Corradi, Valentina; Swanson, Norman R.

Working Paper

Predictive Density Evaluation

Working Paper, No. 2004-19

Provided in Cooperation with:
Department of Economics, Rutgers University

Suggested Citation: Corradi, Valentina; Swanson, Norman R. (2004) : Predictive Density Evaluation, Working Paper, No. 2004-19, Rutgers University, Department of Economics, New Brunswick, NJ

This Version is available at:
https://hdl.handle.net/10419/23196

Predictive Density Evaluation*

Valentina Corradi[1] and Norman R. Swanson[2]

[1]Queen Mary, University of London and [2]Rutgers University

September 2004

Abstract

This chapter discusses estimation, specification testing, and model selection of predictive density models. In particular, predictive density estimation is briefly discussed, and a variety of different specification and model evaluation tests due to various authors including Christoffersen and Diebold (2000), Diebold, Gunther and Tay (1998), Diebold, Hahn and Tay (1999), White (2000), Bai (2003), Corradi and Swanson (2003 and 2004(a),(b),(c)), Hong and Li (2003), and others are reviewed. Extensions of some existing techniques to the case of out-of-sample evaluation are also provided, and asymptotic results associated with these extensions are outlined.

---

* Valentina Corradi, Department of Economics, Queen Mary, University of London, Mile End Road, London E1 4NS, UK, v.corradi@qmul.ac.uk. Norman R. Swanson, Department of Economics, Rutgers University, 75 Hamilton Street, New Brunswick, NJ 08901, USA, nswanson@econ.rutgers.edu. This article was prepared for the Handbook of Economic Forecasting. The authors owe great thanks to Clive W.J. Granger, whose discussions provided much of the impetus for the authors' own research that is reported in this paper. Corradi gratefully acknowledges ESRC grant RES-000-23-0006, and Swanson acknowledges financial support from a Rutgers University Research Council grant.

Outline

Part I: Introduction

1. Estimation, Specification Testing, and Model Evaluation

Table 1. Summary of Selected Specification Testing and Model Evaluation Papers

Part II: Testing for Correct Specification of Conditional Distributions for the Entire or a Given Information Set

2. Specification Testing and Model Evaluation In-Sample

2.1 Diebold, Gunther and Tay Approach - Probability Integral Transform

2.2 Bai Approach - Martingalization

2.3 Hong and Li Approach - A Nonparametric Test

2.4 Corradi and Swanson Approach - A Parametric Test

2.5 Bootstrap Critical Values for the V_{1T} and V_{2T} Tests

2.6 *Other Related Work*

3. Specification Testing and Model Evaluation Out-of-Sample

3.1 Parameter Estimation Error in Recursive and Rolling Estimation Schemes - West as well as West and McCracken Results

3.2 Out-of-Sample Implementation of Bai as well as Hong and Li Tests

3.3 Out-of-Sample Implementation of Corradi and Swanson Tests

3.4 Bootstrap Critical Values for $V_{1P,J}$ and $V_{2P,J}$ Tests Under Recursive Estimation

3.4.1 Recursive PEE Bootstrap

3.4.2 $V_{1P,J}$ and $V_{2P,J}$ Bootstrap Statistics Under Recursive Estimation

3.5 Bootstrap Critical Values for $V_{1P,J}$ and $V_{2P,J}$ Tests Under Rolling Estimation

3.5.1Rolling PEE Bootstrap

3.5.2 $V_{1P,J}$ and $V_{2P,J}$ Bootstrap Statistics Under Rolling Estimation

Part III: Evaluation of (Multiple) Misspecified Predictive Models

4. Pointwise Comparison of (Multiple) Misspecified Predictive Models

4.1 Comparison of Two Nonnested Models: Diebold and Mariano Test

4.2 Comparison of Two Nested Models

4.2.1 Clark and McCracken Tests

4.2.2 Chao, Corradi and Swanson Tests

4.3 Comparison of Multiple Models: The Reality Check

4.3.1 White's Reality Check and Extensions

4.3.2 Hansen's Approach to the Reality Check

4.3.3 The Subsampling Approach Applied to the Reality Check

4.3.4 The False Discovery Rate Approach Applied to the Reality Check

4.4 A Predictive Accuracy Test That is Consistent Against Generic Alternatives

Part I: Introduction

1 Estimation, Specification Testing, and Model Evaluation

The topic of predictive density evaluation has received considerable attention in economics and finance over the last few years, a fact which is not at all surprising when one notes the importance of predictive densities to virtually all public and private institutions involved with the construction and dissemination of forecasts. As a case in point, consider the plethora conditional mean forecasts reported by the news media. These sorts of predictions are not very useful for economic decision making unless confidence intervals are also provided. Indeed, there is a clear need when forming macroeconomic policies and when managing financial risk in the insurance and banking industries to use predictive confidence intervals or entire predictive conditional distributions. One such case is when value at risk measures are constructed in order to assess the amount of capital at risk from small probability events, such as catastrophes (in insurance markets) or monetary shocks that have large impact on interest rates (see Duffie and Pan (1997) for further discussion). In this chapter we shall discuss some of the tools that are useful in such situations, with particular focus on estimation, specification testing, and model evaluation.[1]

There are many important historical precedents for predictive density estimation, testing, and model selection. From the perspective of estimation, the parameters characterizing distributions, conditional distributions and predictive densities can be constructed using innumerable well established techniques, including maximum likelihood, (simulated generalized) methods of moments, and a plethora of other estimation techniques. Additionally, one can specify parametric models, nonparametric models, and semi-parametric models. For example, a random variable of interest, say y_t, may be assumed to have a particular distribution, say $F(u|\theta_0) = P(y \leq u|\theta_0) = \Phi(u) = \int_{-\infty}^{u} f(y)dy$, where $f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(y-\mu)^2}{2\sigma^2}}$. Here, the consistent maximum likelihood estimator of θ_0 is $\widehat{\mu} = n^{-1}\sum_{t=1}^{T} y_t$, and $\widehat{\sigma}^2 = n^{-1}\sum_{t=1}^{T}(y_t - \widehat{\mu})^2$, where T is the sample size. This example corresponds to the case where the variable of interest is a martingale difference sequence and so there is no potentially useful (conditioning) information which may help in prediction. Then, the predictive density for y_t is simply $\widehat{f}(y) = \frac{1}{\widehat{\sigma}\sqrt{2\pi}} e^{\frac{-(y-\widehat{\mu})^2}{2\widehat{\sigma}^2}}$. Alternatively, one may wish to use a nonpara-

[1] In this chapter, the distinction that is made between specification testing and model evaluation (or predictive accuracy testing) is predicated on the fact that specification tests often consider only one model. Such tests usually attempt to ascertain whether the model is misspecified, and they usually assume correct specification under the null hypothesis. On the other hand, predictive accuracy tests compare multiple models and should (in our view) allow for various forms of misspecification, under both hypotheses.

metric estimator. For example, if the functional form of the distribution is unknown, one might choose to construct a kernel density estimator. In this case, one would construct $\widehat{f}(y) = \frac{1}{T\lambda}\sum_{t=1}^{T}\kappa\left(\frac{y_t - y}{\lambda}\right)$, where κ is a kernel function and λ is the bandwidth parameter that satisfies a particular rate condition in order to ensure consistent estimation, such as $\lambda = O\left(T^{-1/5}\right)$. Nonparametric density estimators converge to the true underlying density at a nonparametric (slow) rate. For this reason, a valid alternative is the use of empirical distributions, which instead converge to the cumulative distribution (CDF) at a parametric rate (see e.g. Andrews (1993) for a thorough overview of empirical distributions, and empirical processes in general). In particular, the empirical distribution is crucial in our discussion of predictive density because it is useful in estimation, testing, and model evaluation; and has the property that $\frac{1}{\sqrt{T}}\sum_{i=1}^{T}(1\{y_t \leq u\} - F(u|\theta_0))$ satisfies a central limit theorem.

Of course, in economics it is natural to suppose that better predictions can be constructed by conditioning on other important economic variables, say. Indeed, discussions of predictive density are usually linked to discussions of conditional distribution, where we define conditioning information as $Z^t = (y_{t-1}, ..., y_{t-v}, X_t, ..., X_{t-w})$ with v, w finite, and where X_t may be vector valued. In this context, we could define a parametric model, say $F(u|Z^t, \theta)$ to characterize the conditional distribution $F_0(u|Z^t, \theta_0) = \Pr(Y_t \leq u|Z^t)$. Needless to say, our model would be misspecified, unless $F = F_0$.

Alternatively, one may wish to estimate and evaluate a group of alternative models, say $F_1(u|Z^t, \theta_1^\dagger), ..., F_m(u|Z^t, \theta_m^\dagger)$, where the parameters in these distributions correspond to the probability limits of the estimated parameters, and m is the number of models to be estimated and evaluated. Estimation in this context can be carried out in much the same way as when unconditional models are estimated. For example, one can construct a conditional distribution model by postulating that $y_t|Z^t \sim N(\theta' Z^t, \sigma^2)$, estimate θ by least square, σ^2 using least square residual and then forming predictive confidence intervals or the entire predictive density. The foregoing discussion underscores the fact that there are numerous well established estimation techniques which one can use to estimate predictive density models, and hence which one can use to make associated probabilistic statements such as: "There is 0.9 probability, based on the use of my particular model, that inflation next period will lie between 4 and 5 percent." Indeed, for a discussion of estimation, one need merely pick up any basic or advanced statistics and/or econometrics text. Naturally, and as one might expect, the appropriateness of a particular estimation technique hinges on two factors. The first is the nature of the data. Marketing survey data are quite different from aggregate measures of economic activity, and there are well established literatures describing appropriate models and estimation techniques for these and other varieties of data, from spatial to panel, and from time series to cross sectional. Given that there is already a

huge literature on the topic of estimation, we shall hereafter assume that the reader has at her/his disposal software and know-how concerning model estimation (for some discussion of estimation in cross sectional, panel, and time series models, for example, the reader might refer to Baltagi (1995), Bickel and Doksum (2001), Davidson and MacKinnon (1993), Hamilton (1996), White (1994), and Wooldridge (2002), to name but a very few). The second factor upon which the appropriateness of a particular estimation strategy hinges concerns model specification. In the context of model specification and evaluation, it is crucial to make it clear in empirical settings whether one is assuming that a model is correctly specified (prior to estimation), or whether the model is simply an approximation, possibly from amongst a group of many "approximate models", from whence some "best" predictive density model is to be selected. The reason this assumption is important is because it impacts on the assumed properties of the residuals from the first stage conditional mean regression in the above example, which in turn impacts on the validity and appropriateness of specification testing and model evaluation techniques that are usually applied after a model has been estimated.

The focus in this chapter is on the last two issues, namely specification testing and model evaluation. One reason why we are able to discuss both of these topics in a (relatively) short handbook chapter is that the literature on the subjects is not near so large as that for estimation; although it is currently growing at an impressive rate! The fact that the literature in these areas is still relatively underdeveloped is perhaps surprising, given that the "tools" used in specification testing and model evaluation have been around for so long, and include such important classical contributions as the Kolmogorov-Smirnov test (see e.g. Kolmogorov (1933) and Smirnov (1939)), various results on empirical processes (see e.g. Andrews (1993) and the discussion in chapter 19 of van der Vaart (1998) on the contributions of Glivenko, Cantelli, Doob, Donsker and others), the probability integral transform (see e.g. Rosenblatt (1952)), and the Kullback-Leibler Information Criterion (see e.g. White (1982) and Vuong (1989)). However, the immaturity of the literature is perhaps not so surprising when one considers that many of the contributions in the area depend upon recent advances including results validating the use of the bootstrap (see e.g. Horowitz (2001)) and the invention of crucial tools for dealing with parameter estimation error (see e.g. Khmaladze (1981,1988) and West (1996)), for example.

We start by outlining various contributions which are from the literature on (consistent) specification testing (see e.g. Bierens (1982,1990) and Bierens and Ploberger (1997)). An important feature of such tests is that if one subsequently carries out a series of these tests, such as when one performs a series of specification tests using alternative conditional distributions (e.g. the conditional Kolmogorov-Smirnov test of Andrews

(1997)), then sequential test bias arises (i.e. critical values may be incorrectly sized, and so inference based on such sequential tests may be incorrect). Additionally, it may be difficult in some contexts to justify the assumption under the null that a model is correctly specified, as we may want to allow for possible dynamic misspecification under the null, for example. After all, if two tests for the correct specification of two different models are carried out sequentially, then surely one of the models is misspecified under the null, implying that the critical values of one of the two tests may be incorrect, as we shall shortly illustrate. It is in this sense that the idea of model evaluation in which a group of models are jointly compared, and in which case all models are allowed to be misspecified, is important, particularly from the perspective of prediction. Also, there are many settings for which the objective is not to find the correct model, but rather to select the "best" model (based on a given metric or loss function to be used for predictive evaluation) from amongst a group of models, all of which are approximations to some underlying unknown model. Nevertheless, given that advances in multiple model comparison under misspecification derive to a large extent from earlier advances in (correct) specification testing, and given that specification testing and model evaluation are likely most powerful when used together, we shall discuss tools and techniques in both areas.

Although a more mature literature, there is still a great amount of activity in the area of tests for the correct specification of conditional distributions. One reason for this is that testing for the correct conditional distribution is equivalent to jointly evaluating many conditional features of a process, including the conditional mean, variance, and symmetry. Along these lines, Inoue (1999) constructs tests for generic conditional aspects of a distribution, and Bai and Ng (2001) construct tests for conditional asymmetry. These sorts of tests can be generalized to the evaluation of predictive intervals and predictive densities, too.

One group of tests that we discuss along these lines is that due to Corradi and Swanson (2003). In their paper, they construct Kolmogorov type conditional distribution tests in the presence of both dynamic misspecification and parameter estimation error. As shall be discussed shortly, the approach taken by these authors differs somewhat from much of the related literature because they construct a statistics that allow for dynamic misspecification under both hypotheses, rather than assuming correct dynamic specification under the null hypothesis. This difference can be most easily motivated within the framework used by Diebold, Gunther and Tay (DGT: 1998), Hong (2001), and Bai (2003). In their paper, DGT use the probability integral transform to show that $F_t(y_t|\Im_{t-1}, \theta_0)$ is identically and independently distributed as a uniform random variable on $[0,1]$, where $F_t(\cdot|\Im_{t-1}, \theta_0)$ is a parametric distribution with underlying parameter θ_0, y_t is again our random variable of interest, and $\Im_{t-1}$ is the information set containing all "relevant" past information (see below for further discussion). They thus suggest using the difference between the empirical distribution

of $F_t(y_t|\Im_{t-1}, \widehat{\theta}_T)$ and the $45^\circ-$degree line as a measure of "goodness of fit", where $\widehat{\theta}_T$ is some estimator of θ_0. This approach has been shown to be very useful for financial risk management (see e.g. Diebold, Hahn and Tay (1999)), as well as for macroeconomic forecasting (see e.g. Diebold, Tay and Wallis (1998) and Clements and Smith (2000,2002)). Likewise, Bai (2003) proposes a Kolmogorov type test of $F_t(u|\Im_{t-1}, \theta_0)$ based on the comparison of $F_t(y_t|\Im_{t-1}, \widehat{\theta}_T)$ with the CDF of a uniform on $[0, 1]$. As a consequence of using estimated parameters, the limiting distribution of his test reflects the contribution of parameter estimation error and is not nuisance parameter free. To overcome this problem, Bai (2003) uses a novel approach based on a martingalization argument to construct a modified Kolmogorov test which has a nuisance parameter free limiting distribution. This test has power against violations of uniformity but not against violations of independence (see below for further discussion). Hong (2001) proposes another related interesting test, based on the generalized spectrum, which has power against both uniformity and independence violations, for the case in which the contribution of parameter estimation error vanishes asymptotically. If the null is rejected, Hong (2001) also proposes a test for uniformity robust to non independence, which is based on the comparison between a kernel density estimator and the uniform density. All of these tests are discussed in detail below. In summary, two features differentiate the tests of Corradi and Swanson (CS: 2003) from the tests outlined in the other papers mentioned above. First, CS assume strict stationarity. Second, CS allow for dynamic misspecification under the null hypothesis. The second feature allows CS to obtain asymptotically valid critical values even when the conditioning information set does not contain all of the relevant past history. More precisely, assume that we are interested in testing for correct specification, given a particular information set which may or may not contain all of the relevant past information. This is important when a Kolmogorov test is constructed, as one is generally faced with the problem of defining $\Im_{t-1}$. If enough history is not included, then there may be dynamic misspecification. Additionally, finding out how much information (e.g. how many lags) to include may involve pre-testing, hence leading to a form of sequential test bias. By allowing for dynamic misspecification, such pre-testing is not required.

To be more precise, critical values derived under correct specification given $\Im_{t-1}$ are not in general valid in the case of correct specification given a subset of $\Im_{t-1}$. Consider the following example. Assume that we are interested in testing whether the conditional distribution of $y_t|y_{t-1}$ is $N(\alpha_1^\dagger y_{t-1}, \sigma_1)$. Suppose also that in actual fact the "relevant" information set has $\Im_{t-1}$ including both y_{t-1} and y_{t-2}, so that the true conditional model is $y_t|\Im_{t-1} = y_t|y_{t-1}, y_{t-2} = N(\alpha_1 y_{t-1} + \alpha_2 y_{t-2}, \sigma_2)$, where $\alpha_1^\dagger$ differs from α_1. In this case, correct specification holds with respect to the information contained in y_{t-1}; but there is dynamic misspecification with respect to y_{t-1}, y_{t-2}. Even without taking account of parameter estimation error, the

critical values obtained assuming correct dynamic specification are invalid, thus leading to invalid inference. Stated differently, tests that are designed to have power against both uniformity and independence violations (i.e. tests that assume correct dynamic specification under H_0) will reject; an inference which is incorrect, at least in the sense that the "normality" assumption is *not* false. In summary, if one is interested in the particular problem of testing for correct specification for a given information set, then the CS approach is appropriate, while if one is instead interested in testing for correct specification assuming that $\Im_{t-1}$ is known, that the other tests discussed above are useful - these are some of the tests discussed in the second part of this chapter, and all are based on probability integral transforms and Kolmogorov Smirnov distance measures.

In the third part of this chapter, attention is turned to the case of model evaluation. Much of the development in this area stems from earlier work in the area of point evaluation, and hence various tests of conditional mean models for nested and nonnested models, both under assumption of correct specification, and under the assumption that all models should be viewed as "approximations", are first discussed. These tests include important ones by Diebold and Mariano (1995), West (1996), White (2000), and many others. Attention is then turned to a discussion of predictive density selection. To illustrate the sort of model evaluation tools that are discussed, consider the following. Assume that we are given a group of (possibly) misspecified conditional distributions, $F_1(u|Z^t, \theta_1^\dagger), ..., F_m(u|Z^t, \theta_m^\dagger)$, and assume that the objective is to compare these models in terms of their "closeness" to the true conditional distribution, $F_0(u|Z^t, \theta_0) = \Pr(Y_{t+1} \leq u|Z^t)$. Corradi and Swanson (2004a,b) consider such a problem. If $m > 2$, they follow White (2000), in the sense that a particular conditional distribution model is chosen as the "benchmark" and one tests 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. However, unlike White, they evaluate predictive densities rather than point forecasts. Pairwise comparison of alternative models, in which no benchmark needs to be specified, follows from their results as a special case. In their context, accuracy is measured using a distributional analog of mean square error. More precisely, the squared (approximation) error associated with model i, $i = 1, ..., m$, is measured in terms of $E\left(\left(F_i(u|Z^{t+1}, \theta_i^\dagger) - F_0(u|Z^{t+1}, \theta_0)\right)^2\right)$, where $u \in U$, and U is a possibly unbounded set on the real line. The case of evaluation of multiple conditional confidence interval models is analyzed too.

Another well known measure of distributional accuracy which is also discussed in Part 3 is the Kullback-Leibler Information Criterion (KLIC). The KLIC is useful because the "most accurate" model can shown to be that which minimizes the KLIC (see below for more details). Using the KLIC approach, Giacomini (2002)

suggests a weighted version of the Vuong (1989) likelihood ratio test for the case of dependent observations, while Kitamura (2002) employs a KLIC based approach to select among misspecified conditional models that satisfy given moment conditions. Furthermore, the KLIC approach has been recently employed for the evaluation of dynamic stochastic general equilibrium models (see e.g. Schorfheide (2000), Fernandez-Villaverde and Rubio-Ramirez (2004), and Chang, Gomes and Schorfheide (2002)). For example, Fernandez-Villaverde and Rubio-Ramirez (2004) show that the KLIC-best model is also the model with the highest posterior probability. In general, there is no reason why either of the above two measures of accuracy is more "natural". These tests are discussed in detail in the chapter.

As a further preamble to this chapter, we now present a table which summarizes selected testing and model evaluation papers. The list of papers in the table is undoubtedly incomplete, but nevertheless serves as a rough benchmark to the sorts of papers and results that are discussed in this chapter. The primary reason for including the table is to summarize in a directly comparable manner the assumptions made in the various papers. Later on, assumptions are given as they appear in the original papers.

Table 1: Summary of Selected Specification Testing and Model Evaluation Papers

Paper	*Eval*	*Test*	*Misspec*	*Loss*	*PEE*	*Horizon*	*Nesting*	*CV*
Bai (2003)[1]	S	CD	C	NA	Yes	$h = 1$	NA	Standard
Corradi and Swanson (2003)[2]	S	CD	D	NA	Yes	$h = 1$	NA	Boot
Diebold, Gunther and Tay (1998)[2]	S	CD	C	NA	No	$h = 1$	NA	NA
Hong (2001)	S	CD	C,D,G	NA	No	$h = 1$	NA	Standard
Hong and Li (2003)[1]	S	CD	C,D,G	NA	Yes	$h = 1$	NA	Standard
Chao, Corradi and Swanson (2001)	S	CM	D	D	Yes	$h \geq 1$	NA	Boot
Clark and McCracken (2001,2003)	S,P	CM	C	D	Yes	$h \geq 1$	N,A	Boot,Standard
Corradi and Swanson (2002)[3]	S	CM	D	D	Yes	$h \geq 1$	NA	Boot
Corradi and Swanson (2004a)	M	CD	G	D	Yes	$h \geq 1$	O	Boot
Corradi, Swanson and Olivetti (2001)	P	CM	C	D	Yes	$h \geq 1$	O	Standard
Diebold, Hahn and Tay (1999)	M	CD	C	NA	No	$h \geq 1$	NA	NA
Diebold and Mariano (1995)	P	CM	G	N	No	$h \geq 1$	O	Standard
Giacomini (2002)	P	CD	G	NA	Yes	$h \geq 1$	A	Standard
Giacomini and White (2002)[5]	P	CM	G	D	Yes	$h \geq 1$	A	Standard
Li and Tcakz (2004)	S	CD	C	NA	Yes	$h \geq 1$	NA	Standard
Rossi (2003)	P	CM	C	D	Yes	$h \geq 1$	O	Standard
Thompson (2002)	S	CD	C	NA	Yes	$h \geq 1$	NA	Standard
West (1996)	P	CM	C	D	Yes	$h \geq 1$	O	Standard
White (2000)[4]	M	CM	G	N	Yes	$h \geq 1$	O	Boot

Notes: The table provides a summary of various tests currently available. For completeness, some tests of conditional mean are also included, particularly when they have been, or could be, extended to the case of conditional distribution evaluation. Many tests are considered ancilliary, or have been ommitted due to ignorance. Many other tests are discussed in the papers cited in this table. "NA" entries denote "Not Applicable". Columns and mnemonics used are defined as follows:

* $Eval$ = Evaluation is of: Single Model (S); Pair of Models (P); Multiple Models (M).
* $Test$ = Test is of: Conditional Distribution (CD); Conditional Mean (CM).
* $Misspec$ = Misspecification assumption under H_0: Correct Specification (C); Dynamic Misspecification Allowed (D); General Misspecification Allowed (G).
* $Loss$ = Loss function assumption: Differentiable (D); May be Non-differntiable (N).
* PEE = Parameter estimation error: Accounted for (yes); Not Accounted for (no).
* $Horizon$ = Prediction horizon: 1-step ($h = 1$); Multi-step ($h \geq 1$).
* $Nesting$ = Assumption vis nestedness of models: (At least one) Nonnested Model Required (O); Nested Models (N); Any Combination (A).
* CV = Critical values constructed via: Standard Limiting Distribution or Nuisance Parameter Free Nonstandard Distribution (Standard); Bootstrap or Other Procedure (Boot).

[1] See extension in this paper to out-of-sample case.
[2] Extension to multiple horizon follows straightforwardly if the marginal distribution of the errors is normal, for example; otherwise extension is not always straightforward.
[3] This is the only predictive accuracy test from the listed papers that is consistent against generic (nonlinear) alternatives.
[4] See extention in this paper to predictive density evaluation, allowing for parameter estimation error.
[5] Parameters are estimated using a fixed window of observations, so that parameters do not approach their probability limits, but are instead treated as mixing variables under the null hypothesis.

Part II: Testing for Correct Specification of Conditional Distributions for the Entire or a Given Information Set

2 Specification Testing and Model Evaluation In-Sample

There are several instances in which a "good" model for the conditional mean and/or variance is not 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 these cases, the choice of the best loss function specific model for the conditional mean may not be of too much help.

Important contributions that go beyond the examination of models of conditional mean include assessing the correctness of conditional interval prediction (Christoffersen (1998)) and assessing volatility predictability by comparing unconditional and conditional interval forecasts (Christoffersen and Diebold (2000)).[2] Needless to say, correct specification of the conditional distribution implies correct specification of all conditional aspects of the model. Perhaps in part for this reason, there has been growing interest in recent years in providing tests for the correct specification of conditional distributions. In this section, we analyze the issue of testing for the correct specification of the conditional distribution, distinguishing between the case in which we condition on the entire history and that in which we condition on a given information set, thus allowing for dynamic misspecification. In particular, we illustrate with some detail recent important work by Diebold, Gunther and Tay (1998), based on the probability integral transformation (see also Diebold, Hahn and Tay (1999) and Christoffersen and Diebold (2000)); by Bai (2003), based on Kolmogorov tests and martingalization techniques; by Hong (2001), based on the notion of generalized cross-spectrum; and by Corradi and Swanson (2003), based on Kolmogorov type tests. We begin by considering the in-sample version of the tests, in which the same set of observations is used for both estimation and testing. Further, we provide an out-of-sample version of these tests, in which the first subset of observations is used for estimation and the last subset is used for testing. In the out-of-sample case, parameters are generally estimated using either a recursive or a rolling estimation scheme. Thus, we first review important result by West (1996) and West and McCracken (1998) about the limiting distribution of $m-$estimators and GMM estimators in

[2] Prediction confidence intervals are also discussed in Granger, White and Kamstra (1989), Chatfield (1993), Diebold, Tay and Wallis (1998), Clements and Taylor (2001), and the references cited therein.

the recursive and rolling case, respectively. As pointed in section 2.3.3 below, asymptotic critical values for both the in-sample and out-of-sample versions of the statistic by Corradi and Swanson can be obtained via an application of the bootstrap. While the asymptotic behavior of (full sample) bootstrap $m-$estimators is already well known, see the literature cited below, this is no longer true for the case of bootstrap estimators based on either a recursive or a rolling scheme. This issue is addressed by Corradi and Swanson (2004a, 2004c) and summarized in sections 2.3.4.1 and 2.3.4.2 below.

2.1 Diebold, Gunther and Tay Approach - Probability Integral Transform

In a key paper in the field, Diebold, Gunther and Tay (DGT: 1998) use the probability integral transform (see e.g. Rosenblatt (1952)) to show that $F_t(y_t|\Im_{t-1},\theta_0) = \int_{-\infty}^{y_t} f_t(y|\Im_{t-1},\theta_0)$, is identically and independently distributed as a uniform random variable on $[0,1]$, whenever $F_t(y_t|\Im_{t-1},\theta_0)$ is dynamically correctly specified for the CDF of $y_t|\Im_{t-1}$. Thus, they suggest to use the difference between the empirical distribution of $F_t(y_t|\Im_{t-1},\widehat{\theta}_T)$ and the $45^\circ-$degree line as a measure of "goodness of fit", where $\widehat{\theta}_T$ is some estimator of θ_0. Visual inspection of the plot of this difference gives also some information about the deficiency of the candidate conditional density, and so may suggest some way of improving it. The univariate framework of DGT is extended to a multivariate framework in Diebold, Hahn and Tay (DHT: 1999), in order to allow to evaluate the adequacy of density forecasts involving cross-variable interactions. This approach has been shown to be very useful for financial risk management (see e.g. DGT (1998) and DHT (1999)), as well as for macroeconomic forecasting (see Diebold, Tay and Wallis (1998), where inflation predictions based on professional forecasts are evaluated, and see Clements and Smith (2000), where predictive densities based on nonlinear models of output and unemployment are evaluated). Important closely related work in the area of the evaluation of volatility forecasting and risk management is discussed in Christoffersen and Diebold (2000). Additional tests based on the DGT idea of comparing the empirical distribution of $F_t(y_t|\Im_{t-1},\widehat{\theta}_T)$ with the $45^\circ-$degree line have been suggested by Bai (2003), Hong (2001), Hong and Lee (2003), and Corradi and Swanson (2003).

2.2 Bai Approach - Martingalization

Bai (2003) considers the following hypotheses:

$$H_0 \quad : \quad \Pr(y_t \leq y|\Im_{t-1},\theta_0) = F_t(y|\Im_{t-1},\theta_0), \text{ a.s. for some } \theta_0 \in \Theta \tag{1}$$

$$H_A \quad : \quad \text{the negation of } H_0, \tag{2}$$

where $\Im_{t-1}$ contains all the relevant history up to time $t-1$. In this sense, the null hypotheses corresponds with dynamic correct specification of the conditional distribution.

Bai (2003) proposes a Kolmogorov type test based on the comparison of $F_t(y|\Im_{t-1},\theta_0)$ with the CDF of a uniform random variable on $[0,1]$. In practice, we need to replace the unknown parameters, θ_0, with an estimator, say $\widehat{\theta}_T$. Additionally, we often do not observe the full information set $\Im_{t-1}$, but only a subset of it, say $Z^t \subseteq \Im_{t-1}$. Therefore, we need to approximate $F_t(y|\Im_{t-1},\theta_0)$ with $F_t(y|Z^{t-1},\widehat{\theta}_T)$. Hereafter, for notational simplicity, define

$$\widehat{U}_t = F_t(y_t|Z^{t-1},\widehat{\theta}_T) \tag{3}$$

$$\widetilde{U}_t = F_t(y_t|Z^{t-1},\theta^\dagger) \tag{4}$$

$$U_t = F_t(y_t|\Im_{t-1},\theta_0), \tag{5}$$

where $\theta^\dagger = \theta_0$ whenever Z^{t-1} contains all useful information in $\Im_{t-1}$, so that in this case $\widetilde{U}_t = U_t$. As a consequence of using estimated parameters, the limiting distribution of his test reflects the contribution of parameter estimation error and is not nuisance parameter free. In fact, as shown in his eqs. (1)-(4),

$$\begin{aligned}
\widehat{V}_T(r) &= \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left(1\{\widehat{U}_t \le r\} - r\right) \\
&= \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left(1\{\widetilde{U}_t \le r\} - r\right) + \overline{g}(r)'\sqrt{T}\left(\widehat{\theta}_T - \theta^\dagger\right) + o_P(1), \\
&= \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left(1\{U_t \le r\} - r\right) + \overline{g}(r)'\sqrt{T}\left(\widehat{\theta}_T - \theta_0\right) + o_P(1)
\end{aligned} \tag{6}$$

where the last equality holds only if Z^{t-1} contains all useful information in $\Im_{t-1}$.[3] Here,

$$\overline{g}(r) = \text{plim}_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\frac{\partial F_t}{\partial\theta}(x|Z^{t-1},\theta^\dagger)|_{x=F_t^{-1}(r|Z^{t-1},\theta^\dagger)}.$$

Also, let

$$g(r) = (1, \overline{g}(r)').$$

To overcome the nuisance parameter problem, Bai uses a novel approach based on a martingalization argument to construct a modified Kolmogorov test which has a nuisance parameter free limiting distribution. In

[3]Note that $\widehat{U}_t$ should be defined for $t > s$, where s is the largest lags contained in the information set Z^{t-1}, however for notational simplicity we start all sumamation from $t = 1$, as if $s = 0$.

particular, let $\dot{g}$ be the derivative of g, and let $C(r) = \int_r^1 \dot{g}(\tau)\dot{g}(\tau)'d\tau$. Bai's test statistic (eq. (5), p. 533) is defined as:

$$\widehat{W}_T(r) = \widehat{V}_T(r) - \int_0^r \left(\dot{g}(s)C^{-1}(s)\dot{g}(s)' \int_s^1 \dot{g}(\tau)d\widehat{V}_T(\tau) \right) ds, \tag{7}$$

where the second term may be difficult to compute, depending on the specific application. Several examples, including GARCH models and (self-exciting) threshold autoregressive models are provided in Section IIIB of Bai (2003). The limiting distribution of the statistic in (7) is obtained under the following assumptions, where it is of note that stationarity is not required. (Note also that **BAI4** below rules out non-negligible differences between the information in Z^{t-1} and $\Im_{t-1}$, with respect to the model of interest).

BAI1: $F_t(y_t|Z^{t-1},\theta)$ and its density $f_t(y_t|Z^{t-1},\theta)$ are continuously differentiable in θ. $F_t(y|Z^{t-1},\theta)$ is strictly increasing in y, so that F_t^{-1} is well defined. Also,

$$E \sup_x \sup_\theta f_t(y_t|Z^{t-1},\theta) \leq M_1 < \infty$$

and

$$E \sup_x \sup_\theta \left\| \frac{\partial F_t}{\partial \theta}(x|Z^{t-1},\theta) \right\| \leq M_1 < \infty,$$

where the supremum is taken over all θ, such that $|\theta - \theta^\dagger| \leq MT^{-1/2}$, $M < \infty$.

BAI2: There exists a continuously differentiable function $\overline{g}(r)$, such that for every $M > 0$,

$$\sup_{\substack{u,v \\ |u-\theta^\dagger|<MT^{-1/2}, |v-\theta^\dagger|<MT^{-1/2}}} \left\| \frac{1}{T}\sum_{t=1}^T \frac{\partial F_t}{\partial \theta}(F_t^{-1}(r|u)|v) - \overline{g}(r) \right\| = o_P(1),$$

where the $o_P(1)$ is uniform in $r \in [0,1]$. In addition, $\int_0^1 \left\| \dot{g}(r) \right\| dr < \infty$, $C(r) = \int_r^1 \dot{g}(\tau)\dot{g}(\tau)'d\tau$ is invertible for all r.

BAI3: $\sqrt{T}\left(\widehat{\theta}_T - \theta^\dagger\right) = O_P(1)$.

BAI4: The effect of using Z^{t-1} instead of $\Im_{t-1}$ is negligible. That is,

$$\sup_{u,|u-\theta_0|<MT^{-1/2}} T^{-1/2}\sum_{t=1}^T \left| F_t\left(F_t^{-1}(r|Z^{t-1},u)|\Im_{t-1},\theta_0\right) - F_t\left(F_t^{-1}(r|\Im_{t-1},u)|\Im_{t-1},\theta_0\right)\right| = o_P(1)$$

Given this setup, the following result can be proven.

Theorem 2.1 (from Corollary 1 in Bai (2003)): Let **BAI1-BAI4** hold, then under H_0,

$$\sup_{r\in[0,1]} \left|\widehat{W}_T(r)\right| \xrightarrow{d} \sup_{r\in[0,1]} |W(r)|,$$

where $W(r)$ is a standard Brownian motion. Therefore, the limiting distribution is nuisance parameter free and critical values can be tabulated.

Now, suppose there is dynamic misspecification, so that $\Pr(y_t \leq y|\Im_{t-1}, \theta_0) \neq \Pr(y_t \leq y|Z^{t-1}, \theta^\dagger)$. In this case, critical values relying on the limiting distribution in Theorem 2.1 are no longer valid. However, if $F(y_t|Z^{t-1}, \theta^\dagger)$ is correctly specified for $\Pr(y_t \leq y|Z^{t-1}, \theta^\dagger)$, uniformity still holds, and there is no guarantee that the statistic diverges. Thus, while Bai's test has unit asymptotic power against violations of uniformity, is does not have unit asymptotic power against violations of independence. Note that in the case of dynamic misspecification, assumption **BAI4** is violated. Also, the assumption cannot be checked from the data, in general. In summary, the limiting distribution of Kolmogorov type tests is affected by dynamic misspecification. Critical values derived under correct dynamic specification are not in general valid in the case of correct specification given a subset of the full information set. Consider the following example. Assume that we are interested in testing whether the conditional distribution of $y_t|y_{t-1}$ is $N(\alpha_1^\dagger y_{t-1}, \sigma_1)$. Suppose also that in actual fact the "relevant" information set has Z^{t-1} including both y_{t-1} and y_{t-2}, so that the true conditional model is $y_t|Z^{t-1} = y_t|y_{t-1}, y_{t-2} = N(\alpha_1 y_{t-1} + \alpha_2 y_{t-2}, \sigma_2)$, where $\alpha_1^\dagger$ differs from α_1. In this case, we have correct specification with respect to the information contained in y_{t-1}; but we have dynamic misspecification with respect to y_{t-1}, y_{t-2}. Even without taking account of parameter estimation error, the critical values obtained assuming correct dynamic specification are invalid, thus leading to invalid inference.

2.3 Hong and Li Approach - A Nonparametric Test

As mentioned above, the Kolmogorov test of Bai does not necessarily have power against violations of independence. A test with power against violations of both independence and uniformity has been recently suggested by Hong and Li (2003), who also draw on results by Hong (2001). Their test is based on the comparison of the joint nonparametric density of $\widehat{U}_t$ and $\widehat{U}_{t-j}$, as defined in (3), with the product of two $UN[0,1]$ random variables. In particular, they introduce a boundary modified kernel which ensures a "good" nonparametric estimator, even around 0 and 1. This forms the basis for a test which has power against both non-uniformity and non-independence. For any $j > 0$, define

$$\widehat{\phi}(u_1, u_2) = (n-j)^{-1} \sum_{\tau=j+1}^{n} K_h(u_1, \widehat{U}_\tau) K_h(u_2, \widehat{U}_{\tau-j}), \tag{8}$$

where

$$K_h(x,y) = \begin{cases} h^{-1}(\frac{x-y}{h}) / \int_{-(x/h)}^{1} k(u)du & \text{if } x \in [0,h) \\ h^{-1}(\frac{x-y}{h}) & \text{if } x \in [h, 1-h) \\ h^{-1}(\frac{x-y}{h}) / \int_{-1}^{(1-x)/h} k(u)du & \text{if } x \in [1-h, 1] \end{cases} \tag{9}$$

In the above expression, h defines the bandwidth parameter, although in later sections (where confusion cannot easily arise), h is used to denote forecast horizon. As an example, one might use,

$$k(u) = \frac{15}{16}(1-u^2)^2 1\{|u| \leq 1\}.$$

Also, define

$$\widehat{M}(j) = \int_0^1 \int_0^1 \left(\widehat{\phi}(u_1, u_2) - 1\right)^2 du_1 du_2 \tag{10}$$

and

$$\widehat{Q}(j) = \left((n-j)\widehat{M}(j) - A_h^0\right)/V_0^{1/2}, \tag{11}$$

with

$$A_h^0 = \left((h^{-1}-2)\int_{-1}^1 k^2(u)du + 2\int_0^1 \int_{-1}^b k_b(u)dudb\right)^2 - 1,$$

$$k_b(\cdot) = k(\cdot)/\int_{-1}^b k(v)dv,$$

and

$$V_0 = 2\left(\int_{-1}^1 \left(\int_{-1}^1 k(u+v)k(v)dv\right)^2 du\right)^2.$$

The limiting distribution of $\widehat{Q}(j)$ is obtained by Hong and Li (2003) under the following assumptions:[4]

HL1: (y_t, Z^{t-1}) are strong mixing with mixing coefficients $\alpha(\tau)$ satisfying $\sum_{\tau=0}^{\infty} \alpha(\tau)^{(v-1).v} \leq C < \infty$, with $v > 1$.

HL2: $f_t(y|Z^t, \theta)$ is twice continuously differentiable in θ, in a neighborhood of θ_0, and $\lim_{T\to\infty} \sum_{\tau=1}^n E\left|\frac{\partial U_t}{\partial \theta}\right|^4 \leq C$, $\lim_{T\to\infty} \sum_{\tau=1}^n E \sup_{\theta\in\Theta}\left|\frac{\partial^2 U_t}{\partial\theta\partial\theta'}\right|^2 \leq C$, for some constant C.

HL3: $\sqrt{T}(\widehat{\theta}_T - \theta^\dagger) = O_P(1)$, where $\theta^\dagger$ is the probability limit of $\widehat{\theta}_T$, and is equal to θ_0, under the null in (1).

HL4: The kernel function $k : [-1,1] \to \Re^+$ is a symmetric, bounded, twice continuously differentiable probability density, such that $\int_{-1}^1 k(u)du = 0$ and $\int_{-1}^1 k^2(u)du < \infty$.

Given this setup, the following result can be proven.

Theorem 2.2 (from Theorem 1 in Hong and Li (2003): Let **HL1-HL4** hold. If $h = cT^{-\delta}$, $\delta \in (0, 1/5)$, then under H_0 (i.e. see (1)), for any $j > 0$, $j = o(T^{1-\delta(5-2/v)})$, $\widehat{Q}(j) \xrightarrow{d} N(0,1)$.

[4]Hong et al. specialize their test to the case of testing continuous time models. However, as they point out, it is equally valid for discrete time models.

Once the null is rejected, it remains of interest to know whether the rejection is due to violation of uniformity or to violation of independence (or both). Broadly speaking, violations of independence arises in the case of dynamic misspecification (Z^t does not contain enough information), while violations of uniformity arise when we misspecify the functional form of f_t when constructing $\widehat{U}_t$. Along these lines, Hong (2001) proposes a test for uniformity, which is robust to dynamic misspecification. Define, the hypotheses of interest as:

$$\begin{aligned} H_0 &: \Pr(y_t \leq y | Z^{t-1}, \theta^\dagger) = F_t(y | Z^{t-1}, \theta^\dagger), \ a.s. \text{ for some } \theta_0 \in \Theta \\ H_A &: \text{the negation of } H_0, \end{aligned} \tag{12}$$

where $F_t(y|Z^{t-1}, \theta^\dagger)$ may differ from $F_t(y|\Im_{t-1}, \theta_0)$. The relevant test is based on the comparison of a kernel estimator of the marginal density of $\widehat{U}_t$ with the uniform density, and has a standard normal limiting distribution under the null in (12). Hong (2001) also provides a test for the null of independence, which is robust to violations of uniformity.

Note that the limiting distribution in Theorem 2.2, as well as the limiting distribution of the uniformity (independence) test which is robust to non uniformity (non independence) in Hong (2001) are all asymptotically standard normal, regardless of the fact that we construct the statistic using $\widehat{U}_t$ instead on U_t. This is due to the feature that parameter estimators converge at rate $T^{1/2}$, while the statistics converge at non-parametric rates. The choice of the bandwidth parameter and the slower rate of convergence are thus the prices to be paid for not having to directly account for parameter estimation error.

2.4 Corradi and Swanson Approach

Corradi and Swanson (2003) suggest a test for the null hypothesis of correct specification of the conditional distribution, for a given information set which is, as usual, called Z^t, and which, as above, does not necessarily contain all relevant historical information. The test is again a Kolmogorov type test, and is based on the fact that under the null of correct (but not necessarily dynamically correct) specification of the conditional distribution, U_t is distributed as $[0, 1]$. As with Hong's (2001) test, this test is thus robust to violations of independence. As will become clear below, the advantages of the test relative to that of Hong (2001) is that it converges at a parametric rate and there is no need to choose the bandwidth parameter. The disadvantage is that the limiting distribution is not nuisance parameters free and hence one needs to rely on bootstrap

techniques in order to obtain valid critical values. Define:

$$V_{1T} = \sup_{r\in[0,1]} |V_{1T}(r)|, \tag{13}$$

where,

$$V_{1T}(r) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left(1\{\widehat{U}_t \leq r\} - r \right),$$

and

$$\widehat{\theta}_T = \arg\max_{\theta\in\Theta} \frac{1}{T} \sum_{t=1}^{T} \ln f(y_t|X_t,\theta).$$

Note that the above statistic is similar to that of Bai (2003). However, there is no "extra" term to cancel out the effect of parameter estimation error. The reason is that Bai's martingale transformation argument does not apply to the case in which the score is not a martingale difference process (so that (dynamic) misspecification is not allowed for when using his test).

The standard rationale underlying the above test, which is known to hold when $Z^{t-1} = \Im_{t-1}$, is that under H_0 (given above as (12)), $F(y_t|Z^{t-1},\theta_0)$ is distributed independently and uniformly on $[0,1]$. The uniformity result also holds under dynamic misspecification. To see this, let $c_f^r(Z^{t-1})$ be the $r-th$ critical value of $f(\cdot|Z^{t-1},\theta_0)$, where f is the density associated with $F(\cdot|Z^{t-1},\theta_0)$ (i.e. the conditional distribution under the null)[5]. It then follows that,

$$\begin{aligned} \Pr(F(y_t|Z^{t-1},\theta_0) &\leq r) = \Pr\left(\int_{-\infty}^{y_t} f(y|Z^{t-1},\theta_0)dy \leq r \right) \\ &= \Pr\left(1\{y_t \leq c_f^r(Z^{t-1})\} = 1|Z^{t-1} \right) = r, \text{ for all } r\in[0,1], \end{aligned}$$

if $y_t|Z^{t-1}$ has density $f(\cdot|Z^{t-1},\theta_0)$. Now, if the density of $y_t|\ Z^{t-1}$ is different from $f(\cdot|Z^{t-1},\theta_0)$, then,

$$\Pr\left(1\{y_t \leq c_f^r(Z^{t-1})\} = 1|Z^{t-1} \right) \neq r,$$

for some r with nonzero Lebesgue measure on $[0,1]$. However, under dynamic misspecification, $F(y_t|Z^{t-1},\theta_0)$ is no longer independent (or even martingale difference), in general, and this will clearly affect the covariance structure of the limiting distribution of the statistic. Theorem 2.3 below relies on the following assumptions.

CS1: (y_t, Z^{t-1}), are jointly strictly stationary and strong mixing with size $-4(4+\psi)/\psi$, $0<\psi<1/2$.

CS2: (i) $F(y_t|Z^{t-1},\theta)$ is twice continuously differentiable on the interior of $\Theta \subset R^p$, Θ compact; (ii) $E(\sup_{\theta\in\Theta} |\nabla_\theta F(y_t|Z^t,\theta)_i|^{5+\psi}) \leq C < \infty$, $i = 1,...,p$, where ψ is the same positive constant defined in A1, and $\nabla_\theta F(y_t|Z^{t-1},\theta)_i$ is the $i-$th element of $\nabla_\theta F(y_t|Z^{t-1},\theta)$; (iii) $F(u|Z^{t-1},\theta)$ is twice differentiable on the

[5]For example, if $f(Y|X_t,\theta_0) \sim N(\alpha X_t, \sigma^2)$, then $c_f^{0.95}(X_t) = 1.645 + \sigma\alpha X_t$.

interior of $U \times \Theta$, where U and Θ are compact subsets of $\Re$ and $\Re^p$ respectively; and (iv) $\nabla_\theta F(u|Z^{t-1},\theta)$ and $\nabla_{u,\theta} F(u|Z^{t-1},\theta)$ are jointly continuous on $U \times \Theta$ and $4s-$dominated on $U \times \Theta$ for $s > 3/2$.

CS3: (i) $\theta^\dagger = \arg\max_{\theta\in\Theta} E(\ln f(y_1|Z^0,\theta))$ is uniquely identified, (ii) $f(y_t|Z^{t-1},\theta)$ is twice continuously differentiable in θ in the interior of Θ, (ii) the elements of $\nabla_\theta \ln f(y_t|Z^{t-1},\theta)$ and of $\nabla^2_\theta \ln f(y_t|Z^{t-1},\theta)$ are $4s-$dominated on Θ, with $s > 3/2$, $E\left(-\nabla^2_\theta \ln f(y_t|Z^{t-1},\theta)\right)$ is positive definite uniformly in Θ.[6]

Of note is that **CS2** imposes mild smoothness and moment restrictions on the cumulative distribution function under the null, and is thus easily verifiable. Also, we use **CS2**(i)-(ii) in the study of the limiting behavior of V_{1T} and **CS2**(iii)-(iv) in the study of V_{2T}.

Theorem 2.3 (from Theorem 1 in Corradi and Swanson (2003)): Let **CS1**, **CS2**(i)–(ii) and **CS3** hold. Then: (i) Under H_0, $V_{1T} \Rightarrow \sup_{r\in[0,1]} |V_1(r)|$, where V is a zero mean Gaussian process with covariance kernel $K_1(r,r')$ given by:

$$E(V_1(r)V_1(r')) = K_1(r,r') = E(\sum_{s=-\infty}^{\infty} \left(1\{F(y_1|Z^0,\theta_0) \le r\} - r\right)\left(1\{F(y_s|Z^{s-1},\theta_0) \le r'\} - r'\right))$$

$$+E(\nabla_\theta F(x(r)|Z^{t-1},\theta_0))'A(\theta_0) \sum_{s=-\infty}^{\infty} E(q_1(\theta_0)q_s(\theta_0)')A(\theta_0)E(\nabla_\theta F(x(r')|Z^{t-1},\theta_0))$$

$$-2E(\nabla_\theta F(x(r)|Z^{t-1},\theta_0))'A(\theta_0) \sum_{s=-\infty}^{\infty} E(\left(1\{F(y_1|Z^0,\theta_0) \le r\} - r\right) q_s(\theta_0)'),$$

with $q_s(\theta_0) = \nabla_\theta \ln f_s(y_s|Z^{s-1},\theta_0)$, $x(r) = F^{-1}(r|Z^{t-1},\theta_0)$, and $A(\theta_0) = \left(E\left(\nabla_\theta q_s(\theta_0)\nabla_\theta q_s(\theta_0)'\right)\right)^{-1}$.

(ii) Under H_A, there exists an $\varepsilon > 0$ such that $\lim_{T\to\infty} \Pr(\frac{1}{T^{1/2}} V_{1T} > \varepsilon) = 1$.

Notice that the limiting distribution is a zero mean Gaussian process, with a covariance kernel that reflects both dynamic misspecification as well as the contribution of parameter estimation error. Thus, the limiting distribution is not nuisance parameter free and so critical values cannot be tabulated.

Corradi and Swanson (2003) also suggest another Kolmogorov test, which is no longer based on the probability integral transformation, but can be seen as an extension of the conditional Kolmogorov (CK) test of Andrews (1997) to the case of time series data and possible dynamic misspecification.

In a related important paper, Li and Tkacz (2004) discuss an interesting approach to testing for correct specification of the conditional density which involves comparing a nonparametric kernel estimate of the conditional density with the density implied under the null hypothesis. As in Hong and Li (2003) and Hong (2001), the Tkacz and Li test is characterized by a nonparametric rate. Of further note is that Whang

[6] Let $\nabla_\theta \ln f(y_t|X_t,\theta)_i$ be the $i-th$ element of $\nabla_\theta \ln f(y_t|X_t,\theta)$. For $4s-$domination on Θ, we require $|\nabla_\theta \ln f(y_t|X_t,\theta)_i| \le m(X_t)$, for all i, with $E((m(X_t))^{4s}) < \infty$, for some function m.

(2000,2001) also proposes a version of Andrews' CK test for the correct specification, although his focus is on conditional mean, and not conditional distribution.

This test is constructed by comparing the empirical joint distribution of y_t and Z^{t-1} with the product of the distribution of $y_t|Z^t$ and the empirical CDF of Z^{t-1}. In practice, the empirical joint distribution, say $\widehat{H}_T(u,v) = \frac{1}{T}\sum_{t=1}^{T} 1\{y_t \leq u\}1\{Z^{t-1} < v\}$, and the semi-empirical/semi-parametric analog of $F(u,v,\theta_0)$, say $\widehat{F}_T(u,v,\widehat{\theta}_T) = \frac{1}{T}\sum_{t=1}^{T} F(u|Z^{t-1},\widehat{\theta}_T)1\{Z^{t-1} < v\}$ are used, and the test statistic is:

$$V_{2T} = \sup_{u\times v \in U\times V} |V_{2T}(u,v)|, \tag{14}$$

where U and V are compact subsets of $\Re$ and $\Re^d$, respectively, and

$$V_{2T}(u,v) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left((1\{y_t \leq u\} - F(u|Z^{t-1},\widehat{\theta}_T))1\{Z^{t-1} \leq v\}\right).$$

Note that V_{2T} is given in equation (3.9) of Andrews (1997).[7] Note also that when computing this statistic, a grid search over $U \times V$ may be computationally demanding when V is high-dimensional. To avoid this problem, Andrews shows that when all (u,v) combinations are replaced with (y_t, X_t) combinations, the resulting test is asymptotically equivalent to $V_{2T}(u,v)$.

Theorem 2.4 (from Theorem 2 in Corradi and Swanson (2003)):

Let **CS1**, **CS2**(iii)–(iv) and **CS3** hold. Then: (i) Under H_0, $V_{2T} \Rightarrow \sup_{u\times v\in U\times V}|Z(u,v)|$, where V_{2T} is defined in (14) and Z is a zero mean Gaussian process with covariance kernel $K_2(u,v,u',v')$ given by:

$$E(\sum_{s=-\infty}^{\infty}((1\{y_1 \leq u\} - F(u|Z^0,\theta_0))1\{X_0 \leq v\})((1\{y_s \leq u'\} - F(u|Z^{s-1},\theta_0))1\{X_s \leq v'\}))$$

$$+E(\nabla_\theta F(u|Z^0,\theta_0)'1\{Z^0 \leq v\})A(\theta_0)\sum_{s=-\infty}^{\infty} q_0(\theta_0)q_s(\theta_0)'A(\theta_0)E(\nabla_\theta F(u'|Z^0,\theta_0)1\{Z^0 \leq v'\})$$

$$-2\sum_{s=-\infty}^{\infty}((1\{y_0 \leq u\} - F(u|Z^0,\theta_0))1\{Z^0 \leq v\})E(\nabla_\theta F(u'|Z^0,\theta_0)'1\{Z^0 \leq v'\})A(\theta_0)q_s(\theta_0)).$$

(ii) Under H_A, there exists an $\varepsilon > 0$ such that $\lim_{T\to\infty}\Pr(\frac{1}{T^{1/2}}V_{2T} > \varepsilon) = 1$.

As in Theorem 2.3, the limiting distribution is a zero mean Gaussian process with a covariance kernel that reflects both dynamic misspecification as well as the contribution of parameter estimation error. Thus, the limiting distribution is not nuisance parameter free and so critical values cannot be tabulated. Below, we outline a bootstrap procedure that takes into account the joint presence of parameter estimation error and possible dynamic misspecification.

[7] Andrews (1997), for the case of *iid* observations, actually addresses the more complex situation where U and V are unbounded sets in R and R^d, respectively. We believe that an analogous result for the case of dependent observations holds, but showing this involves proofs for stochastic equicontinuity which are quite demanding.

2.5 Bootstrap Critical Values for the V_{1T} and V_{2T} Tests

Given that the limiting distributions of V_{1T} and V_{2T} are not nuisance parameter free, one approach is to construct bootstrap critical values for the tests. In order to show the first order validity of the bootstrap, it thus remains to obtain the limiting distribution of the bootstrapped statistic and show that it coincides with the limiting distribution of the actual statistic under H_0. Then, a test with correct asymptotic size and unit asymptotic power can be obtained by comparing the value of the original statistic with bootstrapped critical values.

If the data consists of *iid* observations, we should consider proceeding along the lines of Andrews (1997), by drawing B samples of T *iid* observations from the distribution under H_0, conditional on the observed values for the covariates, Z^{t-1}. The same approach could also be used in the case of dependence, if H_0 were correct dynamic specification, (i.e. if $Z^{t-1} = \Im_{t-1}$); in fact, in that case we could use a parametric bootstrap and draw observations from $F(y_t|Z^t, \widehat{\theta}_T)$. However, if instead $Z^{t-1} \subset \Im_{t-1}$, using the parametric bootstrap procedure based on drawing observations from $F(y_t|Z^{t-1}, \widehat{\theta}_T)$ does not ensure that the long run variance of the resampled statistic properly mimics the long run variance of the original statistic; thus leading in general to the construction of invalid asymptotic critical values.

The approach used by Corradi and Swanson (2003) involves comparing the empirical CDF of the resampled series, evaluated at the bootstrap estimator, with the empirical CDF of the actual series, evaluated at the estimator based on the actual data. For this, they use the overlapping block resampling scheme of Künsch (1989), as follows:[8] At each replication, draw b blocks (with replacement) of length l from the sample $W_t = (y_t, Z^{t-1})$, where $T = lb$. Thus, the first block is equal to $W_{i+1}, ..., W_{i+l}$, for some i, with probability $1/(T-l+1)$, the second block is equal to $W_{i+1}, ..., W_{i+l}$, for some i, with probability $1/(T-l+1)$, and so on for all blocks. More formally, let I_k, $k = 1, ..., b$ be *iid* discrete uniform random variables on $[0, 1, ..., T-l]$, and let $T = bl$. Then, the resampled series, $W_t^* = (y_t^*, X_t^*)$, is such that $W_1^*, W_2^*, ..., W_l^*, W_{l+1}^*, ..., W_T^* = W_{I_1+1}, W_{I_1+2}, ..., W_{I_1+l}, W_{I_2}, ..., W_{I_b+l}$, and so a resampled series consists

[8] Alternatively, one could use the stationary bootstrap of Politis and Romano (1994(a)(b)). The main difference between the block bootstrap and the stationary bootstrap of Politis and Romano (PR: 1994a) is that the former uses a deterministic block length, which may be either overlapping as in Künsch (1989) or non-overlapping as in Carlstein (1986), while the latter resamples using blocks of random length. One important feature of the PR bootstrap is that the resampled series, conditional on the sample, is stationary, while a series resampled from the (overlapping or non overlapping) block bootstrap is nonstationary, even if the original sample is strictly stationary. However, Lahiri (1999) shows that all block boostrap methods, regardless of whether the block length is deterministic or random, have a first order bias of the same magnitude, but the bootstrap with deterministic block length has a smaller first order variance. In addition, the overlapping block boostrap is more efficient than the non overlapping block bootstrap.

of b blocks that are discrete iid uniform random variables, conditional on the sample. Also, let $\widehat{\theta}_T^*$ be the estimator constructed using the resampled series. For V_{1T}, the bootstrap statistic is:

$$V_{1T}^* = \sup_{r\in[0,1]} |V_{1T}^*(r)|,$$

where

$$V_{1T}^*(r) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left(1\{F(y_t^*|Z^{*,t-1}, \widehat{\theta}_T^*) \leq r\} - 1\{F(y_t|Z^{t-1}, \widehat{\theta}_T) \leq r\} \right), \tag{15}$$

and

$$\widehat{\theta}_T^* = \arg\max_{\theta\in\Theta} \frac{1}{T} \sum_{t=1}^{T} \ln f(y_t^*|Z^{*,t-1}, \theta).$$

The rationale behind the choice of (15) is the following. By a mean value expansion it can be shown that,

$$\begin{aligned} V_{1T}^*(r) &= \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left(1\{F(y_t^*|Z^{*,t-1}, \theta^\dagger) \leq r\} - 1\{F(y_t|Z^{t-1}, \theta^\dagger) \leq r\} \right) \\ &\quad - \frac{1}{T} \sum_{t=1}^{T} \nabla_\theta F(y_t|Z^{t-1}, \theta^\dagger) \sqrt{T}(\widehat{\theta}_T^* - \widehat{\theta}_T) + o_{P^*}(1), \ \Pr - P, \end{aligned} \tag{16}$$

where P^* denotes the probability law of the resampled series, conditional on the sample; P denotes the probability law of the sample; and where "$o_{P^*}(1)$, $\Pr - P$", means a term approaching zero according to P^*, conditional on the sample and for all samples except a set of measure approaching zero. Now, the first term on the RHS of (16) can be treated via the empirical process version of the block bootstrap, suggesting that the term has the same limiting distribution as $\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left(1\{F(y_t|Z^{t-1}, \theta^\dagger) \leq r\} - E\left(1\{F(y_t|Z^{t-1}, \theta^\dagger) \leq r\}\right) \right)$, where $E\left(1\{F(y_t|X_t, \theta^\dagger) \leq r\}\right) = r$ under H_0, and is different from r under H_A, conditional of the sample. If $\sqrt{T}(\widehat{\theta}_T^* - \widehat{\theta}_T)$ has the same limiting distribution as $\sqrt{T}(\widehat{\theta}_T - \theta^\dagger)$, conditionally on the sample and for all samples except a set of measure approaching zero, then the second term on the RHS of (16) will properly capture the contribution of parameter estimation error to the covariance kernel. For the case of dependent observations, the limiting distribution of $\sqrt{T}(\widehat{\theta}_T^* - \widehat{\theta}_T)$ for a variety of quasi maximum likelihood (QMLE) and GMM estimators has been examined in numerous papers in recent years.

For example, Hall and Horowitz (1996) and Andrews (2002) show that the block bootstrap provides improved critical values, in the sense of asymptotic refinement, for "studentized" GMM estimators and for tests of overidentifying restrictions, in the case where the covariance across moment conditions is zero after a given number of lags.[9] In addition, Inoue and Shintani (2004) show that the block bootstrap provides

[9] Andrews (2002) shows first order validity and asymptotic refinements of the equivalent $k-$step estimator of Davidson and MacKinnon (1999), which only requires the construction of a closed form expression at each bootstrap replication, thus avoiding nonlinear optimization at each replication.

asymptotic refinements for linear overidentified GMM estimators for general mixing processes. In the present context, however, one cannot "studentize" the statistic, and we are thus unable to show second order refinement, as mentioned above. Instead, and again as mentioned above, the approach of Corradi and Swanson (2003) is to show first order validity of $\sqrt{T}(\widehat{\theta}_T^* - \widehat{\theta}_T)$. An important recent contribution which is useful in the current context is that of Goncalves and White (2002,2004) who show that for QMLE estimators, the limiting distribution of $\sqrt{T}(\widehat{\theta}_T^* - \widehat{\theta}_T)$ provides a valid first order approximation to that of $\sqrt{T}(\widehat{\theta}_T - \theta^\dagger)$ for heterogeneous and near epoch dependent series.

Theorem 2.5 (from Theorem 3 of Corradi and Swanson (2003)): Let **CS1**, **CS2**(i)–(ii) and **CS3** hold, and let $T = bl$, with $l = l_T$, such that as $T \to \infty$, $l_T^2/T \to 0$. Then,

$$P\left(\omega : \sup_{x\in\Re}\left|P^*\left[V_{1T}^*(\omega) \le u\right] - P\left[\sup_{r\in[0,1]} \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left(1\{F(y_t|Z^{t-1},\widehat{\theta}_T) \le r\} - E\left(1\{F(y_t|Z^{t-1},\theta^\dagger) \le r\}\right)\right) \le x\right]\right| > \varepsilon\right)$$
$$\to \quad 0.$$

Thus, V_{1T}^* has a well defined limiting distribution under both hypotheses, which under the null coincides with the same limiting distribution of V_{1T} , Pr - P, as $E(1\{F(y_t|Z^{t-1},\theta^\dagger) \le r\}) = r$. Now, define $V_{2T}^* = \sup_{u\times v\in U\times V}|V_{2T}^*(u,v)|$, where

$$V_{2T}^*(u,v) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left((1\{y_t^* \le u\} - F(u|Z^{*,t-1},\widehat{\theta}_T^*))1\{Z^{*,t-1} \le v\} - (1\{y_t \le u\} - F(u|Z^{t-1},\widehat{\theta}_T))1\{Z^{t-1} \le v\}\right).$$

Theorem 2.6 (from Theorem 4 of Corradi and Swanson (2003): Let **CS1**, **CS2**(iii)–(iv) and **CS3** hold, and let $T = bl$, with $l = l_T$, such that as $T \to \infty$, $l_T^2/T \to 0$. Then,

$$P\Bigg(\omega : \sup_{x\in\Re}|P^*[V_{2T}^*(\omega) \le x]$$
$$P\left[\sup_{u\times v\in U\times V}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}((1\{y_t \le u\} - F(u|Z^{t-1},\widehat{\theta}_T))1\{Z^{t-1} \le v\}\right.$$
$$\left.\left.-E((1\{y_t \le u\} - F(u|Z^{t-1},\theta^\dagger))1\{Z^{t-1} \le v\}))\le x\right] > \varepsilon\Big|\Bigg)$$
$$\to \quad 0$$

In summary, from Theorems 2.5 and 2.6, we know that $V_{1T}^*(\omega)$ (resp. $V_{2T}^*(\omega)$) has a well defined limiting distribution, conditional on the sample and for all samples except a set of probability measure approaching zero. Furthermore, the limiting distribution coincides with that of V_{1T} (resp. V_{2T}), under H_0. The above results suggest proceeding in the following manner. For any bootstrap replication, compute the bootstrapped statistic, V_{1T}^* (resp. V_{2T}^*). Perform B bootstrap replications (B large) and compute the percentiles of the empirical distribution of the B bootstrapped statistics. Reject H_0 if V_{1T} (V_{2T}) is greater than the

$(1-\alpha)th$-percentile. Otherwise, do not reject H_0. Now, for all samples except a set with probability measure approaching zero, V_{1T} (V_{2T}) has the same limiting distribution as the corresponding bootstrapped statistic, under H_0. Thus, the above approach ensures that the test has asymptotic size equal to α. Under the alternative, V_{1T} (V_{2T}) diverges to infinity, while the corresponding bootstrap statistic has a well defined limiting distribution. This ensures unit asymptotic power. Note that the validity of the bootstrap critical values is based on an infinite number of bootstrap replications, although in practice we need to choose B. Andrews and Buchinsky (2000) suggest an adaptive rule for choosing B, Davidson and MacKinnon (2000) suggest a pretesting procedure ensuring that there is a "small probability" of drawing different conclusions from the ideal bootstrap and from the bootstrap with B replications, for a test with a given level. However, in the current context, the limiting distribution is a functional of a Gaussian process, so that the explicit density function is not known; and thus one cannot directly apply the approaches suggested in the papers above. In Monte Carlo experiments, Corradi and Swanson (2003) show that finite sample results are quite robust to the choice of B. For example, they find that even for values of B as small as 100, the bootstrap has good finite sample properties.

Needless to say, if the parameters are estimated using T observations, and the statistic is constructed using only R observations, with $R = o(T)$, then the contribution of parameter estimation error to the covariance kernel is asymptotically negligible. In this case, it is not necessary to compute $\widehat{\theta}_T^*$. For example, when bootstrapping critical values for a statistic analogous to V_{1T}, but constructed using R observations, say V_{1R}, one can instead construct V_{1R}^* as follows:

$$V_{1R}^* = \sup_{r\in[0,1]} \frac{1}{\sqrt{R}} \sum_{t=1}^{R} \left(1\{F(y_t^*|Z^{*,t-1},\widehat{\theta}_T) \leq r\} - 1\{F(y_t|Z^{t-1},\widehat{\theta}_T) \leq r\} \right). \tag{17}$$

The intuition for this statistic is that $\sqrt{R}(\widehat{\theta}_T - \theta^\dagger) = o_p(1)$, and so the bootstrap estimator of θ is not needed in order to mimic the distribution of $\sqrt{T}(\widehat{\theta}_T - \theta^\dagger)$. Analogs of V_{1R} and V_{1R}^* can similarly be constructed for V_{2T}. However, Corradi and Swanson (2003) do not suggest using this approach because of the cost to finite sample power, and also because of the lack of an adaptive, data-driven rule for choosing R.

2.6 Other Related Work

Most of the test statistics described above are based on testing for the uniformity on $[0,1]$ and/or independence of $F_t(y_t|Z^{t-1},\theta_0) = \int_{-\infty}^{y_t} f_t(y|Z^{t-1},\theta_0)$. Needless to say, if $F_t(y_t|Z^{t-1},\theta_0)$ is $iid\ UN[0,1]$, then $\Phi^{-1}\left(F_t(y_t|Z^{t-1},\theta_0)\right)$, where Φ denotes the CDF of a standard normal, is $iidN(0,1)$.

Berkowitz (2001) proposes a likelihood ratio test for the null of (standard) normality against autoregressive alternatives. The advantage of his test is that is easy to implement and has standard limiting distribution, while the disadvantage is that it only has unit asymptotic power against fixed alternatives.

Recently, Bontemps and Meddahi (BM: 2003a,b) introduce a novel approach to testing distributional assumptions. More precisely, they derive set of moment conditions which are satisfied under the null of a particular distribution. This leads to a GMM type test. Of interest is the fact that, the tests suggest by BM do not suffer of the parameter estimation error issue, as the suggested moment condition ensure that the contribution of estimation uncertainty vanishes asymptotically. Furthermore, if the null is rejected, by looking at which moment condition is violated one can get some guidance on how to "improve" the model. Interestingly, BM (2003b) point out that, a test for the normality of $\Phi^{-1}\left(F_t(y_t|Z^{t-1},\theta_0)\right)$ is instead affected by the contribution of estimation uncertainty, because of the double transformation. Finally, other tests for normality have been recently suggested by Bai and Ng (2004) and by Duan (2003).

3 Specification Testing and Model Selection Out-of-Sample

In the previous section we discussed in-sample implementation of tests for the correct specification of the conditional distribution for the entire or for a given information set. Thus, the same set of observations were to be used for both estimation and model evaluation. In this section, we outline out of sample versions of the same tests, where the sample is split into two parts, and the latter portion is used for validation. Indeed, going back at least as far as Granger (1980) and Ashley, Granger and Schmalensee (1980), it has been noted that interest focuses on assessing the predictive accuracy of different models, it should be of interest to evaluate them in an out of sample manner - namely by looking at predictions and associated prediction errors. This is particularly true if all models are assumed to be approximations of some "true" underlying unknown model (i.e. if all models may be misspecified). In addition, it has been stressed that sample evaluation may lead to model overfitting, a problem that can easily be avoided if out of sample evaluation is used. On the other hand, when the null is that of correct dynamic specification, or correct specification for given information set, then there is no clear consensus about whether use an in-sample or an out-of-sample version of a given test. For example, Inoue and Kilian (2004)) claim that in-sample tests are more powerful than out of sample variants thereof. In many cases, however, their analysis is based in part upon assuming correct specification under the null hypothesis - a practice which is not always appropriate when assessing forecasting models. Furthermore, the probability integral transform approach has been frequently used in an

out of sample fashion (see e.g. the empirical applications in DGT (1998) and Hong (2001)), and hence the tests discussed above (which are based on the probability integral transform approach of DGT) should be of interest from the persepctive of out of sample evaluation. For this reason, and for sake of completeness, in this section we provide out of sample versions of the test statistics in Subsection 2.2.2-2.2.4. This requires some preliminary results on the asymptotic behavior of recursive and rolling estimators, as these results are not available elsewhere.

3.1 Estimation and Parameter Estimation Error in Recursive and Rolling Estimation Schemes - West as well as West and McCracken Results

In out of sample model evaluation, the sample of T observations is split into R observations to be used for estimation, and P observations to be used for forecast construction, predictive density evaluation, and generally for model validation and selection. In this context, it is assumed that $T = R + P$. In out of sample contexts, parameters are usually estimated using either recursive or rolling estimation schemes. In both cases, one constructs a sequence of P estimators, which are in turn used in the construction of P $h-$step ahead predictions and prediction errors, where h is the forecast horizon.

In the recursive estimation scheme, one constructs the first estimator using the first R observations, say $\widehat{\theta}_R$, the second using observations up to $R+1$, say $\widehat{\theta}_{R+1}$, and so on until one has a sequence of P estimators, $(\widehat{\theta}_R, \widehat{\theta}_{R+1}, ..., \widehat{\theta}_{R+P-1})$. In the sequel, we consider the generic case of extremum estimators, or $m-$estimators, which include ordinary least squares, nonlinear least squares, and (quasi) maximum-likelihood estimators. Define the recursive estimator as:[10]

$$\widehat{\theta}_{t,rec} = \arg\min_{\theta \in \Theta} \frac{1}{t} \sum_{j=1}^{t} q(y_j, Z^{j-1}, \theta), \quad t = R, R+1, ...R+P-1, \tag{18}$$

where $q(y_j, Z^{j-1}, \theta_i)$ denotes the objective function (i.e. in (quasi) MLE, $q(y_j, Z^{j-1}, \theta_i) = -\ln f(y_j, Z^{j-1}, \theta_i)$, with f denoting the (pseudo) density of y^t given Z^{t-1}).[11]

In the rolling estimation scheme, one constructs a sequence of P estimators using a rolling window of R observations. That is, the first estimator is constructed using the first R observations, the second using observations from 2 to $R+1$, and so on, with the last estimator being constructed using observations from

[10] For notational simplicity, we begin all summations at $t = 1$. Note, however, that in general if Z^{t-1} contains information up to the s^{th} lag, say, then summation should be initiated at $t = s + 1$.

[11] Generalized method of moments (GMM) estimators can be treated in an analogous manner. As one is often interested in comparing misspecified models, we avoid using overidentified GMM estimators in our discussion. This is because, as pointed out by Hall and Inoue (2003), one cannot obtain asymptotic normality for overidentified GMM in the misspecified case.

$T-R$ to $T-1$, so that we have a sequence of P estimators, $(\widehat{\theta}_{R,R}, \widehat{\theta}_{R+1,R}, ..., \widehat{\theta}_{R+P-1,R})$.[12]

In general, it is common to assume that P and R grow as T grows. This assumption is maintained in the sequel. Notable exceptions to this approach are Giacomini and White (2003)[13], who propose using a rolling scheme with a fixed window that does not increase with the sample size, so that estimated parameters are treated as mixing variables, and Pesaran and Timmermann (2003, 2004) who suggest rules for choosing the window of observations, in order to take into account possible structure breaks.

Turning now to the rolling estimation scheme, define the relevant estimator as:

$$\widehat{\theta}_{t,rol} = \arg\min_{\theta\in\Theta} \frac{1}{R} \sum_{j=t-R+1}^{t} q(y_j, Z^{j-1}, \theta), \quad R \leq t \leq T-1. \tag{19}$$

In the case of in sample model evaluation, the contribution of parameter estimation error is summarized by the limiting distribution of $\sqrt{T}\left(\widehat{\theta}_T - \theta^\dagger\right)$, where $\theta^\dagger$ is the probability limit of $\widehat{\theta}_T$. This is clear, for example, from the proofs of Theorems 2.3 and 2.4 above, which are given in Corradi and Swanson (2003). On the other hand, in the case of recursive and rolling estimation schemes, the contribution of parameter estimation error is summarized by the limiting distribution of $\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widehat{\theta}_{t,rec} - \theta^\dagger\right)$ and $\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widehat{\theta}_{t,rol} - \theta^\dagger\right)$ respectively. Under mild conditions, because of the central limit theorem, $\left(\widehat{\theta}_{t,rec} - \theta^\dagger\right)$ and $\left(\widehat{\theta}_{t,rol} - \theta^\dagger\right)$ are $O_P(R^{-1/2})$. Thus, if P grows at a slower rate than R (i.e. if $P/R \to 0$, as $T \to \infty$), then $\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widehat{\theta}_{t,rec} - \theta^\dagger\right)$ and $\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widehat{\theta}_{t,rol} - \theta^\dagger\right)$ are asymptotically negligible. In other words, if the in sample portion of the data used for estimation is "much larger" than the out of sample portion of the data to be used for predictive accuracy testing and generally for model evaluation, then the contribution of parameter estimation error is asymptotically negligible.

A key result which is used in all of the subsequent limiting distribution results discussed in this chapter is the derivation of the limiting distribution of $\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widehat{\theta}_{t,rec} - \theta^\dagger\right)$ (see West (1996)) and of $\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widehat{\theta}_{t,rol} - \theta^\dagger\right)$ (see West and McCracken (1998)). Their results follow, given the following assumptions.

W1: (y_t, Z^{t-1}), with y_t scalar and Z^{t-1} an $R^\zeta-$valued $(0 < \zeta < \infty)$ vector, is a strictly stationary and absolutely regular $\beta-$mixing process with size $-4(4+\psi)/\psi$, $\psi > 0$.

W2: (i) $\theta^\dagger$ is uniquely identified (i.e. $E(q(y_t, Z^{t-1}, \theta)) > E(q(y_t, Z^{t-1}, \theta_i^\dagger))$ for any $\theta \neq \theta^\dagger$); (ii) q is twice continuously differentiable on the interior of Θ, and for Θ a compact subset of R^ϱ; (iii) the elements of $\nabla_\theta q$

[12] Here, for simplicity, we have assumed that in sample estimation ends with period $T-R$ to $T-1$. Thus, we are implicity assuming that $h=1$, so that P out of sample predictions *and* prediction errors can be constructed.

[13] The Giacomini and White (2003) test is designed for conditional mean evaluation, although it can likely be easily extended to the case of conditional density evaluation. One important advantage of this test is that it is valid for both nested and nonnested models (see below for further discussion).

and $\nabla^2_\theta q$ are $p-$dominated on Θ, with $p > 2(2+\psi)$, where ψ is the same positive constant as defined in **W1**; and (iii) $E\left(-\nabla^2_\theta q(\theta)\right)$ is negative definite uniformly on Θ.

Theorem 3.1 (from Lemma 4.1 and Theorem 4.1 in West (1996)):

Let **W1** and **W2** hold. Also, as $T \to \infty$, $P/R \to \pi$, $0 < \pi < \infty$. Then,

$$\frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(\widehat{\theta}_{t,rec} - \theta^\dagger \right) \xrightarrow{d} N(0, 2\Pi A^\dagger C_{00} A^\dagger),$$

where $\Pi = (1 - \pi^{-1} \ln(1+\pi))$, $C_{00} = \sum_{j=-\infty}^{\infty} E\left(\left(\nabla_\theta q(y_{1+s}, Z^s, \theta^\dagger) \right) \left(\nabla_\theta q(y_{1+s+j}, Z^{s+j}, \theta^\dagger) \right)' \right)$, and $A^\dagger = E\left(-\nabla^2_{\theta_i} q(y_t, Z^{t-1}, \theta^\dagger) \right)$.

Theorem 3.2 (from Lemmas 4.1 and 4.2 in West (1996) and McCracken (1998)):

Let **W1** and **W2** hold. Also, as $T \to \infty$, $P/R \to \pi$, $0 < \pi < \infty$. Then,

$$\frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(\widehat{\theta}_{t,rol} - \theta^\dagger \right) \xrightarrow{d} N(0, 2\Pi C_{00}),$$

where for $\pi \leq 1$, $\Pi = \pi - \frac{\pi^2}{3}$ and for $\pi > 1$, $\Pi = 1 - \frac{1}{3\pi}$. Also, C_{00} and $A^\dagger$ defined as in Theorem 3.1.

3.2 Out-of-Sample Implementation of Bai as well as Hong and Li Tests

We begin by analyzing the out of sample versions of Bai's (2003) test. Define the out of sample version of the statistic in (6) for the recursive case, as

$$\widehat{V}_{P,rec} = \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(1\{F_{t+1}(y_{t+1}|Z^t, \widehat{\theta}_{t,rec}) \leq r\} - r \right), \tag{20}$$

and for the rolling case as

$$\widehat{V}_{P,rol} = \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(1\{F_{t+1}(y_{t+1}|Z^t, \widehat{\theta}_{t,rol}) \leq r\} - r \right), \tag{21}$$

where $\widehat{\theta}_{t,rec}$ and $\widehat{\theta}_{t,rol}$ are defined as in (18) and (19), respectively. Also, define

$$\widehat{W}_{P,rec}(r) = \widehat{V}_{P,rec}(r) - \int_0^r \left(\dot{g}(s) C^{-1}(s) \dot{g}(s)' \int_s^1 \dot{g}(\tau) d\widehat{V}_{P,rec}(\tau) \right) ds$$

and

$$\widehat{W}_{P,rol}(r) = \widehat{V}_{P,rol}(r) - \int_0^r \left(\dot{g}(s) C^{-1}(s) \dot{g}(s)' \int_s^1 \dot{g}(\tau) d\widehat{V}_{P,rol}(\tau) \right) ds$$

Let **BAI1**, **BAI2** and **BAI4** be as given in Section 2.1, and modify **BAI3** as follows:

BAI3': $\left(\widehat{\theta}_{t,rec} - \theta_0 \right) = O_P(P^{-1/2})$, uniformly in t.[14]

[14]Note that BAI3' is satisfied under mild conditions, provided $P/R \to \pi$ with $\pi < \infty$. In particular,

$$P^{1/2} \left(\widehat{\theta}_t - \theta_0 \right) = \left(\frac{1}{t} \sum_{j=1}^{t} \nabla^2_\theta q_j(\bar{\theta}_t) \right)^{-1} \left(\frac{P^{1/2}}{t} \sum_{j=1}^{t} \nabla_\theta q_j(\theta_0) \right)$$

BAI3": $\left(\widehat{\theta}_{t,rol} - \theta_0\right) = O_P(P^{-1/2})$, uniformly in t.[15]

Given this setup, the following proposition holds.

Proposition 3.2: Let **BAI1**,**BAI2,BAI4** hold and assume that as $T \to \infty$, $P/R \to \pi$, with $\pi < \infty$. Then,

(i) If **BAI3'** hold, under the null hypothesis in (1), $\sup_{r\in[0,1]} \widehat{W}_{P,rec}(r) \xrightarrow{d} \sup_{r\in[0,1]} W(r)$.

(ii) If **BAI3"** hold, under the null hypothesis in (1), $\sup_{r\in[0,1]} \widehat{W}_{P,rol}(r) \xrightarrow{d} \sup_{r\in[0,1]} W(r)$.

Proof: See Appendix.

Turning now to an out of sample version of the Hong and Li test, note that these tests can be defined as in equations (8)-(11) above, by replacing $\widehat{U}_t$ in (8) with $\widehat{U}_{t,rec}$ and $\widehat{U}_{t,rol}$, respectively, where

$$\widehat{U}_{t+1,rec} = F_{t+1}(y_{t+1}|Z^t, \widehat{\theta}_{t,rec}), \text{ and } \widehat{U}_{t+1,rol} = F_{t+1}(y_{t+1}|Z^t, \widehat{\theta}_{t,rol}), \tag{22}$$

with $\widehat{\theta}_{t,rec}$ and $\widehat{\theta}_{t,rol}$ defined as in (18) and (19). Thus, for the recursive estimation case, it follows that

$$\widehat{\phi}_{rec}(u_1, u_2) = (P-j)^{-1} \sum_{\tau=R+j+1}^{T-1} K_h(u_1, \widehat{U}_{\tau,rec}) K_h(u_2, \widehat{U}_{\tau-j,rec}),$$

where $n = T = R + P$. For the rolling estimation case, it follows that

$$\widehat{\phi}_{rol}(u_1, u_2) = (P-j)^{-1} \sum_{\tau=R+j+1}^{T-1} K_h(u_1, \widehat{U}_{\tau,rol}) K_h(u_2, \widehat{U}_{\tau-j,rol}).$$

Also, define

$$\widehat{M}_{rec}(j) = \int_0^1 \int_0^1 \left(\widehat{\phi}_{rec}(u_1, u_2) - 1\right)^2 du_1 du_2, \ \widehat{M}_{rol}(j) = \int_0^1 \int_0^1 \left(\widehat{\phi}_{rol}(u_1, u_2) - 1\right)^2 du_1 du_2$$

and

$$\widehat{Q}_{rec}(j) = \left((n-j)\widehat{M}_{rec}(j) - A_h^0\right)/V_0^{1/2}, \ \widehat{Q}_{rol}(j) = \left((n-j)\widehat{M}_{rol}(j) - A_h^0\right)/V_0^{1/2}.$$

The following proposition then holds.

Proposition 3.3: Let **HL1-HL4** hold. If $h = cP^{-\delta}$, $\delta \in (0, 1/5)$, then under the null in (1), and for any $j > 0$, $j = o(P^{1-\delta(5-2/v)})$, if as $P, R \to \infty$, $P/R \to \pi$, $\pi < \infty$, $\widehat{Q}_{rec}(j) \xrightarrow{d} N(0,1)$ and $\widehat{Q}_{rol}(j) \xrightarrow{d} N(0,1)$.

Now, by uniform law of large numbers, $\left(\frac{1}{t}\sum_{j=1}^t \nabla_\theta^2 q_j(\bar{\theta}_t)\right)^{-1} - \left(\frac{1}{t}\sum_{j=1}^t E\left(\nabla_\theta^2 q_j(\theta_0)\right)\right)^{-1} \xrightarrow{pr} 0$. Let $t = [Tr]$, with $(1+\pi)^{-1} \leq r \leq 1$. Then,

$$\frac{P^{1/2}}{[Tr]} \sum_{j=1}^{[Tr]} \nabla_\theta q_j(\theta_0) = \sqrt{\frac{P}{T}} \frac{1}{r} \frac{1}{\sqrt{T}} \sum_{j=1}^{[Tr]} \nabla_\theta q_j(\theta_0).$$

For any r, $\frac{1}{r}\frac{1}{\sqrt{T}}\sum_{j=1}^{[Tr]} \nabla_\theta q_j(\theta_0)$ satisfies a CLT and so is $O_P(T^{-1/2})$ and so $O(P^{-1/2})$. As r is bounded away from zero, and because of stochastic equicontinuity in r, $\sup_{r\in[(1+\pi)^{-1},1]} \sqrt{\frac{P}{T}}\frac{1}{r}\frac{1}{\sqrt{T}}\sum_{j=1}^{[Tr]} \nabla_\theta q_j(\theta_0) = O_P(P^{-1/2})$.

[15] BAI3" is also satisfied under mild assumptions, by the same arguments used in the footnote above.

The statement in the proposition above follows straightforwardly by the same arguments used in the proof of Theorem 1 in Hong and Li (2003). Additionally, and as noted above, the contribution of parameter estimation error is of order $O_P(P^{1/2})$, while the statistic converges at a nonparametric rate, depending on the bandwidth parameter. Therefore, regardless of the estimation scheme used, the contribution of parameter estimation error is asymptotically negligible.

3.3 Out-of-Sample Implementation of Corradi and Swanson Tests

We now outline out of sample versions of the Corradi and Swanson (2003) tests. First, redefine the statistics using the above out of sample notation as

$$V_{1P,rec} = \sup_{r\in[0,1]} |V_{1P,rec}(r)|,\ V_{1P,rol} = \sup_{r\in[0,1]} |V_{1P,rol}(r)|$$

where

$$V_{1P,rec}(r) = \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(1\{\widehat{U}_{t+1,rec} \leq r\} - r\right)$$

and

$$V_{1P,rol}(r) = \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(1\{\widehat{U}_{t+1,rol} \leq r\} - r\right),$$

with $\widehat{U}_{t,rec}$ and $\widehat{U}_{t,rol}$ defined as in (22). Further, define

$$V_{2P,rec} = \sup_{u\times v\in U\times V} |V_{2P,rec}(u,v)|\ V_{2P,rol} = \sup_{u\times v\in U\times V} |V_{2P,rol}(u,v)|,$$

where

$$V_{2P,rec}(u,v) = \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left((1\{y_{t+1} \leq u\} - F(u|Z^t, \widehat{\theta}_{t,rec}))1\{Z^t \leq v\}\right)$$

and

$$V_{2P,rol}(u,v) = \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left((1\{y_{t+1} \leq u\} - F(u|Z^t, \widehat{\theta}_{t,rol}))1\{Z^t \leq v\}\right).$$

Hereafter, let $V_{1P,J} = V_{1P,rec}$ when $J = 1$ and $V_{1P,J} = V_{1P,rol}$ when $J = 2$ and similarly, $V_{2P,J} = V_{2P,rec}$ when $J = 1$ and $V_{2P,J} = V_{2P,rol}$ when $J = 2$. The following propositions then hold.

Proposition 3.4: Let **CS1**, **CS2**(i)–(ii) and **CS3** hold. Also, as $P, R \to \infty$, $P/R \to \pi$, $0 < \pi < \infty$.[16] Then for $J = 1, 2$: (i) Under H_0, $V_{1P,J} \Rightarrow \sup_{r\in[0,1]} |V_{1,J}(r)|$, where $V_{1,J}$ is a zero mean Gaussian process with covariance kernel $K_{1,J}(r,r')$ given by:

$$K_{1,J}(r,r') = E(\sum_{s=-\infty}^{\infty} \left(1\{F(y_1|Z^0,\theta_0) \leq r\} - r\right)\left(1\{F(y_s|Z^{s-1},\theta_0) \leq r'\} - r'\right))$$

[16]Note that for $\pi = 0$, the contribution of parameter estimation error is asymptotically negligible, and so the covariance kernel is the same as that given in Theorem 2.3.

$$+\Pi_J E(\nabla_\theta F(x(r)|Z^{t-1},\theta_0))' A(\theta_0) \sum_{s=-\infty}^{\infty} E(q_1(\theta_0)q_s(\theta_0)')A(\theta_0)E(\nabla_\theta F(x(r')|Z^{t-1},\theta_0))$$

$$-2C\Pi_J E(\nabla_\theta F(x(r)|Z^{t-1},\theta_0))' A(\theta_0) \sum_{s=-\infty}^{\infty} E\left(\left(1\{F(y_1|Z^0,\theta_0)\le r\}-r\right)q_s(\theta_0)'\right)$$

with $q_s(\theta_0) = \nabla_\theta \ln f_s(y_s|Z^{s-1},\theta_0)$, $x(r) = F^{-1}(r|Z^{t-1},\theta_0)$, $A(\theta_0) = \left(E\left(\nabla_\theta q_s(\theta_0)\nabla_\theta q_s(\theta_0)'\right)\right)^{-1}$, $\Pi_1 = 2(1-\pi^{-1}\ln(1+\pi))$, and $C\Pi_1 = (1-\pi^{-1}\ln(1+\pi))$. For $J=2$, $j=1$ and $P \le R$, $\Pi_2 = \left(\pi - \frac{\pi^2}{3}\right)$, $C\Pi_2 = \frac{\pi}{2}$, and for $P > R$, $\Pi_2 = \left(1-\frac{1}{3\pi}\right)$ and $C\Pi_2 = \left(1-\frac{1}{2\pi}\right)$.

(ii) Under H_A, there exists an $\varepsilon > 0$ such that $\lim_{T\to\infty} \Pr(\frac{1}{P^{1/2}} V_{1T,J} > \varepsilon) = 1$, $J = 1,2$.

Proof: See Appendix.

Proposition 3.5: Let **CS1**, **CS2**(iii)–(iv) and **CS3** hold. Also, as $P, R \to \infty$, $P/R \to \pi$, $0 < \pi < \infty$. Then for $J = 1,2$: (i) Under H_0, $V_{2P,J} \Rightarrow \sup_{u\times v \in U\times V} |Z_J(u,v)|$, where $V_{2P,J}$ is defined as in (14) and Z is a zero mean Gaussian process with covariance kernel $K_{2,J}(u,v,u',v')$ given by:

$$E(\sum_{s=-\infty}^{\infty} ((1\{y_1 \le u\} - F(u|Z^0,\theta_0))1\{X_0 \le v\})((1\{y_s \le u'\} - F(u|Z^{s-1},\theta_0))1\{X_s \le v'\}))$$

$$+\Pi_J E(\nabla_\theta F(u|Z^0,\theta_0)' 1\{Z^0 \le v\})A(\theta_0) \sum_{s=-\infty}^{\infty} q_0(\theta_0)q_s(\theta_0)'A(\theta_0)E(\nabla_\theta F(u'|Z^0,\theta_0)1\{Z^0 \le v'\})$$

$$-2C\Pi_J \sum_{s=-\infty}^{\infty} ((1\{y_0 \le u\} - F(u|Z^0,\theta_0))1\{Z^0 \le v\})E(\nabla_\theta F(u'|Z^0,\theta_0)' 1\{Z^0 \le v'\})A(\theta_0)q_s(\theta_0)).$$

where Π_J and $C\Pi_J$ are defined as in the statement of Proposition 3.4.

(ii) Under H_A, there exists an $\varepsilon > 0$ such that $\lim_{T\to\infty} \Pr(\frac{1}{T^{1/2}} V_{2T} > \varepsilon) = 1$.

Proof: See Appendix.

It is immediate to see that the limiting distributions in Propositions 3.4 and 3.5 differ from the ones in Theorems 2.3 and 2.4 only up terms Π_j and $C\Pi_j$, $j = 1,2$. On the other hand, we shall see that valid asymptotic critical values cannot be obtained by directly following the bootstrap procedure described in Section 2.5. Below, we outline how to obtain valid bootstrap critical values in the recursive and in the rolling estimation cases, respectively.

3.4 Bootstrap Critical for the $V_{1P,J}$ and $V_{2P,J}$ Tests Under Recursive Estimation

When forming the block bootstrap for recursive m-estimators, 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, say $\widehat{\theta}^*_{t,rec}$, which is constructed as a direct analog of $\widehat{\theta}_{t,rec}$, 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, we suggest a re-centering of the bootstrap score which ensures that the new bootstrap estimator, which is no longer the direct analog of $\widehat{\theta}_{t,rec}$, 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, 2004), and Inoue and Shintani (2004)). 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)). However, in the case of recursive $m-$estimators the bias term is instead of order $T^{-1/2}$, and so it does contribute to the limiting distribution. This points to a need for re-centering when using recursive estimation schemes, and such re-centering is discussed in the next subsection.

3.4.1 The Recursive PEE Bootstrap

We now show how Künsch (1989) block bootstrap can be used in the context of a recursive estimation scheme.[17] At each replication, draw b blocks (with replacement) of length l from the sample $W_t = (y_t, Z^{t-1})$, where $bl = T - 1$. Thus, the first block is equal to $W_{i+1}, ..., W_{i+l}$, for some $i = 0, ..., T - l - 1$, with probability $1/(T - l)$, the second block is equal to $W_{i+1}, ..., W_{i+l}$, again for some $i = 0, ..., T - l - 1$, with probability $1/(T - l)$, and so on, for all blocks. More formally, let I_k, $k = 1, ..., b$ be iid discrete uniform random variables on $[0, 1, ..., T - l + 1]$. Then, the resampled series, $W^*_t = (y^*_t, Z^{*,t-1})$, is such that $W^*_1, W^*_2, ..., W^*_l, W^*_{l+1}, ..., W^*_T = W_{I_1+1}, W_{I_1+2}, ..., W_{I_1+l}, W_{I_2}, ..., W_{I_b+l}$, and so a resampled series consists of b blocks that are discrete iid uniform random variables, conditional on the sample.

[17] The main difference between the block bootstrap and the stationary bootstrap of Politis and Romano (PR:1994) is that the former uses a deterministic block length, which may be either overlapping as in Künsch (1989) or non-overlapping as in Carlstein (1986), while the latter resamples using blocks of random length. One important feature of the PR bootstrap is that the resampled series, conditional on the sample, is stationary, while a series resampled from the (overlapping or non overlapping) block bootstrap is nonstationary, even if the original sample is strictly stationary. However, Lahiri (1999) shows that all block boostrap methods, regardless of whether the block length is deterministic or random, have a first order bias of the same magnitude, but the bootstrap with deterministic block length has a smaller first order variance. In addition, the overlapping block boostrap is more efficient than the non overlapping block bootstrap.

Suppose we define the bootstrap estimator, $\widehat{\theta}^*_{t,rec}$, to be the direct analog of $\widehat{\theta}_{t,rec}$. Namely,

$$\widehat{\theta}^*_{t,rec} = \arg\min_{\theta\in\Theta} \frac{1}{t}\sum_{j=1}^{t} q(y^*_j, Z^{*,j-1}, \theta), \;\; R \leq t \leq T-1. \tag{23}$$

By first order conditions, $\frac{1}{t}\sum_{j=1}^{t} \nabla_\theta q(y^*_j, Z^{*,j-1}, \widehat{\theta}^*_{t,rec}) = 0$, and via a mean value expansion of $\frac{1}{t}\sum_{j=1}^{t} \nabla_\theta q(y^*_j, Z^{*,j-1}, \widehat{\theta}^*_{t,r}$ around $\widehat{\theta}_{t,rec}$, after a few simple manipulations, we have that

$$\begin{aligned}
&\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widehat{\theta}^*_{t,rec} - \widehat{\theta}_{t,rec}\right) \\
= \;& \frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\left(\frac{1}{t}\sum_{j=1}^{t}\nabla^2_\theta q(y^*_j, Z^{*,j-1}, \overline{\theta}^*_{t,rec})\right)^{-1}\frac{1}{t}\sum_{j=1}^{t}\nabla_\theta q(y^*_j, Z^{*,j-1}, \widehat{\theta}_{t,rec})\right) \\
= \;& A^\dagger_i \frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\frac{1}{t}\sum_{j=1}^{t}\nabla_\theta q(y^*_j, Z^{*,j-1}, \widehat{\theta}_{t,rec})\right) + o_{P^*}(1) \;\; \Pr - P \\
= \;& A^\dagger_i \frac{a_{R,0}}{\sqrt{P}}\sum_{t=1}^{R}\nabla_\theta q(y^*_j, Z^{*,j-1}, \widehat{\theta}_{t,rec}) + A^\dagger_i \frac{1}{\sqrt{P}}\sum_{j=1}^{P-1} a_{R,j}\nabla_\theta q(y^*_{R+j}, Z^{*,R+j-1}, \widehat{\theta}_{t,rec}) \\
& + o_{P^*}(1) \;\; \Pr - P,
\end{aligned} \tag{24}$$

where $\overline{\theta}^*_{t,rec} \in \left(\widehat{\theta}^*_{t,rec}, \widehat{\theta}_{t,rec}\right)$, $A^\dagger = E\left(\nabla^2_\theta q(y_j, Z^{j-1}, \theta^\dagger)\right)^{-1}$, $a_{R,j} = \frac{1}{R+j} + \frac{1}{R+j+1} + ... + \frac{1}{R+P-1}$, $j = 0, 1, ..., P-1$, and where the last equality on the right hand side of (24) follows immediately, using the same arguments as those used in Lemma A5 of West (1996). Analogously,

$$\begin{aligned}
&\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widehat{\theta}_{t,rec} - \theta^\dagger\right) \\
= \;& A^\dagger \frac{a_{R,0}}{\sqrt{P}}\sum_{t=s}^{R}\nabla_\theta q(y_j, Z^{j-1}, \theta^\dagger) + A^\dagger \frac{1}{\sqrt{P}}\sum_{j=1}^{P-1} a_{R,j}\nabla_\theta q(y_{R+j}, Z^{R+j-1}, \theta^\dagger) + o_P(1).
\end{aligned} \tag{25}$$

Now, given the definition of $\theta^\dagger$, $E\left(\nabla_\theta q(y_j, Z^{j-1}, \theta^\dagger)\right) = 0$ for all j, and $\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widehat{\theta}_{t,rec} - \theta^\dagger\right)$ has a zero mean normal limiting distribution (see Theorem 4.1 in West (1996)). On the other hand, as any block of observations has the same chance of being drawn,

$$E^*\left(\nabla_\theta q(y^*_j, Z^{*,j-1}, \widehat{\theta}_{t,rec})\right) = \frac{1}{T-1}\sum_{k=1}^{T-1}\nabla_\theta q(y_k, Z^{k-1}, \widehat{\theta}_{t,rec}) + O\left(\frac{l}{T}\right) \;\; \Pr - P, \tag{26}$$

where the $O\left(\frac{l}{T}\right)$ term arises because the first and last l observations have a lesser chance of being drawn (see e.g. Fitzenberger (1997)).[18] Now, $\frac{1}{T-1}\sum_{k=1}^{T-1}\nabla_\theta q(y_k, Z^{k-1}, \widehat{\theta}_{t,rec}) \neq 0$, and is instead of order $O_P\left(T^{-1/2}\right)$. Thus, $\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\frac{1}{T-1}\sum_{k=1}^{T-1}\nabla_\theta q(y_k, Z^{k-1}, \widehat{\theta}_{t,rec}) = O_P(1)$, and does not vanish in probability. This clearly

[18] In fact, the first and last observation in the sample can appear only at the beginning and end of the block, for example.

contrasts with the full sample case, in which $\frac{1}{T-1}\sum_{k=1}^{T-1}\nabla_\theta q(y_k, Z^{k-1}, \widehat{\theta}_T) = 0$, because of the first order conditions. Thus, $\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widehat{\theta}^*_{t,rec} - \widehat{\theta}_{t,rec}\right)$ cannot have a zero mean normal limiting distribution, but is instead characterized by a location bias that can be either positive or negative depending on the sample. Given (26), our objective is thus to have the bootstrap score centered around $\frac{1}{T-1}\sum_{k=1}^{T-1}\nabla_\theta q(y_k, Z^{k-1}, \widehat{\theta}_{t,rec})$. Hence, define a new bootstrap estimator, $\widetilde{\theta}^*_{t,rec}$, as:

$$\widetilde{\theta}^*_{t,rec} = \arg\min_{\theta\in\Theta}\frac{1}{t}\sum_{j=1}^{t}\left(q(y^*_j, Z^{*,j-1}, \theta) - \theta'\left(\frac{1}{T}\sum_{k=1}^{T-1}\nabla_\theta q(y_k, Z^{k-1}, \widehat{\theta}_{t,rec})\right)\right), \tag{27}$$

$R \leq t \leq T-1$.[19]

Given first order conditions, $\frac{1}{t}\sum_{j=1}^{t}\left(\nabla_\theta q(y^*_j, Z^{*,j-1}, \widetilde{\theta}^*_{t,rec}) - \left(\frac{1}{T}\sum_{k=1}^{T-1}\nabla_\theta q(y_k, Z^{k-1}, \widehat{\theta}_{t,rec})\right)\right) = 0$, and via a mean value expansion of $\frac{1}{t}\sum_{j=1}^{t}\nabla_\theta q(y^*_j, Z^{*,j-1}, \widetilde{\theta}^*_{t,rec})$ around $\widehat{\theta}_{t,rec}$, after a few simple manipulations, we have that

$$\begin{aligned}
&\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widetilde{\theta}^*_{t,rec} - \widehat{\theta}_{t,rec}\right) \\
= \;& A^\dagger\frac{1}{\sqrt{P}}\sum_{t=R}^{T}\left(\frac{1}{t}\sum_{j=s}^{t}\left(\nabla_\theta q(y^*_j, Z^{*,j-1}, \widehat{\theta}_{t,rec}) - \left(\frac{1}{T}\sum_{k=s}^{T-1}\nabla_\theta q(y_k, Z^{k-1}, \widehat{\theta}_{t,rec})\right)\right)\right) \\
&+ o_{P^*}(1) \;\Pr - P.
\end{aligned}$$

Given (26), it is immediate to see that the bias associated with $\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widetilde{\theta}^*_{t,rec} - \widehat{\theta}_{t,rec}\right)$ is of order $O\left(lT^{-1/2}\right)$, conditional on the sample, and so it is negligible for first order asymptotics, as $l = o(T^{1/2})$. The following result pertains given the above setup.

Theorem 3.6 (from Theorem 1 in Corradi and Swanson (2004c): Let **CS1** and **CS3** hold. Also, assume that as $T \to \infty$, $l \to \infty$, and that $\frac{l}{T^{1/4}} \to 0$. Then, as T, P and $R \to \infty$,

$$P\left(\omega : \sup_{v\in\Re^{p(i)}}\left|P^*_T\left(\frac{1}{\sqrt{P}}\sum_{t=R}^{T}\left(\widetilde{\theta}^*_{t,rec} - \theta^\dagger\right) \leq v\right) - P\left(\frac{1}{\sqrt{P}}\sum_{t=R}^{T}\left(\widehat{\theta}_{t,rec} - \theta^\dagger\right) \leq v\right)\right| > \varepsilon\right) \to 0,$$

where P^*_T denotes the probability law of the resampled series, conditional on the (entire) sample.

Broadly speaking, Theorem 3.6 states that $\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widetilde{\theta}^*_{t,rec} - \theta^\dagger\right)$ has the same limiting distribution as $\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widehat{\theta}_{t,rec} - \theta^\dagger\right)$, conditional on sample, and for all samples except a set with probability measure approaching zero. As outlined in the following sections, application of Theorem 1 allows us to capture the

[19] More precisely, we should define

$$\widetilde{\theta}^*_{i,t} = \arg\min_{\theta_i\in\Theta_i}\frac{1}{t-s}\sum_{j=s}^{t}\left(q_i(y^*_j, Z^{*,j-1}, \theta_i) - \theta_i'\left(\frac{1}{T-s}\sum_{k=s}^{T-1}\nabla_{\theta_i}q_i(y_k, Z^{k-1}, \widehat{\theta}_{i,t})\right)\right)$$

However, for notational simplicity we approximate $\frac{1}{t-s}$ and $\frac{1}{T-s}$ with $\frac{1}{t}$ and $\frac{1}{T}$.

contribution of (recursive) parameter estimation error to the covariance kernel of the limiting distribution of various statistics.

3.4.2 $V_{1P,J}$ and $V_{2P,J}$ Bootstrap Statistics Under Recursive Estimation

One can apply the results above to provide a bootstrap statistic for the case of the recursive estimation scheme. Define,

$$V^*_{1P,rec} = \sup_{r\in[0,1]} |V^*_{1P,rec}(r)|,$$

where

$$V^*_{1P,rec}(r) = \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(1\{F(y^*_{t+1}|Z^{*,t}, \widetilde{\theta}^*_{t,rec}) \leq r\} - \frac{1}{T}\sum_{j=1}^{T-1} 1\{F(y_{j+1}|Z^j, \widehat{\theta}_{t,rec}) \leq r\} \right) \tag{28}$$

Also define,

$$V^*_{2P,rec} = \sup_{u\times v\in U\times V} V^*_{2P,rec}(u,v)$$

where

$$\begin{aligned} V^*_{2P,rec}(u,v) &= \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \Big((1\{y^*_{t+1} \leq u\} - F(u|Z^{*,t}, \widetilde{\theta}^*_{t,rec}))1\{Z^{*,t} \leq v\} \\ &\quad - \frac{1}{T}\sum_{j=1}^{T-1} (1\{y_{j+1} \leq u\} - F(u|Z^j, \widehat{\theta}_{t,rec}))1\{Z^j \leq v\} \Big) \end{aligned} \tag{29}$$

Note that bootstrap statistics in (28) and (29) are different from the "usual" bootstrap statistics, which are defined as the difference between the statistic computed over the sample observations and over the bootstrap observations. For brevity, just consider $V^*_{1P,rec}$. Note that each bootstrap term, say $1\{F(y^*_{t+1}|Z^{*,t}, \widetilde{\theta}^*_{t,rec}) \leq r\}$, $t \geq R$, is recentered around the (full) sample mean $\frac{1}{T}\sum_{j=1}^{T-1} 1\{F(y_{j+1}|Z^j, \widehat{\theta}_{t,rec}) \leq r\}$. This is necessary as the bootstrap statistic is constructed using the last P resampled observations, which in turn have been resampled from the full sample. In particular, this is necessary regardless of the ratio P/R. If $P/R \to 0$, then we do not need to mimic parameter estimation error, and so could simply use $\widehat{\theta}_{1,t,\tau}$ instead of $\widetilde{\theta}^*_{1,t,\tau}$, but we still need to recenter any bootstrap term around the (full) sample mean. This leads to the following proposition.

Proposition 3.7: Let **CS1**, **CS2**(i)–(ii) and **CS3** hold. Also, assume that as $T \to \infty$, $l \to \infty$, and that $\frac{l}{T^{1/4}} \to 0$. Then, as T, P and $R \to \infty$,

$$P\left(\omega : \sup_{x\in\Re} \left| P^*\left[V^*_{1P,rec}(\omega) \leq u\right] - P\left[\sup_{r\in[0,1]} \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \Big(1\{F(y_{t+1}|Z^t, \theta^\dagger) \leq r\} - E\left(1\{F(y_{t+1}|Z^t, \theta^\dagger) \leq r\}\right)\Big) \leq x \right] \right| > \varepsilon \right)$$

$$\rightarrow 0.$$

Proof: See Appendix.

Proposition 3.8: Let **CS1**, **CS2**(iii)–(iv) and **CS3** hold. Also, assume that as $T \rightarrow \infty$, $l \rightarrow \infty$, and that $\frac{l}{T^{1/4}} \rightarrow 0$. Then, as T, P and $R \rightarrow \infty$,

$$\begin{aligned} & P\Bigg(\omega : \sup_{x \in \Re} \Big| P^*[V^*_{2P,rec}(\omega) \leq x] \\ & P\Bigg[\sup_{u \times v \in U \times V} \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} ((1\{y_{t+1} \leq u\} - F(u|Z^t, \theta^\dagger))1\{Z^t \leq v\} \\ & -E((1\{y_{t+1} \leq u\} - F(u|Z^t, \theta^\dagger))1\{Z^t \leq v\})) \leq x \Bigg] > \varepsilon \Big| \Bigg) \\ \rightarrow \quad & 0 \end{aligned}$$

Proof: See Appendix.

The same remarks given below Theorems 2.5 and 2.6 apply here.

3.5 Bootstrap Critical for the $V_{1P,J}$ and $V_{2P,J}$ Tests Under Rolling Estimation

In the rolling estimation scheme, observations in the middle of the sample are used more frequently than observation at either the beginning or the end of the sample. As in the recursive case, this introduces a location bias to the usual block bootstrap, as under standard resampling with replacement, any block from the original sample has the same probability of being selected. Also, the bias term varies across samples and can be either positive or negative, depending on the specific sample. In the sequel, we shall show how to properly recenter the objective function in order to obtain a bootstrap rolling estimator, say $\widetilde{\theta}^*_{t,rol}$ such that $\frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(\widetilde{\theta}^*_{t,rol} - \widehat{\theta}_{t,rol} \right)$ has the same limiting distribution as $\frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(\widehat{\theta}_{t,rol} - \theta^\dagger \right)$, conditionally on the sample.

Resample b overlapping blocks of length l from $W_t = (y_t, Z^{t-1})$, as in the recursive case and define the rolling bootstrap estimator as,

$$\widetilde{\theta}^*_{t,rol} = \arg \max_{\theta_i \in \Theta_i} \frac{1}{R} \sum_{j=t-R+1}^{t} \left(q(y^*_j, Z^{*,j-1}, \theta) - \theta' \left(\frac{1}{T} \sum_{k=s}^{T-1} \nabla_\theta q(y_k, Z^{k-1}, \widehat{\theta}_{t,rol}) \right) \right).$$

Theorem 3.9 (from Proposition 2 in Corradi and Swanson (2004c)): Let **CS1** and **CS3** hold. Also, assume that as $T \rightarrow \infty$, $l \rightarrow \infty$, and that $\frac{l}{T^{1/4}} \rightarrow 0$. Then, as T, P and $R \rightarrow \infty$,

$$P\left(\omega : \sup_{v \in \Re^{p(i)}} \left| P^*_T \left(\frac{1}{\sqrt{P}} \sum_{t=R}^{T} \left(\widetilde{\theta}^*_{t,rol} - \widehat{\theta}_{t,rol} \right) \leq v \right) - P\left(\frac{1}{\sqrt{P}} \sum_{t=R}^{T} \left(\widehat{\theta}_{t,rol} - \theta^\dagger \right) \leq v \right) \right| > \varepsilon \right) \rightarrow 0,$$

Finally note that in the rolling case, $V^*_{1P,rol}, V^*_{2P,rol}$ can be constructed as in (28) and (29), $\widetilde{\theta}^*_{t,rec}$ and $\widehat{\theta}_{t,rec}$ with $\widetilde{\theta}^*_{t,rol}$ and $\widehat{\theta}_{t,rol}$, and the same statement as in Propositions 3.7 and 3.8 hold.

Part III: Evaluation of (Multiple) Misspecified Predictive Models

4 Pointwise Comparison of (Multiple) Misspecified Predictive Models

In the previous two sections we discussed several in sample and out of sample tests for the null of either correct dynamic specification of the conditional distribution or for the null of correct conditional distribution for given information set. Needless to say, the correct (either dynamically, or for a given information set) conditional distribution is the best predictive density. However, it is often sensible to account for the fact that all models may be approximations, and so may be misspecified. The literature on point forecast evaluation does indeed acknowledge that the objective of interest is often to choose a model which provides the best (loss function specific) out of sample predictions, from amongst a set of potentially misspecified models, and not just from amongst models that may only be dynamically misspecified, as is the case with some of the tests discussed above. In this section we outline several popular tests for comparing the relative out of sample accuracy of misspecified models in the case of point forecasts. We shall distinguish among three main group of tests: (i) tests for comparing two nonnested models, (ii) tests for comparing two (or more) nested models; and (iii) tests for comparing multiple models, where at least one model is non-nested. In the next section, we broaden the scope by considering tests for comparing misspecified predictive density models.

4.1 Comparison of Two Nonnested Models: Diebold and Mariano Test

Diebold and Mariano (DM: 1995) propose a test for the null hypothesis of equal predictive ability that is based part on the pairwise model comparison test discussed in Granger and Newbold (1986). The Diebold and Mariano test allows for nondifferentiable loss functions, but does not explicitly account for parameter estimation error, instead relying on the assumption that the in-sample estimation period is growing more quickly than the out-of-sample prediction period, so that parameter estimation error vanishes asymptotically.

West (1996) takes the more general approach of explicitly allowing for parameter estimation error, although at the cost of assuming that the loss function used is differentiable. Let $u_{0,t+h}$ and $u_{1,t+h}$ be the $h-$step ahead prediction error associated with predictions of y_{t+h}, using information available up to time t. For example, for $h=1$, $u_{0,t+1}=y_{t+1}-\kappa_0(Z_0^{t-1},\theta_0^{\dagger})$, and $u_{1,t+1}=y_{t+1}-\kappa_1(Z_1^{t-1},\theta_1^{\dagger})$, where Z_0^{t-1} and Z_1^{t-1} contain past values of y_t and possibly other conditioning variables. Assume that the two models be nonnested (i.e. Z_0^{t-1} not a subset of Z_1^{t-1} -and vice-versa- and/or $\kappa_1 \neq \kappa_0$). As lucidly pointed out by Granger and Pesaran (2000), when comparing misspecified models, the ranking of models based on their predictive accuracy depends on the loss function used. Hereafter, denote the loss function as g, and as usual let $T=R+P$, where only the last P observations are used for model evaluation. Under the assumption that $u_{0,t}$ and $u_{1,t}$ are strictly stationary, the null hypothesis of equal predictive accuracy is specified as:

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

and

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

In practice, we do not observe $u_{0,t+1}$ and $u_{1,t+1}$, but only $\widehat{u}_{0,t+1}$ and $\widehat{u}_{1,t+1}$, where $\widehat{u}_{0,t+1}=y_{t+1}-\kappa_0(Z_0^t,\widehat{\theta}_{0,t})$, and where $\widehat{\theta}_{0,t}$ is an estimator constructed using observations from 1 up to t, $t \geq R$, in the recursive estimation case, and between $t-R+1$ and t in the rolling case. For brevity, in this subsection we just consider the recursive scheme. Therefore, for notational simplicity, we simply denote the recursive estimator for model i, $\widehat{\theta}_{0,t}$, $\widehat{\theta}_{0,t,rec}$. Note that the rolling scheme can be treated in an analogous manner. Of crucial importance is the loss function used for estimation. In fact, as we shall show below if we use the same loss function for estimation and model evaluation, the contribution of parameter estimation error is asymptotically negligible, regardless the limit of the ratio P/R as $T \to \infty$. Here, for $i=0,1$

$$\widehat{\theta}_{i,t} = \arg\min_{\theta_i \in \Theta_i} \frac{1}{t}\sum_{j=1}^{t} q(y_j - \kappa_i(Z_i^{j-1},\theta_i)),\ t \geq R$$

In the sequel, we rely on the assumption that g is continuously differentiable. The case of non-differentiable loss functions is treated by McCracken (2000,2003). Now,

$$\begin{aligned}
\frac{1}{\sqrt{P}}\sum_{t=R}^{T-h} g(\widehat{u}_{i,t+1}) &= \frac{1}{\sqrt{P}}\sum_{t=R}^{T-1} g(u_{i,t+1}) + \frac{1}{\sqrt{P}}\sum_{t=R}^{T-1} \nabla g(\overline{u}_{i,t+1})\left(\widehat{\theta}_{i,t}-\theta_i^{\dagger}\right) \\
&= \frac{1}{\sqrt{P}}\sum_{t=R}^{T-1} g(u_{i,t+1}) + E\left(\nabla g(u_{i,t+1})\right)\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widehat{\theta}_{i,t}-\theta_i^{\dagger}\right) + o_P(1). \qquad (30)
\end{aligned}$$

It is immediate to see that if $g=q$ (i.e. the same loss is used for estimation and model evaluation), then $E\left(\nabla g(u_{i,t+1})\right)=0$ because of the first order conditions. Of course, another case in which the second term

on the RHS of (30) vanishes is when $P/R \to 0$ (these are the cases DM consider). The limiting distribution of the RHS in (30) is given in Section 3.1. The Diebold and Mariano test is

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

where

$$\begin{aligned}
&\frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(g(\widehat{u}_{0,t+1}) - g(\widehat{u}_{1,t+1}) \right) \\
&\xrightarrow{d} N \left(0, S_{gg} + 2\Pi F_0' A_0 S_{h0h_0} A_0 F_0 \right. \\
&+2\Pi F_1' A_1 S_{h_1 h_1} A_1 F_1 - \Pi (S'_{gh_0} A_0 F_0 + F_0' A_0 S_{gh_0}) \\
&-2\Pi \left(F_1' A_1 S_{h_1 h_0} A_0 F_0 + F_0' A_0 S_{h_0 h_1} A_1 F_1 \right) \\
&\left. +\Pi (S'_{gh_1} A_1 F_1 + F_1' A_1 S_{gh_1}) \right),
\end{aligned}$$

with

$$\begin{aligned}
\widehat{\sigma}_P^2 &= \widehat{S}_{gg} + 2\Pi \widehat{F}_0' \widehat{A}_0 \widehat{S}_{h_0 h_0} + +2\Pi \widehat{F}_1' \widehat{A}_1 S_{h_1 h_1} \widehat{A}_1 \widehat{F}_1 \\
&\quad -2\Pi \left(\widehat{F}_1' \widehat{A}_1 \widehat{S}_{h_1 h_0} \widehat{A}_0 \widehat{F}_0 + \widehat{F}_0' \widehat{A}_0 \widehat{S}_{h_0 h_1} \widehat{A}_1 \widehat{F}_1 \right) + \Pi (\widehat{S}'_{gh_1} \widehat{A}_1 \widehat{F}_1 + \widehat{F}_1' \widehat{A}_1 \widehat{S}_{gh_1}),
\end{aligned}$$

where for $i, l = 0, 1$, $\Pi = \Pi = 1 - \pi^{-1} \ln(1+\pi)$, and $q_t(\widehat{\theta}_{i,t}) = q(y_t - \kappa_i(Z_i^{t-1}, \widehat{\theta}_{i,t})$,

$$\widehat{S}_{h_i h_l} = \frac{1}{P} \sum_{\tau=-l_P}^{l_P} w_\tau \sum_{t=R+l_P}^{T-l_P} \nabla_\theta q_t(\widehat{\theta}_{i,t}) \nabla_\theta q_{t+\tau}(\widehat{\theta}_{l,t})'$$

$$\begin{aligned}
\widehat{S}_{fh_i} &= \frac{1}{P} \sum_{\tau=-l_P}^{l_P} w_\tau \sum_{t=R+l_P}^{T-l_P} \\
&\quad \left((g(\widehat{u}_{0,t}) - g(\widehat{u}_{1,t})) - \frac{1}{P} \sum_{t=R}^{T-1} (g(\widehat{u}_{0,t+1}) - g(\widehat{u}_{1,t+1})) \right) \\
&\quad \times \nabla_\beta q_{t+\tau}(\widehat{\theta}_{i,t})'
\end{aligned}$$

$$\begin{aligned}
\widehat{S}_{gg} &= \frac{1}{P} \sum_{\tau=-l_P}^{l_P} w_\tau \sum_{t=R+l_P}^{T-l_P} \\
&\quad \left(g(\widehat{u}_{0,t}) - g(\widehat{u}_{1,t}) - \frac{1}{P} \sum_{t=R}^{T-1} (g(\widehat{u}_{0,t+1}) - g(\widehat{u}_{1,t+1})) \right) \\
&\quad \left(g(\widehat{u}_{0,t+\tau}) - g(\widehat{u}_{1,t+\tau}) - \frac{1}{P} \sum_{t=R}^{T-1} (g(\widehat{u}_{0,t+1}) - g(\widehat{u}_{1,t+1})) \right)
\end{aligned}$$

with $w = 1 - \left(\frac{\tau}{l_P+1}\right)$, and where

$$\widehat{F}_i = \frac{1}{P}\sum_{t=R}^{T-1} \nabla_{\theta_i} g(\widehat{u}_{i,t+1}),\ \widehat{A}_i = \left(-\frac{1}{P}\sum_{t=R}^{T-1} \nabla^2_{\theta_i} q(\widehat{\theta}_{i,t})\right)^{-1}$$

Proposition 4.1 (from Theorem 4.1 in West (1996)): Let **W1-W2** hold. Also, assume that g is continuously differentiable, then, if as $P \to \infty$, $l_p \to \infty$ and $l_P/P^{1/4} \to 0$, then as $P, R \to \infty$, under H_0, $DM_P \xrightarrow{d} N(0,1)$ and under H_A, $\Pr\left(P^{-1/2}|DM_P| > \varepsilon\right) \to 1$, for any $\varepsilon > 0$.

Recall that it is immediate to see that if either $g = q$ or $P/R \to 0$, then the estimator of the long-run variance collapses to $\widehat{\sigma}^2_P = \widehat{S}_{gg}$. The proposition is valid for the case of short-memory series. Corradi, Swanson and Olivetti (2001) consider DM tests in the context of cointegrated series, and Rossi (2003) in the context of processes with roots local to unity.

The proposition above has been stated in terms of one-step ahead prediction errors. All results carry over to the case of $h > 1$. However, in the multistep ahead case, one needs to decide whether to compute "direct" $h-$step ahead forecast errors (i.e. $\widehat{u}_{i,t+h} = y_{t+h} - \kappa_i(Z_i^{t-h}, \widehat{\theta}_{i,t})$) or to compute iterated $h-$ahead forecast errors (i.e. first predict y_{t+1} using observations up to time t, and then use this predicted value in order to predict y_{t+2}, and so on). Within the context of VAR models, Marcellino, Stock and Watson (2004) conduct an extensive and careful Monte Carlo study in order to examine the properties of these direct and indirect approaches to prediction.

Finally, note that when the two models are nested, so that $u_{0,t} = u_{1,t}$ under H_0, both the numerator of the DM_P statistic and $\widehat{\sigma}_P$ approach zero in probability at the same rate, if $P/R \to 0$, so that the DM_P statistic no longer has a normal limiting distribution under the null. The asymptotic distribution of the Diebold-Mariano statistic in the nested case has been recently provided by McCracken (2004), who shows that the limiting distribution is a functional over Brownian motions. Comparison of nested models in the subject of the next subsection.

4.2 Comparison of Two Nested Models

In several instances we may be interested in comparing nested models, such as when forming out of sample Granger causality tests. Also, in the empirical international finance literature, an extensively studied issue concerns comparing the relative accuracy of models driven by fundamentals against random walk models. Since the seminal paper by Meese and Rogoff (1983), who find that no economic models can beat a random walk in terms of their ability to predict exchange rates, several papers have tried to challenge that view, a partial list of which includes Mark (1995), Kilian (1999a), Clarida, Sarno and Taylor (2003), Kilian and

Taylor (2003), Rossi (2003), Clark and West (2004), and McCracken and Sapp (2004). Indeed, the debate about predictability of exchange rates was one of the driving force behind the literature on out of sample comparison of nested models.

4.2.1 Clark and McCracken Tests

Within the context of nested linear models, Clark and McCracken (CMa: 2001) propose some easy to implement tests, under the assumption of martingale difference prediction errors (these tests thus rule out the possibility of dynamic misspecification under the null model). Such tests are thus tailored for the case of one-step ahead prediction. This is because $h-$step ahead prediction errors follow an $MA(h-1)$ process. For the case where $h > 1$, Clark and McCracken (CMb: 2003) propose a different set tests. We begin by outlining the CMa tests.

Consider the following two nested models. The restricted model is

$$y_t = \sum_{j=1}^{q} \beta_j y_{t-j} + \epsilon_t \tag{31}$$

and the unrestricted model is

$$y_t = \sum_{j=1}^{q} \beta_j y_{t-j} + \sum_{j=1}^{k} \alpha_j x_{t-j} + u_t \tag{32}$$

The null and the alternative hypotheses are formulated as:

$$H_0 : E(\epsilon_t^2) - E(u_t^2) = 0$$

$$H_A : E(\epsilon_t^2) - E(u_t^2) > 0,$$

so that it is implicitly assumed that the smaller model cannot outperform the larger. This is actually the case when the loss function is quadratic and when parameters are estimated by LS, which is the case considered by CMa. Note that under the null hypothesis, $u_t = \epsilon_t$, and so DM tests are not applicable in the current context. We use the following assumptions in the sequel of this section.

CM1: (y_t, x_t) are strictly stationary, strong mixing processes, with size $\frac{-4(4+\delta)}{\delta}$, for some $\delta > 0$, and $E(y_t)^8 < \infty, E(x_t)^8$.

CM2: Let $z_t = (y_{t-1}, ..., y_{t-q}, x_{t-1}, ..., x_{t-q})$ and $E(z_t u_t | \Im_{t-1}) = 0$, where $\Im_{t-1}$ contains all the information at time $t-1$ generated by all the past of x_t and y_t. Also, $E(u_t^2 | \Im_{t-1}) = \sigma_u^2$.

Note that **CM2** requires that the larger model is dynamically correctly specified, and requires u_t to be conditionally homoskedastic. The three different tests proposed by CMa are

$$ENC-T=(P-1)^{1/2}\frac{\overline{c}}{(P^{-1}\sum_{t=R}^{T-1}(c_{t+1}-\overline{c}))^{1/2}},$$

where $c_{t+1}=\widehat{\epsilon}_{t+1}(\widehat{\epsilon}_{t+1}-\widehat{u}_{t+1})$, $\overline{c}=P^{-1}\sum_{t=R}^{T-1}c_{t+1}$, and where $\widehat{\epsilon}_{t+1}$ and $\widehat{u}_{t+1}$ are residuals from the LS estimation. Additionally,

$$ENC-REG=(P-1)^{1/2}\frac{P^{-1}\sum_{t=R}^{T-1}\left(\widehat{\epsilon}_{t+1}\left(\widehat{\epsilon}_{t+1}-\widehat{u}_{t+1}\right)\right)}{\left(P^{-1}\sum_{t=R}^{T-1}\left(\widehat{\epsilon}_{t+1}-\widehat{u}_{t+1}\right)^2P^{-1}\sum_{t=R}^{T-1}\widehat{\epsilon}_{t+1}^2-\overline{c}^2\right)^{1/2}},$$

and

$$ENC-NEW=P\frac{\overline{c}}{P^{-1}\sum_{t=1}\widehat{u}_{t+1}^2}$$

Proposition 4.2 (from Theorems 3.1, 3.2, 3.3 in CMa): Let **CM1-CM2** hold. Then under the null, (i) If as $T\to\infty$, $P/R\to\pi>0$, then $ENC-T$ and $ENC-REG$ converge in distribution to Γ_1/Γ_2 where $\Gamma_1=\int_{(1+\pi)^{-1}}^1 s^{-1}W'(s)dW(s)$ and $\Gamma_2=\int_{(1+\pi)^{-1}}^1 s^{-2}W'(s)W(s)ds$. Here, $W(s)$ is a standard $k-$dimensional Brownian motion (note that k is the number of restrictions or the number of extra regressors in the larger model). Also, $ENC-NEW$ converges in distribution to Γ_1, and
(ii) If as $T\to\infty$, $P/R\to\pi=0$, then $ENC-T$ and $ENC-REG$ converge in distribution to $N(0,1)$, and $ENC-NEW$ converges to 0 in probability.

Thus, for $\pi>0$ all three tests have non-standard limiting distributions, although the distributions are nuisance parameter free. Critical values for these statistics under $\pi>0$ have been tabulated by CMa for different values of k and π.

It is immediate to see that **CM2** is violated in the case of multiple step ahead prediction errors. For the case of $h>1$, CMb provide modified versions of the above tests in order to allow for MA(h-1) errors. Their modification essentially consists of using a robust covariance matrix estimator in the context of the above tests.[20] Their new version of the $ENC-T$ test is

$$ENC-T'=(P-h+1)^{1/2}\frac{\frac{1}{P-h+1}\sum_{t=R}^{T-h}\widehat{c}_{t+h}}{\frac{1}{P-h+1}\sum_{j=-\overline{j}}^{\overline{j}}\sum_{t=R+j}^{T-h}K\left(\frac{j}{M}\right)\left(\widehat{c}_{t+h}-\overline{c}\right)\left(\widehat{c}_{t+h-j}-\overline{c}\right)},\tag{33}$$

where $\widehat{c}_{t+h}=\widehat{\epsilon}_{t+h}\left(\widehat{\epsilon}_{t+h}-\widehat{u}_{t+h}\right)$, $\overline{c}=\frac{1}{P-h+1}\sum_{t=R}^{T-\tau}\widehat{c}_{t+h}$, $K\left(\cdot\right)$ is a kernel (such as the Bartlett kernel), and $0\leq K\left(\frac{j}{M}\right)\leq 1$, with $K(0)=1$, and $M=o(P^{1/2})$. Note that $\overline{j}$ does not grow with the sample size. Therefore, the denominator in $ENC-T'$ is a consistent estimator of the long run variance only when

[20] The tests are applied to the problem of comparing linear economic models of exchange rates in McCracken and Sapp (2004).

$E\left(c_t c_{t+|k|}\right) = 0$ for all $|k| > h$ (see Assumption A3 in CMb). Thus, the statistic takes into account the moving average structure of the prediction errors, but still does not allow for dynamic misspecification under the null. Another statistic suggested by CMb is a rescaled version of the Diebold Mariano statistic. Namely

$$MSE - T = (P-h+1)^{1/2} \frac{\frac{1}{P-h+1}\sum_{t=R}^{T-h} \widehat{d}_{t+h}}{\frac{1}{P-h+1}\sum_{j=-\overline{j}}^{\overline{j}} \sum_{t=R+j}^{T-h} K\left(\frac{j}{M}\right)\left(\widehat{d}_{t+h} - \overline{d}\right)\left(\widehat{d}_{t+h-j} - \overline{d}\right)},$$

where $\widehat{d}_{t+h} = \widehat{\epsilon}^2_{t+h} - \widehat{u}^2_{t+h}$, and $\overline{d} = \frac{1}{P-h+1}\sum_{t=R}^{T-\tau} \widehat{d}_{t+h}$.

The limiting distributions of the $ENC-T'$ and $MSE-T$ statistics are given in Theorems 3.1 and 3.2 in CMb, and for $h > 1$ contain nuisance parameters so their critical values cannot be directly tabulated. CMb suggest using a modified version of the bootstrap in Kilian (1999b) to obtain critical values.[21]

4.2.2 Chao, Corradi and Swanson Tests

A limitation of the tests above is that they rule out possible dynamic misspecification under the null. A test which does not require correct dynamic specification and/or conditional homoskedasticity is proposed by Chao, Corradi, and Swanson (2001). Of note, however, is that the Clark and McCracken tests are one-sided while the Chao, Corradi and Swanson test are two-sided, and so may be less powerful in small samples. The test statistic is

$$m_P = P^{-1/2} \sum_{t=R}^{T-1} \widehat{\epsilon}_{t+1} X_t, \tag{34}$$

where $\widehat{\epsilon}_{t+1} = y_{t+1} - \sum_{j=1}^{p-1} \widehat{\beta}_{t,j} y_{t-j}$, $X_t = (x_t, x_{t-1}, \ldots x_{t-k-1})'$. We shall formulate the null and the alternative as

$$\begin{aligned} \widetilde{H}_0 &: E(\epsilon_{t+1} x_{t-j}) = 0, j = 0, 1, \ldots k-1 \\ \widetilde{H}_A &: E(\epsilon_{t+1} x_{t-j}) \neq 0 \text{ for some } j, j = 0, 1, \ldots k-1. \end{aligned}$$

The idea underlying the test is very simple, if $\alpha_1 = \alpha_2 = ... = \alpha_k = 0$, then ϵ_t is uncorrelated with the past of X. Thus, models including lags of X_t do not "outperform" the smaller model. In the sequel we shall require the following assumption.

CCS: (y_t, x_t) are strictly stationary, strong mixing processes, with size $\frac{-4(4+\delta)}{\delta}$, for some $\delta > 0$, and $E(y_t)^8 < \infty, E(x_t)^8 < \infty, E(\epsilon_t y_{t-j}) = 0$, $j = 1, 2, \ldots q$.[22]

[21] For the case of $h = 1$, the limit distribution of $ENC - T'$ corresponds with that of $ENC - T$, given in Proposition 4.2, and the limiting dustribution is derived by McCracken (2000).

[22] Note that the requirement $E(\epsilon_t y_{t-j}) = 0$, $j = 1, 2, \ldots p$ is equivalent to the requirement that $E(y_t | y_{t-1}, ..., y_{t-p}) = \sum_{j=1}^{p-1} \beta_j y_{t-j}$. However, we allow to dynamic misspecification under the null.

Proposition 4.3 (from Theorem 1 in Chao, Corradi and Swanson (2001)): Let **CCS** hold. As $T \to \infty$, $P, R \to \infty, P/R \to \pi, 0 \leq \pi < \infty$,

(i) Under $\widetilde{H}_0$, for $0 < \pi < \infty$,

$$m_P \xrightarrow{d} N\left(0, S_{11} + 2(1 - \pi^{-1}\ln(1+\pi))F'MS_{22}MF \right.$$
$$\left. -(1 - \pi^{-1}\ln(1+\pi))(F'MS_{12} + S_{12}'MF))\right)$$

In addition, for $\pi = 0, m_P \xrightarrow{d} N(0, S_{11})$, where $F = E(Y_t X_t')$, M =plim $\left(\frac{1}{t}\sum_{j=q}^{t} Y_j Y_j'\right)^{-1}$, and $Y_j = (y_{j-1}, \ldots y_{j-q})'$, so that M is a $q \times q$ matrix, F is a $q \times k$ matrix, Y_j is a $k \times 1$ vector, S_{11} is a $k \times k$ matrix, S_{12} is a $q \times k$ matrix, and S_{22} is a $q \times q$ matrix, with

$$S_{11} = \sum_{j=-\infty}^{\infty} E\left((X_t\varepsilon_{t+1} - \mu)(X_{t-j}\varepsilon_{t+1-j} - \mu)'\right),$$

where $\mu = E(X_t\epsilon_{t+1})$, $S_{22} = \sum_{j=-\infty}^{\infty} E\left((Y_{t-1}\varepsilon_t)(Y_{t-1-j}\varepsilon_{t-j})'\right)$ and

$S_{12}' = \sum_{j=-\infty}^{\infty} E\left((\epsilon_{t+1}X_t - \mu)(Y_{t-1-j}\epsilon_{t-j})'\right)$.

(ii) Under $\widetilde{H}_A$, $\lim_{P\to\infty} \Pr\left(\left|\frac{m_p}{P^{1/2}}\right| > 0\right) = 1$.

Corollary 4.4 (from Corollary 2 in Chao, Corradi and Swanson (2001)): Let Assumption **CCS** hold. As $T \to \infty$, $P, R \to \infty, P/R \to \pi, 0 \leq \pi < \infty$, $l_T \to \infty, l_T/T^{1/4} \to 0$,

(i) Under $\widetilde{H}_0$, for $0 < \pi < \infty$,

$$m_p' \left(\widehat{S}_{11} + 2(1 - \pi^{-1}\ln(1+\pi))\widehat{F}'\widehat{M}\widehat{S}_{22}\widehat{M}\widehat{F} \right.$$
$$\left. -(1 - \pi^{-1}\ln(1+\pi))(\widehat{F}'\widehat{M}\widehat{S}_{12} + \widehat{S}_{12}'\widehat{M}\widehat{F}))^{-1}\right)^{-1} m_P$$
$$\xrightarrow{d} \chi_k^2 \tag{35}$$

where $\widehat{F} = \frac{1}{P}\sum_{t=R}^{T} Y_t X_t'$, $\widehat{M} = \left(\frac{1}{P}\sum_{t=R}^{T-1} Y_t Y_t'\right)^{-1}$, and $\widehat{S}_{11} =$

$$\frac{1}{P}\sum_{t=R}^{T-1}(\widehat{\epsilon}_{t+1}X_t - \widehat{\mu}_1)(\widehat{\epsilon}_{t+1}X_t - \widehat{\mu}_1)'$$
$$+\frac{1}{P}\sum_{t=\tau}^{l_T} w_\tau \sum_{t=R+\tau}^{T-1}(\widehat{\epsilon}_{t+1}X_t - \widehat{\mu}_1)(\widehat{\epsilon}_{t+1-\tau}X_{t-\tau} - \widehat{\mu}_1)'$$
$$+\frac{1}{P}\sum_{t=\tau}^{l_T} w_\tau \sum_{t=R+\tau}^{T-1}(\widehat{\epsilon}_{t+1-\tau}X_{t-\tau} - \widehat{\mu}_1)(\widehat{\epsilon}_{t+1}X_t - \widehat{\mu}_1)',$$

where $\widehat{\mu}_1 = \frac{1}{P}\sum_{t=R}^{T-1}\widehat{\epsilon}_{t+1}X_t$,

$$\begin{aligned}\widehat{S}'_{12} &= \frac{1}{P}\sum_{\tau=0}^{l_T} w_\tau \sum_{t=R+\tau}^{T-1} \left(\widehat{\epsilon}_{t+1-\tau} X_{t-\tau} - \widehat{\mu}_1\right)\left(Y_{t-1}\widehat{\epsilon}_t\right)' \\ &\quad + \frac{1}{P}\sum_{\tau=1}^{l_T} w_\tau \sum_{t=R+\tau}^{T-1} \left(\widehat{\epsilon}_{t+1} X_t - \widehat{\mu}_1\right)\left(Y_{t-1-\tau}\widehat{\epsilon}_{t-\tau}\right)',\end{aligned}$$

and

$$\begin{aligned}\widehat{S}_{22} &= \frac{1}{P}\sum_{t=R}^{T-1} \left(Y_{t-1}\widehat{\epsilon}_t\right)\left(Y_{t-1}\widehat{\epsilon}_t\right)' + \\ &\quad \frac{1}{P}\sum_{\tau=1}^{l_T} w_\tau \sum_{t=R+\tau}^{T-1} \left(Y_{t-1}\widehat{\epsilon}_t\right)\left(Y_{t-1-\tau}\widehat{\epsilon}_{t-\tau}\right)' \\ &\quad + \frac{1}{P}\sum_{\tau=1}^{l_T} w_\tau \sum_{t=R+\tau}^{T-1} \left(Y_{t-1-\tau}\widehat{\epsilon}_{t-\tau}\right)\left(Y_{t-1}\widehat{\epsilon}_t\right)',\end{aligned}$$

with $w_\tau = 1 - \frac{\tau}{l_T+1}$.

In addition, for $\pi = 0$, $m_p'\widehat{S}_{11}m_p \xrightarrow{d} \chi^2_k$.

(ii) Under $\widetilde{H}_A$, $m_p'\widehat{S}_{11}^{-1}m_p$ diverges at rate P.

Two final remarks: (i) note that the test can be easily applied to the case of multistep-ahead prediction, it just suffices to replace "1" with "h" above. (ii) linearity of neither the null or the larger model is not required. In fact the test, can be equally applied using residuals from a nonlinear model and using a nonlinear function of X_t as a test function.

4.3 Comparison of Multiple Models: The Reality Check

In the previous subsection, we considered the issue of choosing between two competing models. However, in a lot of situations many different competing models are available and we want to be able to choose the best model from amongst them. When we estimate and compare a very large number of models using the same data set, the problem of data mining or data snooping is prevalent. Broadly speaking, the problem of data snooping is that a model may appear to be superior by chance and not because of its intrinsic merit (recall also the problem of sequential test bias). In other words, if we keep testing the null hypothesis of efficient markets, using the same data set, eventually we shall find a model that results in rejection. The data snooping problem is particularly serious when there is no economic theory supporting an alternative hypothesis. For example, the data snooping problem in the context of evaluating trading rules has been pointed out by Brock, Lakonishok and LeBaron (1992), as well as Sullivan, Timmerman and White (1999,2001).

4.3.1 White's Reality Check and Extensions

White (2000) proposes a novel approach for dealing with the issue of choosing amongst many different models. Suppose there are m models, and we select model 1 as our benchmark (or reference) model. Models $i = 2, ..., m$ are called the competitor (alternative) models. Typically, the benchmark model is either a simple model, our favorite model, or the most commonly used model. Given the benchmark model, the objective is to answer the following question: "Is there any model, amongst the set of $m-1$ competitor models, that yields more accurate predictions (for the variable of interest) than the benchmark?".

In this section, let the generic forecast error be $u_{i,t+1} = y_{t+1} - \kappa_i(Z^t, \theta_i^\dagger)$, and let $\widehat{u}_{i,t+1} = y_{t+1} - \kappa_i(Z^t, \widehat{\theta}_{i,t})$, where $\kappa_i(Z^t, \widehat{\theta}_{i,t})$ is the conditional mean function under model i, and $\widehat{\theta}_{i,t}$ is defined as in Section 3.1. Assume that the set of regressors may vary across different models, so that Z^t is meant to denote the collection of all potential regressors. Following White (2000), define the statistic

$$S_P = \max_{k=2,...,m} S_P(1,k),$$

where

$$S_P(1,k) = \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(g(\widehat{u}_{1,t+1}) - g(\widehat{u}_{k,t+1}) \right), \ k = 2, ..., m,$$

The hypotheses are formulated as

$$H_0 : \max_{k=2,...,m} E(g(u_{1,t+1}) - g(g_{k,t+1})) \leq 0$$

$$H_A : \max_{k=2,...,m} E(g(u_{1,t+1}) - g(u_{k,t+1})) > 0,$$

where $u_{k,t+1} = y_{t+1} - \kappa_k(Z^t, \theta_{k,t}^\dagger)$, and $\theta_{k,t}^\dagger$ denotes the probability limit of $\theta_{i,t}$.

Thus, under the null hypothesis, no competitor model, amongst the set of the $m-1$ alternatives, can provide a more (loss function specific) accurate prediction than the benchmark model. On the other hand, under the alternative, at least one competitor (and in particular, the best competitor) provides more accurate predictions than the benchmark. Now, let **W1** and **W2** be as given in Section 3.1, and also assume the following.

WH: (i) κ_i is twice continuously differentiable on the interior of Θ_i and the elements of $\nabla_{\theta_i} \kappa_i(Z^t, \theta_i)$ and $\nabla^2_{\theta_i} \kappa_i(Z^t, \theta_i)$ are $p-$dominated on Θ_i, for $i = 2, ..., m$, with $p > 2(2 + \psi)$, where ψ is the same positive constant defined in **W1**; (ii) g is positive valued, twice continuously differentiable on Θ_i, and g, g' and g'' are $p-$dominated on Θ_i with p defined as in (i); and (iii) let $c_{kk} =$
$\lim_{T\to\infty} Var\left(\frac{1}{\sqrt{T}} \sum_{t=s}^{T} \left(g(u_{1,t+1}) - g(u_{k,t+1}) \right) \right)$, $k = 2, ..., m$, define analogous covariance terms, $c_{j,k}$, $j, k = 2, ..., m$, and assume that $[c_{j,k}]$ is positive semi-definite.

It is important to stress that for this test, at least one of the competitor models has to be nonnested with the benchmark model.[23] This is ensured by Assumption **WH.**

Proposition 4.5: (Parts (i) and (iii) are from Proposition 2.2 in White (2000)): Let **W1-W2** and **WH** hold. Then, under H_0,

$$\max_{k=2,...,m} \left(S_P(1,k) - \sqrt{P} E\left(g(u_{1,t+1}) - g(u_{k,t+1})\right)\right) \xrightarrow{d} \max_{k=2,...,m} S(1,k), \tag{36}$$

where $S = (S(1,2),...,S(1,n))$ is a zero mean Gaussian process with covariance kernel given by V, with V a $m \times m$ matrix, and

(i) If parameter estimation error vanishes (i.e. if either P/R goes to zero and/or the same loss function is used for estimation and model evaluation, $g = q$, where q is again the objective function), then for $i = 1,...,m-1$, $V = [v_{i,i}] = S_{g_i g_i}$, and

(ii) If parameter estimation error does not vanish (i.e. if $P/R \to 0$ and $g \neq q$), then for $i, j = 1,...,m-1$

$V = [v_{i,i}] = S_{g_i g_i} + 2\Pi\mu_1' A_1^\dagger C_{11} A_1^\dagger \mu_1 + 2\Pi\mu_i' A_i^\dagger C_{ii} A_i^\dagger \mu_i - 4\Pi\mu_1' A_1^\dagger C_{1i} A_i^\dagger \mu_i + 2\Pi S_{g_i q_1} A_1^\dagger \mu_1 - 2\Pi S_{g_i q_i} A_i^\dagger \mu_i,$

where $S_{g_i g_i} = \sum_{\tau=-\infty}^{\infty} E\left((g(u_{1,1}) - g(u_{i,1}))\left(g(u_{1,1+\tau}) - g(u_{i,1+\tau})\right)\right),$

$C_{ii} = \sum_{\tau=-\infty}^{\infty} E\left(\left(\nabla_{\theta_i} q_i(y_{1+s}, Z^s, \theta_i^\dagger)\right)\left(\nabla_{\theta i} q_i(y_{1+s+\tau}, Z^{s+\tau}, \theta_i^\dagger)\right)'\right),$

$S_{g_i q_i} = \sum_{\tau=-\infty}^{\infty} E\left((g(u_{1,1}) - g(u_{i,1}))\left(\nabla_{\theta i} q_i(y_{1+s+\tau}, Z^{s+\tau}, \theta_i^\dagger)\right)'\right),$

$B_i^\dagger = \left(E\left(-\nabla_{\theta i}^2 q_i(y_t, Z^{t-1}, \theta_i^\dagger)\right)\right)^{-1}$, $\mu_i = E\left(\nabla_{\theta_i} g(u_{i,t+1})\right)$, and $\Pi = 1 - \pi^{-1}\ln(1+\pi)$.

(iii) Under H_A, $\Pr\left(\frac{1}{\sqrt{P}}|S_P| > \varepsilon\right) \to 1$, as $P \to \infty$.

Proof: For the proof of part (ii), see the Appendix.

Note that under the null, the least favorable case arises when $E\left(g(u_{1,t+1}) - g(u_{k,t+1})\right) = 0$, $\forall\ k$. In this case, the distribution of S_P coincides with that of $\max_{k=2,...,m}\left(S_P(1,k) - \sqrt{P}E\left(g(u_{1,t+1}) - g(u_{k,t+1})\right)\right)$, so that S_P has the above limiting distribution, which is a functional of a Gaussian process with a covariance kernel that reflects uncertainty due to dynamic misspecification and possibly to parameter estimation error. Additionally, when all competitor models are worse than the benchmark, the statistic diverges to minus infinity at rate $\sqrt{P}$. Finally, when only some competitor models are worse than the benchmark, the limiting distribution provides a conservative test, as S_P will always be smaller than

$\max_{k=2,...,m}\left(S_P(1,k) - \sqrt{P}E\left(g(u_{1,t+1}) - g(u_{k,t+1})\right)\right)$, asymptotically. Of course, when H_A holds, the statistic diverges to plus infinity at rate $\sqrt{P}$.

We now outline how to obtain valid asymptotic critical values for the limiting distribution on the RHS of (36), regardless whether the contribution of parameter estimation error vanishes or not. As noted above,

[23] This is for the same reasons as discussed in the context of the Diebold and Mariano test.

such critical values are conservative, except for the least favorable case under the null. We later outline two ways of alleviating this problem, one suggested by Hansen (2004a) and another, based on subsampling, suggested by Linton, Maasoumi and Whang (2004).

Recall that the maximum of a Gaussian process is not Gaussian in general, so that standard critical values cannot be used to conduct inference on S_P. As pointed out by White (2000), one possibility in this case is to first estimate the covariance structure and then draw 1 realization from an $(m-1)$-dimensional normal with covariance equal to the estimated covariance structure. From this realization, pick the maximum value over $k = 2, \dots, n$. Repeat this a large number of times, form an empirical distribution using the maximum values over $k = 2, \dots, m$, and obtain critical values in the usual way. A drawback to this approach is that we need to rely on an estimator of the covariance structure based on the available sample of observations, which in many cases may be small relative to the number of models being compared. Furthermore, whenever the forecasting errors are not martingale difference sequences (as in our context), heteroskedasticity and autocorrelation consistent covariance matrices should be estimated, and thus a lag truncation parameter must be chosen. Another approach which avoids these problems involves using the stationary bootstrap of Politis and Romano (1994). This is the approach used by White (2000). In general, bootstrap procedures have been shown to perform well in a variety of finite sample contexts (see e.g. Diebold and Chen (1996)). White's suggested bootstrap procedure is valid for the case in which parameter estimation error vanishes asymptotically. His bootstrap statistic is given by:

$$S_P^{**} = \max_{k=2,\dots m} \left| S_P^{**}(1,k) \right|, \tag{37}$$

where

$$S_P^{**}(1,k) = \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(\left(g(\widehat{u}_{1,t+1}^{**}) - g(\widehat{u}_{1,t+1}) \right) - \left(g(\widehat{u}_{k,t+1}^{**}) - g(\widehat{u}_{k,t+1}) \right) \right),$$

and $\widehat{u}_{k,t+1}^{**} = y_{t+1}^{**} - \kappa_k(Z^{**,t}, \widehat{\theta}_{k,t})$, where y_{t+1}^{**} $Z^{**,t}$ denoted the resampled series. White uses the stationary bootstrap by Politis and Romano (1994), but both the block bootstrap and stationary bootstrap deliver the same asymptotic critical values. Note that the bootstrap statistics "contains" only estimators based on the original sample: this is because in White's context PEE vanishes. Our approach to handling PEE is to apply the recursive PEE bootstrap outlined in Section 3.3 in order to obtain critical values which are asymptotically valid in the presence of non vanishing PEE.

Define the bootstrap statistic as:

$$S_P^{*} = \max_{k=2,\dots,m} S_P^{*}(1,k),$$

where

$$S_P^*(1,k) = \frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left[\left(g(y_{t+1}^* - \kappa_1(Z^{*,t},\widetilde{\theta}_{1,t}^*)) - g(y_{t+1}^* - \kappa_k(Z^{*,t},\widetilde{\theta}_{k,t}^*))\right)\right.$$
$$\left. - \left\{\frac{1}{T}\sum_{j=s}^{T-1}\left(g(y_{j+1} - \kappa_1(Z^j,\widehat{\theta}_{1,t})) - g(y_{j+1} - \kappa_k(Z^j,\widehat{\theta}_{k,t}))\right)\right\}\right]. \quad (38)$$

Proposition 4.6: ((i) from Corollary 2.6 in White (2000), (ii) from Proposition 3 in Corradi and Swanson (2004c)).

Let W1-W2 and WH hold.

(i) If $P/R \to 0$ and/or $g = q$, then as $P, R \to \infty$

$$P\left(\omega : \sup_{v\in\Re}\left|P_{R,P}^*\left(\max_{k=2,\ldots,n} S_P^{**}(1,k) \le v\right) - P\left(\max_{k=2,\ldots n} S_P^{\mu}(1,k) \le v\right)\right| > \varepsilon\right) \to 0,$$

(ii) Let Assumptions A1-A4 hold. Also, assume that as $T \to \infty$, $l \to \infty$, and that $\frac{l}{T^{1/4}} \to 0$. Then, as T, P and $R \to \infty$,

$$P\left(\omega : \sup_{v\in\Re}\left|P_T^*\left(\max_{k=2,\ldots,n} S_P^{*}(1,k) \le v\right) - P\left(\max_{k=2,\ldots n} S_P^{\mu}(1,k) \le v\right)\right| > \varepsilon\right) \to 0,$$

and

$$S_P^{\mu}(1,k) = S_P(1,k) - \sqrt{P}E\left(g(u_{1,t+1}) - g(u_{k,t+1})\right),$$

The above result suggests proceeding in the following manner. For any bootstrap replication, compute the bootstrap statistic, S_P^*. Perform B bootstrap replications (B large) and compute the quantiles of the empirical distribution of the B bootstrap statistics. Reject H_0, if S_P is greater than the $(1-\alpha)th$-percentile. Otherwise, do not reject. Now, for all samples except a set with probability measure approaching zero, S_P has the same limiting distribution as the corresponding bootstrapped statistic when $E\left(g(u_{1,t+1}) - g(u_{k,t+1})\right) = 0 \ \forall \ k$, ensuring asymptotic size equal to α. On the other hand, when one or more competitor models are strictly dominated by the benchmark, the rule provides a test with asymptotic size between 0 and α (see above discussion). Under the alternative, S_P diverges to (plus) infinity, while the corresponding bootstrap statistic has a well defined limiting distribution, ensuring unit asymptotic power.

In summary, this application shows that the block bootstrap for recursive m-estimators can be readily adapted in order to provide asymptotically valid critical values that are robust to parameter estimation error as well as model misspecification. In addition, the bootstrap statistics are very easy to construct, as no complicated adjustment terms involving possibly higher order derivatives need be included.

4.3.2 Hansen's Approach Applied to the Reality Check

As mentioned above, the critical values obtained via the empirical distribution of S_P^{**} or S_P^* are upper bounds whenever some competing models are strictly dominated by the benchmark. The issue of conservativeness is particularly relevant when a large number of dominated (bad) models are included in the analysis. In fact, such models do not contribute to the limiting distribution, but drive up the reality check $p-$values, which are obtained for the least favorable case under the null hypothesis. The idea of Hansen (2004a)[24] is to eliminate the models which are dominated, while paying careful attention to not eliminate relevant models. In summary, Hansen defines the statistic

$$\widetilde{S}_P = \max \left\{ \max_{k=2,\ldots,m} \frac{S_P(1,k)}{\left(\widehat{var} \frac{1}{P} \sum_{t=R}^{T-1} \left(g(\widehat{u}_{1,t+1}) - g(\widehat{u}_{k,t+1}) \right) \right)^{1/2}}, 0 \right\},$$

where $\widehat{var} \frac{1}{P} \sum_{t=R}^{T-1} \left(g(\widehat{u}_{1,t+1}) - g(\widehat{u}_{k,t+1}) \right)$ is defined in (39) below. In this way, the modified reality check statistic does not take into account strictly dominated models.

The idea of Hansen is also to impose the "entire" null (not only the least favorable component of the null) when constructing the bootstrap statistic. For this reason, he adds a recentering term. Define,

$$\widehat{\mu}_k = \frac{1}{P} \sum_{t=R}^{T-1} \left(g(\widehat{u}_{1,t+1}) - g(\widehat{u}_{k,t+1}) \right) 1\{ g(\widehat{u}_{1,t+1}) - g(\widehat{u}_{k,t+1}) \geq A_{T,k} \},$$

where $A_{T,k} = \frac{1}{4} T^{-1/4} \sqrt{\widehat{var} \frac{1}{P} \sum_{t=R}^{T-1} \left(g(\widehat{u}_{1,t+1}) - g(\widehat{u}_{k,t+1}) \right)}$,

with

$$\begin{aligned} & \widehat{var} \frac{1}{P} \sum_{t=R}^{T-1} \left(g(\widehat{u}_{1,t+1}) - g(\widehat{u}_{k,t+1}) \right) \\ = \; & B^{-1} \sum_{b=1}^{B} \left(\frac{1}{P} \sum_{t=R}^{T-1} \left(\left(g(\widehat{u}_{1,t+1}) - g(\widehat{u}_{k,t+1}) \right) - \left(g(\widehat{u}^*_{1,t+1}) - g(\widehat{u}^*_{k,t+1}) \right) \right)^2 \right), \end{aligned} \tag{39}$$

and where B denotes the number of bootstrap replications. Hansen's bootstrap statistic is then defined as

$$\widetilde{S}^*_P = \max_{k=2,\ldots,m} \frac{\frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left[\left(g(\widehat{u}^*_{1,t+1}) - g(\widehat{u}^*_{k,t+1}) \right) - \widehat{\mu}_k \right]}{\left(\widehat{var} \frac{1}{P} \sum_{t=R}^{T-1} \left(g(\widehat{u}_{1,t+1}) - g(\widehat{u}_{k,t+1}) \right) \right)^{1/2}}$$

P-values are then computed in terms of the number of times the statistic is smaller than the bootstrap statistic, and H_0 is rejected if, say, $\frac{1}{B} \sum_{b=1}^{B} 1 \left\{ \widetilde{S}_P \leq \widetilde{S}^*_P \right\}$ is below α. This procedure is valid, provided that the effect of parameter estimation error vanishes.

[24] A careful analysis of testing in the presence of composite null hypotheses is given in Hansen (2004b).

4.3.3 The Subsampling Approach Applied to the Reality Check

The idea of subsampling is based on constructing a sequence of statistics using a (sub)sample of size b, where b grows with the sample size, but at a slower rate. Critical values are constructed using the empirical distribution of the sequence of statistics (see e.g. the book by Politis, Romano and Wolf (1999)). In the current context, let the subsampling size to be equal to b, where as $P \to \infty$, $b \to \infty$ and $b/P \to 0$. Define

$$S_{P,a,b} = \max_{k=2,...,m} S_{P,a,b}(1,k),\ a = R, ...T-b-1$$

where

$$S_{P,a,b}(1,k) = \frac{1}{\sqrt{b}} \sum_{t=a}^{a+b-1} \left(g(\widehat{u}_{1,t+1}) - g(\widehat{u}_{k,t+1})\right),\ k = 2, ..., m.$$

Compute the empirical distribution of $S_{P,a,b}$ using $T-b-1$ statistics constructed using b observations. The rule is to reject if we get a value for S_P larger then the $(1-\alpha)-$critical value of the (subsample) empirical distribution, and do not reject otherwise. If $\max_{k=2,...,m} E(g(u_{1,t+1}) - g(u_{k,t+1})) = 0$, then this rule gives a test with asymptotic size equal to α, while if $\max_{k=2,...,m} E(g(u_{1,t+1}) - g(u_{k,t+1})) < 0$ (i.e. if all models are dominated by the benchmark), then the rule gives a test with asymptotic size equal to zero. Finally, under the alternative, $S_{P,a,b}$ diverges at rate $\sqrt{b}$, ensuring unit asymptotic power, provided that $b/P \to 0$. The advantage of subsampling over the block bootstrap, is that the test then has correct size when $\max_{k=2,...,m} E(g(u_{1,t+1}) - g(u_{k,t+1})) = 0$, while the bootstrap approach gives conservative critical values, whenever $E(g(u_{1,t+1}) - g(u_{k,t+1})) < 0$ for some k. Note that the subsampling approach is valid also in the case of non vanishing parameter estimation error. This is because each subsample statistic properly mimics the distribution of the actual statistic. On the other hand the subsampling approach has two drawbacks. First, subsampling critical values are based on a sample of size b instead of P. Second, the finite sample power may be rather low, as the subsampling quantiles under the alternative diverge at rate $\sqrt{b}$, while bootstrap quantiles are bounded under both hypotheses. In a recent paper, Linton, Maasoumi and Whang (2004) apply the subsampling approach to the problem of testing for stochastic dominance; a problem characterized by a composite null, as in the reality check case.

4.3.4 The False Discovery Rate Approach Applied to the Reality Check

Another way to avoid sequential testing bias is to rely on bounds, such as (modified) Bonferroni bounds. However, a well known drawback of such an approach is that it is conservative, particularly when we compare a large number of models. Recently, a new approach, based on the false discovery rate (FDR) has been suggested by Benjamini and Hochberg (1995), for the case of independent statistics. Their approach has

been extended to the case of dependent statistics by Benjamini and Yekutieli (2001).[25] The FDR approach allows one to select among alternative groups of models, in the sense that one can assess which group(s) contribute to the rejection of the null. The FDR approach has the objective of controlling the expected number of false rejections, and in practice one computes p-values associated with m hypotheses, and orders these p-values in increasing fashion, say $P_1 \leq ... \leq P_i \leq .... \leq P_m$. Then, all hypotheses characterized by $P_i \leq (1-(i-1)/m)\alpha$ are rejected, where α is a given significance level. Such an approach, though less conservative than Hochberg's (1988) approach, is still conservative as it provides bounds on p-values. More recently, Storey (2003) introduces the $q-$value of a test statistic, which is defined as the minimum possible false discovery rate for the null is rejected. McCracken and Sapp (2004) implement the $q-$value approach for the comparison of multiple exchange rate models. Overall, we think that a sound practical strategy could be to first implement the above reality check type tests. These tests can then be complemented by using a multiple comparison approach, yielding a better overall understanding concerning which model(s) contribute to the rejection of the null, if it is indeed rejected. If the null is not rejected, then one simply chooses the benchmark model. Nevertheless, even in this case, it may not hurt to see whether some of the individual hypotheses in their joint null hypothesis are rejected via a multiple test comparison approach.

4.4 A Predictive Accuracy Test That is Consistent Against Generic Alternatives

So far we have considered tests for comparing one model against a fixed number of alternative models. Needless to say, such tests have power only against a given alternative. However, there may clearly be some other model with greater predictive accuracy. This is a feature of predictive ability tests which has already been addressed in the consistent specification testing literature (see e.g. Bierens (1982, 1990), Bierens and Ploberger (1997), de Jong (1996), Hansen (1996), Lee, Granger and White (1993), Stinchcombe and White (1998)).

Corradi and Swanson (2002) draw on both the consistent specification and predictive accuracy testing literatures, and propose a test for predictive accuracy which is consistent against generic nonlinear alternatives, and which is designed for comparing nested models. The test is based on an out-of-sample version of the integrated conditional moment (ICM) test of Bierens (1982,1990) and Bierens and Ploberger (1997).

Summarizing, assume that the objective is to test whether there exists any unknown alternative model

[25] Benjamini and Yekutieli (2001) show that the Benjamini and Hochberg (1995) FDR is valid when the statistics have positive regression dependency. This condition allows for multivariate test statistics with a non diagonal correlation matrix.

that has better predictive accuracy than a given benchmark model, for a given loss function. A typical example is the case in which the benchmark model is a simple autoregressive model and we want to check whether a more accurate forecasting model can be constructed by including possibly unknown (non)linear functions of the past of the process or of the past of some other process(es).[26] Although this is the case that we focus on, the benchmark model can in general be any (non)linear model. One important feature of this test is that the same loss function is used for in-sample estimation and out-of-sample prediction (see Granger (1993) and Weiss (1996)).

Let the benchmark model be

$$y_t = \theta^\dagger_{1,1} + \theta^\dagger_{1,2} y_{t-1} + u_{1,t}, \tag{40}$$

where $\theta^\dagger_1 = (\theta^\dagger_{1,1}, \theta^\dagger_{1,2})' = \arg\min_{\theta_1 \in \Theta_1} E(q(y_t - \theta_{1,1} - \theta_{1,2} y_{t-1}))$, $\theta_1 = (\theta_{1,1}, \theta_{1,2})'$, y_t is a scalar, $q = g$, as the same loss function is used both for in-sample estimation and out-of-sample predictive evaluation, and everything else is defined above. The generic alternative model is:

$$y_t = \theta^\dagger_{2,1}(\gamma) + \theta^\dagger_{2,2}(\gamma) y_{t-1} + \theta^\dagger_{2,3}(\gamma) w(Z^{t-1}, \gamma) + u_{2,t}(\gamma), \tag{41}$$

where $\theta^\dagger_2(\gamma) = (\theta^\dagger_{2,1}(\gamma), \theta^\dagger_{2,2}(\gamma), \theta^\dagger_{2,3}(\gamma))' = \arg\min_{\theta_2 \in \Theta_2} E(q(y_t - \theta_{2,1} - \theta_{2,2} y_{t-1} - \theta_{2,3} w(Z^{t-1}, \gamma)))$, $\theta_2(\gamma) = (\theta_{2,1}(\gamma), \theta_{2,2}(\gamma), \theta_{2,3}(\gamma))'$, and $\theta_2 \in \Theta_2$, where Γ is a compact subset of $\Re^d$, for some finite d. The alternative model is called "generic" because of the presence of $w(Z^{t-1}, \gamma)$, which is a generically comprehensive function, such as Bierens' exponential, a logistic, or a cumulative distribution function (see e.g. Stinchcombe and White (1998) for a detailed explanation of generic comprehensiveness). One example has $w(Z^{t-1}, \gamma) = \exp(\sum_{i=1}^{s} \gamma_i \Phi(X_{t-i}))$, where Φ is a measurable one to one mapping from $\Re$ to a bounded subset of $\Re$, so that here $Z^t = (X_t, ..., X_{t-s+1})$, and we are thus testing for nonlinear Granger causality. The hypotheses of interest are:

$$\begin{aligned} H_0 &: E(g(u_{1,t+1}) - g(u_{2,t+1}(\gamma))) = 0 \\ H_A &: E(g(u_{1,t+1}) - g(u_{2,t+1}(\gamma))) > 0. \end{aligned} \tag{42}$$

Clearly, the reference model is nested within the alternative model, and given the definitions of $\theta^\dagger_1$ and $\theta^\dagger_2(\gamma)$, the null model can never outperform the alternative. For this reason, H_0 corresponds to equal predictive accuracy, while H_A corresponds to the case where the alternative model outperforms the reference model, as

[26] For example, Swanson and White (1997) compare the predictive accuracy of various linear models against neural network models using both in-sample and out-of-sample model selection criteria.

long as the errors above are loss function specific forecast errors. It follows that H_0 and H_A can be restated as:

$$H_0 : \theta^{\dagger}_{2,3}(\gamma) = 0 \text{ versus } H_A : \theta^{\dagger}_{2,3}(\gamma) \neq 0,$$

for $\forall \gamma \in \Gamma$, except for a subset with zero Lebesgue measure. Now, given the definition of $\theta^{\dagger}_2(\gamma)$, note that

$$E\left(g'(y_{t+1} - \theta^{\dagger}_{2,1}(\gamma) - \theta^{\dagger}_{2,2}(\gamma) y_t - \theta^{\dagger}_{2,3}(\gamma) w(Z^t, \gamma)) \times \begin{pmatrix} -1 \\ -y_t \\ -w(Z^t, \gamma) \end{pmatrix} \right) = 0,$$

where g' is defined as above. Hence, under H_0 we have that $\theta^{\dagger}_{2,3}(\gamma) = 0$, $\theta^{\dagger}_{2,1}(\gamma) = \theta^{\dagger}_{1,1}$, $\theta^{\dagger}_{2,2}(\gamma) = \theta^{\dagger}_{1,2}$, and $E(g'(u_{1,t+1}) w(Z^t, \gamma)) = 0$. Thus, we can once again restate H_0 and H_A as:

$$H_0 : E(g'(u_{1,t+1}) w(Z^t, \gamma)) = 0 \text{ versus } H_A : E(g'(u_{1,t+1}) w(Z^t, \gamma)) \neq 0, \tag{43}$$

for $\forall \gamma \in \Gamma$, except for a subset with zero Lebesgue measure. Finally, define $\widehat{u}_{1,t+1} = y_{t+1} - \begin{pmatrix} 1 & y_t \end{pmatrix} \widehat{\theta}_{1,t}$. The test statistic is:

$$M_P = \int_{\Gamma} m_P(\gamma)^2 \phi(\gamma) d\gamma, \tag{44}$$

and

$$m_P(\gamma) = \frac{1}{P^{1/2}} \sum_{t=R}^{T-1} g'(\widehat{u}_{1,t+1}) w(Z^t, \gamma), \tag{45}$$

where $\int_{\Gamma} \phi(\gamma) d\gamma = 1$, $\phi(\gamma) \geq 0$, and $\phi(\gamma)$ is absolutely continuous with respect to Lebesgue measure. In the sequel, we need:

NV1: (i) (y_t, Z^t) is a strictly stationary and absolutely regular strong mixing sequence with size $-4(4+\psi)/\psi$, $\psi > 0$, (ii) g is three times continuously differentiable in θ, over the interior of B, and $\nabla_\theta g$, $\nabla^2_\theta g$, $\nabla_\theta g'$, $\nabla^2_\theta g'$ are $2r-$dominated uniformly in Θ, with $r \geq 2(2+\psi)$, (iii) $E\left(-\nabla^2_\theta g_t(\theta)\right)$ is negative definite, uniformly in Θ, (iv) w is a bounded, twice continuously differentiable function on the interior of Γ and $\nabla_\gamma w(z^t, \gamma)$ is bounded uniformly in Γ and (v) $\nabla_\gamma \nabla_\theta g'_t(\theta) w(Z^{t-1}, \gamma)$ is continuous on $\Theta \times \Gamma$, Γ a compact subset of R^d and is $2r-$dominated uniformly in $\Theta \times \Gamma$, with $r \geq 2(2+\psi)$.

NV2: (i) $E(g'(y_t - \theta_{1,1} - \theta_{1,2} y_{t-1})) > E(g'(x_t - \theta^{\dagger}_{1,1} - \theta^{\dagger}_{1,2} x_{t-1}))$, $\forall \theta \neq \theta^{\dagger}$ and
(ii) $E(g'(y_t - \theta_{2,1} - \theta_{2,2} x_{t-1} - \theta_{2,3} w(Z^{t-1}, \gamma))) > \inf_\gamma E(g'(y_t - \theta^{\dagger}_{2,1}(\gamma) - \theta^{\dagger}_{2,2}(\gamma) y_{t-1} - \theta^{\dagger}_{2,3}(\gamma) w(Z^{t-1}, \gamma)))$ for $\theta \neq \theta^{\dagger}(\gamma)$.

NV3: $T = R + P$, and as $T \to \infty$, $\frac{P}{R} \to \pi$, with, $0 \leq \pi < \infty$.

NV4: For any t, s; $\forall\ i, j, k = 1, 2$; and for $\Delta < \infty$:

(i) $E\left(\sup_{\theta\times\gamma\times\gamma^+\in\Theta\times\Gamma\times\Gamma}\left|g_t'(\theta)w(Z^{t-1},\gamma)\nabla_\theta^k g_s'(\theta)w(Z^{s-1},\gamma^+)\right|^4\right)<\Delta$,

where $\nabla_\theta^k(\cdot)$ denotes the $k-$th element of the derivative of its argument with respect to θ.

(ii) $E\left(\sup_{\theta\in\Theta}\left|\left(\nabla_\theta^k(\nabla_\theta^i g_t(\theta))\nabla_\theta^j g_s(\theta)\right)\right|^4\right)<\Delta$, and

(iii) $E\left(\sup_{\theta\times\gamma\in\Theta\times\Gamma}\left|\left(g_t'(\theta)w(Z^{t-1},\gamma)\nabla_\theta^k(\nabla_\theta^j g_s(\theta))\right)\right|^4\right)<\Delta$.

Theorem 4.7 (from Theorem 1 in Corradi and Swanson (2002)): Let **NV1-NV3** hold. Then, the following results hold:

(i) Under H_0,

$$M_P=\int_\Gamma m_P(\gamma)^2\phi(\gamma)d\gamma\xrightarrow{d}\int_\Gamma Z(\gamma)^2\phi(\gamma)d\gamma,$$

where $m_P(\gamma)$ is defined in equation (45) and Z is a Gaussian process with covariance kernel given by:

$$\begin{aligned}K(\gamma_1,\gamma_2) &= S_{gg}(\gamma_1,\gamma_2)+2\Pi\mu_{\gamma_1}'A^\dagger S_{hh}A^\dagger\mu_{\gamma_2}+\Pi\mu_{\gamma_1}'A^\dagger S_{gh}(\gamma_2)\\ &\quad+\Pi\mu_{\gamma_2}'A^\dagger S_{gh}(\gamma_1),\end{aligned}$$

with $\mu_{\gamma_1}=E(\nabla_{\theta_1}(g_{t+1}'(u_{1,t+1})w(Z^t,\gamma_1)))$, $A^\dagger=(-E(\nabla_{\theta_1}^2 q_1(u_{1,t})))^{-1}$,

$S_{gg}(\gamma_1,\gamma_2)=\sum_{j=-\infty}^{\infty}E(g'(u_{1,s+1})w(Z^s,\gamma_1)g'(u_{1,s+j+1})w(Z^{s+j},\gamma_2))$,

$S_{hh}=\sum_{j=-\infty}^{\infty}E(\nabla_{\theta_1}q_1(u_{1,s})\nabla_{\theta_1}q_1(u_{1,s+j})')$,

$S_{gh}(\gamma_1)=\sum_{j=-\infty}^{\infty}E(g'(u_{1,s+1})w(Z^s,\gamma_1)\nabla_{\theta_1}q_1(u_{1,s+j})')$, and γ, γ_1, and γ_2 are generic elements of Γ.

$\Pi=1-\pi^{-1}\ln(1+\pi)$, for $\pi>0$ and $\Pi=0$ for $\pi=0$, $z^q=(z_1,...,z_q)'$, and γ, γ_1, γ_2 are generic elements of Γ.

(ii) Under H_A, for $\varepsilon>0$ and $\delta<1$,

$$\lim_{P\to\infty}\Pr\left(\frac{1}{P^\delta}\int_\Gamma m_P(\gamma)^2\phi(\gamma)d\gamma>\varepsilon\right)=1.$$

Thus, the limiting distribution under H_0 is a Gaussian process with a covariance kernel that reflects both the dependence structure of the data and, for $\pi>0$, the effect of parameter estimation error. Hence, critical values are data dependent and cannot be tabulated.

Valid asymptotic critical values have been obtained via a conditional P-value approach by Corradi and Swanson (2002, Theorem 2). Basically, they have extended Inoue's (2001) to the case of non vanishing parameter estimation error. In turn, Inoue (2001) has extended this approach to allow for non-martingale difference score functions. A drawback of the conditional P-values approach is that the simulated statistic is of order $O_P(l)$, where l plays the same role of the block length in the block bootstrap, under the alternative.

This may lead to a loss in power, specially with small and medium size samples. A valid alternative is provided by the block bootstrap for recursive estimation scheme.

Define,

$$\widetilde{\theta}^*_{1,t} = (\widetilde{\theta}^*_{1,1,t}, \widetilde{\theta}^*_{1,2,t})' = \arg\min_{\theta_1 \in \Theta_1} \frac{1}{t}\sum_{j=2}^{t} [g(y^*_j - \theta_{1,1} - \theta_{1,2} y^*_{j-1}) - \theta'_1 \frac{1}{T}\sum_{i=2}^{T-1} \nabla_\theta g(y_i - \widehat{\theta}_{1,1,t} - \widehat{\theta}_{1,2,t} y_{i-1})] \quad (46)$$

Also, define $\widetilde{u}^*_{1,t+1} = y^*_{t+1} - \begin{pmatrix} 1 & y^*_t \end{pmatrix} \widetilde{\theta}^*_{1,t}$. The bootstrap test statistic is:

$$M^*_P = \int_\Gamma m^*_P(\gamma)^2 \phi(\gamma) d\gamma,$$

where,

$$m^*_P(\gamma) = \frac{1}{P^{1/2}} \sum_{t=R}^{T-1} \left(g'\left(y^*_{t+1} - \begin{pmatrix} 1 & y^*_t \end{pmatrix} \widetilde{\theta}^*_{1,t}\right) w(Z^{*,t}, \gamma) - \frac{1}{T}\sum_{i=1}^{T-1} g'\left(y_{i+1} - \begin{pmatrix} 1 & y_i \end{pmatrix} \widehat{\theta}_{1,t}\right) w(Z^i, \gamma) \right.$$
$$\left. - \frac{1}{T}\sum_{i=1}^{T-1} g'\left(y_{i+1} - \begin{pmatrix} 1 & y_i \end{pmatrix} \widehat{\theta}_{1,t}\right) w(Z^i, \gamma) \right) \quad (47)$$

Theorem 4.8: (from Proposition 5 in Corradi and Swanson (2004c))

Let Assumptions A1-A3 and A5 hold. Also, assume that as $T \to \infty$, $l \to \infty$, and that $\frac{l}{T^{1/4}} \to 0$. Then, as T, P and $R \to \infty$,

$$P\left(\omega : \sup_{v \in \Re} \left| P^*_T \left(\int_\Gamma m^*_P(\gamma)^2 \phi(\gamma) d\gamma \leq v \right) - P\left(\int_\Gamma m^\mu_P(\gamma)^2 \phi(\gamma) d\gamma \leq v \right) \right| > \varepsilon \right) \to 0,$$

where $m^\mu_P(\gamma) = m_P(\gamma) - \sqrt{P} E\left(g'(u_{1,t+1}) w(Z^t, \gamma)\right)$.

The above result suggests proceeding the same way as in the first application. For any bootstrap replication, compute the bootstrap statistic, M^*_P. Perform B bootstrap replications (B large) and compute the percentiles of the empirical distribution of the B bootstrap statistics. Reject H_0 if M_P is greater than the $(1-\alpha)th$-percentile. Otherwise, do not reject. Now, for all samples except a set with probability measure approaching zero, M_P has the same limiting distribution as the corresponding bootstrap statistic under H_0, thus ensuring asymptotic size equal to α. Under the alternative, M_P diverges to (plus) infinity, while the corresponding bootstrap statistic has a well defined limiting distribution, ensuring unit asymptotic power.

5 Comparison of (Multiple) Misspecified Predictive Density Models

In Section 2 we outlined several tests for the null hypothesis of correct specification of the conditional distribution (some of which allowed for dynamic misspecification). Nevertheless, and as discussed above, most models are approximations of reality and therefore they are typically misspecified, and not just dynamically! In Section 4, we have seen that much of the recent literature on evaluation of point forecast models has already acknowledged the fact that models are typically misspecified. The purpose of this section is to merge these two strands of the literature and discuss recent tests for comparing misspecified conditional distribution models.

5.1 The Kullback-Leibler Information Criterion Approach

A well known measure of distributional accuracy is the Kullback-Leibler Information Criterion (KLIC), according to which we choose the model which minimizes the KLIC (see e.g. White (1982), Vuong (1989), Giacomini (2002), and Kitamura (2002)). In particular, choose model 1 over model 2, if

$$E(\log f_1(Y_t|Z^t,\theta_1^\dagger) - \log f_2(Y_t|Z^t,\theta_2^\dagger)) > 0.$$

For the *iid* case, Vuong (1989) suggests a likelihood ratio test for choosing the conditional density model that is closer to the "true" conditional density in terms of the KLIC. Giacomini (2002) suggests a weighted version of the Vuong likelihood ratio test for the case of dependent observations, while Kitamura (2002) employs a KLIC based approach to select among misspecified conditional models that satisfy given moment conditions.[27] Furthermore, the KLIC approach has recently been employed for the evaluation of dynamic stochastic general equilibrium models (see e.g. Schorfheide (2000), Fernandez-Villaverde and Rubio-Ramirez (2004), and Chan, Gomes and Schorfheide (2002)). For example, Fernandez-Villaverde and Rubio-Ramirez (2004) show that the KLIC-best model is also the model with the highest posterior probability.

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. Also, it leads to simple likelihood ratio type tests which have a standard limiting distribution and are not affected by problems associated with accounting for PEE.

However, it should be noted that if one is interested in measuring accuracy over a specific region, or in measuring accuracy for a given conditional confidence interval, say, this cannot be done in as straightforward

[27] Of note is that White (1982) shows that quasi maximum likelihood estimators minimize the KLIC, under mild conditions.

manner using the KLIC. For example, if we want to evaluate the accuracy of different models for approximating the probability that the rate of inflation tomorrow, given the rate of inflation today, will be between 0.5% and 1.5%, say, we can do so quite easily using the square error criterion, but not using the KLIC.

5.2 A Predictive Density Accuracy Test for Comparing Multiple Misspecified Models

Corradi and Swanson (2004a,b) introduce a measure of distributional accuracy, which can be interpreted as a distributional generalization of mean square error. In addition, Corradi and Swanson (2004b) apply this measure to the problem of selecting amongst multiple misspecified predictive density models. In this section we discuss these contributions to the literature.

5.2.1 A Mean Square Error Measure of Distributional Accuracy

As usual, consider forming parametric conditional distributions for a scalar random variable, y_t, given Z^t, where $Z^t = (y_{t-1}, ..., y_{t-s_1}, X_t, ..., X_{t-s_2+1})$ with s_1, s_2 finite. Define the group of conditional distribution models from which one is to select a "best" model as $F_1(u|Z^t, \theta_1^\dagger), ..., F_m(u|Z^t, \theta_m^\dagger)$, and define the true conditional distribution as

$$F_0(u|Z^t, \theta_0) = \Pr(y_{t+1} \leq u|Z^t).$$

Hereafter, assume that $\theta_i^\dagger \in \Theta_i$, where Θ_i is a compact set in a finite dimensional Euclidean space, and let $\theta_i^\dagger$ be the probability limit of a quasi maximum likelihood estimator (QMLE) of the parameters of the conditional distribution under model i. If model i is correctly specified, then $\theta_i^\dagger = \theta_0$. If $m > 2$, follow White (2000). Namely, 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. Needless to say, pairwise comparison of alternative models, in which no benchmark need be specified, follows as a special case. In this context, measure accuracy using the above distributional analog of mean square error. More precisely, define the mean square (approximation) error associated with model i, $i = 1, ..., m$, in terms of the average over U of $E\left(\left(F_i(u|Z^t, \theta_i^\dagger) - F_0(u|Z^t, \theta_0)\right)^2\right)$, where $u \in U$, and U is a possibly unbounded set on the real line, and the expectation is taken with respect to the conditioning variables. In particular, model 1 is more accurate than model 2, if

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

where $\int_U \phi(u)du = 1$ and $\phi(u) \geq 0$, for all $u \in U \subset \Re$. This measure essentially integrates over different quantiles of the conditional distribution. For any given evaluation point, this measure defines a norm and it implies a standard goodness of fit measure. Note, that this measure of accuracy leads to straightforward evaluation of distributional accuracy over a given region of interest, as well as to straightforward evaluation of specific quantiles.

A conditional confidence interval version of the above condition which is more natural to use in applications involving predictive interval comparison follows immediately, and can be written as

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

5.2.2 The Tests Statistic and Its Asymptotic Behavior

In this section, $F_1(\cdot|\cdot,\theta_1^\dagger)$ is taken as the benchmark model, and the objective is to test whether some competitor model can provide a more accurate approximation of $F_0(\cdot|\cdot,\theta_0)$ than the benchmark. The null and the alternative hypotheses are:

$$H_0 : \max_{k=2,\ldots,m} \int_U E\left(\left(F_1(u|Z^t,\theta_1^\dagger) - F_0(u|Z^t,\theta_0)\right)^2 - \left(F_k(u|Z^t,\theta_k^\dagger) - F_0(u|Z^t,\theta_0)\right)^2\right)\phi(u)du \leq 0 \tag{48}$$

versus

$$H_A : \max_{k=2,\ldots,m} \int_U E\left(\left(F_1(u|Z^t,\theta_1^\dagger) - F_0(u|Z^t,\theta_0)\right)^2 - \left(F_k(u|Z^t,\theta_k^\dagger) - F_0(u|Z^t,\theta_0)\right)^2\right)\phi(u)du > 0, \tag{49}$$

where $\phi(u) \geq 0$ and $\int_U \phi(u) = 1$, $u \in U \subset \Re$, U possibly unbounded. Note that for a given u, we compare conditional distributions in terms of their (mean square) distance from the true distribution. We then average over U. As discussed above, a possibly more natural version of the above hypotheses is in terms of conditional confidence intervals evaluation, so that the objective is to "approximate" $\Pr(\underline{u} \leq Y_{t+1} \leq \overline{u}|Z^t)$, and hence to evaluate a region of the predictive density. In that case, the null and alternative hypotheses can be stated as:

$$H_0' : \max_{k=2,\ldots,m} E\left(\left(\left(F_1(\overline{u}|Z^t,\theta_1^\dagger) - F_1(\underline{u}|Z^t,\theta_1^\dagger)\right) - \left(F_0(\overline{u}|Z^t,\theta_0) - F_0(\underline{u}|Z^t,\theta_0)\right)\right)^2 \right.$$
$$\left. - \left(\left(F_k(\overline{u}|Z^t,\theta_k^\dagger) - F_k(\underline{u}|Z^t,\theta_k^\dagger)\right) - \left(F_0(\overline{u}|Z^t,\theta_0) - F_0(\underline{u}|Z^t,\theta_0)\right)\right)^2\right) \leq 0.$$

versus

$$H_A' : \max_{k=2,\ldots,m} E\left(\left(\left(F_1(\overline{u}|Z^t,\theta_1^\dagger) - F_1(\underline{u}|Z^t,\theta_1^\dagger)\right) - \left(F_0(\overline{u}|Z^t,\theta_0) - F_0(\underline{u}|Z^t,\theta_0)\right)\right)^2\right.$$

$$- \left(\left(F_k(\overline{u}|Z^t, \theta_k^\dagger) - F_k(\underline{u}|Z^t, \theta_k^\dagger) \right) - \left(F_0(\overline{u}|Z^t, \theta_0) - F_0(\underline{u}|Z^t, \theta_0) \right) \right)^2 \Bigg) > 0.$$

Alternatively, if interest focuses on testing the null of equal accuracy of two conditional distribution models, say F_1 and F_k, we can simply state the hypotheses as:

$$H_0'' : \int_U E\left(\left(F_1(u|Z^t, \theta_1^\dagger) - F_0(u|Z^t, \theta_0) \right)^2 - \left(F_k(u|Z^t, \theta_k^\dagger) - F_0(u|Z^t, \theta_0) \right)^2 \right) \phi(u) du = 0$$

versus

$$H_A'' : \int_U E\left(\left(F_1(u|Z^t, \theta_1^\dagger) - F_0(u|Z^t, \theta_0) \right)^2 - \left(F_k(u|Z^t, \theta_k^\dagger) - F_0(u|Z^t, \theta_0) \right)^2 \right) \phi(u) du \neq 0,$$

or we can write the predictive density (interval) version of these hypotheses.

Needless to say, we do not know $F_0(u|Z^t)$. However, it is easy to see that

$$\begin{aligned} & E\left(\left(F_1(u|Z^t, \theta_1^\dagger) - F_0(u|Z^t, \theta_0) \right)^2 - \left(F_k(u|Z^t, \theta_k^\dagger) - F_0(u|Z^t, \theta_0) \right)^2 \right) \\ = \; & E\left(\left(1\{y_{t+1} \leq u\} - F_1(u|Z^t, \theta_1^\dagger) \right)^2 \right) - E\left(\left(1\{y_{t+1} \leq u\} - F_k(u|Z^t, \theta_k^\dagger) \right)^2 \right), \end{aligned} \tag{50}$$

where the RHS of (50) does not require the knowledge of the true conditional distribution.

The intuition behind equation (50) is very simple. First, 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, $1\{y_{t+1} \leq u\} - F_k(u|Z^t, \theta_k^\dagger)$ can be interpreted as an "error" term associated with computation of the conditional expectation under F_i. Now, $j = 1, ..., m$:

$$\begin{aligned} & \mu_k^2(u) = E\left(\left(1\{y_{t+1} \leq u\} - F_k(u|Z^t, \theta_k^\dagger) \right)^2 \right) \\ = \; & E\left(\left(\left(1\{y_{t+1} \leq u\} - F_0(u|Z^t, \theta_0) \right) - \left(F_k(u|Z^t, \theta_k^\dagger) - F_0(u|Z^t, \theta_0) \right) \right)^2 \right) \\ = \; & E\left((1\{y_{t+1} \leq u\} - F_0(u|Z^t, \theta_0))^2 \right) + E\left(\left(F_k(u|Z^t, \theta_k^\dagger) - F_0(u|Z^t, \theta_0) \right)^2 \right), \end{aligned}$$

given that the expectation of the cross product is zero (which follows because $1\{y_{t+1} \leq u\} - F_0(u|Z^t, \theta_0)$ is uncorrelated with any measurable function of Z^t). Therefore,

$$\mu_1^2(u) - \mu_k^2(u) = E\left(\left(F_1(u|Z^t, \theta_1^\dagger) - F_0(u|Z^t, \theta_0) \right)^2 \right) - E\left(\left(F_k(u|Z^t, \theta_k^\dagger) - F_0(u|Z^t, \theta_0) \right)^2 \right). \tag{51}$$

The statistic of interest is

$$Z_{P,j} = \max_{k=2,...,m} \int_U Z_{P,u,j}(1,k) \phi(u) du, \; j = 1, 2 \tag{52}$$

where for $j = 1$ (rolling estimation scheme),

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

and for $j = 2$ (recursive estimation scheme),

$$Z_{P,u,2}(1,k) = \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(\left(1\{y_{t+1} \leq u\} - F_1(u|Z^t, \widehat{\theta}_{1,rec}) \right)^2 - \left(1\{y_{t+1} \leq u\} - F_k(u|Z^t, \widehat{\theta}_{k,t,rec}) \right)^2 \right), \tag{53}$$

where $\widehat{\theta}_{i,t,rol}$ and $\widehat{\theta}_{i,t,rec}$ are defined as in (19) and in (18) in Section 3.1.

As shown above and in Corradi and Swanson (2004b), the hypotheses of interest can be restated as:

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

versus

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

where $\mu_i^2(u) = E\left(\left(1\{y_{t+1} \leq u\} - F_i(u|Z^t, \theta_i^\dagger) \right)^2 \right)$. In the sequel, we require:

MD1: (y_t, X_t), with y_t scalar and X_t an $R^\zeta-$valued $(0 < \zeta < \infty)$ vector, is a strictly stationary and absolutely regular $\beta-$mixing process with size $-4(4+\psi)/\psi$, $\psi > 0$.

MD2: (i) $\theta_i^\dagger$ is uniquely identified (i.e. $E(\ln f_i(y_t, Z^{t-1}, \theta_i)) < E(\ln f_i(y_t, Z^{t-1}, \theta_i^\dagger))$ for any $\theta_i \neq \theta_i^\dagger$); (ii) $\ln f_i$ is twice continuously differentiable on the interior of Θ_i, for $i = 1, ..., m$, and for Θ_i a compact subset of $R^{\varrho(i)}$; (iii) the elements of $\nabla_{\theta_i} \ln f_i$ and $\nabla^2_{\theta_i} \ln f_i$ are $p-$dominated on Θ_i, with $p > 2(2+\psi)$, where ψ is the same positive constant as defined in Assumption A1; and (iii) $E\left(-\nabla^2_{\theta_i} \ln f_i(\theta_i) \right)$ is positive definite uniformly on Θ_i.

MD3: $T = R + P$, and as $T \to \infty$, $P/R \to \pi$, with $0 < \pi < \infty$.

MD4: (i) $F_i(u|Z^t, \theta_i)$ is continuously differentiable on the interior of Θ_i and $\nabla_{\theta_i} F_i(u|Z^t, \theta_i^\dagger)$ is $2r$-dominated on Θ_i, uniformly in u, $r > 2$, $i = 1, ..., m$;[28] and (ii) let $v_{kk}(u) = \text{plim}_{T\to\infty}$
$Var\left(\frac{1}{\sqrt{T}} \sum_{t=s}^{T} \left(\left(\left(1\{y_{t+1} \leq u\} - F_1(u|Z^t, \theta_1^\dagger) \right)^2 - \mu_1^2(u) \right) - \left(\left(1\{y_{t+1} \leq u\} - F_k(u|Z^t, \theta_k^\dagger) \right)^2 - \mu_k^2(u) \right) \right) \right)$, $k = 2, ..., m$, define analogous covariance terms, $v_{j,k}(u)$, $j, k = 2, ..., m$, and assume that $[v_{j,k}(u)]$ is positive semi-definite, uniformly in u.

Assumptions **MD1** and **MD2** are standard memory, moment, smoothness and identifiability conditions. **MD1** requires (y_t, X_t) to be strictly stationary and absolutely regular. The memory condition is stronger than $\alpha-$mixing, but weaker than (uniform) $\phi-$mixing. Assumption **MD3** requires that R and P grow at

[28] We require that for $j = 1, ..., p_i$, $\left(E\left(\nabla_\theta F_i(u|Z^t, \theta_i^\dagger) \right) \right)_j \leq D_t(u)$, with $\sup_t \sup_{u \in \Re} E(D_t(u)^{2r}) < \infty$.

the same rate. Of course, if R grows faster than P, then parameter estimation error vanishes in probability (as discussed above), and there is no need to capture the contribution of parameter estimation error when constructing bootstrap critical values. Assumptions **MD4(i)** states standard smoothness and domination conditions imposed on the conditional distributions of the models, and assumption **MD4(ii)** states that at least one of the competing models, $F_2(\cdot|\cdot,\theta_1^\dagger),...,F_n(\cdot|\cdot,\theta_n^\dagger)$, has to be nonnested with (and non nesting) the benchmark.

Proposition 4.9 (from Proposition 1 in Corradi and Swanson (2004a)): Let **MD1-MD4** hold. Then,

$$\max_{k=2,...,m}\int_U \left(Z_{P,u,j}(1,k)-\sqrt{P}\left(\mu_1^2(u)-\mu_k^2(u)\right)\right)\phi_U(u)du \xrightarrow{d} \max_{k=2,...,m}\int_U Z_{1,k,j}(u)\phi_U(u)du,$$

where $Z_{1,k,j}(u)$ is a zero mean Gaussian process with covariance $C_{k,j}(u,u')$ ($j=1$ corresponds to rolling and $j=2$ to recursive estimation schemes), equal to:

$$E\left(\sum_{j=-\infty}^{\infty}\left(\left(1\{y_{s+1}\leq u\}-F_1(u|Z^s,\theta_1^\dagger)\right)^2-\mu_1^2(u)\right)\left(\left(1\{y_{s+j+1}\leq u'\}-F_1(u'|Z^{s+j},\theta_1^\dagger)\right)^2-\mu_1^2(u')\right)\right)$$

$$+E\left(\sum_{j=-\infty}^{\infty}\left(\left(1\{y_{s+1}\leq u\}-F_k(u|Z^s,\theta_k^\dagger)\right)^2-\mu_k^2(u)\right)\left(\left(1\{y_{s+j+1}\leq u'\}-F_k(u'|Z^{s+j},\theta_k^\dagger)\right)^2-\mu_k^2(u')\right)\right)$$

$$-2E\left(\sum_{j=-\infty}^{\infty}\left(\left(1\{y_{s+1}\leq u\}-F_1(u|Z^s,\theta_1^\dagger)\right)^2-\mu_1^2(u)\right)\left(\left(1\{y_{s+j+1}\leq u'\}-F_k(u'|Z^{s+j},\theta_k^\dagger)\right)^2-\mu_k^2(u')\right)\right)$$

$$+4\Pi_j m_{\theta_1^\dagger}(u)'A(\theta_1^\dagger)E\left(\sum_{j=-\infty}^{\infty}\nabla_{\theta_1}\ln f_1(y_{s+1}|Z^s,\theta_1^\dagger)\nabla_{\theta_1}\ln f_1(y_{s+j+1}|Z^{s+j},\theta_1^\dagger)'\right)A(\theta_1^\dagger)m_{\theta_1^\dagger}(u')$$

$$+4\Pi_j m_{\theta_k^\dagger}(u)'A(\theta_k^\dagger)E\left(\sum_{j=-\infty}^{\infty}\nabla_{\theta_k}\ln f_k(y_{s+1}|Z^s,\theta_k^\dagger)\nabla_{\theta_k}\ln f_k(y_{s+j+1}|Z^{s+j},\theta_k^\dagger)'\right)A(\theta_k^\dagger)m_{\theta_k^\dagger}(u')$$

$$-4\Pi_j m_{\theta_1^\dagger}(u,)'A(\theta_1^\dagger)E\left(\sum_{j=-\infty}^{\infty}\nabla_{\theta_1}\ln f_1(y_{s+1}|Z^s,\theta_1^\dagger)\nabla_{\theta_k}\ln f_k(y_{s+j+1}|Z^{s+j},\theta_k^\dagger)'\right)A(\theta_k^\dagger)m_{\theta_k^\dagger}(u')$$

$$-4C\Pi_j m_{\theta_1^\dagger}(u)'A(\theta_1^\dagger)E\left(\sum_{j=-\infty}^{\infty}\nabla_{\theta_1}\ln f_1(y_{s+1}|Z^s,\theta_1^\dagger)\left(\left(1\{y_{s+j+1}\leq u\}-F_1(u|Z^{s+j},\theta_1^\dagger)\right)^2-\mu_1^2(u)\right)\right)$$

$$+4C\Pi_j m_{\theta_1^\dagger}(u)'A(\theta_1^\dagger)E\left(\sum_{j=-\infty}^{\infty}\nabla_{\theta_1}\ln f_1(y_{s+1}|Z^s,\theta_1^\dagger)\left(\left(1\{y_{s+j+1}\leq u\}-F_k(u|Z^{s+j},\theta_k^\dagger)\right)^2-\mu_k^2(u)\right)\right)$$

$$-4C\Pi_j m_{\theta_k^\dagger}(u)' A(\theta_k^\dagger) E\left(\sum_{j=-\infty}^{\infty} \nabla_{\theta_k} \ln f_k(y_{s+1}|Z^s,\theta_k^\dagger)' \left(\left(1\{y_{s+j+1}\leq u\} - F_k(u|Z^{s+j},\theta_k^\dagger)\right)^2 - \mu_k^2(u)\right)\right)$$
$$+4C\Pi_j m_{\theta_k^\dagger}(u)' A(\theta_k^\dagger) E\left(\sum_{j=-\infty}^{\infty} \nabla_{\theta_k} \ln f_k(y_{s+1}|Z^s,\theta_k^\dagger)' \left(\left(1\{y_{s+j+1}\leq u\} - F_1(u|Z^{s+j},\theta_1^\dagger)\right)^2 - \mu_1^2(u)\right)\right), \tag{54}$$

with $m_{\theta_i^\dagger}(u)' = E\left(\nabla_{\theta_i} F_i(u|Z^t,\theta_i^\dagger)' \left(1\{y_{t+1}\leq u\} - F_i(u|Z^t,\theta_i^\dagger)\right)\right)$ and $A(\theta_i^\dagger) = A_i^\dagger = \left(E\left(-\nabla^2_{\theta_i} \ln f_i(y_{t+1}|Z^t,\theta_i^\dagger)\right)\right)^{-1}$, and for $j=1$ and $P \leq R$, $\Pi_1 = \left(\pi - \frac{\pi^2}{3}\right)$, $C\Pi_1 = \frac{\pi}{2}$, and for $P > R$, $\Pi_1 = \left(1 - \frac{1}{3\pi}\right)$ and $C\Pi_1 = \left(1 - \frac{1}{2\pi}\right)$. Finally, for $j=2$, $\Pi_2 = 2\left(1 - \pi^{-1}\ln(1+\pi)\right)$ and $C\Pi_2 = 0.5\Pi_2$.

From this proposition, note that when all competing models provide an approximation to the true conditional distribution that is as (mean square) accurate as that provided by the benchmark (i.e. when $\int_U \left(\mu_1^2(u) - \mu_k^2(u)\right)\phi(u)du = 0, \forall k$), then the limiting distribution is a zero mean Gaussian process with a covariance kernel which is not nuisance parameter free. Additionally, when all competitor models are worse than the benchmark, the statistic diverges to minus infinity at rate $\sqrt{P}$. Finally, when only some competitor models are worse than the benchmark, the limiting distribution provides a conservative test, as Z_P will always be smaller than $\max_{k=2,\ldots,m} \int_U \left(Z_{P,u}(1,k) - \sqrt{P}\left(\mu_1^2(u) - \mu_k^2(u)\right)\right)\phi(u)du$, asymptotically. Of course, when H_A holds, the statistic diverges to plus infinity at rate $\sqrt{P}$.

For the case of evaluation of multiple conditional confidence intervals, consider the statistic:

$$V_{P,\tau} = \max_{k=2,\ldots,m} V_{P,\underline{u},\overline{u},\tau}(1,k) \tag{55}$$

where

$$\begin{aligned} V_{P,\underline{u},\overline{u},\tau}(1,k) \;=\; & \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(\left(1\{\underline{u} \leq y_{t+1} \leq \overline{u}\} - \left(F_1(\overline{u}|Z^t,\widehat{\theta}_{1,t,\tau}) - F_1(\underline{u}|Z^t,\widehat{\theta}_{1,t,\tau})\right)\right)^2 \right. \\ & \left. - \left(1\{\underline{u} \leq y_{t+1} \leq \overline{u}\} - \left(F_k(\overline{u}|Z^t,\widehat{\theta}_{k,t,\tau}) - F_k(\underline{u}|Z^t,\widehat{\theta}_{k,t,\tau})\right)\right)^2\right) \end{aligned} \tag{56}$$

where $s = \max\{s_1, s_2\}$, $\tau = 1, 2$, $\widehat{\theta}_{k,t,\tau} = \widehat{\theta}_{k,t,rol}$ for $\tau = 1$, and $\widehat{\theta}_{k,t,\tau} = \widehat{\theta}_{k,t,rec}$ for $\tau = 2$.

We then have the following result.

Proposition 4.10 (from Proposition 1b in Corradi and Swanson (2004a)).
Let Assumptions **MD1-MD4** hold. Then for $\tau = 1$,

$$\max_{k=2,\ldots,m} \left(V_{P,\underline{u},\overline{u},\tau}(1,k) - \sqrt{P}\left(\mu_1^2 - \mu_k^2\right)\right) \xrightarrow{d} \max_{k=2,\ldots,m} V_{P,k,\tau}(\underline{u},\overline{u}),$$

where $V_{P,k,\tau}(\underline{u},\overline{u})$ is a zero mean normal random variable with covariance $c_{kk} = v_{kk} + p_{kk} + cp_{kk}$, where v_{kk} denotes the component of the long-run variance matrix we would have in absence of parameter estimation error, p_{kk} denotes the contribution of parameter estimation error and and cp_{kk} denotes the covariance across the two components. In particular:

$$\begin{aligned} v_{kk} &= E\sum_{j=-\infty}^{\infty}\left(\left(\left(1\{\underline{u}\le y_{s+1}\le\overline{u}\}-\left(F_1(\overline{u}|Z^s,\theta_1^\dagger)-F_1(\underline{u}|Z^s,\theta_1^\dagger)\right)\right)^2-\mu_1^2\right)\right. \\ &\quad \left.\left(\left(1\{\underline{u}\le y_{s+1+j}\le\overline{u}\}-\left(F_1(\overline{u}|Z^{s+j},\theta_1^\dagger)-F_1(\underline{u}|Z^{s+j},\theta_1^\dagger)\right)\right)^2-\mu_1^2\right)\right) \end{aligned} \tag{57}$$

$$\begin{aligned} &+E\sum_{j=-\infty}^{\infty}\left(\left(\left(1\{\underline{u}\le y_{s+1}\le\overline{u}\}-\left(F_k(\overline{u}|Z^s,\theta_k^\dagger)-F_k(\underline{u}|Z^s,\theta_k^\dagger)\right)\right)^2-\mu_k^2\right)\right. \\ &\left.\left(\left(1\{\underline{u}\le y_{s+1+j}\le\overline{u}\}-\left(F_k(\overline{u}|Z^{s+j},\theta_k^\dagger)-F_k(\underline{u}|Z^{s+j},\theta_k^\dagger)\right)\right)^2-\mu_k^2\right)\right) \end{aligned} \tag{58}$$

$$\begin{aligned} &-2E\sum_{j=-\infty}^{\infty}\left(\left(\left(1\{\underline{u}\le y_{s+1}\le\overline{u}\}-\left(F_1(\overline{u}|Z^s,\theta_1^\dagger)-F_1(\underline{u}|Z^s,\theta_1^\dagger)\right)\right)^2-\mu_1^2\right)\right. \\ &\left.\left(\left(1\{\underline{u}\le y_{s+1+j}\le\overline{u}\}-\left(F_k(\overline{u}|Z^{s+j},\theta_k^\dagger)-F_k(\underline{u}|Z^{s+j},\theta_k^\dagger)\right)\right)^2-\mu_k^2\right)\right) \end{aligned} \tag{59}$$

$$p_{kk} = 4m_{\theta_1^\dagger}{}'A(\theta_1^\dagger)E\left(\sum_{j=-\infty}^{\infty}\nabla_{\theta_1}\ln f_1(y_{s+1}|Z^s,\theta_1^\dagger)\nabla_{\theta_1}\ln f_1(y_{s+1+j}|Z^{s+j},\theta_1^\dagger)'\right)A(\theta_1^\dagger)m_{\theta_1^\dagger} \tag{60}$$

$$+4m'_{\theta_k^\dagger}A(\theta_k^\dagger)E\left(\sum_{j=-\infty}^{\infty}\nabla_{\theta_k}\ln f_k(y_{s+1}|Z^s,\theta_k^\dagger)\nabla_{\theta_k}\ln f_k(y_{s+1+j}|Z^{s+j},\theta_k^\dagger)'\right)A(\theta_k^\dagger)m_{\theta_k^\dagger} \tag{61}$$

$$-8m'_{\theta_1^\dagger}A(\theta_1^\dagger)E\left(\sum_{j=-\infty}^{\infty}\nabla_{\theta_1}\ln f_1(y_{s+1}|Z^s,\theta_1^\dagger)\nabla_{\theta_k}\ln f_k(y_{s+1+j}|Z^{s+j},\theta_k^\dagger)'\right)A(\theta_k^\dagger)m_{\theta_k^\dagger} \tag{62}$$

$$\begin{aligned} cp_{kk} &= -4m'_{\theta_1^\dagger}A(\theta_1^\dagger)E\left(\sum_{j=-\infty}^{\infty}\nabla_{\theta_1}\ln f_1(y_{s+1}|Z^s,\theta_1^\dagger)\right. \\ &\quad \left.\left(\left(1\{\underline{u}\le y_{s+j}\le\overline{u}\}-\left(F_1(\overline{u}|Z^{s+j},\theta_1^\dagger)-F_1(\underline{u}|Z^{s+j},\theta_1^\dagger)\right)\right)^2-\mu_1^2\right)\right) \\ &\quad +8m_{\theta_1^\dagger}{}'A(\theta_1^\dagger)E\left(\sum_{j=-\infty}^{\infty}\nabla_{\theta_1}\ln f_1(y_s|Z^s,\theta_1^\dagger)\right. \\ &\quad \left.\left(\left(1\{\underline{u}\le y_{s+1+j}\le\overline{u}\}-\left(F_k(\overline{u}|Z^{s+j},\theta_k^\dagger)-F_k(\underline{u}|Z^{s+j},\theta_k^\dagger)\right)\right)^2-\mu_k^2\right)\right) \end{aligned}$$

$$-4m'_{\theta_k^\dagger}A(\theta_k^\dagger)E\left(\sum_{j=-\infty}^{\infty}\nabla_{\theta_k}\ln f_k(y_{s+1}|Z^s,\theta_k^\dagger)\left(\left(1\{\underline{u}\leq y_{s+j}\leq\overline{u}\}-\left(F_k(\overline{u}|Z^{s+j},\theta_k^\dagger)-F_k(\underline{u}|Z^{s+j},\theta_k^\dagger)\right)\right)^2-\mu_k^2\right)\right)$$

with $m_{\theta_i^\dagger}{}' = E\left(\nabla_{\theta_i}\left(F_i(\overline{u}|Z^t,\theta_i^\dagger)-F_i(\overline{u}|Z^t,\theta_i^\dagger)\right)\left(1\{\underline{u}\leq y_t\leq\overline{u}\}-\left(F_i(\overline{u}|Z^t,\theta_i^\dagger)-F_i(\overline{u}|Z^t,\theta_i^\dagger)\right)\right)\right)$ and $A(\theta_i^\dagger) = \left(E\left(-\ln\nabla^2_{\theta_i}f_i(y_t|Z^t,\theta_i^\dagger)\right)\right)^{-1}$. An analogous result holds for the case where $\tau = 2$, and is omitted for the sake of brevity.

5.2.3 Bootstrap Critical Values for the Density Accuracy Test

Turning now to the construction of critical values for the above test, note that using the bootstrap sampling procedures defined in Sections 3.4.1 and 3.5.1 or in Sections 3.4.2 and 3.5.2, one first constructs appropriate bootstrap samples. Thereafter, form bootstrap statistics as follows

$$Z^*_{P,\tau} = \max_{k=2,\ldots,m}\int_U Z^*_{P,u,\tau}(1,k)\phi(u)du,$$

where for $\tau = 1$ (rolling estimation scheme), and for $\tau = 2$ (recursive estimation scheme):

$$\begin{aligned}Z^*_{P,u,\tau}(1,k) &= \frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\left(\left(1\{y^*_{t+1}\leq u\}-F_1(u|Z^{*,t},\widetilde{\theta}^*_{1,t,\tau})\right)^2-\left(1\{y^*_{t+1}\leq u\}-F_k(u|Z^{*,t},\widetilde{\theta}^*_{k,t,\tau})\right)^2\right)\right.\\ &\left.-\frac{1}{T}\sum_{j=s+1}^{T-1}\left(\left(1\{y_{j+1}\leq u\}-F_1(u|Z^i,\widehat{\theta}_{1,t,\tau})\right)^2-\left(1\{y_{j+1}\leq u\}-F_k(u|Z^j,\widehat{\theta}_{k,t,\tau})\right)^2\right)\right)\end{aligned}$$

Note that each bootstrap term, say $1\{y^*_{t+1}\leq u\}-F_i(u|Z^{*,t},\widetilde{\theta}^*_{i,t,\tau})$, $t\geq R$, is recentered around the (full) sample mean $\frac{1}{T}\sum_{j=s+1}^{T-1}\left(1\{y_{j+1}\leq u\}-F_i(u|Z^i,\widehat{\theta}_{i,t,\tau})\right)^2$. This is necessary as the bootstrap statistic is constructed using the last P resampled observations, which in turn have been resampled from the full sample. In particular, this is necessary regardless of the ratio P/R. If $P/R\to 0$, then we do not need to mimic parameter estimation error, and so could simply use $\widehat{\theta}_{1,t,\tau}$ instead of $\widetilde{\theta}^*_{1,t,\tau}$, but we still need to recenter any bootstrap term around the (full) sample mean.

For the confidence interval case, define:

$$V^*_{P,\tau} = \max_{k=2,\ldots,m}V^*_{P,\underline{u},\overline{u},\tau}(1,k)$$

$$V^*_{P,\underline{u},\overline{u},\tau}(1,k) = \frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\left(\left(1\{\underline{u}\leq y^*_{t+1}\leq\overline{u}\}-\left(F_1(\overline{u}|Z^{*t},\widetilde{\theta}^*_{1,t,\tau})-F_1(\underline{u}|Z^{*t},\widetilde{\theta}^*_{1,t,\tau})\right)\right)^2\right.\right.$$
$$\left.-\left(1\{\underline{u}\leq y^*_{t+1}\leq\overline{u}\}-\left(F_k(\overline{u}|Z^{*t},\widetilde{\theta}^*_{k,t,\tau})-F_1(\underline{u}|Z^{*t},\widetilde{\theta}^*_{k,t,\tau})\right)\right)^2\right)$$
$$-\frac{1}{T}\sum_{j=s+1}^{T-1}\left(\left(1\{\underline{u}\leq y_{i+1}\leq\overline{u}\}-\left(F_1(\overline{u}|Z^j,\widehat{\theta}_{1,t,\tau})-F_1(\underline{u}|Z^j,\widehat{\theta}_{1,t,\tau})\right)\right)^2\right.$$
$$\left.\left.-\left(1\{\underline{u}\leq y_{j+1}\leq\overline{u}\}-\left(F_k(\overline{u}|Z^j,\widehat{\theta}_{k,t,\tau})-F_1(\underline{u}|Z^j,\widehat{\theta}_{k,t,\tau})\right)\right)^2\right)\right),$$

where, as usual, $\tau = 1, 2$. The following results then hold.

Proposition 4.11 (from Proposition 6 in Corradi and Swanson (2004a)):

Let Assumptions **MD1-MD4** hold. Also, assume that as $T\to\infty$, $l\to\infty$, and that $\frac{l}{T^{1/4}}\to 0$. Then, as T, P and $R\to\infty$, for $\tau = 1, 2$:

$$P\left(\omega : \sup_{v\in\Re}\left|P^*_T\left(\max_{k=2,\ldots,m}\int_U Z^*_{P,u,\tau}(1,k)\phi(u)du\leq v\right)-P\left(\max_{k=2,\ldots,m}\int_U Z^{\mu}_{P,u,\tau}(1,k)\phi(u)du\leq v\right)\right|>\varepsilon\right)\to 0,$$

where $Z^{\mu}_{P,u,\tau}(1,k) = Z_{P,u,\tau}(1,k)-\sqrt{P}\left(\mu^2_1(u)-\mu^2_k(u)\right)$, and where $\mu^2_1(u)-\mu^2_k(u)$ is defined as in equation (51).

Proposition 4.12 (from Proposition 7 in Corradi and Swanson (2004a)):

Let Assumptions **MD1-MD4** hold. Also, assume that as $T\to\infty$, $l\to\infty$, and that $\frac{l}{T^{1/4}}\to 0$. Then, as T, P and $R\to\infty$, for $\tau = 1, 2$:

$$P\left(\omega : \sup_{v\in\Re}\left|P^*_T\left(\max_{k=2,\ldots,m}V^*_{P,\underline{u},\overline{u},\tau}(1,k)\leq v\right)-P\left(\max_{k=2,\ldots,m}V^*_{P,\underline{u},\overline{u},\tau}(1,k)\leq v\right)\right|>\varepsilon\right)\to 0,$$

where $V^{\mu}_{P,j}(1,k) = V_{P,j}(1,k)-\sqrt{P}\left(\mu^2_1-\mu^2_k\right)$, and where $\mu^2_1-\mu^2_k$ is defined as in equation (**??**).

The above results suggest proceeding in the following manner. For brevity, just consider the case of $Z^*_{P,\tau}$. For any bootstrap replication, compute the bootstrap statistic, $Z^*_{P,\tau}$. Perform B bootstrap replications (B large) and compute the quantiles of the empirical distribution of the B bootstrap statistics. Reject H_0, if $Z_{P,\tau}$ is greater than the $(1-\alpha)th$-percentile. Otherwise, do not reject. Now, for all samples except a set with probability measure approaching zero, $Z_{P,\tau}$ has the same limiting distribution as the corresponding bootstrapped statistic when $E\left(\mu^2_1(u)-\mu^2_k(u)\right) = 0$, $\forall\ k$, ensuring asymptotic size equal to α. On the other hand, when one or more competitor models are strictly dominated by the benchmark, the rule provides a test with asymptotic size between 0 and α. Under the alternative, $Z_{P,\tau}$ diverges to (plus) infinity, while the corresponding bootstrap statistic has a well defined limiting distribution, ensuring unit asymptotic power. From the above discussion, we see that the bootstrap distribution provides correct asymptotic critical values

only for the least favorable case under the null hypothesis; that is, when all competitor models are as good as the benchmark model. When $\max_{k=2,\ldots,m} \int_U \left(\mu_1^2(u) - \mu_k^2(u)\right) \phi(u)du = 0$, but $\int_U \left(\mu_1^2(u) - \mu_k^2(u)\right) \phi(u)du < 0$ for some k, then the bootstrap critical values lead to conservative inference. An alternative to our bootstrap critical values in this case is the construction of critical values based on subsampling (see e.g. Politis, Romano and Wolf (1999), Ch. 3). Heuristically, construct $T - 2b_T$ statistics using subsamples of length b_T, where $b_T/T \to 0$. The empirical distribution of these statistics computed over the various subsamples properly mimics the distribution of the statistic. Thus, subsampling provides valid critical values even for the case where $\max_{k=2,\ldots,m} \int_U \left(\mu_1^2(u) - \mu_k^2(u)\right) \phi(u)du = 0$, but $\int_U \left(\mu_1^2(u) - \mu_k^2(u)\right) \phi(u)du < 0$ for some k. This is the approach used by Linton, Maasoumi and Whang (2003), for example, in the context of testing for stochastic dominance. Needless to say, one problem with subsampling is that unless the sample is very large, the empirical distribution of the subsampled statistics may yield a poor approximation of the limiting distribution of the statistic. An alternative approach for addressing the conservative nature of our bootstrap critical values is suggested in Hansen (2001). Hansen's idea is to recenter the bootstrap statistics using the sample mean, whenever the latter is larger than (minus) a bound of order $\sqrt{2T \log\log T}$. Otherwise, do not recenter the bootstrap statistics. In the current context, his approach leads to correctly sized inference when $\max_{k=2,\ldots,m} \int_U \left(\mu_1^2(u) - \mu_k^2(u)\right) \phi(u)du = 0$, but $\int_U \left(\mu_1^2(u) - \mu_k^2(u)\right) \phi(u)du < 0$ for some k. Additionally, his approach has the feature that if all models are characterized by a sample mean below the bound, the null is "accepted" and no bootstrap statistic is constructed.

Part IV: Appendix and References

6 Appendix

Proof of Proposition 3.2:

For brevity, we just consider the case of recursive estimation. The case of rolling estimation schemes can be treated in an analogous way.

$$\begin{aligned}
\widehat{W}_{P,rec} &= \frac{1}{\sqrt{P}}\sum_{t=R+1}^{T}\Big(1\{F_t(y_t|Z^{t-1},\widehat{\theta}_{t,rec})\leq r\}-r\Big) \\
&= \frac{1}{\sqrt{P}}\sum_{t=R+1}^{T}\Big(1\{F_t(y_t|Z^{t-1},\theta_0)\leq F(F^{-1}(r|Z^{t-1},\widehat{\theta}_{t,rec})|Z^{t-1},\theta_0)\}-r\Big) \\
&= \frac{1}{\sqrt{P}}\sum_{t=R+1}^{T}\Big(1\{F_t(y_t|Z^{t-1},\theta_0)\leq F(F^{-1}(r|Z^{t-1},\widehat{\theta}_{t,rec})|Z^{t-1},\theta_0)\}-F(F^{-1}(r|Z^{t-1},\widehat{\theta}_{t,rec})|Z^{t-1},\theta_0)\Big) \\
&\quad +\frac{1}{\sqrt{P}}\sum_{t=R+1}^{T}\Big(F(F^{-1}(r|Z^{t-1},\widehat{\theta}_t)|Z^{t-1},\theta_0)-r\Big) \\
&= I_P+II_P.
\end{aligned}$$

We first want to show that:

(i) $I_P=\frac{1}{\sqrt{P}}\sum_{t=R+1}^{T}\left(1\{F_t(y_t|Z^{t-1},\theta_0)\leq r\}-r\right)+o_P(1)$, uniformly in r, and

(ii) $II_P=\overline{g}(r)\frac{1}{\sqrt{P}}\sum_{t=R+1}^{T}\left(\widehat{\theta}_{t,rec}-\theta_0\right)+o_P(1)$, uniformly in r.

Given BAI2, (ii) follows immediately. For (i), we need to show that

$$\begin{aligned}
&\frac{1}{\sqrt{P}}\sum_{t=R+1}^{T}(1\left\{F_t(y_t|Z^{t-1},\theta_0)\leq r+\frac{\partial F_t}{\partial\theta}\left(F_t^{-1}(r|\overline{\theta}_{t,rec}),\theta_0\right)\left(\widehat{\theta}_{t,rec}-\theta_0\right)\right\}\\
&-\left(r+\frac{\partial F_t}{\partial\theta}\left(F_t^{-1}(r|\overline{\theta}_{t,rec}),\theta_0\right)\left(\widehat{\theta}_t-\theta_0\right)\right))\\
=\ &\frac{1}{\sqrt{P}}\sum_{t=R+1}^{T}(1\{F_t(y_t|\Omega_{t-1},\theta_0)\leq r\}-r)+o_P(1),\ \text{uniformly in } r
\end{aligned}$$

Given BAI3', the equality above follows by the same argument as that used in the proof of Theorem 1 in Bai (2003). Given (i) and (ii), it follows that

$$\widehat{V}_{P,rec}=\frac{1}{\sqrt{P}}\sum_{t=R+1}^{T}(1\{F_t(y_t|\Omega_{t-1},\theta_0)\leq r\}-r)+\overline{g}(r)\frac{1}{\sqrt{P}}\sum_{t=R+1}^{T}\left(\widehat{\theta}_{t,rec}-\theta_0\right)+o_P(1), \tag{63}$$

uniformly in r, where $\overline{g}(r) = \text{plim}\frac{1}{P}\sum_{t=R+1}^{T}\frac{\partial F_t}{\partial \theta}\left(F_t^{-1}(r|\overline{\theta}_{t,rec}),\theta_0\right)$, $\overline{\theta}_{t,rec} \in \left(\widehat{\theta}_{t,rec},\theta_0\right)$.

The desired outcome follows if the martingalization argument applies also in the recursive estimation case and the parameter estimation error component cancel out in the statistic. Now, equation A4 in Bai (2003) holds in the form of eq. (63) above. Also,

$$\widehat{W}_{P,rol}(r) = \widehat{V}_{P,rol}(r) - \int_0^r \left(\dot{g}(s)C^{-1}(s)\dot{g}(s)'\int_s^1 \dot{g}(\tau)d\widehat{V}_{P,rol}(\tau)\right)ds. \tag{64}$$

It remain to show that the parameter estimation error term, which enters into both $\widehat{V}_{P,rol}(r)$ and $d\widehat{V}_{P,rol}(\tau)$, cancels out, as in the fixed estimation scheme. Notice that $g(r)$ is defined as in the fixed scheme. Now, it suffices to define the term c, which appears at the bottom of p. 543 (below equation A6 in Bai (2003)) as:

$$c = \frac{1}{\sqrt{P}}\sum_{t=R+1}^{T}\left(\widehat{\theta}_{t,rec} - \theta_0\right).$$

Then, the same argument used by Bai (2003) on p. 544 applies here, and the term $\frac{1}{\sqrt{P}}\sum_{t=R+1}^{T}\left(\widehat{\theta}_{t,rec} - \theta_0\right)$ on the RHS in (64) cancels out.

Proof of Proposition 3.4: (i) We begin by considering the case of recursive estimation. Given CS1 and CS3, $\widehat{\theta}_{t,rec} \overset{a.s.}{\to} \theta^{\dagger}$, with $\theta^{\dagger} = \theta_0$, under H_0. Given A2(i), and following Bai (2003, p. 545-546), we have that:

$$\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}(1\{F(y_{t+1}|Z^t,\widehat{\theta}_{t,rec}) \leq r\} - r) = \frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(1\left\{F(y_{t+1}|Z^t,\theta_0) \leq F\left(F^{-1}\left(r|Z^t,\widehat{\theta}_{t,rec}\right)|Z^t,\theta_0\right)\right\} - r\right)$$

$$\begin{aligned} = & \frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(1\{F(y_{t+1}|Z^t,\theta_0) \leq F\left(F^{-1}\left(r|Z^t,\widehat{\theta}_{t,rec}\right)|Z^t,\theta_0\right)\} - F\left(F^{-1}\left(r|Z^t,\theta_0\right)|Z^t,\theta_0\right)\right) \\ & -\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\nabla_\theta F(F^{-1}\left(r|Z^t,\overline{\theta}_{t,rec}\right)|Z^t,\theta_0)(\widehat{\theta}_{t,rec} - \theta_0), \end{aligned} \tag{65}$$

with $\overline{\theta}_{t,rec} \in (\widehat{\theta}_{t,rec},\theta_0)$. Given CS1 and CS3, $(\widehat{\theta}_{t,rec} - \theta_0) = O_P(1)$, uniformly in t. Thus, the first term on the RHS of (65) can be treated by the same argument as that used in the proof of Theorem 1 in Corradi and Swanson (2003). With regard to the last term on the RHS of (65), note that by the uniform law of large numbers for mixing processes,

$$\begin{aligned} & \frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\nabla_\theta F(F^{-1}\left(r|Z^t,\overline{\theta}_{t,rec}\right)|Z^t,\theta_0)(\widehat{\theta}_{t,rec} - \theta_0) \\ = & E(\nabla_\theta F(x(r)|Z^{t-1},\theta_0))'\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}(\widehat{\theta}_{t,rec} - \theta_0) + o_P(1), \end{aligned} \tag{66}$$

where the $o_P(1)$ term is uniform in r. The limiting distribution of $\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}(\widehat{\theta}_{t,rec} - \theta_0)$, and so the key contribution of parameter estimation error, comes from Theorem 4.1 and Lemma 4.1 in West (1996). With

regard to the rolling case, the same argument as above applies, with $\widehat{\theta}_{t,rec}$ replaced by $\widehat{\theta}_{t,rol}$. The limiting distribution of $\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}(\widehat{\theta}_{t,rec}-\theta_0)$ is given by Lemma 4.1 and 4.2 in West and McCracken (1998).

Proof of Proposition 3.5: The proof is straightforward upon combining the proof of Theorem 2 in Corradi and Swanson (2003) and the proof of Proposition 3.4.

Proof of Proposition 3.7: Note that:

$$\begin{aligned}
&\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(1\{F(y_{t+1}^*|Z^{*,t},\widetilde{\theta}_{t,rec}^*)\leq r\}-\frac{1}{T}\sum_{j=1}^{T-1}1\{F(y_{j+1}|Z^j,\widehat{\theta}_{t,rec})\leq r\}\right)\\
=\;&\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(1\{F(y_{t+1}^*|Z^{*,t},\widehat{\theta}_{t,rec})\leq r\}-\frac{1}{T}\sum_{j=1}^{T-1}1\{F(y_{j+1}|Z^j,\widehat{\theta}_{t,rec})\leq r\}\right)\\
&-\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\nabla_\theta F(F^{-1}\left(r|Z^t,\overline{\theta}_{t,rec}^*\right)|Z^t,\theta_0)(\widetilde{\theta}_{t,rec}^*-\widehat{\theta}_{t,rec}),
\end{aligned}\tag{67}$$

where $\overline{\theta}_{t,rec}^*\in\left(\widetilde{\theta}_{t,rec}^*,\widehat{\theta}_{t,rec}\right)$. Now, the first term on the RHS of (67) has the same limiting distribution as $\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(1\{F(y_{t+1}|Z^t,\theta^\dagger)\leq r\}-E\left(1\{F(y_{j+1}|Z^j,\theta^\dagger)\leq r\}\right)\right)$, conditional on the sample. Furthermore, given Theorem 3.6, the last term on the RHS of (67) has the same limiting distribution as

$$E(\nabla_\theta F(x(r)|Z^{t-1},\theta_0))'\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widehat{\theta}_{t,rec}-\theta^\dagger\right),$$

conditional on the sample. The rolling case follows directly, by replacing $\widetilde{\theta}_{t,rec}^*$ and $\widehat{\theta}_{t,rec}$ with $\widetilde{\theta}_{t,rol}^*$ and $\widehat{\theta}_{t,rol}$, respectively.

Proof of Proposition 3.8: The proof is similar to the proof of Proposition 3.7.

Proof of Proposition 4.5 (ii): Note that, via a mean value expansion, and given A1,A2,

$$\begin{aligned}
S_P(1,k) &= \frac{1}{P^{1/2}}\sum_{t=R}^{T-1}(g(\widehat{u}_{1,t+1})-g(\widehat{u}_{k,t+1}))\\
&= \frac{1}{P^{1/2}}\sum_{t=R}^{T-1}(g(u_{1,t+1})-g(u_{k,t+1}))\\
&\quad+\frac{1}{P}\sum_{t=R}^{T-1}g'(\overline{u}_{1,t+1})\nabla_{\theta_1}\kappa_1(Z^t,\overline{\theta}_{1,t})P^{1/2}\left(\widehat{\theta}_{1,t}-\theta_1^\dagger\right)\\
&\quad-\frac{1}{P}\sum_{t=R}^{T-1}g'(\overline{u}_{k,t+1})\nabla_{\theta_k}\kappa_k(Z^t,\overline{\theta}_{k,t})P^{1/2}\left(\widehat{\theta}_{k,t}-\theta_k^\dagger\right)\\
&= \frac{1}{P^{1/2}}\sum_{t=R}^{T-1}(g(u_{1,t+1})-g(u_{k,t+1}))\\
&\quad+\mu_1\frac{1}{P^{1/2}}\sum_{t=R}^{T-1}\left(\widehat{\theta}_{1,t}-\theta_1^\dagger\right)-\mu_k\frac{1}{P^{1/2}}\sum_{t=R}^{T-1}\left(\widehat{\theta}_{k,t}-\theta_k^\dagger\right)+o_P(1),
\end{aligned}$$

where $\mu_1 = E\left(g'(u_{1,t+1})\nabla_{\theta_1}\kappa_1(Z^t,\theta_1^\dagger)\right)$, and μ_k is defined analogously. Now, when all competitors have the same predictive accuracy as the benchmark model, by the same argument as that used in Theorem 4.1 in West (1996),

$$(S_P(1,2),...,S_P(1,n)) \xrightarrow{d} N(0,V),$$

where V is the $n \times n$ matrix defined in the statement of the proposition.

Proof of Proposition 4.6(ii): For brevity, we just analyze model 1. In particular, note that:

$$\frac{1}{P^{1/2}}\sum_{t=R}^{T-1}\left(g(\widehat{u}^*_{1,t+1}) - g(\widehat{u}_{1,t+1})\right) = \frac{1}{P^{1/2}}\sum_{t=R}^{T-1}\left(g(u^*_{1,t+1}) - g(u_{1,t+1})\right)$$

$$+\frac{1}{P^{1/2}}\sum_{t=R}^{T-1}\left(\nabla_{\theta_1}g(\overline{u}^*_{1,t+1})\left(\widehat{\theta}^*_{1,t} - \theta_1^\dagger\right) - \nabla_{\theta_1}g(\overline{u}_{1,t+1})\left(\widehat{\theta}_{1,t} - \theta_1^\dagger\right)\right), \tag{68}$$

where $\overline{u}^*_{1,t+1} = y_{t+1} - \kappa_1(Z^{*,t},\overline{\theta}^*_{1,t})$, $\overline{u}_{1,t+1} = y_{t+1} - \kappa_1(Z^t,\overline{\theta}_{1,t})$, $\overline{\theta}^*_{1,t} \in (\widehat{\theta}^*_{1,t},\theta_1^\dagger)$ and $\overline{\theta}_{1,t} \in (\widehat{\theta}_{1,t},\theta_1^\dagger)$. As an almost straightforward consequence of Theorem 3.5 in Künsch (1989), the first term on the RHS of (68) has the same limiting distribution as $P^{-1/2}\sum_{t=R}^{T-1}\left(g(u_{1,t+1}) - E(g(u_{1,t+1}))\right)$. Additionally, the second line in (68) can be written as:

$$\begin{aligned}
&\frac{1}{P^{1/2}}\sum_{t=R}^{T-1}\nabla_{\theta_1}g(\overline{u}^*_{1,t+1})\left(\widehat{\theta}^*_{1,t} - \widehat{\theta}_{1,t}\right) - \frac{1}{P^{1/2}}\sum_{t=R}^{T-1}\left(\nabla_{\theta_1}g(\overline{u}^*_{1,t+1}) - \nabla_{\theta_1}g(\overline{u}_{1,t+1})\right)\left(\widehat{\theta}_{1,t} - \theta_1^\dagger\right)\\
= &\frac{1}{P^{1/2}}\sum_{t=R}^{T-1}\nabla_{\theta_1}g(\overline{u}^*_{1,t+1})\left(\widehat{\theta}^*_{1,t} - \widehat{\theta}_{1,t}\right) + o^*_P(1),\ \Pr - P\\
= &\mu_1 B_1^\dagger\frac{1}{P^{1/2}}\sum_{t=R}^{T-1}\left(h^*_{1,t} - h_{1,t}\right) + o^*_P(1),\ \Pr - P,
\end{aligned} \tag{69}$$

where $h^*_{1,t+1} = \nabla_{\theta_1}q_1(y^*_{t+1},Z^{*,t},\theta_1^\dagger)$ and $h_{1,t+1} = \nabla_{\theta_1}q_1(y_{t+1},Z^t,\theta_1^\dagger)$. Also, the last line in (69) can be written as:

$$\begin{aligned}
&\mu_1 B_1^\dagger\left(a^2_{R,0}\frac{1}{P^{1/2}}\sum_{t=1}^{R}\left(h^*_{1,t} - h_{1,t}\right) + \frac{1}{P^{1/2}}\sum_{i=1}^{P-1}a_{R,i}\left(h^*_{1,R+i} - \overline{h}_{1,P}\right)\right)\\
&-\mu_1 B_1^\dagger\frac{1}{P^{1/2}}\sum_{i=1}^{P-1}a_{R,i}\left(h_{1,R+i} - \overline{h}_{1,P}\right) + o^*_P(1),\ \Pr - P,
\end{aligned} \tag{70}$$

where $\overline{h}_{1,P}$ is the sample average of $h_{1,t}$ computed over the last P observations. Given Lemma A3, by the same argument used in the proof of Theorem 1, the first line in (70) has the same limiting distribution as $\frac{1}{P^{1/2}}\sum_{t=R}^{T-1}\left(\widehat{\theta}_{1,t} - \theta_1^\dagger\right)$, conditional on sample. Therefore we need to show that the correction term for model 1 offsets the second line in (70), up to an $o(1)$ $\Pr - P$ term. Let $h_{1,t+1}\left(\widehat{\theta}_{1,T}\right) = \nabla_{\theta_1}q_1(y_{t+1},Z^t,\widehat{\theta}_{1,T})$ and let

$\overline{h}_{1,P}\left(\widehat{\theta}_{1,T}\right)$ be the sample average of $h_{1,t+1}\left(\widehat{\theta}_{1,T}\right)$, over the last P observations. Now, by the uniform law of large numbers

$$\frac{1}{T}\sum_{t=s}^{T-1}\nabla_{\theta_1}g(\overline{u}^*_{1,t+1})\left(\frac{1}{T}\sum_{t=s}^{T-1}\nabla^2_{\theta_1}q_1(y^*_t,Z^{*,t-1},\widehat{\theta}_{1,T})\right)^{-1}-\mu_1 B_1^{\dagger}=o^*_P(1),\ \ \Pr-P.$$

Also, by the same argument used in the proof of Theorem 1, it follows that,

$$\frac{1}{P^{1/2}}\sum_{i=1}^{P-1}a_{R,i}\left(h_{1,R+i}-\overline{h}_{1,P}\right)-\frac{1}{P^{1/2}}\sum_{i=1}^{P-1}a_{R,i}\left(h_{1,R+i}\left(\widehat{\theta}_{1,T}\right)-\overline{h}_{1,P}\left(\widehat{\theta}_{1,T}\right)\right)=o(1),\ \ \Pr-P.$$

7 References

Andrews, D.W.K., (1993), An Introduction to Econometric Applications of Empirical Process Theory for Dependent Random Variables, *Econometric Reviews,* 12, 183-216.

Andrews, D.W.K., (1997), A Conditional Kolmogorov Test, *Econometrica,* 65, 1097-1128.

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 M., Buchinsky, (2000), A Three Step Method for Choosing the Number of Bootstrap Replications, *Econometrica,* 68, 23-52.

Ashley, R., C.W.J., Granger and R. Schmalensee, (1980), Advertising and Aggregate Consumption: An Analysis of Causality, *Econometrica,* 48, 1149-1167.

Bai, J., (2003), Testing Parametric Conditional Distributions of Dynamic Models, *Review of Economics and Statistics,* 85, 531-549.

Bai, J. and S. Ng, (2001), A Consistent test for Conditional Symmetry in Time Series Models, *Journal of Econometrics,* 103, 225-258.

Bai, J. and S. Ng, (2004), Tests for Skewness, Kurtosis and Normality in Time Series Data, *Journal of Business and Economic Statistics,* forthcoming.

Baltagi, B.H., (1995), *Econometric Analysis of Panel Data*, Wiley, New York.

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 Series B,* 57, 289-300.

Benjamini, Y. and Y. Yekutieli, (2001), The Control of the False Discovery Rate in Multiple Testing Under Dependency, *Annals of Statistics,* 29, 1165-1188.

Berkowitz, J., (2001), Testing Density Forecasts with Applications to Risk Management, *Journal of Business and Economic Statistics, 19, 465-474.*

Bickel, P.J. and K.A. Doksum, (1977), *Mathematical Statistics*, Prentice Hall, Englewood Cliffs.

Bierens, H.J., (1982), Consistent Model-Specification Tests, *Journal of Econometrics,* 20, 105-134.

Bierens, H.J., (1990), A Consistent Conditional Moment Test of Functional Form, *Econometrica*, 58, 1443-1458.

Bierens, H.J. and W., Ploberger, (1997), Asymptotic Theory of Integrated Conditional Moments Tests, *Econometrica,* 65, 1129-1151.

Bontemps, C., and N. Meddahi, (2003a), Testing Normality: a GMM Approach, *Journal of Econometrics*, forthcoming.

Bontemps, C., and N. Meddahi, (2003b), Testing Distributional Assumptions: a GMM Approach, Working Paper, University of Montreal.

Brock, W., J. Lakonishok and B. LeBaron, (1992), Simple Technical Trading Rules and the Stochastic Properties of Stock Returns, *Journal of Finance*, 47, 1731-1764.

Chao, J.C., V. Corradi and N.R. Swanson, (2001), An Out of Sample Test for Granger Causality", *Macroeconomic Dynamics,* 5, 598-620

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

Clarida, R.H., L. Sarno and M.P. Taylor, (2003), The Out-of-Sample Success of Term Structure Models as Exchange-Rate Predictors: A Step Beyond, *Journal of International Economics*, 60, 61-83.

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

Clark, T.E. and M.W. McCracken, (2003), Evaluating Long-Horizon Forecasts, Working Paper, University of Missouri-Columbia.

Clark, T.E. and K.J. West, (2004), Using Out-of-Sample Mean Squared Prediction Errors to Test the Martingale Difference Hypothesis, Working Papers, University of Wisconsin.

Clements, M.P. and J. Smith, (2000), Evaluating the Forecast Densities of Linear and Nonlinear Models: Applications to Output Growth and Unemployment, *Journal of Forecasting*, 19, 255-276.

Clements, M.P. and J. Smith, (2002), Evaluating Multivariate Forecast Densities: A Comparison of Two Approaches, *International Journal of Forecasting,* 18, 397-407.

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

Corradi, V. and N.R. Swanson and C. Olivetti, (2001), Predictive Ability with Cointegrated Variables", *Journal of Econometrics,* 104, 315-358.

Corradi, V. and N.R., Swanson, (2002), A Consistent Test for Out of Sample Nonlinear Predictive Ability", *Journal of Econometrics*, 110, 353-381.

Corradi, V. and N.R. Swanson, (2003), Bootstrap Conditional Distribution Tests in the Presence of Dynamic Misspecification", *Journal of Econometrics*, forthcoming.

Corradi. V. and N.R. Swanson, (20004a), Predictive Density and Conditional Confidence Interval Accuracy Tests, Working Paper, Rutgers University.

Corradi. V. and N.R. Swanson, (2004b), A Test for Comparing Multiple Misspecified Conditional Distributions, Working Paper, Rutgers University.

Corradi. V. and N.R. Swanson, (2004c), Bootstrap Procedures for Recursive Estimation Schemes, with Applications to Forecast Model Evaluation, Working Paper, Rutgers University.

Davidson, R. and J.G. MacKinnon, (1993), *Estimation and Inference in Econometrics*, Oxford University Press, New York.

Davidson, R. and J.G. MacKinnon, (1999), Bootstrap Testing in Nonlinear Models, *International Economic Review,* 40, 487-508.

Davidson, R. and J.G. MacKinnon, (2000), Bootstrap Tests: How Many Bootstraps, *Econometric Reviews, 19, 55-68.*

DeJong, R.M., (1996), The Bierens Test Under Data Dependence, *Journal of Econometrics*, 72, 1-32.

Diebold, F.X. and C. Chen, (1996), Testing Structural Stability with Endogenous Breakpoint: A Size Comparison of Analytical and Bootstrap Procedures, *Journal of Econometrics*, 70, 221-241.

Diebold, F.X., T. Gunther and A.S. Tay, (1998), Evaluating Density Forecasts with Applications to Finance and Management, *International Economic Review,* 39, 863-883.

Diebold, F.X., J. Hahn and A.S. Tay, (1999), Multivariate Density Forecast Evaluation and Calibration in Financial Risk Management: High Frequency Returns on Foreign Exchange, *Review of Economics and Statistics*, 81, 661-673.

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

Diebold, F.X., A.S. Tay and K.D. Wallis, (1998), Evaluating Density Forecasts of Inflation: The Survey of Professional Forecasters, *in Festschrift in Honor of C.W.J. Granger*, eds. R.F. Engle and H. White, Oxford University Press, Oxford.

Duan, J.C., (2003), A Specification Test for Time Series Models by a Normality Transformation, Working Paper, University of Toronto.

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

Fernandez-Villaverde, J. and J.F. Rubio-Ramirez, (2004), Comparing Dynamic Equilibrium Models to Data, *Journal of Econometrics*, 123, 153-187.

Giacomini, R., (2002), Comparing Density Forecasts via Weighted Likelihood Ratio Tests: Asymptotic and Bootstrap Methods, Working Paper, University of California, San Diego.

Giacomini, R. and H. White, (2003), Conditional Tests for Predictive Ability, Working Paper, University of California, San Diego.

Goncalves, S. and H. White, (2002), The Bootstrap of the Mean for Dependent Heterogeneous Arrays, *Econometric Theory*, 18, 1367-1384.

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

Granger, C.W.J., (1980), Testing for Causality: A Personal Viewpoint, *Journal of Economics and Dynamic Control*, 2, 329-352

Granger, C.W.J., (1993), On the Limitations on Comparing Mean Squared Errors: A Comment, *Journal of Forecasting,* 12, 651-652

Granger, C.W.J. and P. Newbold, (1986), *Forecasting Economic Time Series,* Academic Press, San Diego.

Granger, C.W.J. and M.H. Pesaran, (1993), Economic and Statistical Measures of Forecast Accuracy, *Journal of Forecasting,* 19, 537-560.

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

Hall, A.R. and A. Inoue, (2003), The Large Sample Behavior of the Generalized Method of Moments Estimator in Misspecified Models, *Journal of Econometrics,* 361-394.

Hamilton, J.D., (1994), *Time Series Analysis*, Princeton University Press, Princeton.

Hansen, B.E., (1996), Inference when a Nuisance Parameter is Not Identified Under the Null Hypothesis, *Econometrica, 64, 413-430.*

Hansen, P.R., (2004a), A Test for Superior Predictive Ability, Working Paper, Brown University.

Hansen, P.R., (2004b), Asymptotic Tests of Composite Hypotheses, Working Paper, Brown University.

Hong, Y., (2001), Evaluation of Out of Sample Probability Density Forecasts with Applications to S&P 500 Stock Prices, Working Paper, Cornell University.

Hong, Y.M., H. Li (2003), Out of Sample Performance of Spot Interest Rate Models, *Review of Financial Studies,* forthcoming.

Horowitz, J., (2001), The Bootstrap, in: *Handbook of Econometrics, Volume 5,* ed. JJ. Heckman and E. Leamer, Elsevier, Amsterdam.

Inoue, (2001), Testing for Distributional Change in Time Series, *Econometric Theory*, 17, 156-187.

Inoue, A. and L., Kilian (2004), In-Sample or Out-of-Sample Tests of Predictability: Which One Should We Use? *Econometric Reviews,* forthcoming.

Inoue, A. and M. Shintani, (2004), Bootstrapping GMM Estimators for Time Series, *Journal of Econometrics*, forthcoming.

Khmaladze, E., (1981), Martingale Approach in the Theory of Goodness of Fit Tests, *Theory of Probability and Its Applications*, 20, 240-257.

Khmaladze, E., (1988), An Innovation Approach to Goodness of Fit Tests in R^m, *Annals of Statistics*, 100, 789-829.

Kilian, L., (1999a), Exchange Rate and Monetary Fundamentals: What Do We Learn from Long-Horizon Regressions? *Journal of Applied Econometrics*, 14, 491-510.

Kilian, L., (1999b), Finite Sample Properties of Percentile and Percentile t-Bootstrap Confidence Intervals for Impulse Responses, *Review of Economics and Statistics*, 81, 652-660.

Kilian, L. and M.P., Taylor, (2003), Why is it so Difficult to Beat the Random Walk Forecast of Exchange Rates? *Journal of International Economics, 60, 85-107.*

Kitamura, Y., (2002), Econometric Comparisons of Conditional Models, Working Paper, University of Pennsylvania.

Kolmogorov A.N., (1933), Sulla Determinazione Empirica di una Legge di Distribuzione, *Giornale dell'Istituto degli Attuari,* 4, 83-91.

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

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

Lee, T.H., H. White, and C.W.J. Granger, (1993), Testing for Neglected Nonlinearity in Time Series Models: A Comparison of Neural Network Methods and Alternative Tests, *Journal of Econometrics,* 56, 269-290.

Li, F. and G. Tkacz, (2004), A Consistent Test for Conditional Density Functions with Time Dependent Data, *Journal of Econometrics,* forthcoming.

Linton, O., E. Maasoumi and Y.J., Whang, (2004), Testing for Stochastic Dominance: A subsampling Approach, Working Paper, London School of Economics.

Marcellino, M., J. Stock and M. Watson, (2004), A Comparison of Direct and Iterated AR Methods for Forecasting Macroeconomic Series h-Steps Ahead, Working Paper, Princeton University.

Mark, N.C., (1995), Exchange Rates and Fundamentals: Evidence on Long-Run Predictability, *American Economic Review*, 85, 201-218.

McCracken, M.W., (2000), Robust Out-of-Sample Inference, *Journal of Econometrics*, 99, 195-223.

McCracken, M.W., (2004), Asymptotics for Out of Sample Tests of Granger Causality, Working Paper, University of Missouri-Columbia.

McCracken, M.W., (2003), Parameter Estimation Error and Tests of Equal Forecast Accuracy Between Non-nested Models, *International Journal of Forecasting*, forthcoming.

McCracken. M.W., and S. Sapp, (2004), Evaluating the Predictive Ability of Exchange Rates Using Long Horizon Regressions: Mind your p's and q's. *Journal of Money, Credit and Banking*, forthcoming.

Meese, R.A., and K. Rogoff, (1983), Empirical Exchange Rate Models of the Seventies: Do they Fit Out-of-Sample? *Journal of International Economics, 14, 3-24.*

Pesaran M. H., and A. Timmerman, (2003), How Costly is to Ignore Breaks when Forecasting the Direction of a Time Series? *International Journal of Forecasting,* forthcoming.

Pesaran M. H., and A. Timmerman, (2004), Selection of Estimation Window for Strictly Exogenous Regressors, Working Paper, Cambridge University and University of California, San Diego.

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

Politis, D.N. and J.P. Romano, (1994b), Limit Theorems for Weakly Dependent Hilbert Space Valued Random Variables with Application to the Stationary Bootstrap, *Statistica Sinica*, 4, 461-476.

Politis, D.N., J.P. Romano and M. Wolf, (1999), *Subsampling,* Springer and Verlag, New York.

Rosenblatt, M., (1952), Remarks on a Multivariate Transformation, *Annals of Mathematical Statistics*, 23, 470-472.

Rossi, B., (2003), Testing Long-Horizon Predictive Ability with High Persistence and the Meese-Rogoff Puzzle, *International Economic Review*, forthcoming.

Schörfheide, F., (2000), Loss Function Based Evaluation of DSGE Models, *Journal of Applied Econometrics,* 15, 645-670.

Smirnov N., (1939), On the Estimation of the Discrepancy Between Empirical Curves of Distribution for Two Independent Samples, *Bulletin Mathematique de l'Universite' de Moscou,* 2, fasc. 2.

Storey, J.D., (2003), The positive False Discovery Rate: A Bayesian Interpretation and the q-value, *Annals of Statistics*, forthcoming.

Stinchcombe, M.B. and H., White, (1998), Consistent Specification Testing with Nuisance Parameters Present Only Under the Alternative, *Econometric Theory*, 14, 295-325.

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

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

Swanson, N.R., and H. White, (1997), A Model Selection Approach to Real-Time Macroeconomic Forecasting Using Linear Models and Artificial Neural Networks, *Review of Economic Statistics*, 59, 540-550.

Thompson, S.B., (2002), Evaluating the Goodness of Fit of Conditional Distributions, with an Application to Affine Term Structure Models, Working Paper, Harvard University.

van der Vaart, A.W., (1998), *Asymptotic Statistics*, Cambridge, New York.

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

Weiss, A., (1996), Estimating Time Series Models Using the Relevant Cost Function, *Journal of Applied Econometrics*, 11, 539-560.

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

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

Whang, Y.J., (2000), Consistent Bootstrap Tests of Parametric Regression Functions, *Journal of Econometrics,* 27-46.

Whang, Y.J., (2001), Consistent Specification Testing for Conditional Moment Restrictions, *Economics Letters,* 71, 299-306.

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

White, H., (1994), *Estimation, Inference and Specification Analysis*, Cambridge University Press, Cambridge.

White, H., (2000), A Reality Check for Data Snooping, *Econometrica*, 68, 1097-1126.

Wooldridge, J.M., (2002), *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge.

Zheng, J.X., (2000), A Consistent Test of Conditional Parametric Distribution, *Econometric Theory,* 16, 667-691.