 in Andrews and Guggenberger (2009) delivers a test for superior predictive ability, ensuring a false discovery of at most $\alpha$ and controlling for irrelevant models. A natural choice is the maximum over the positive moment conditions, i.e.

$$S_{P,Max} = \max_{k=1,\ldots,K} \left\{ \sqrt{P} \left( \widehat{m}_{k,P}^{+} / \sqrt{\widehat{v}_{k,k}} \right) \right\},$$

which satisfies Assumptions 1-4 in Andrews and Guggenberger (2009) and is equivalent to

Hansen's $T_P$ in (4).[5] Under the null, as $P \to \infty$,

$$S_{P,Max} \xrightarrow{d} \max_{k=1,\ldots,k} \left\{ \left( \sum_{j=1}^{K} \omega_{k,j} Z_j + h_k \right)^{+} \right\}$$

and bootstrap critical values can be obtained from the quantiles of the empirical distribution of

$$\begin{aligned} S^*_{P,Max} &= \max_{k=1,\ldots,K} \left\{ \sqrt{P} \left( \frac{\widehat{m}^*_{k,P} - \widehat{m}_{k,P}}{\sqrt{\widehat{v}^*_{k,k}}} \right)^{+} 1_{\left\{ \widehat{m}_{k,P} \geq -\sqrt{\widehat{v}_{k,k}}\sqrt{2\ln\ln P/P} \right\}} \right\} \\ &= \max \left\{ 0, \max_{k=1,\ldots,K} \left\{ \sqrt{P} \left( \frac{\widehat{m}^*_{k,P} - \widehat{m}_{k,P}}{\sqrt{\widehat{v}^*_{k,k}}} \right) 1_{\left\{ \widehat{m}_{k,P} \geq -\sqrt{\widehat{v}_{k,k}}\sqrt{2\ln\ln P/P} \right\}} \right\} \right\}. \end{aligned}$$

Note that Hansen's bootstrap statistic in (5) writes as

$$\begin{aligned} T^*_P &= \max \left\{ 0, \max_{k=1,\ldots,K} \left\{ \sqrt{P} \left( \frac{\widehat{m}^*_{k,P} - \widehat{m}_{k,P}}{\sqrt{\widehat{v}_{k,k}}} \right) 1_{\left\{ \widehat{m}_{k,P} \geq -\sqrt{\widehat{v}_{k,k}}\sqrt{2\ln\ln P/P} \right\}} \right\}, \right. \\ &\left. \max_{k=1,\ldots,K} \left\{ \sqrt{P} \frac{\widehat{m}^*_{k,P}}{\sqrt{\widehat{v}_{k,k}}} 1_{\left\{ \widehat{m}_{k,P} < -\sqrt{\widehat{v}_{k,k}}\sqrt{2\ln\ln P/P} \right\}} \right\} \right\}. \end{aligned} \qquad (10)$$

By comparing $S^*_{P,Max}$ and $T^*_P$, we note two differences. First, the scaling factor in $S^*_{P,Max}$ is $\widehat{v}^*_{k,k}$ while that in $T^*_P$ is $\widehat{v}_{k,k}$. Second, $S^*_{P,Max}$ is defined as the maximum over two objects, while $T^*_P$ as the maximum over three objects. More precisely, $S^*_{P,Max}$ does not take into account $\max_{k=1,\ldots,K} \left\{ \sqrt{P} \frac{\widehat{m}^*_{k,P}}{\widehat{v}_{k,k}} 1_{\left\{ \widehat{m}_{k,P} < -\sqrt{\widehat{v}_{k,k}}\sqrt{2\ln\ln P/P} \right\}} \right\}$. Nevertheless, $S^*_{P,Max} - T^*_P = o_{p^*}(1)$, where $o_{p^*}(1)$ denotes a term converging to zero under $P^*$, the probability law of the resampled series. $\widehat{v}^*_{k,k} - \widehat{v}_{k,k} = o_{p^*}(1)$, and the contribution of the term in the second line of (10) is asymptotically negligible. In fact, up to a vanishing error, $\mathrm{E}^* \left( \widehat{m}^*_{k,P} \right) = \widehat{m}_{k,P}$, where $\mathrm{E}^*$ denotes the expectation under $P^*$, conditional on the sample. Therefore, if $\widehat{m}_{k,P} < -\sqrt{\widehat{v}_{k,k}}\sqrt{2\ln\ln P/P}$, it is unlikely

[5] It suffices to note that $\max_{k=1,\ldots,K} \left\{ \sqrt{P} \left( \widehat{m}^+_{k,P} / \sqrt{\widehat{v}_{k,k}} \right) \right\} = \max \left\{ 0, \max_{k=1,\ldots,K} \sqrt{P} \left( \widehat{m}^+_{k,P} / \sqrt{\widehat{v}_{k,k}} \right) \right\}$.

that $\widehat{m}^*_{k,P} > 0$. More formally, by the law of the iterated logarithm,

$$\begin{aligned} & P^*\left(\max_{k=1,\dots,K}\left\{\sqrt{P}\frac{\widehat{m}^*_{k,P}}{\sqrt{\widehat{v}_{k,k}}}1_{\left\{\widehat{m}_{k,P}<-\sqrt{\widehat{v}_{k,k}}\sqrt{2\ln\ln P/P}\right\}}\right\}>0\right) \\ = \; & P^*\left(\max_{k=1,\dots,K}\left\{\sqrt{P}\frac{\widehat{m}^*_{k,P}-\widehat{m}_{k,P}}{\sqrt{\widehat{v}_{k,k}}}\right\}>\sqrt{2\ln\ln P}\right)\to 0, \end{aligned}$$

and so $\max_{k=1,\dots,K}\left\{\sqrt{P}\frac{\widehat{m}^*_{k,P}}{\widehat{v}_{k,k}}1_{\left\{\widehat{m}_{k,P}<-\sqrt{\widehat{v}_{k,k}}\sqrt{2\ln\ln P/P}\right\}}\right\}$ does not contribute to the bootstrap critical values of $T^*_P$. Hansen's SPA test can therefore be seen as a member of the class of superior predictive ability tests.

It is worthwhile to point out that superior predictive ability tests provide a remedy to the problem of $p$-values artificially inflated by inclusion of irrelevant models, but they do not completely solve the issue of having conservative $p$-values. This is because, as a selection rule to eliminate irrelevant models, they use the law of iterated logarithm and eliminate only moment conditions diverging to minus infinity at a slow rate. To have a test of superior predictive ability with exact asymptotic level $\alpha$, one should eliminate all models whose moment conditions are smaller than a given negative constant. Andrews and Jia (2008) show how to find the value for this constant maximizing the average power of the test, and how to implement a size correction to obtain a test with asymptotic size equal to $\alpha$.

# 5 Controlling for FDR

Table 1 summarizes the different possible outcomes that occur when testing $K$ hypotheses. Bonferroni's and Holm's approaches described in the Introduction controlled for the FWER, defined as $\Pr(V \geq 1) \leq \alpha$. But, since the probability of at least a false discovery increases very quickly with $K$, controlling for the FWER entails a loss of power, and the loss of power increases with $K$. To cope with this problem, Benjamini and Hochberg (1995) propose to

Table 1: Outcomes when testing $K$ hypotheses

| | Do not Reject | Reject | Total |
|---|---|---|---|
| Null true | $U$ | $V$ | $K_0$ |
| Alternative true | $T$ | $S$ | $K_1$ |
| | $W$ | $R$ | $K$ |

control for the ratio of false discoveries to total discoveries, that is to control for

$$\mathrm{FDR} = \mathrm{E}\left(\frac{V}{R}\right) = \mathrm{E}\left(\frac{V}{R}\middle| R\right)\Pr(R > 0).$$

They provide a sequential method, based on the obtained $p$-values from the individual tests, to control for this quantity; they prove its validity under the assumption that the $p$-values are independent. After ordering the $p$-values $p_{(1)} \leq \ldots \leq p_{(K)}$, calculate $k_{max} = \max\left\{1 \leq k \leq K : \; p_{(k)} \leq \alpha k/K\right\}$ and reject hypotheses $\mathrm{H}_{0,(1)}, \ldots, \mathrm{H}_{0,(k_{max})}$. If no such $k$ exists, reject no hypothesis. FDR criterion is much less strict than FWER and, as a consequence, leads to a substantial increase of power. FDR can be interpreted as "the expected proportion of false discoveries to total discoveries times the probability of making at least one discovery". Therefore, when controlling for FDR $\leq \gamma$ and discoveries have occurred, this method really controls for FDR at a level $\gamma/\Pr(R > 0)$. For small values of $\alpha$, $\Pr(R > 0)$ can be small, and so this is an unfortunate characteristic of the method.[6] This observation has lead Storey (2003) to propose an alternative quantity to control, called "positive FDR", defined by

$$\mathrm{pFDR} = \mathrm{E}\left(\frac{V}{R}\middle| R\right),$$

[6] At the same time, as $K$ increases, then $\Pr(R > 0)$ tends to 1, so that FDR and pFDR are asymptotically equivalent.

which does not suffer from the problem above. pFDR can be interpreted as “the expected proportion of false discoveries to total discoveries”.

## 5.1 Estimating FDR

Suppose that all the single hypotheses are tested at the same level of significance, $\alpha$. Storey (2002) proposes the following estimators for $\mathrm{FDR}(\alpha)$ and $\mathrm{pFDR}(\alpha)$

$$\begin{aligned} \widehat{\mathrm{FDR}}_\lambda(\alpha) &= \frac{\widehat{w}_0(\lambda)\alpha}{\max\{R(\alpha),1\}/K}, && (11)\\ \widehat{\mathrm{pFDR}}_\lambda(\alpha) &= \frac{\widehat{w}_0(\lambda)\alpha}{\max\{R(\alpha),1\}(1-(1-\alpha)^K)/K}, && (12)\end{aligned}$$

where $R(\alpha)$ denotes the total number of rejections at the confidence level $\alpha$ and is calculated as $R(\alpha) = \sum_{i=1}^K 1_{\{p_i \leq \alpha\}}$. $\widehat{w}_0(\lambda)$ is an estimator of $w_0$, the percentage of times in which the null hypothesis is true. For a well chosen tuning parameter $\lambda$ (within the interval $[0,1]$), a natural estimator is given by

$$\widehat{w}_0(\lambda) = \frac{W(\lambda)}{(1-\lambda)K} = \frac{K-R(\lambda)}{(1-\lambda)K}. \qquad (13)$$

The intuition behind equation (13) is the following. Recalling that under the null hypothesis the $p$-values are uniformly distributed over the interval $[0,1]$, for any given $0 < \lambda < 1$, as $T, K \to \infty$,

$$\frac{(K-R(\lambda))}{K} \xrightarrow{p} (1-\lambda)w_0, \quad \text{hence} \quad \widehat{w}_0(\lambda) \xrightarrow{p} w_0.$$

Note that we need $T \to \infty$ to ensure that the estimated $p$-values converge to the true ones, and $K \to \infty$ to ensure that the empirical distribution of the null $p$-values approaches a uniform distribution.

Similarly, the denominator of (12) has the following interpretation. Since the null hypothesis is rejected when the $p$-value is less than or equal to $\alpha$, we can define the rejection region

$\Gamma = (0, \alpha]$. Given that the denominator of (12) denotes the total number of rejections out of the $K$ tests with critical region $\Gamma$, as $T, K \to \infty$,

$$\frac{\max\{R(\alpha), 1\}(1-(1-\alpha)^K)}{K} \xrightarrow{p} w_0(\text{Type I error of } \Gamma) + (1-w_0)(1 - \text{Type II error of } \Gamma).$$

Therefore, combining the previous results, for any $0 < \lambda < 1$, as $T, K \to \infty$,

$$\begin{aligned}\widehat{\text{pFDR}}_\lambda(\alpha) &\xrightarrow{p} \frac{w_0(\text{Type I error of } \Gamma)}{w_0(\text{Type I error of } \Gamma) + (1-w_0)(1 - \text{Type II error of } \Gamma)} \\ &= \frac{w_0(\text{Size}(\Gamma))}{w_0(\text{Size}(\Gamma)) + (1-w_0)(\text{Power}(\Gamma))} \\ &= \Pr\left(\text{H}_0 \text{ true} \,|\, \text{H}_0 \text{ rejected}\right).\end{aligned}$$

Finally, it is useful to introduce a statistical measure, the $q$-value, which plays the same role to pFDR as the $p$-value does to the type I error in testing statistical hypotheses. The $q$-value gives us an error measure for each observed test statistic with respect to pFDR. In order to highlight its meaning, it is useful to define the $p$-value associated with an observed test statistic $s$ as

$$p(s) = \min_{\{\Gamma : s \in \Gamma\}} \Pr(s \in \Gamma | \text{H}_0 \text{ true}).$$

Analogously, the $q$-value associated with $s$ is

$$q(s) = \inf_{\{\Gamma : s \in \Gamma\}} \text{pFDR}(\Gamma).$$

In other words, the $q$-value is the minimum pFDR that can occur when rejecting a statistic with value $s$ for a rejection region $\Gamma$. McCracken and Sapp (2005) use $q$-values to evaluate the out-of-sample performance of exchange rate models.

## 5.2 Controlling for FDR with a given $\lambda$

Controlling for FDR, following the implementation suggested by Storey, Taylor and Siegmund (2004), we are able to identify the forecasting models genuinely outperforming the benchmark. This procedure allows us to fix *a priori* an acceptable level of false discoveries, that is the proportion of falsely rejected null hypotheses. The implementation of the method requires the use of a tuning parameter $\lambda$. In this subsection, we outline the procedure for a given $\lambda$. The implementation requires three sequential Steps.

**Step 1** Fix $\gamma$ as the level at which we want to control FDR (i.e. FDR $\leq \gamma$).

**Step 2** For any rejection region as $(0,\alpha]$, construct $\widehat{\mathrm{FDR}}_\lambda(\alpha)$.

**Step 3** Define

$$t_\gamma\left(\widehat{\mathrm{FDR}}_\lambda\right) = \sup_{\alpha\in[0,1]}\left\{\widehat{\mathrm{FDR}}_\lambda(\alpha) \leq \gamma\right\}.$$

**Step 4** Reject the null hypotheses characterized by $p_i \leq t_\gamma\left(\widehat{\mathrm{FDR}}_\lambda\right)$.

Notice that, in contrast to some of the methods described in previous Sections, this is an example of a *stepup* method. It starts with the least significant hypothesis and then moves up. If the $p$-values associated with the various hypotheses are independent, the rule ensures that the percentage of falsely rejected null hypotheses is bounded above by $(1-\lambda^{w_0 K})\gamma \leq \gamma$ (Storey, Taylor and Siegmund, 2004). Hence, it controls FDR conservatively at level $\gamma$.

Storey, Taylor and Siegmund (2004) prove that this procedure controls the FDR asymptotically, under a weak dependence assumption on the $p$-values that does not cover the case of constant cross-sectional correlation. In this case, their procedure can be quite liberal. Romano, Shaikh and Wolf (2008a) propose resampling techniques, which allow to control asymptotically for FDR under dependence.[7]

[7] They propose both a bootstrap and subsampling solution. The latter works under less restrictive assumptions.

Finally, note that the FDR is the mean of the false discovery proportion (FDP). Hence, even if FDR is controlled for at a given level, its realization can be quite different. Most of the procedures explained in this Chapter can be modified to control for FDP (for details, see Romano, Shaikh and Wolf, 2008b).

### 5.3 Optimal choice of $\lambda$

Inspection of (13) reveals that, for any $0 < \lambda < 1$, $\widehat{w}_0(\lambda)$ consistently estimates $w_0$. However, for finite $K$, the bias of $\widehat{w}_0(\lambda)$ is decreasing in $\lambda$, while its variance is increasing in $\lambda$. Similarly to any situation characterized by a bias/variance trade-off, the natural solution is to select $\lambda$ using a data-driven approach, in order to minimize the mean squared error of $\widehat{w}_0(\lambda)$.

An adaptive method for choosing $\lambda$ has been proposed by Storey, Taylor and Siegmund (2004). The rule they suggest is optimal for the case of independence among $p$-values. This is because it involves a standard i.i.d. resampling of the $p$-values, thus ignoring the possible correlation among them.

## 6 The Model Confidence Set Approach

The tests outlined in the previous Sections consider the null hypothesis that no candidate model outperforms a given benchmark. If the null is rejected, one gets useful information about better alternative models by looking at the specific moment conditions which contribute most to the rejection. This is a natural approach when the benchmark model stands out as a clear candidate, e.g. when the benchmark is the simplest, or the most frequently used model, or embeds the prediction of some economic model of interest. Notice that failure to reject the null is consistent with either the fact that the benchmark is a superior model or with the fact that it is performing as well as (some, or even all) the competing models.

In the latter case, one may be interested in extracting the subset of models with equal predictive ability, and may be combine them in order to produce forecasts. Indeed, if two or more models have the same predictive ability, and we have no particular interest in choosing one over the other, forecast combination increases efficiency and reduces model risk (see, e.g., Timmermann, 2006).

A formal approach for sequentially eliminating the worst performing model and retain best model(s) is the Model Confidence Set (MCS) approach of Hansen, Lunde and Nason (2009). Let $\mathcal{M}_k$ denote a collection of $k$ out the $K$ available models. Hence, $\mathcal{M}_K \supset \mathcal{M}_{K-1} \supset \ldots \supset \mathcal{M}_1$. Define

$$\widehat{m}_{i,j,P} = \frac{1}{P-\tau+1} \sum_{t=R+\tau}^{T} \left( f\left(\widehat{u}_{i,t}\right) - f\left(\widehat{u}_{j,t}\right) \right)$$

The Model Confidence Set, $\mathcal{M}^\dagger$ is defined as

$$\mathcal{M}^\dagger = \left\{ i: \text{ there is no } j \neq i \text{ such that } m_{i,j,P} > 0 \right\},$$

where $m_{i,j,P} = \mathrm{E}\left( f\left(u_{i,t}\right) - f\left(u_{j,t}\right) \right)$.[8] Therefore, no model belonging to $\mathcal{M}^\dagger$ can be "inferior" to a model belonging to its complement $\mathcal{M}^{\dagger c}$. Hansen, Lunde and Nason (2009) provide an estimator of $\mathcal{M}^\dagger$, say $\widehat{\mathcal{M}_P^\dagger}$, such that, as $P \to \infty$, $\widehat{\mathcal{M}_P^\dagger} \supseteq \mathcal{M}^\dagger$ with probability larger or equal than $1-\alpha$. Furthermore, their procedure ensures that, as $P \to \infty$, the probability that $\mathcal{M}^{\dagger c} \subseteq \widehat{\mathcal{M}_P^\dagger}$ tends to zero.

The MCS methodology is implemented as follows. Each step of the procedure consists of two parts: a test for equal predictive ability between the $k$ survived models and, if the null is rejected, elimination of the worst model. The procedure terminates the first time one fails to reject the null of equal predictive ability among the survived models. The MCS then contains the final surviving models. More formally, after the initialization $\mathcal{M} = \mathcal{M}_K$, the MCS algorithm

[8] We now need two subscripts, $i, j$ to denote that we are comparing model $i$ versus model $j$.

is based on two sequential steps:

(i) Test of $\mathrm{H}_{0,\mathcal{M}} : m_{i,j} = 0$ for all $i, j \in \mathcal{M}$, using the statistic $Z_{P,\mathcal{M}} = \max_{i,j\in\mathcal{M}} t_{i,j}$, where $t_{i,j} = \sqrt{P}\left(\widehat{m}_{i,j,P}/\sqrt{\widehat{\sigma}_{i,j}}\right)$ and $\widehat{\sigma}_{i,j}$ is a consistent estimator of $\lim_{P\to\infty} \operatorname{var}\left(\sqrt{P} m_{i,j,P}\right)$.

(ii) Let $p$ be the $p$-value associated with $\mathrm{H}_{0,\mathcal{M}}$. If $p \geq \alpha$, all models in $\mathcal{M}$ are considered to be equally "good" models and $\mathcal{M}$ is the MCS. Otherwise, the worst model is eliminated from $\mathcal{M}$, and one goes back to (1).

Note that we require $\widehat{\sigma}_{i,j} \xrightarrow{p} \sigma_{i,j} > 0$ for all $i, j$. Broadly speaking, we need that no models is nested with or nesting any other model. The worst model is defined by the rule $\arg\max_{i\in\mathcal{M}_k} k^{-1}\sum_{j=1}^{k} t_{i,j}$, i.e. it is the model that on average has the worst performance relatively to all other models.

The limiting distribution of the equal predictive ability statistic $Z_{P,\mathcal{M}}$ is not nuisance parameters free, because it depends on the cross-sectional correlation of $\widehat{m}_{i,j,P}$. A natural solution is to construct bootstrap $p$-values. Let $\widehat{m}^{*}_{i,j,P}$ be the bootstrap counterpart of $\widehat{m}_{i,j,P}$. Hansen, Lunde and Nason (2009), suggest to compute the following bootstrap analog of $Z_{P,\mathcal{M}}$, for $b = 1, \ldots, B$

$$Z^{*(b)}_{P,\mathcal{M}} = \max_{i,j\in\mathcal{M}} \left\{ \sqrt{P} \left| \frac{\widehat{m}^{*(b)}_{i,j,P} - \widehat{m}_{i,j,P}}{\sqrt{\widehat{\sigma}^{*}_{i,j}}} \right| \right\},$$

where $\widehat{\sigma}^{*}_{i,j} = \frac{P}{B}\sum_{b=1}^{B}\left(\widehat{m}^{*(b)}_{i,j,P} - \widehat{m}_{i,j,P}\right)^2$. Bootstrap $p$-values are readily obtained by

$$p^{*} = \frac{1}{B}\sum_{b=1}^{B} 1_{\left\{Z^{*(b)}_{P,\mathcal{M}} > Z_{P,\mathcal{M}}\right\}}.$$

Therefore, according to step (2) above, if $p^{*} < \alpha$, the worst model is eliminated and one goes back to step (1), otherwise $\mathcal{M}$ is selected as the MCS, i.e. $\widehat{\mathcal{M}}^{\dagger}_{P} = \mathcal{M}$, and the procedure is terminated.

Hansen, Lunde and Nason (2009) show that, as $P \to \infty$, $\widehat{\mathcal{M}}^{\dagger}_{P} \supseteq \mathcal{M}^{\dagger}$ with probability larger than $1-\alpha$, and $\widehat{\mathcal{M}}_P^{\dagger} \supseteq \mathcal{M}^{\dagger c}$ with probability approaching zero.

## 6.1 A Simple Doubly Consistent Alternative to Model Confidence Set

We now outline a simple procedure which provides an estimate of $\mathcal{M}^{\dagger}$, say $\widetilde{\mathcal{M}}_P^{\dagger}$, such that as $P \to \infty$, $\widetilde{\mathcal{M}}_P^{\dagger} \supseteq \mathcal{M}^{\dagger}$ with probability one, $\widetilde{\mathcal{M}}_P^{\dagger} \supseteq \mathcal{M}^{\dagger c}$ with probability zero. This ensures that both the probabilities of keeping irrelevant models and eliminating relevant models are asymptotically zero. In this sense, it is a doubly consistent procedure.

We proceed sequentially. First we pick one out of the $K$ models, say model 1, and we compare all models against 1. All models performing worse than 1 are eliminated. If there is at least one model beating 1, the latter is eliminated as well. We then repeat the same procedure between the models survived in the first step, until convergence. As it will become clear below, ordering of the models for pairwise comparisons is irrelevant.

Suppose we first pick model 1 between the available $k$ models; the rule is to eliminate model $j$ when

$$\widehat{m}_{1,j,P}/\sqrt{\widehat{\sigma}_{1,j}} < -\sqrt{2\ln\ln P/P}$$

and to eliminate model 1 if there is at least a $j \neq 1$, such that

$$\widehat{m}_{1,j,P}/\sqrt{\widehat{\sigma}_{1,j}} > \sqrt{2\ln\ln P/P}.$$

Now suppose model 2 survived to the first stage, and is selected for comparison. We repeat the same procedure outlined above, until either no model is eliminated or we remain with only one model left. The set of models which survived all the steps form $\widetilde{\mathcal{M}}_P^{\dagger}$. For all $i, j$, by the law of the iterated logarithm,

$$\Pr\left(\limsup_{P\to\infty}\sqrt{P}\left(\frac{m_{i,j,P}-m_{i,j}}{\sqrt{\sigma_{i,j}}}\right)=\sqrt{2\ln\ln P}\right)=1,$$

and

$$\Pr\left(\liminf_{P\to\infty}\sqrt{P}\left(\frac{m_{ij,P}-m_{ij}}{\sqrt{\sigma_{i,j}}}\right)=-\sqrt{2\ln\ln P}\right)=1.$$

Provided $\widehat{\sigma}_{i,j}-\sigma_{i,j}=o_p(1)$, $\widehat{m}_{i,j,P}-m_{i,j,P}=o_p(1)$, it follows that

$$\lim_{P\to\infty}\Pr\left(\widetilde{\mathcal{M}_P^{\dagger}}\supseteq\mathcal{M}^{\dagger}\right)=1,\quad \lim_{P\to\infty}\Pr\left(\widetilde{\mathcal{M}_P^{\dagger}}\supseteq\mathcal{M}^{\dagger c}\right)=0, \tag{14}$$

because, at each stage, a model is eliminated only if it is dominated by at least another model. Note that the order according to which we choose the reference model at each stage is irrelevant, in the sense it affects only the order in which models are eliminated, but not which models are eliminated. This is because a model is eliminated if and only if there is at least a competitor exhibiting better predictive ability. Importantly, as the asymptotic size is zero at each step, there is no sequential size distortion, and the order in which the procedure is implemented is irrelevant. Given the definition of $\mathcal{M}^{\dagger}$ and given (14), it is immediate to see that as $P$ gets large $\lim_{P\to\infty}\Pr\left(\widetilde{\mathcal{M}_P^{\dagger}}\supseteq\widehat{\mathcal{M}_P^{\dagger}}\right)=1$.

# References

Andrews, D.W.K. (1999). Estimation When a Parameter is on the Boundary. *Econometrica*, 67, 1341-1383.

Andrews, D.W.K. and P. Guggenberger (2009). Validity of Subsampling and "Plug-In Asymptotic" Inference for Parameters Defined by Moment Inequalities. *Econometric Theory*, 25, 669-709.

Andrews, D.W.K. and P. Jia (2008). Inference for Parameters Defined by Moment Inequalities: A Recommended Moment Selection Procedure. Cowles Foundation Discussion Paper 1676.

Andrews, D.W.K. and G. Soares (2010). Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection. *Econometrica*, 78, 119-158.

Avramov, D., L. Barras and R. Kosowski (2009). Hedge Fund Predictability Under the Magnifying Glass: Forecasting Individual Fund Returns Using Multiple Predictors. Working Paper.

Awartani, B.M. and V. Corradi (2005). Predicting the Volatility of the S&P500 Stock Index: the Role of Asymmetries. *International Journal of Forecasting*, 21, 167-193.

Bajgrowicz, P. and O. Scaillet (2008). Technical Trading Revisited: Persistence Tests, Transaction Costs, and False Discoveries. Working Paper.

Barras, L., O. Scaillet and R. Wermers (2005). False Discoveries in Mutual Fund Performance: Measuring Luck in Estimated Alphas. Manuscript.

Beg, A.B.M.R., M. Silvapulle and P. Silvapulle (2001). Tests Against Inequality Constraints When Some Nuisance Parameters are Present only under the Alternative: Test of ARCH in ARCH-M Models. *Journal of Business & Economic Statistics*, 19, 245-253.

Benjamini, Y., and Y. Hochberg (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society, B*, 57, 289-300.

Benjamini, Y., and Y. Hochberg (2000). The adaptive control of the false discovery rate in multiple hypotheses testing. *Journal of Educational and Behavioral Statistics*, 25, 68-83.

Bugni, F. (2008). Bootstrap Inference in Partially Identified Models. *Econometrica*, forthcoming.

Chan, L.K., N. Jegadeesh, and J. Lakonishok (1996). Momentum Strategies. *Journal of Finance*, LI, 1681-1713.

Chernozhukov, V., H. Hong and E. Tamer (2007). Estimation and Confidence Regions for Parameters Sets in Econometric Models. *Econometrica*, 75, 1243-1284.

Clark, T.E. and M.W. McCraken (2001). Test for Equal Forecast Accuracy and Encompassing for Nested Models. *Journal of Econometrics*, 105, 85-110.

Clark, T.E. and K.D. West (2007). Approximately Normal Tests for Equal Predictive Accuracy in Nested Models. *Journal of Econometrics*, 138, 291-3111.

Corradi, V. and N.R. Swanson (2006a). Predictive Density and Conditional Confidence Interval Accuracy Tests. *Journal of Econometrics*, 135, 187-228.

Corradi, V., and N.R. Swanson (2006b). Predictive Density Evaluation. In *Handbook of Economic Forecasting*, eds. C.W.J. Granger, G. Elliott and A. Timmermann, Elsevier, Amsterdam, pp. 197-284.

Corradi, V. and N.R. Swanson (2007). Nonparametric Bootstrap Procedures for Predictive Inference Based on Recursive Estimation Schemes. *International Economic Review*, 48, 67-109.

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

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

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

Hansen, P.R. and A. Lunde (2005). A Forecast Comparison of Volatility Models: Does Anything Beat a GARCH(1,1)? *Journal of Applied Econometrics*, 20, 873-889.

Hansen, P.R. and A. Lunde (2006). Consistent Ranking of Volatility Models. *Journal of Econometrics*, 131, 97-121.

Hansen, P.R., A. Lunde and J.M. Nason (2009). The Model Confidence Set. Federal Reserve of Atlanta, WP 2005-7a.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6, 65–70.

McCracken, M.W. (2007). Asymptotics for Out of Sample Tests of Granger Causality. *Journal of Econometrics*, 140, 719-752.

McCracken, M.W. and S.G. Sapp (2005). Evaluating the Predictive Ability of Exchange Rates using Long Horizon Regressions: Mind Your $p$'s and $q$'s! *Journal of Money, Credit and Banking*, 37, 473-94.

Romano, J.P. and A.M. Shaikh (2010). Inference for the Identified Set in Partially Identified Econometric Models. *Econometrica,* 78, 169-212.

Romano, J.P., A.M. Shaikh and M. Wolf (2008a). Control of the false discovery rate under dependence using the bootstrap and subsampling. *TEST*, 17, 417-442 (Invited Paper with discussion).

Romano, J.P., A.M. Shaikh and M. Wolf (2008b). Formalized Data Snooping Based On Generalized Error Rates. *Econometric Theory*, 24, 404–447.

Romano, J.P. and M. Wolf (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, 73, 1237-1282.

Rosen, A.M. (2008). Confidence Sets for Partially Identified Parameters that Satisfy a Finite Number of Moment Inequalities. *Journal of Econometrics*, 146, 107-117.

Storey, J.D. (2002). A Direct Approach to False Discovery Rates. *Journal of the Royal Statistical Society, B*, 64, 479-498.

Storey, J.D. (2003). The Positive False Discovery Rate: a Bayesan Interpretation and the $q$-value. *Annals of Statistics*, 31, 2013-2035.

Storey, J.D., and R. Tibshirani (2003). Statistical Significance for Genomewide Studies. *Proceedings of the National Academy of Science*, 100, 9440-9445.

Storey, J.D., J.E. Taylor, and D. Siegmund (2004). Strong Control, Conservative Point Estimation and Simultaneous Conservative Consistency of False Discovery Rates: A Unified Approach. *Journal of the Royal Statistical Society, B*, 66, 187-205.

Sullivan, R., A. Timmermann, and H. White (1999). Data-Snooping, Technical Trading Rule Performance, and the Bootstrap. *Journal of Finance*, LIV, 1647-1691.

Sullivan, R., A. Timmermann, and H. White (2001). The Danger of Data Mining: The Case of Calendar Effect in Stock Returns. *Journal of Econometrics*, 105, 249-286.

Timmermann, A. (2006). Forecast Combination. In *Handbook of Economic Forecasting*, eds. C.W.J. Granger, G. Elliott and A. Timmermann, Elsevier, Amsterdam, pp. 135–194.

West, K.D. (1996). Asymptotic Inference About Predictive Ability. *Econometrica*, 64, 1067-1084.

West, K.D. (2006). Forecast Evaluation. In *Handbook of Economic Forecasting*, eds. C.W.J. Granger, G. Elliott and A. Timmermann, Elsevier, Amsterdam, pp. 100-134.

West, K.D. and M.W. McCracken (1998), Regression Tests for Predictive Ability. *International Economic Review*, 39, 817-840.

White, H. (2000). A Reality Check For Data Snooping. *Econometrica*, 68, 1097-1127.

Wolf, M. and D. Wunderli (2009). Fund-of-Funds Construction by Statistical Multiple Testing Methods. Working Paper.