A Survey of Recent Advances in Forecast Accuracy Comparison Testing, with an Extension to Stochastic Dominance*

Valentina Corradi[1] and Norman R. Swanson[2]
[1]University of Warwick and [2]Rutgers University

February 2012

Abstract

In recent years, an impressive body or research on predictive accuracy testing and model comparison has been published in the econometrics discipline. Key contributions to this literature include the paper by Diebold and Mariano (DM: 1995) that sets the groundwork for much of the subsequent work in the area, West (1996) who considers a variant of the DM test that allows for parameter estimation error in certain contexts, and White (2000) who develops testing methodology suitable for comparing many models. In this chapter, we begin by reviewing various key testing results in the extant literature, both under vanishing and non-vanishing parameter estimation error, with focus on the construction of valid bootstrap critical values in the case of non-vanishing parameter estimation error, under recursive estimation schemes, drawing on Corradi and Swanson (2007a). We then review recent extensions to the evaluation of multiple confidence intervals and predictive densities, for both the case of a known conditional distribution (Corradi and Swanson 2006a,b) and of an unknown conditional distribution (Corradi and Swanson 2007b). Finally, we introduce a novel approach in which forecast combinations are evaluated via the examination of the quantiles of the expected loss distribution. More precisely, we compare models looking at cumulative distribution functions (CDFs) of prediction errors, for a given loss function, via the principle of stochastic dominance; and we choose the model whose CDF is stochastically dominated, over some given range of interest.

JEL classification: C22, C51.
Keywords: block bootstrap, recursive estimation scheme, reality check, parameter estimation error.

---

* Valentina Corradi, Department of Economics, University of Warwick, Department of Economics, Coventry CV4 7AL, UK, email: v.corradi@warwick.ac.uk. Norman R. Swanson, Department of Economics, Rutgers University, 75 Hamilton Street, New Brunswick, NJ 08901, USA, nswanson@econ.rutgers.edu. This chapter has been prepared for the Festschrift in honor of Halbert L. White in the event of the conference celebrating his 60th birthday, entitled "Causality, Prediction, and Specification Analysis: Recent Advances and Future Directions", and held at the University of California, San Diego on May 6-7, 2011. Swanson thanks the Rutgers University Research Council for financial support.

1 Introduction

One of the key contributions permeating the econometric research of Halbert White is the development of statistical tools for specification, estimation and inference with possibly misspecified models. His main message is that, even though models are merely (crude) approximations to reality, important things can be learned from carrying out inference and generally analyzing "wrong" models. Certainly, the notion of misspecification is absolutely crucial in the context of out-of-sample prediction. After all, if one is carrying out a predictive accuracy assessment in order to "choose" between two competing models, then, at the very least, one of the models is probably misspecified.

In this chapter, we begin by assuming that we are given multiple predictions, arising from multiple different models. Our objective is either to select the model(s) producing the more accurate predictions, for a given loss function, or alternatively, to eliminate the models giving the least accurate predictions. Furthermore, in many such situations, we can choose a benchmark or reference model. This can be a model suggested by economic theory, can be the winner of past competitions, or can simply be a model commonly used by practitioners. The key challenge in this case, is to asses whether there exists a competing model that outperforms the benchmark. However, if we sequentially compare the reference model with each of its competitors, we may run into problems. In fact, as the number of competitors increases, the probability of picking an alternative model just by "luck", and not because of its intrinsic merit, increases and eventually will reach one. This is the well known problem of data-mining or data snooping.

The starting point for our discussion is Diebold and Mariano (1995), who develop the "workhorse" of predictive accuracy tests. Two models are compared by assessing their relative predictive losses, given a particular loss function. Assuming that parameter estimation error vanishes asymptotically and that the models are nonnested ensures that the DM test is asymptotically normally distributed, regardless of whether or not the loss function is differentiable.[1] West (1996) allows for non-vanishing parameter estimation error in the DM test, although at the cost of assuming differentiability. In White (2000), a sequence of DM tests are constructed, and the supremum thereof (called the reality check) is used to test whether a given "benchmark" model is at least as accurate as all competitors. The null hypothesis is thus that no competing model can produce a

[1]For a discussion of nested models in the current context, see Clark and McCracken (2001) and Corradi and Swanson (2006b).

more accurate prediction than the benchmark model, for a given loss function. The key contribution of White (2000) is that he recognizes the importance of sequential test bias when comparing many (rather than two, say) models, and he develops the asymptotic theory allowing for the valid construction of critical values for his reality check, using for example, block bootstrap and related bootstrap techniques. In related work, Corradi and Swanson (2006a,b; 2007a,b) extend the reality check version of the DM test to the evaluation of confidence intervals and predictive densities (rather than focussing on the evaluation of point predictive loss measures). They additionally develop bootstrap techniques for addressing parameter estimation error, and allow for the evaluation of conditional distributions of both known and unknown functional form. By discussing all of the above papers, we undertake to construct a path describing developments in the predictive accuracy testing literature.

Of note is that if any of the above tests fail to reject the null hypothesis that no competitor outperforms the benchmark model, the obvious consequence is to base prediction on only the benchmark model. The tests, thus, are of a "model selection" variety. This is somewhat in contrast with the alternative approach of using forecast combination (see e.g. Elliott and Timmermann, 2004) to construct "optimal" predictions. In light of this observation, we conclude this chapter by proposing a new stochastic dominance type test that combines features of DM and reality check tests with forecast combination. In particular, we suggest a model selection method for selecting amongst alternative combination forecasts constructed from panel of forecasters. More broadly, we close by arguing that the notions of stochastic dominance discussed in this context may have a variety of uses in the predictive accuracy testing literature.

Before turning to our discussion of the above tests, it is worth making two comments that further underscore the sense in which the results of the above papers build on one another. First, recall that the prediction errors used to construct DM type tests arise in at least two ways. First, there are situations in which we have series of prediction errors, although we do not know the models used to generate the underlying predictions. For example, this situation arises when we have forecasts from different agents, or professional forecasters. Alternatively, we may have a sequence of Sharpe ratios or returns from different trading rules, as in the financial applications of Sullivan, Timmermann and White (1999, 2001). Second, there are situations in which we are interested in comparing estimated models. For example, we may want to decide whether to predict tomorrow's inflation rate using an autoregressive model, a threshold model, or a Markov switching

model. The parameters of these models are generally estimated. If the number of observations used to estimate the model is larger than the number of observations used for forecast evaluation, or if the same loss function is used for in-sample estimation and out-of-sample prediction (e.g., estimation by ordinary least squares (OLS) and a quadratic loss function), then the contribution of estimated parameters can be ignored. Otherwise, it has to be taken into account. Corradi and Swanson (2006a and 2007a) develop bootstrap procedures which properly capture the contribution of parameter estimation error in the case of rolling or recursive estimation schemes, respectively.

Second, and as mentioned above, DM and reality check type tests compare point forecasts (and forecast errors) from two or multiple models, respectively. For example, we may want to pick the model producing the most accurate point predictions of the inflation rate. However, there are situations in which we are instead interested in finding the model producing the most accurate interval predictions (e.g. that inflation will be within a given interval). Predictive interval accuracy is particularly important in the management of financial risk, in the insurance and banking industries, where confidence intervals or entire conditional distributions are often examined. Evaluation of Value at Risk and Expected Shortfall are two main examples (see Duffie and Pan (1997) for further discussion). Corradi and Swanson (2005, 2006a, 2006b, 2007b) extend the DM and reality check tests to the case of intervals and conditional distributions, using both simulated and historical data.

The rest of the chapter is organized as follows. In Section 2 we discuss the DM and reality check tests; and outline how to construct valid bootstrap p-values in the case of non-vanishing parameter estimation error, with both recursive and rolling estimation schemes. In Section 3 we extend the DM and reality check tests to the evaluation of multiple confidence intervals and predictive densities. Finally, in Section 4 we outline a new technique that draws together concepts of forecast combination with multiple model evaluation. Namely, we introduce a stochastic dominance type approach in which forecast combinations are evaluated via the examination of the quantiles of the expected loss distribution. More precisely, we compare models by prediction error CDFs, for given loss functions, via the principle of stochastic dominance; and we choose the model whose CDF is stochastically dominated, over some given range of interest.

2 DM and Reality Check Tests

2.1 The Case of Vanishing Estimation Error

We begin by outlining the DM (1995) and White (2000) tests, when parameter estimation error is asymptotically negligible. Consider a collection of $K+1$ models, where model 0 is treated as the benchmark or reference model and models $k = 1, \ldots, K$ compose the set of competing models. For the DM test, $K = 1$. For the reality check, $K > 1$. The $h-$step ahead forecast error associated with model k, is $u_{i,t+h} = y_{t+h} - \phi_k(Z^t, \theta_k^\dagger)$. As $\theta_k^\dagger$ is unknown, we don't observe the prediction error $u_{k,t+h}$, but we only observe $\widehat{u}_{k,t+h} = y_{t+h} - \phi_k(Z^t, \widehat{\theta}_{k,t})$, where $\widehat{\theta}_{k,t}$ is an estimator of $\theta_k^\dagger$ based on observations available at time t.

The common practice in out-of-sample prediction is to split the total sample of T observations into two sub samples of length R and P, with $R + P = T$. One uses the first R observations to estimate a candidate model, and construct the first $h-$step ahead prediction error. Then, one uses $R+1$ observations to re-estimate the model and compute the second $h-$step ahead prediction error, and so on, until one has a sequence of $(P - h + 1)$ $h-$step ahead prediction errors.[2] If we use this recursive estimation scheme, at each step the estimated parameters are given by

$$\widehat{\theta}_{k,t} = \arg\max_{\theta_k} \left\{ \frac{1}{t} \sum_{j=1}^{t} q_{k,j}\left(X_{k,t}, \theta_k\right) \right\} \text{ for } t \geq R, \tag{1}$$

where $q_{k,j}$ can be thought of as the quasi-likelihood function associated with model k.[3] Under stationarity, $\theta_k^\dagger = \arg\max_{\theta_i} E(q_{k,j}\left(X_{k,t}, \theta_k\right)$.

Hereafter, for notational simplicity, we consider only the case of $h = 1$.

For a given loss function, g, the reality check tests the following hypotheses[4]:

$$H_0 : E\left(g(u_{0,t+1}) - g(u_{1,t+1})\right) = 0$$

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

[3]If we instead use a rolling estimation scheme, then

$$\widetilde{\theta}_{k,t} = \arg\max_{\theta_k} \left\{ \frac{1}{R} \sum_{j=t-R+1}^{t} q_{k,j}\left(X_{k,t}, \theta_k\right) \right\} \quad R \leq t \leq T.$$

[4]See Christoffersen and Diebold (1996,1997) and Elliott and Timmermann (2004,2005) for a detailed discussion of loss functions used in predictive evaluation.

versus

$$H_A : E\left(g(u_{0,t+1}) - g(u_{1,t+1})\right) \neq 0.$$

If $R \to \infty$ at a faster rate than $P \to \infty$, as $T \to \infty$, then, assuming that models "0" and "1" are nonnested, the limiting distribution of

$$\widehat{DM}_P = \left(\frac{1}{P}\sum_{t=R}^{T-1}\left(g(\widehat{u}_{0,t+1}) - g(\widehat{u}_{1,t+1})\right)/\widehat{\sigma}_S\right.$$

is $N(0,1)$, when scaled appropriately by $\widehat{\sigma}_S$, heteroscedasticity and autocorrelation consistent (HAC) estimation of the variance of $\frac{1}{P}\sum_{t=R}^{T-1} g(\widehat{u}_{0,t+1}) - g(\widehat{u}_{1,t+1})$. Evidently, $\widehat{DM}_P$ is the HAC t-statistic associated with the intercept in a regression of the loss differential series, $g(\widehat{u}_{0,t+1}) - g(\widehat{u}_{1,t+1})$, on a constant. For a discussion of the limit distribution of this test statistic when the two forecasting models are nested, see Clark and McCracken (2001). Note also that g need not be differentiable, unless one wishes to adjust the limit distribution for the effect of parameter estimation error in cases where $P/R \to \pi$, as $P, R, T- > \infty$, $0 < \pi < \infty$, as in West (1996). Moreover, even if $P/R \to \pi$, as $P, R, T- > \infty$, $0 < \pi < \infty$, parameter estimation error is asymptotic negligible whenever we use the same loss for in-sample estimation and out-of-sample prediction (see below for further discussion).

Now, for a given loss function, g, the reality check tests the following hypotheses:

$$H_0 : \max_{k=1,\ldots,K} E\left(g(u_{0,t+1}) - g(u_{k,t+1})\right) \leq 0$$

versus

$$H_A : \max_{k=1,\ldots,K} E\left(g(u_{0,t+1}) - g(u_{k,t+1})\right) > 0.$$

The null hypothesis is that no competing model outperforms the benchmark (i.e., model "0"), for a given loss function, while the alternative is that at least one competitor outperforms the benchmark. By jointly considering all competing models, the reality check controls the family-wise error rate (FWER), and circumvents so-called "data snooping" problems. In fact, the test is designed to ensure that the probability of rejecting the null when it is false is smaller than or equal to a fixed nominal level, α. The reality check statistic is given by:

$$\widehat{S}_P = \max_{k=1,\ldots,K} \widehat{S}_P(0,k), \tag{2}$$

where

$$\widehat{S}_P(0,k) = \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(g(\widehat{u}_{0,t+1}) - g(\widehat{u}_{k,t+1}) \right), \ k = 1, ..., K.$$

Letting $S_P(0,k) = \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(g(u_{0,t+1}) - g(u_{k,t+1}) \right)$, it is immediate to see that,

$$\begin{aligned} \widehat{S}_P(0,k) - S_P(0,k) &= \mathrm{E}\left(\nabla_{\theta_0} g(u_{0,t+1}) \right) \frac{1}{\sqrt{P}} \sum_{t=R+1}^{T} \left(\widehat{\theta}_{0,t} - \theta_0^{\dagger} \right) \\ &\quad - \mathrm{E}\left(\nabla_{\theta_k} g(u_{k,t+1}) \right) \frac{1}{\sqrt{P}} \sum_{t=R+\tau}^{T} \left(\widehat{\theta}_{k,t} - \theta_k^{\dagger} \right) + o_p(1). \end{aligned} \tag{3}$$

Now, if $g = q_k$, then by the first order conditions, $\mathrm{E}\left(\nabla_{\theta_k} g(u_{k,t+1}) \right) = 0$. Thus, if we use the same loss function for estimation and prediction (e.g. we estimate the model by OLS and use a quadratic loss function), then parameter estimation error is asymptotically negligible. Furthermore, if $P/R \to 0$, as $P, R, T- > \infty$ (i.e., the sample used for estimation grows at a faster rate than the sample used for forecast evaluation), then parameter estimation is again asymptotically negligible. Otherwise, it has to be taken into account.

Proposition 2.2 in White (2000) establishes that

$$\max_{k=1,\dots,K} \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(\left(g(\widehat{u}_{0,t+1}) - g(\widehat{u}_{k,t+1}) \right) - \mu_k \right) \xrightarrow{d} \max_{k=1,\dots,K} Z_k,$$

where $\mu_k = \mathrm{E}\left(g(\widehat{u}_{0,t+1}) - g(\widehat{u}_{k,t+1}) \right)$, $Z = (\ Z_1, \dots,\ Z_k)^{\top}$ is distributed as $\mathrm{N}(0, V)$ and V has typical element

$$v_{j,k} = \lim_{P \to \infty} \mathrm{Cov} \left(\frac{1}{\sqrt{P}} \sum_{t=R+\tau}^{T} \left(g(\widehat{u}_{0,t+1}) - g(\widehat{u}_{k,t+1}) \right), \frac{1}{\sqrt{P}} \sum_{t=R+\tau}^{T} \left(g(\widehat{u}_{0,t+1}) - g(\widehat{u}_{k,t+1}) \right) \right). \tag{4}$$

Because the maximum of a Gaussian process is not a Gaussian process, the construction of p-values for the limiting distribution above 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 V, say $\widehat{V}$. Then, for each simulation $s = 1, \dots, S$, we construct

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

where $\left(\eta_1^{(s)}, \dots, \eta_K^{(s)} \right)^{\top}$ is drawn from a $\mathrm{N}(0, \mathbf{I}_K)$. Then, we compute $\max_{k=1,\dots,K} \left| \widehat{d}_P^{(s)} \right|$, and the $(1-\alpha)-$percentile of its empirical distribution. This simulation based approach requires the estimation

of V. Note that we can use an estimator of V which captures the contribution of parameter estimation error, along the lines of West (1996) and West and McCracken (1998). However, 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 is asymptotically negligible. In this case, we resample blocks of $g(\widehat{u}_{0,t+1}) - g(\widehat{u}_{k,t+1})$ and, for each bootstrap replication $b = 1, \ldots, B$, calculate

$$\widehat{S}_P^{*(b)}(0,k) = \frac{1}{\sqrt{P}} \sum_{t=R+\tau}^{T} \left(g^*(\widehat{u}_{0,t+1}) - g^*(\widehat{u}_{k,t+1}) \right).$$

Then, we compute the bootstrap statistic as $\max_{k=1,\ldots,K} \left| \widehat{S}_P^{*(b)}(0,k) - \widehat{S}_P(0,k) \right|$ and the $(1-\alpha)$-percentile of the empirical distribution of B statistics is used for inference. Evidently, the same approach discussed above can be used for the DM test, although such is clearly not needed, given the earlier results discussed, in cases where parameter estimation error vanishes asymptotically.

Before turning to the issue of constructing DM and reality check p-values in the case of non-vanishing parameter estimation error, it is worthwhile to review some other recent developments in the reality check literature.[5]

2.1.1 Controlling for Irrelevant Models

From the statistic in (2), it is immediate to see that any model which is strictly dominated by the benchmark does not contribute to the limiting distribution, simply because it does not contribute to the maximum. On the other hand, all models contribute to the limiting distribution of either the simulated or the bootstrap statistic. Thus, by introducing irrelevant models, the overall p-value increases. In fact, for a given level α, the probability of rejecting the null when it is false is α when all models are as good as the benchmark (i.e. when $E\left(g(u_{0,t+1}) - g(u_{k,t+1})\right) = 0$ for $k = 1, ..., K$), otherwise the probability of rejecting the null is smaller than α, and decreases as the number of irrelevant models increases. While the reality check is able to control the family-wise error rate, and so avoids the issue of data snooping, it may thus be rather conservative.

[5] In the sequel, for ease of notation, the version of the DM test that we discuss will be $\widehat{S}_P(0,k)$, with $k = 1$.

For this reason, attempts have been made to modify the reality check in such a way as to control for both the family-wise error rate and the inclusion of irrelevant models. Hansen (2005) suggests a variant of the reality check, called the SPA (Superior Predictive Ability) test, which is less sensitive to the inclusion of poor models and thus less conservative. The SPA statistic is given by

$$T_P = \max\left\{0, \max_{k=1,\dots,K} \frac{\frac{1}{\sqrt{P}}\sum_{t=R+\tau}^{T} \widehat{d}_{k,t}}{\sqrt{\widehat{v}_{k,k}}}\right\},$$

where $\widehat{d}_{k,t} = (g(\widehat{u}_{0,t+1}) - g(\widehat{u}_{k,t+1}))$ and $\widehat{v}_{k,k}$ is defined as in (4). The bootstrap counterpart to T_P at replication b, $T_P^{*(b)}$ is given by

$$T_P^{*(b)} = \max\left\{0, \max_{k=1,\dots,K}\left\{\frac{\frac{1}{\sqrt{P}}\sum_{t=R+\tau}^{T}\left(\widehat{d}_{k,t}^{*(b)} - \widehat{d}_{k,t}1_{\left\{\widehat{m}_{k,P} > -\widehat{v}_{k,k}\sqrt{2\ln\ln P/P}\right\}}\right)}{\sqrt{\widehat{v}_{k,k}}}\right\}\right\}.$$

Here, 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 a model is too slack, and so it does not contribute to T_P, the corresponding bootstrap moment condition is not recentered, and so the bootstrap statistic is also not affected by the irrelevant model. The fact that very poor models do not contribute to the bootstrap p-values makes the SPA p-values less conservative than the reality check p-values. Nevertheless, it cannot be established that the SPA test is uniformly more powerful than the reality check test. Corradi and Distaso (2011), using the generalized moment selection approach of Andrews and Soares (2010), derive a general class of superior predictive accuracy tests, which control for FWER and the contribution of irrelevant models. They show that Hansen's SPA belongs to this class. Additionally, Romano and Wolf (2005) suggest a multiple step extension of the reality check which ensures tighter control of irrelevant models. A review of alternative ways of controlling for the overall error rate is provided in Corradi and Distaso (2011), and references contained therein.

2.1.2 Conditional Predictive Ability

In the Diebold-Mariano framework, as well as in the reality check framework, model k and model 0 are considered equally good, in terms of a given loss function, g, if $\mathrm{E}(g(u_{t,0}) - g(u_{t,k})) = 0$. This is a statement about forecasting models. In fact, the null hypothesis is evaluated at the "pseudo-true" value for the parameters. Giacomini and White (GW: 2006) propose a novel approach in which

model k and model 0 are considered equally good if $\mathrm{E}(g(\widehat{u}_{t,0}) - g(\widehat{u}_{t,k})|\mathcal{G}_t) = 0$, where $\mathcal{G}_t$ is an information set, containing (part of) the history available up to time t. The two key differences between unconditional and conditional predictive accuracy tests are: (i) model comparison is based on estimated parameters in the GW approach, rather than on their probability limits, and (ii) models in the GW approach are evaluated according to the expected loss conditional on a given information set $\mathcal{G}_t$, rather than unconditionally. The above is a statement about forecasting methods rather than forecasting models. The notion is that not only the model, but also the way it is estimated matters. Needless to say, if a large number of observations is used for estimation, the estimated parameters get close to their probability limits. For this reason, GW suggest using relatively short observation windows, whose length is fixed and does increase with the sample size. In this way, estimated parameters can be treated as strong mixing random variables.

Recall also that the $\widehat{DM}_P$ is the HAC t-statistic associated with the intercept in a regression of the loss differential series, $g(\widehat{u}_{0,t+1}) - g(\widehat{u}_{1,t+1})$, on a constant. Evidently, DM and subsequent tests are easily made conditional by including other conditioning variables in the HAC regression.

2.2 Bootstrap Critical Values for Recursive Estimation Schemes

Whenever $g \neq q_k$, for at least some k, and $P/R \to \pi \neq 0$, then parameter estimation error contributes to the variance of the limiting distribution of the DM and reality check tests. One reason for using a different loss for estimation and prediction occurs when, for example, we use OLS for estimation, but then we want to use an asymmetric loss function which penalizes positive and negative errors in a different manner, when comparing predictive accuracy (see Zellner, 1986 and Christoffersen and Diebold, 1997). More specifically, when parameter estimation error does not vanish, we need to take into account the contribution of $\frac{1}{\sqrt{P}} \sum_{t=R+\tau}^{T} \left(\widehat{\theta}_{k,t} - \theta_k^{\dagger}\right)$ to the asymptotic variance in (4). Hence, we need a bootstrap procedure which is valid for recursive m-estimators, in the sense that its use suffices to mimic the limiting distribution of $\frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(\widehat{\theta}_{k,t} - \theta_k^{\dagger}\right)$.

One approach to the above issue of parameter estimation error is to use the block bootstrap for recursive m-estimators for constructing critical values. In this context, it is important to note that earlier observations are used more frequently than temporally subsequent observations, when forming test statistics. On the other hand, in the standard block bootstrap, all blocks from the original sample have the same probability of being selected, regardless of the dates of the observations in the blocks. Thus, the bootstrap estimator which is constructed as a direct analog

of $\widehat{\theta}_t$, is characterized by a location bias that can be either positive or negative, depending on the sample that we observe. In order to circumvent this problem, Corradi and Swanson (2007a) suggest a re-centering of the bootstrap score which ensures that the new bootstrap estimator, which is no longer the direct analog of $\widehat{\theta}_{k,t}$, is asymptotically unbiased. It should be noted that the idea of re-centering is not new in the bootstrap literature for the case of full sample estimation. In fact, re-centering is necessary, even for first order validity, in the case of overidentified generalized method of moments (GMM) estimators (see e.g. Hall and Horowitz, 1996, Andrews, 2002, and Inoue and Shintani, 2006). This is due to the fact that, in the overidentified case, the bootstrap moment conditions are not equal to zero, even if the population moment conditions are. However, in the context of $m-$estimators using the full sample, re-centering is needed only for higher order asymptotics, but not for first order validity, in the sense that the bias term is of smaller order than $T^{-1/2}$ (see e.g. Andrews, 2002 and Goncalves and White, 2004). In the case of recursive $m-$estimators, on the other hand, the bias term is instead of order $T^{-1/2}$, so that it does contribute to the limiting distribution. This points to a need for re-centering when using recursive estimation schemes.

To keep notation simple, suppose that we want to predict, y_t using one of its past lags, and one lag of vector of additional variables, X_t, and let $Z_t = (y_t, X_t)$. Using the overlapping block resampling scheme of Künsch (1989), at each replication, we draw b blocks (with replacement) of length l from the sample $W_t = (y_t, Z_{t-1})$, where $bl = T - 1$. Let $W_t^* = (y_t^*, Z_{t-1}^*)$ denote the resampled observations. As a bootstrap counterpart to $\widehat{\theta}_{k,t}$, Corradi and Swanson (2007a) suggest constructing $\widehat{\theta}_{k,t}^*$, defined as follows:

$$\widehat{\theta}_{k,t}^* = \arg\min_{\theta_k} \frac{1}{t}\sum_{j=1}^{t}\left(q_k(y_j^*, Z_{j-1}^*, \theta_k) - \theta_k'\left(\frac{1}{T}\sum_{h=1}^{T-1}\nabla_{\theta_k} q_k(y_h, Z_{h-1}, \widehat{\theta}_{k,t})\right)\right), \tag{5}$$

where $R \leq t \leq T-1$, $k = 0, 1, ..., K$.

Note that $\widehat{\theta}_{k,t}^*$ is not the direct analog of $\widehat{\theta}_{k,t}$ in (1). Heuristically, the additional recentering term in (5) has the role of offsetting the bias that arises due to the fact in the that earlier observations have the same chance of being drawn as

temporally subsequent observations. Theorem 1 in Corradi and Swanson (2007a) establishes that the limiting distribution of $\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widehat{\theta}_{k,t}^* - \widehat{\theta}_{k,t}\right)$ is the same as that of $\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widehat{\theta}_{k,t} - \theta_k^\dagger\right)$, conditional on the sample, and for all samples except a set with probability measure approaching

zero. We can easily see how this result allows for the construction of valid bootstrap critical values for the reality check. Let $\widehat{u}_{k,t+1} = y_{t+1} - \phi_k\left(Z_t, \widehat{\theta}_{k,t}\right)$ and $\widehat{u}^*_{k,t+1} = y^*_{t+1} - \phi_k\left(Z^*_t, \widehat{\theta}^*_{k,t}\right)$, so that the reality check statistic $\widehat{S}_P$ is defined as in (2). The bootstrap counterpart of $\widehat{S}_P$ is given by

$$\widehat{S}^*_P = \max_{k=1,\ldots,K} S^*_P(0,k),$$

where

$$\begin{aligned}\widehat{S}^*_P(0,k) &= \frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left[\left(g(y^*_{t+1} - \phi_0(Z^*_t, \widehat{\theta}^*_{0,t})) - g(y^*_{t+1} - \phi_k(Z^*_t, \widehat{\theta}^*_{k,t}))\right)\right. \\ &\quad \left. - \left\{\frac{1}{T}\sum_{j=1}^{T-1}\left(g(y_{j+1} - \phi_0(Z_j, \widehat{\theta}_{0,t})) - g(y_{j+1} - \phi_k(Z_j, \widehat{\theta}_{k,t}))\right)\right\}\right]. \qquad (6)\end{aligned}$$

It is important to notice that the bootstrap statistic in (6) is different from the "usual" bootstrap statistic, which is defined as the difference between the statistic computed over the sample observations and over the bootstrap observations. In fact, in $\widehat{S}^*_P(0,k)$, the bootstrap (resampled) component is constructed only over the last P observations, while the sample component is constructed over all T observations. The percentiles of the empirical distribution of $\widehat{S}^*_P$ can be used to construct valid bootstrap critical values for $\widehat{S}_P$, in the case of non vanishing parameter estimation error. Their first order validity is established in Proposition 2 in Corradi and Swanson (2007a). Valid bootstrap critical values for the rolling estimation case are outlined in Corradi and Swanson (2006a).

3 Extending the DM and Reality Check Tests to Forecast Interval Evaluation

3.1 The Case of Known Distribution Function

Thus far, we have discussed pointwise predictive accuracy testing (i.e. wherein models are evaluated on the basis of selecting the most accurate pointwise forecasts of a given variable). However, there are several instances in which merely having a "good" model for the conditional mean and/or variance may not be adequate for the task at hand. For example, financial risk management involves tracking the entire distribution of a portfolio, or measuring certain distributional aspects, such as value at risk (see e.g. Duffie and Pan (1997)). In such cases, models of conditional mean

and/or variance may not be satisfactory. A very small subset of important contributions that go beyond the examination of models of conditional mean and/or variance include papers which: assess the correctness of conditional interval predictions (see e.g. Christoffersen (1998)); assess volatility predictability by comparing unconditional and conditional interval forecasts (see e.g. Christoffersen and Diebold (2000)); and assess conditional quantiles (see e.g. Giacomini and Komunjer (2005)). A thorough review of the literature on predictive interval and predictive density evaluation is given in Corradi and Swanson (2006b).

Corradi and Swanson (2006a) extend the DM and reality check tests to predictive density evaluation, and outline a procedure for assessing the relative out-of-sample predictive accuracy of multiple misspecified conditional distribution models, that can be used with rolling and recursive estimation schemes. The objective is to compare these models in terms of their closeness to the true conditional distribution, $F_0(u|Z^t, \theta_0) = \Pr(y_{t+1} \leq u|Z^t)$.[6] In the spirit of White (2000), we choose a particular conditional distribution model as the "benchmark" and test the null hypothesis that no competing model can provide a more accurate approximation of the "true" conditional distribution, against the alternative that at least one competitor outperforms the benchmark model. Following Corradi and Swanson (2005), accuracy is measured using a distributional analog of mean square error. More precisely, the squared (approximation) error associated with model k, $k = 1, ..., K$, is measured in terms of the average over U of $E\left(\left(F_k(u|Z^t, \theta_k^\dagger) - F_0(u|Z^t, \theta_0)\right)^2\right)$, where $u \in U$, and U is a possibly unbounded set on the real line. Additionally, integration over u in the formation of the actual test statistic is governed by $\phi(u) \geq 0$, where $\int_U \phi(u) = 1$. Thus, one can control not only the range of u, but also the weights attached to different values of u, so that more weight can be attached to important tail events, for example. We also consider tests based on an analogous conditional confidence interval version of the above measure. Namely, $E(\left(\left(F_k(\overline{u}|Z^t, \theta_k^\dagger) - F_k(\underline{u}|Z^t, \theta_k^\dagger)\right) - (F_0(\overline{u}|Z^t, \theta_0) - F_0(\underline{u}|Z^t, \theta_0))\right)^2)$, where $\underline{u}$ and $\overline{u}$ are "lower" and "upper" bounds on the confidence interval to be evaluated. For notational simplicity, in the sequel we focus on conditional forecast interval comparison, and set $\underline{u} = -\infty$ and $\overline{u} = u$. For example, we say that model 1 is more accurate than model 2, if

$$E\left(\left(F_1(u|Z^t, \theta_1^\dagger) - F_0(u|Z^t, \theta_0)\right)^2 - \left(F_2(u|Z^t, \theta_2^\dagger) - F_0(u|Z^t, \theta_0)\right)^2\right) < 0.$$

[6] With a slight abuse of notation, in this section the subscript 0 denotes the "true" conditional distribution model, rather than the benchmark model; and the subscript 1 thus now denotes the benchmark model.

This measure defines a norm and it implies a standard goodness of fit measure.

Another measure of distributional accuracy available in the literature is the Kullback-Leibler Information Criterion, KLIC (see e.g. White, 1982, Vuong, 1989, Fernandez-Villaverde and Rubio-Ramirez, 2004, Amisano and Giacomini, 2007, and Kitamura, 2004). According to the KLIC approach, we should choose Model 1 over Model 2 if

$$E(\log f_1(y_{t+1}|Z^t, \theta_1^\dagger) - \log f_2(y_{t+1}|Z^t, \theta_2^\dagger)) > 0.$$

The KLIC is a sensible measure of accuracy, as it chooses the model which on average gives higher probability to events which have actually occurred. The drawback is the KLIC approach cannot be easily generalized to compare conditional intervals.

The hypotheses of interest are formulated as:

$$H_0 : \max_{k=2,...,K} \left(\mu_1^2(u) - \mu_k^2(u)\right) \leq 0$$

versus

$$H_A : \max_{k=2,...,K} \left(\mu_1^2(u) - \mu_k^2(u)\right) > 0,$$

where $\mu_k^2(u) = E\left(\left(1\{y_{t+1} \leq u\} - F_k(u|Z^t, \theta_i^\dagger)\right)^2\right)$, $k = 1, ..., K$. Note that for any given u, $E(1\{y_{t+1} \leq u\}|Z^t) = \Pr(y_{t+1} \leq u|Z^t) = F_0(u|Z^t, \theta_0)$. Thus, $1\{y_{t+1} \leq u\} - F_k(u|Z^t, \theta_i^\dagger)$ can be interpreted as an "error" term associated with computation of the conditional expectation under F_k.

The statistic is:

$$Z_P = \max_{k=2,...,K} Z_{P,u,\tau}(1, k), \tag{7}$$

with

$$Z_{P,u}(1, k) = \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(\left(1\{y_{t+1} \leq u\} - F_1(u|Z^t, \widehat{\theta}_{1,t})\right)^2 - \left(1\{y_{t+1} \leq u\} - F_k(u|Z^t, \widehat{\theta}_{k,t})\right)^2\right),$$

where, as usual, $R + P = T$, and $\widehat{\theta}_{k,t}$ can be either a recursive or a rolling estimator. The limiting distribution of (7) is established in Proposition 1(a) in Corradi and Swanson (2006a), who also suggest how to construct valid bootstrap critical values, for both the recursive and the rolling estimation cases.

3.2 The Case of Unknown Distribution Function

There are cases in which the distribution function is not known in closed form. This problem typically arises when the variable we want to predict is generated by highly nonlinear dynamic models. Very important examples are Dynamic Stochastic General Equilibrium (DSGE) Models, which generally cannot be solved in closed form (see Bierens, 2007, for a discussion of different ways of approximating DSGEs). Since the seminal papers by Kydland and Prescott (1982), Long and Plosser (1983) and King, Plosser and Rebelo (KPR: 1988a,b), there has been substantial attention given to the problem of reconciling the dynamic properties of data simulated from DSGE models, and in particular from real business cycle (RBC) models, with the historical record. A partial list of advances in this area includes: (i) the examination of how RBC simulated data reproduce the covariance and autocorrelation functions of actual time series (see e.g. Watson, 1993); (ii) the comparison of DSGE and historical spectral densities (see e.g. Diebold, Ohanian and Berkowitz, 1998); (iii) the evaluation of the difference between the second order time series properties of vector autoregression (VAR) predictions and out-of-sample predictions from DSGE models (see e.g. Schmitt-Grohe, 2000); (iv) the construction of Bayesian odds ratios for comparing DSGE models with unrestricted VAR models (see e.g. Chang, Gomes and Schorfheide, 2002 and Fernandez-Villaverde and Rubio-Ramirez, 2004); (v) the comparison of historical and simulated data impulse response functions (see e.g. Cogley and Nason, 1995); (vi) the formulation of "Reality" bounds for measuring how close the density of an DSGE model is to the density associated with an unrestricted VAR model (see e.g. Bierens and Swanson, 2000); and (vii) loss function based evaluation of DSGE models (see e.g. Schorfheide, 2000).

The papers cited above evaluate the ability of a given DSGE model to reproduce a particular characteristic of the data. Corradi and Swanson (2007b) use a DM (reality check) approach to evaluate DSGEs in terms of their ability to match (with historical data) the joint distribution of the variables of interest, and provide an empirical application in terms of the comparison of several variants of the stochastic growth model of Christiano (1988). As the distribution function is not known in closed form, we replace it with its simulated counterpart.

To keep notation simple, as above, we consider the case of confidence intervals, setting $\underline{u} = -\infty$, and $u = \infty$. Hereafter, F represents the joint distribution of a variable of interest, say Y_t (e.g. output

growth and hours worked). The hypotheses are:

$$H_0 : \max_{k=2,\ldots,K} \left(\left(F_0(u;\theta_0) - F_1(u;\theta_1^\dagger) \right)^2 - \left(F_0(u;\theta_0) - F_k(u;\theta_k^\dagger) \right)^2 \right) \leq 0$$

$$H_A : \max_{k=2,\ldots,K} \left(\left(F_0(u;\theta_0) - F_1(u;\theta_1^\dagger) \right)^2 - \left(F_0(u;\theta_0) - F_k(u;\theta_k^\dagger) \right)^2 \right) > 0.$$

Thus, under H_0, no model can provide a better approximation of the joint CDF than model 1. In order to test H_0 versus H_A, the relevant test statistic is $\sqrt{T} Z_{T,S}$, where

$$Z_{T,S} = \max_{k=2,\ldots,K} \sqrt{T} Z_{k,T,S}(u), \tag{8}$$

$$\begin{aligned} Z_{k,T,S}(u) &= \frac{1}{T}\sum_{t=1}^{T} \left(1\{Y_t \leq u\} - \frac{1}{S}\sum_{n=1}^{S} 1\{Y_{1,n}(\widehat{\theta}_{1,T}) \leq u\} \right)^2 \\ &\quad - \frac{1}{T}\sum_{t=1}^{T} \left(1\{Y_t \leq u\} - \frac{1}{S}\sum_{n=1}^{S} 1\{Y_{k,n}(\widehat{\theta}_{k,T}) \leq u\} \right)^2, \end{aligned}$$

and $Y_{k,n}(\widehat{\theta}_{k,T})$ represents simulated counterparts of Y_t (i.e. the variables simulated under model k at simulation n, using the estimated parameters $\widehat{\theta}_{k,T}$). Heuristically, if S grows sufficiently fast with respect to T, then $\frac{1}{S}\sum_{n=1}^{S} 1\{Y_{k,n}(\widehat{\theta}_{k,T}) \leq u\}$ can be treated as the "true" distribution of the data simulated under model k. Broadly speaking, we are comparing different DSGE models, on the basis of their ability to match a given simulated joint CDF with that of the historical data. As we are comparing joint CDFs, the statistic in (8) provides an in-sample test.

When constructing the bootstrap counterpart of $Z_{k,T,S}$, we need to distinguish between the case in which $T/S \to 0$ and that in which $T/S \to \delta \neq 0$. Whenever $T/S \to 0$, simulation error is asymptotically negligible, and thus there is no need to resample the simulated observations. In this case, the bootstrap statistic is given by $\max_{k=2,\ldots,K} \sqrt{T} Z^*_{k,T,S}(u)$, where

$$\begin{aligned} & Z^*_{k,T,S}(u) \\ = \; & \frac{1}{T}\sum_{t=1}^{T} \left(\left(1\{Y_t^* \leq u\} - \frac{1}{S}\sum_{n=1}^{S} 1\{Y_{1,n}(\widehat{\theta}^*_{1,T}) \leq u\} \right)^2 - \left(1\{Y_t \leq u\} - \frac{1}{S}\sum_{n=1}^{S} 1\{Y_{1,n}(\widehat{\theta}_{1,T}) \leq u\} \right)^2 \right) \\ & - \frac{1}{T}\sum_{t=1}^{T} \left(\left(1\{Y_t^* \leq u\} - \frac{1}{S}\sum_{n=1}^{S} 1\{Y_{k,n}(\widehat{\theta}^*_{k,T}) \leq u\} \right)^2 - \left(1\{Y_t \leq u\} - \frac{1}{S}\sum_{n=1}^{S} 1\{Y_{k,n}(\widehat{\theta}_{k,T}) \leq u\} \right)^2 \right) \end{aligned} \tag{9}$$

On the other hand, whenever $T/S \to \delta \neq 0$, then simulation error contributes to the limiting distribution. In this case, one has to additionally resample the simulated observations, and thus

$Y_{1,n}(\widehat{\theta}^*_{1,T})$ and $Y_{k,n}(\widehat{\theta}^*_{k,T})$ in (9) should be replaced by $Y^*_{1,n}(\widehat{\theta}^*_{1,T})$ and $Y^*_{k,n}(\widehat{\theta}^*_{k,T})$. In both cases, the validity of bootstrap critical values is been established in Proposition 2 of Corradi and Swanson (2007b).

4 Stochastic Dominance: Predictive Evaluation Based on Distribution of Loss

In this section, we discuss a predictive accuracy testing approach based on distributional loss, as in the previous sections. However, rather than focusing on DM and reality check type approaches, we incorporate notions of stochastic dominance in our analysis. Namely, we introduce a criterion that is designed to include cases of generic predictive accuracy testing, forecast model selection, and forecast combination. The criterion is constructed via evaluation of error loss distributions using basic principles of stochastic dominance, wherein one examines whether or not one CDF lies "above" another, for example. In our discussion, we are concerned only with the evaluation of alternative panels or combinations of forecasts, such as are available when analyzing the Survey of Professional Forecasters (SPF) dataset available on the webpages of the Federal Reserve Bank of Philadelphia. Moreover, we consider first order stochastic dominance. Evidently, the ideas presented here can be adapted to many varieties of predictive accuracy testing, and extension to second and higher order stochastic dominance will also play an important role in such applications. These and related issues are left to future research; and our example below is meant as a starting point in this sort of analysis.

4.1 Motivation

Central Banks and financial institutions have regular access to panels of forecasts for key macroeconomic variables that are made by professional forecasters. A leading example is the SPF. Using this dataset, much focus has centered on how to combine predictions (see e.g. Capistran and Timmermann (2009)) and how to assess forecast rationality (see e.g. Elliott, Komunjer and Timmermann (2008)). With regard to forecast combination, Capistran and Timmermann (2009), as well as Elliott and Timmermann (2004, 2005), estimate combination weights by minimizing a given loss function, ensuring that the weights converge to those minimizing expected loss. Wallis (2005) proposes combining forecasts using a finite mixture distribution, and Smith and Wallis (2009) suggest the use

of simple averages. With regard to rationality assessment, Elliott, Komunjer and Timmermann (2008) test whether forecasters taking part in the SPF are rational for some parameterization of a flexible loss function. This is clearly an important approach when testing for rationality. However, in many instances, users already have a given loss function in mind, and only assess the accuracy of available forecasts under this loss function. Here, we take the loss function as given, and discuss predictive combination and accuracy assessment of datasets such as the SPF. However, this is done via analysis of cumulative loss distributions rather than synthetic measures of loss accuracy such as mean square error and mean absolute error.

More specifically, the objective is to introduce an alternative criterion for predictive evaluation which measures accuracy via examination of the quantiles of the expected loss distribution. The criterion is based on comparing empirical CDFs of predictive error loss, using the principle of stochastic dominance. The heuristic argument underpinning our approach is that the preferred model is one for which the error loss CDF is stochastically dominated by the error loss CDF of every competing model, at all evaluation points. In this sense, a model that has smaller quantiles at all regions of the loss distribution is selected, rather than a model that minimizes a single criterion, such as the mean square error. If a model is not strictly dominated, then our approach allows us to pinpoint the region of the loss distribution for which one model is preferred to another.

As alluded to above, applications for which the criterion is designed include: generic predictive accuracy testing; forecast model selection; and forecast combination. For example, in the context of the SPF, a panel of N_t forecasts for a given variable are made by professionals at each point in time, t. Both the number of individuals taking part in the survey, as well as the specific individuals generally change, from period to period. In this context, the criterion can be applied as follows. Assume that objective is to select and combine forecasts from the SPF. A set of rules, including for example, the simple mean or median across all forecasters, and quantile based weighted combinations across forecasts are defined. Then, the loss function of the forecast errors implied by the rules are evaluated using tests based on the stochastic dominance criterion.

4.2 Set-Up

In each period t, we have a panel of N_t forecasts. The objective is to choose among k possible combinations of the available forecasts, under a given loss function, $g(\cdot)$. In order to allow for frequent possible entry and exit into the panel, the combinations are simple rules, which are applied

each period, regardless of the composition of the panels. Examples are: (i) simple average, (ii) simple average over a given range, such as the 25th-75th percentiles, or (iii) assigning different weights to different interquantile groups from the panel, such as a weight of 0.75 for the average over the 25th-75th percentile and 0.125 for the average over the first and last quartiles.

Define $e_{i,t} = y_t - y^f_{t,h,i}$, $i = 1, ..., k$, to be the forecast error associated with the $h-$step ahead prediction constructed using combination i. Let $g_{i,t} = g(e_{i,t})$, where $g(\cdot)$ is a generic loss function. Also, let $F_{g,i}(x)$ be the empirical distribution of $g(e_{i,t})$ evaluated at x, and let $\widehat{F}_{g,i,T}(x)$ be its sample analog, i.e.

$$\widehat{F}_{g,i,T}(x) = \frac{1}{T}\sum_{t=1}^{T} 1\{g(e_{i,t}) \leq x\}.$$

The hypotheses of interest are:

$$H_0 : \max_{i>1} \inf_{x \in X} (F_{g,1}(x) - F_{g,i}(x)) \geq 0$$

versus

$$H_A : \max_{i>1} \inf_{x \in X} (F_{g,1}(x) - F_{g,i}(x)) < 0.$$

For the sake of simplicity suppose that $k = 2$. If $F_{g,1}(x) - F_{g,2}(x) \geq 0$ for all x, then the CDF associated with rule 1 always lies above the CDF associated with rule 2. Then, heuristically, $g(e_{1,t})$ is (first order) stochastically dominated by $g(e_{2,t})$ and rule 1 is the preferred combination. This is because all of the quantiles of $g(e_{1,t})$ are smaller than the corresponding quantiles of $g(e_{2,t})$. More formally, for a given x, suppose that

$$F_{g,1}(x) = \theta_1 \text{ and } F_{g,2}(x) = \theta_2,$$

then we choose rule 1 if $\theta_1 > \theta_2$. This is because x is the θ_1-quantile under $F_{g,1}$ and the θ_2-quantile under $F_{g,2}$ and, as $\theta_1 > \theta_2$, the θ_2-quantile under $F_{g,1}$ is smaller than under $F_{g,2}$. Thus, for all evaluation points smaller than x, $g(e_{1,t})$ has more probability mass associated with smaller values than $g(e_{2,t})$ does.

It follows that if we fail to reject the null, rule 1 is selected. On the other hand, rejection of the null does not imply that rule 1 should be discarded. Instead, further analysis is required in order to select a rule. First, one needs to discriminate between the cases for which the various CDFs do not cross, and those for which they do cross. This is accomplished by proceeding sequentially as

follows. For all $i \neq j$ $i, j = 1, ..., k$, sequentially test

$$H_0^{i,j} : \sup_{x \in X} \left(F_{g,i}(x) - F_{g,j}(x) \right) \leq 0 \tag{10}$$

versus its negation. Eliminate rule i, if $H_0^{i,j}$ is not rejected. Otherwise, retain rule i. There are two possible outcomes.

I: If there is a rule which is stochastically dominated by all other rules, we eventually discard all the "dominating" rules and remain with only the dominated one. This is always the case when no CDFs cross; but also clearly occurs in cases when various CDFs cross, as long as the dominated CDF cross no other CDF.

II: Otherwise, we remain with a subset of rules, all of which have crossing CDFs, and all of which are stochastically dominated by the eliminated rules.

Note that the logic underlying the outlined sequential procedure is reminiscent of the idea underlying the Model Confidence Set approach of Hansen, Lunde and Nason (2010), in which the worst models are eliminated in a sequential manner, and one remains with a set of models that are roughly equally good, according to the given evaluation criterion.

In the case where there are crossings, further investigation is needed. In particular, in this case, some rules are clearly dominant over certain ranges of loss, and are dominated over others. At this point, one might choose to plot the relevant CDFs, and examine their crossing points. Then, one has to make a choice. For example, one can choose a rule which is dominant over small values of x and is dominated over large values of x. This is the case in which one is concerned about making larger losses than would be incurred, were the other rule used, in a region where losses are large; while not being concerned with the fact that they are making larger losses, relative to those that would be incurred, were the other rule used, when losses are relatively small. Needless to say, one can also use a model averaging approach over the various survivor rules.

4.3 Statistic

In order to test H_0 versus H_A construct the following statistic:

$$L_{g,T} = -\max_{i>1} \inf_{x \in X} \sqrt{T} \left(\widehat{F}_{g,1,T}(x) - \widehat{F}_{g,i,T}(x) \right),$$

where $\widehat{F}_{g,j,T}(x)$, $j \geq 1$, is defined above; and where the negative sign in front of the statistic ensures that the statistic does not diverge to minus infinity under the null hypothesis. On the other hand,

in order to test H_0^{ij}, we instead suggest the following statistic,

$$L_{g,T}^{i,j} = -\sup_{x \in X} \sqrt{T} \left(\widehat{F}_{g,i,T}(x) - \widehat{F}_{g,j,T}(x) \right).$$

In the context of testing for stochastic dominance, Linton, Maasoumi and Whang (2005) construct critical values via subsampling. Here we instead use the "m out of n" bootstrap.[7] Proceed to construct critical values as follows:

(i) We have T observations. Set $\Upsilon < T$.

(ii) Draw b blocks of length l, where $bl = \Upsilon$. One block consists, simultaneously, of draws on the actual data as well as the rule based combination forecasts. Thus, if there are two rules, say, and the block length is 5, then a "block" consists of a 3x1vector of length 5. This yields one bootstrap sample, which is used to construct a bootstrap statistic,

$$L_{g,\Upsilon}^{*} = -\max_{i>1} \inf_{x \in X} \sqrt{\Upsilon} \left(\widehat{F}_{g,1,\Upsilon}^{*}(x) - \widehat{F}_{g,i,\Upsilon}^{*}(x) \right)$$

where

$$\widehat{F}_{g,i,\Upsilon}^{*}(x) = \frac{1}{\Upsilon} \sum_{t=1}^{\Upsilon} 1 \left\{ g^{*}\left(e_{i,t}\right) \leq x \right\}$$

$$g^{*}\left(e_{i,t}\right) = g\left(y_t^{*} - y_{t,h,i}^{*f} \right)$$

.

(iii) Construct B bootstrap statistics and then compute their empirical distribution. The sample statistic is then compared against the percentile of this empirical distribution.

5 Concluding Remarks

In this chapter, we have reviewed the extant literature on Diebold and Mariano (1995) type predictive accuracy testing. We discuss pairwise and multiple model comparison (i.e., DM and reality check type predictive accuracy tests) using differentiable pointwise prediction accuracy measures such as mean square forecast error, as well as using non-differentiable loss functions. We also discuss valid inference under both asymptotically negligible and non-negligible parameter estimation

[7] The basic difference between subsampling and "m out of n" bootstrap is that in the latter case we resample overlapping blocks.

error. Extensions to pairwise and multiple model comparison using predictive densities, distributions, intervals, and conditional distributions are then outlined, with emphasis on inference using these more complicated varieties of DM and reality check type tests. Finally, extension and generalization of all of these testing approaches using notions of stochastic dominance are introduced; and future research directions, including the use of second and higher order stochastic dominance are outlined.

6 References

Amisano, G. and R. Giacomini, 2007, Comparing Density Forecasts via Weighted Likelihood Ratio Tests. Journal of Business and Economic Statistics 25, 177-190.

Andrews, D.W.K., 2002, Higher-Order Improvements of a Computationally Attractive $k-$step Bootstrap for Extremum Estimators, Econometrica 70, 119-162.

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

Bierens, H.J., 2007, Econometric Analysis of Linearized Singular Dynamic Stochastic General Equilibrium Models, Journal of Econometrics, 136, 595-627.

Bierens, H.J., and N.R. Swanson, 2000, The Econometric Consequences of the Ceteris Paribus Condition in Theoretical Economics, Journal of Econometrics, 95, 223-253.

Capistran, C. and A. Timmermann, 2009, Disagreement and Biases in Inflation Expectations, Journal of Money, Credit and Banking 41, 365-396.

Chang, Y.S., J.F. Gomes, and F. Schorfheide, 2002, Learning-by-Doing as a Propagation Mechanism, American Economic Review 92, 1498-1520.

Christiano, L.J., 1988, Why Does Inventory Investment Fluctuate So Much, Journal of Monetary Economics, 21, 247-280.

Christoffersen, P., 1998, Evaluating Interval Forecasts, International Economic Review 39, 841-862.

Christoffersen, P. and F.X. Diebold, 1996, Further Results on Forecasting and Model Selection under Asymmetric Loss, Journal of Applied Econometrics, 11, 561-572.

Christoffersen, P. and F.X. Diebold, 1997, Optimal Prediction Under Asymmetric Loss, Econometric Theory 13, 808-817.

Christoffersen, P. and F.X. Diebold, 2000, How Relevant is Volatility Forecasting for Financial Risk Management?, Review of Economics and Statistics 82, 12-22.

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

Cogley, T., and J.M. Nason, 1995, Output Dynamics for Real Business Cycles Models, American Economic Review, 85, 492-511.

Corradi, V. and N.R. Swanson, 2005, A test for comparing multiple misspecified conditional intervals. Econometric Theory 21, 991-1016.

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: C.W.J. Granger, G. Elliot and A. Timmermann, (Eds.), Handbook of economic forecasting Elsevier, Amsterdam, pp. 197-284.

Corradi, V. and N.R. Swanson, 2007a, Nonparametric bootstrap procedures for predictive inference based on recursive estimation schemes. International Economic Review 48, 67-109.

Corradi, V. and N.R. Swanson, 2007b, Evaluation of dynamic stochastic general equilibrium models based on distributional comparison of simulated and historical data. Journal of Econometrics 136, 699-723.

Corradi, V. and W. Distaso, 2011, Multiple Forecast Evaluation, Oxford Handbook of Economic Forecasting, eds. D.F. Hendry and M.P. Clements, Oxford University Press, Oxford.

Diebold, F.X. and R.S. Mariano, 1995, Comparing predictive accuracy. Journal of Business and Economic Statistics 13, 253-263.

Diebold, F.X., L.E. Ohanian, and J. Berkowitz, 1998, Dynamic Equilibrium Economies: A Framework for Comparing Models and Data, Review of Economic Studies 65, 433-451.

Duffie, D. and J. Pan, 1997, An Overview of Value at Risk, Journal of Derivatives 4, 7-49.

Elliott, G. and A. Timmermann, 2004, Optimal Forecast Combinations Under General Loss Functions and Forecast Error Distributions, Journal of Econometrics 122, 47-79.

Elliott, G. and A. Timmermann, 2005, Optimal Forecast Combination Under Regime Switching, International Economic Review 46, 1081-1102.

Elliott, G., Komunjer I., and A. Timmermann, 2008, Biases in Macroeconomic Forecasts: Irrationality or Asymmetric Loss? 2008, Journal of the European Economic Association, 6, 122-157

Fernandez-Villaverde, J. and J.F. Rubio-Ramirez, 2004, Comparing Dynamic Equilibrium Models to Data, Journal of Econometrics 123, 153-180.

Giacomini, R. and I. Komunjer, 2005, Evaluation and Combination of Conditional Quantile Forecasts, Journal of Business and Economic Statistics 23, 416-431.

Giacomini, R., and H. White, 2006, Conditional Tests for Predictive Ability, Econometrica 74, 1545-1578.

Goncalves, S., and H. White, 2004, Maximum Likelihood and the Bootstrap for Nonlinear Dynamic Models, Journal of Econometrics, 119, 199-219.

Hall, P., and J.L. Horowitz, 1996, Bootstrap Critical Values for Tests Based on Generalized Method of Moments Estimators. Econometrica 64, 891-916.

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

Hansen, P.R., A. Lunde and J.M. Nason, 2010, The Model Confidence Set, Econometrica, forthcoming.

Inoue, A., and M. Shintani, 2006, Bootstrapping GMM Estimators for Time Series, Journal of Econometrics 133, 531-555.

King, R.G., C.I. Plosser, and S.T. Rebelo, 1988a, Production, Growth and Business Cycles 1: The Basic Neoclassical Model, Journal of Monetary Economics 21, 195-232.

King, R.G., C.I. Plosser, and S.T. Rebelo, 1988b, Production, Growth and Business Cycles 2: New Directions, Journal of Monetary Economics 21, 309-341.

Kitamura, Y., 2004, Econometric Comparisons of Conditional Models, Working Paper, Yale University.

Künsch H.R., 1989, The Jackknife and the Bootstrap for General Stationary Observations. Annals of Statistics 17, 1217-1241.

Kydland, F.E., and E.C. Prescott, 1982, Time to Build and Aggregate Fluctuations, Econometrica 50, 1345-1370.

Linton, O., E. Maasoumi and Y.J. Whang, 2005, Consistent Testing for Stochastic Dominance Under General Sampling Schemes, Review of Economic Studies 72, 735-765.

Long, J.B. and C.I. Plosser, 1983, Real Business Cycles, Journal of Political Economy 91, 39-69.

Romano, J.P. and M. Wolf, 2005, Stepwise Multiple Testing as Formalized Data Snooping, Econometrica 73, 1237-1282.

Schorfheide, F., 2000, Loss Function Based Evaluation of DSGE Models, Journal of Applied Econometrics 15, 645-670.

Schmitt-Grohe, S., 2000, Endogenous Business Cycles and the Dynamics of Output, Hours and Consumption, American Economic Review 90, 1136-1159.

Smith, J. and K.F. Wallis, 2009, A Simple Explanation of the Forecast Combination Puzzle, Oxford Bullettin of Economics and Statistics 71, 331-355.

Sullivan R., A. Timmermann, and H. White, 1999, Data-Snooping, Technical Trading Rule Performance, and the Bootstrap, Journal of Finance 54, 1647-1691.

Sullivan, R., A. Timmermann, and H. White, 2001, Dangers of Data-Driven Inference: The Case of Calendar Effects in Stock Returns, Journal of Econometrics 105, 249– 286.

Vuong, Q., 1989, Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses, Econometrica 57, 307-333.

Wallis, K.F., 2005, Combining Interval and Density Forecast: A Modest Proposal, Oxford Bullettin of Economics and Statistics 67, 983-994.

Watson, M.W., 1993, Measure of Fit for Calibrated Models, Journal of Political Economy 101, 1011-1041.

West, K.F., 1996, Asymptotic Inference About Predictive Ability, Econometrica 64, 1067-1084.

West, K.F., and M.W. McCracken, 1998, Regression Based Tests for Predictive Ability, International Economic Review 39, 817-840.

White, H., 1982, Maximum Likelihood Estimation of Misspecified Models, Econometrica 50, 1-25.

White, H., 2000, A reality check for Data Snooping, Econometrica 68, 1097-1126.

Zellner, A., 1986, Bayesian Estimation and Prediction Using Asymmetric Loss Function, Journal of the American Statistical Association 81, 446-451.