A Test for Comparing Multiple Misspecified Conditional Interval Models*

Valentina Corradi
Queen Mary, University of London

Norman R. Swanson
Rutgers University

March 2003
this version: January 2005

Abstract

This paper introduces a test for the comparison of multiple misspecifed conditional interval models, for the case of dependent observations. Model accuracy is measured using a distributional analog of mean square error, in which the approximation error associated with a given model, say model i, for a given interval, is measured by the expected squared difference between the conditional confidence interval under model i and the "true" one.

When comparing more than two models, a "benchmark" model is specified, and the test is constructed along the lines of the "reality check" of White (2000). Valid asymptotic critical values are obtained via a version of the block bootstrap which properly captures the effect of parameter estimation error. The results of a small Monte Carlo experiment indicate that the test does not have unreasonable finite sample properties, given small samples of 60 and 120 observations, although the results do suggest that larger samples should likely be used in empirical applications of the test.

JEL classification: C22, C52.
Keywords: block bootstrap, conditional confidence intervals, data snooping, misspecified distributional models.

*Valentina Corradi, Department of Economics, Queen Mary-University of London, Mile End, London E1 4NS, U.K., v.corradi@qmul.ac.uk; and Norman R. Swanson, Department of Economics, Rutgers University, New Brunswick, NJ 08901-1248, U.S.A., nswanson@econ.rutgers.edu. The authors would like to express their gratitude to Don Andrews and an anonymous referee for providing numerous useful suggestions, all of which we feel have been instrumental in improving earlier drafts of this paper. The authors would also like to thank Russell Davidson, Clive Granger, Lutz Kilian, Christelle Viaroux and seminar participants at the 2002 UK Econometrics Group meeting in Bristol, the 2002 European Econometric Society meetings, the 2002 University of Pennsylvania NSF-NBER time series conference, the 2002 EC^2 Conference in Bologna, Cornell University, the State University of New York at Stony Brook and the University of California at Davis for many helpful comments and suggestions on previous versions of this paper.

1 Introduction

There are several instances in which merely having a "good" model for the conditional mean and/or variance may not be adequate for the task at hand. For example, financial risk management involves tracking the entire distribution of a portfolio, or measuring certain distributional aspects, such as value at risk (see e.g. Duffie and Pan (1997)). In such cases, models of conditional mean and/or variance may not be satisfactory for the task at hand.

A very small subset of important contributions that go beyond the examination of models of conditional mean and/or variance include papers which: assess the correctness of conditional interval predictions (see e.g. Christoffersen (1998)); assess volatility predictability by comparing unconditional and conditional interval forecasts (see e.g. Christoffersen and Diebold (2000)); and assess conditional quantiles (see e.g. Giacomini and Komunjer (2003)).[1] 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. One contribution in this direction is the conditional Kolmogorov (CK) test of Andrews (1997), which is based on the comparison of the empirical joint distribution of y_t and X_t with the product of a given distribution of $y_t|X_t$ and the empirical CDF of X_t. Other contributions in this direction include, for example, Zheng (2000), who suggests a nonparametric test based on a first-order, linear, expansion of the Kullback Leibler Information Criterion (KLIC), Altissimo and Mele (2002) and Li and Tkacz (2004), who propose a test based on the comparison of a nonparametric kernel estimate of the conditional density with the density implied under the null hypothesis.[2] Following a different route based on use of the probability integral transform, Diebold, Gunther and Tay (1998) suggest a simple and effective means by which predictive densities can be evaluated (see also Bai (2003), Diebold, Hahn and Tay (1999), Hong (2001) and Hong and Li (2005)).

All of the papers cited in the preceding paragraph consider a null hypothesis of correct dynamic specification of the conditional distribution or of a given conditional confidence interval.[3] However,

[1] 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.

[2] Whang (2000,2001) proposes a CK type test for the correct specification of the conditional mean.

[3] One exception is the approach taken by Corradi and Swanson (2005a), who consider testing the null of correct specification of the conditional distribution for a given information set, thus allowing for dynamic misspecification

a reasonable assumption in the context of model selection may instead be that all models are approximations of the truth, and hence all models are likely misspecified. Along these lines, it is our objective in this paper to provide a test that allows for the joint comparison of multiple misspecified conditional interval models, for the case of dependent observations.

Assume that the object of interest is a conditional interval model for a scalar random variable, Y_t, given a (possibly vector valued) conditioning set, Z^t, where Z^t contains lags of Y_t and/or other variables. In particular, given a group of (possibly) misspecified conditional interval models, say $\left(F_1(\overline{u}|Z^t,\theta_1^\dagger)-F_1(\underline{u}|Z^t,\theta_1^\dagger),...,F_m(\overline{u}|Z^t,\theta_m^\dagger)-F_m(\underline{u}|Z^t,\theta_m^\dagger)\right)$, assume that the objective is to compare these models in terms of their closeness to the true conditional interval model, $F_0(\overline{u}|Z^t,\theta_0)-F_0(\overline{u}|Z^t,\theta_0)=\Pr(\underline{u}\leq Y_t\leq\overline{u}|Z^t)$. If $m>2$, we follow White (2000). Namely, we choose a particular model as the "benchmark" and test the null hypothesis that no competing model can provide a more accurate approximation of the "true" model, against the alternative that at least one competitor outperforms the benchmark. Needless to say, pairwise comparison of alternative models, in which no benchmark need be specified, follows as a special case. In our 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 $E\left(\left(F_i(\overline{u}|Z^t,\theta_i^\dagger)-F_i(\underline{u}|Z^t,\theta_i^\dagger)\right)-\left(F_0(\overline{u}|Z^t,\theta_0^\dagger)-F_0(\underline{u}|Z^t,\theta_0^\dagger)\right)\right)^2$, where $\underline{u},\overline{u}\in U$, and U is a possibly unbounded set on the real line.

It should be pointed out that one well known measure of distributional accuracy is the Kullback-Leibler Information Criterion (KLIC), in the sense that the "most accurate" model can be shown to be that which minimizes the KLIC (see Section 2 for a more detailed discussion). For the iid case, Vuong (1989) suggests a likelihood ratio test for choosing the conditional density model which is closest to the "true" conditional density, in terms of the KLIC. Additionally, 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.[4] 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

under both hypotheses.

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

Schorfheide (2002)). For example, Fernandez-Villaverde and Rubio-Ramirez (2001) show that the KLIC-best model is also the model with the highest posterior probability. However, as we outline in the next section, problems concerning the comparison of conditional confidence intervals may be difficult to address using the KLIC, but can be handled quite easily using our generalized mean square measure of accuracy.

The rest of the paper is organized as follows. Section 2 states the hypothesis of interest and describes the test statistic which will be examined in the sequel. In Section 3.1, it is shown that the limiting distribution of the statistic (properly recentered) is a functional of a zero mean Gaussian process, with a covariance kernel that reflects both the contribution of parameter estimation error and the effect of (dynamic) misspecification. Section 3.2 discusses the construction of asymptotically valid critical values. This is done via an extension of White's (2000) bootstrap approach to the case of non-vanishing parameter estimation error. The results of a small Monte Carlo experiment are collected in Section 4, and concluding remarks are given in Section 5. Proofs of results stated in the text are given in the Appendix.

Hereafter, P^* denotes the probability law governing the resampled series, conditional on the sample, E^* and Var^* are the mean and variance operators associated with P^*, $o_P^*(1)$ $\Pr -P$ denotes a term converging to zero in P^*-probability, conditional on the sample, and for all samples except a subset with probability measure approaching zero, and $O_P^*(1)$ $\Pr -P$ denotes a term which is bounded in P^*-probability, conditional on the sample, and for all samples except a subset with probability measure approaching zero. Analogously, $O_{a.s.^*}(1)$ and $o_{a.s.^*}(1)$ denote terms that are almost surely bounded and terms that approach zero almost surely, according the the probability law P^*, and conditional on the sample.

2 Set-Up and Test Statistics

Our objective is to select amongst alternative conditional confidence interval models by using 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. Note that although we assume s_1 and s_2 to be finite, we do not require (Y_t, X_t) to be Markovian. In fact, Z^t might not contain the entire (relevant) history, and all models may be dynamically misspecified.

Define the group of conditional interval models from which one is to make a selection as

$\left(F_1(\overline{u}|Z^t,\theta_1^\dagger) - F_1(\underline{u}|Z^t,\theta_1^\dagger), ..., F_m(\overline{u}|Z^t,\theta_m^\dagger) - F_m(\underline{u}|Z^t,\theta_m^\dagger)\right)$, and define the true conditional interval as

$$F_0(\overline{u}|Z^t,\theta_0) - F_0(\underline{u}|Z^t,\theta_0) = \Pr(\underline{u} \leq Y_t \leq \overline{u}|Z^t).$$

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 $\theta_i^\dagger = \theta_0$. As mentioned in the introduction, accuracy is measured in terms of a distributional analog of mean square error. In particular, we say that model 1 is more accurate than model 2, if

$$\begin{aligned} & E\left(\left((F_1(\overline{u}|Z^t,\theta_1^\dagger) - F_1(\underline{u}|Z^t,\theta_1^\dagger)) - \left(F_0(\overline{u}|Z^t,\theta_0) - F_0(\underline{u}|Z^t,\theta_0)\right)\right)^2\right) \\ < \quad & E\left(\left((F_2(\overline{u}|Z^t,\theta_2^\dagger) - F_2(\underline{u}|Z^t,\theta_2^\dagger)) - \left(F_0(\overline{u}|Z^t,\theta_0) - F_0(\underline{u}|Z^t,\theta_0)\right)\right)^2\right) \end{aligned}$$

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

As mentioned above, a very well known measure of distributional accuracy which is already available in the literature is the KLIC (see e.g. White (1982), Vuong (1989), Giacomini (2002), and Kitamura (2002)), according to which we should choose Model 1 over Model 2 if

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

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. Interestingly, Fernandez-Villaverde and Rubio-Ramirez (2004) have shown that the best model under the KLIC is also the model with the highest posterior probability. However, if we are interested in measuring accuracy for a given conditional confidence interval, this cannot be easily done 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, this cannot be done in a straightforward manner using the KLIC. On the other hand, our approach gives an easy way of addressing question of this type. In this sense, we believe that our approach provides a reasonable alternative to the KLIC.

In the sequel, model 1 is taken as the benchmark model, and the objective is to test whether some competitor model can provide a more accurate approximation of $F_0(\underline{u}|\cdot,\theta_0) - F_0(\overline{u}|\cdot,\theta_0)$ than

the benchmark. The null and the alternative hypotheses are:

$$H_0 : \max_{k=2,\ldots,m} E\left(\Big((F_1(\overline{u}|Z^t,\theta_1^\dagger) - F_1(\underline{u}|Z^t,\theta_1^\dagger)) - \big(F_0(\overline{u}|Z^t,\theta_0) - F_0(\underline{u}|Z^t,\theta_0)\big)\Big)^2\right.$$
$$\left. - \Big((F_k(\overline{u}|Z^t,\theta_k^\dagger) - F_k(\underline{u}|Z^t,\theta_k^\dagger)) - \big(F_0(\overline{u}|Z^t,\theta_0) - F_0(\underline{u}|Z^t,\theta_0)\big)\Big)^2\right) \leq 0.$$

versus

$$H_A : \max_{k=2,\ldots,m} E\left(\Big((F_1(\overline{u}|Z^t,\theta_1^\dagger) - F_1(\underline{u}|Z^t,\theta_1^\dagger)) - \big(F_0(\overline{u}|Z^t,\theta_0) - F_0(\underline{u}|Z^t,\theta_0)\big)\Big)^2\right.$$
$$\left. - \Big((F_k(\overline{u}|Z^t,\theta_k^\dagger) - F_k(\underline{u}|Z^t,\theta_k^\dagger)) - \big(F_0(\overline{u}|Z^t,\theta_0) - F_0(\underline{u}|Z^t,\theta_0)\big)\Big)^2\right) > 0.$$

Alternatively, if interest focuses on testing the null of equal accuracy of two conditional confidence interval models, say model 1 and 2, we can simply state the hypotheses as:

$$\begin{aligned} H_0' \quad &: \quad E\left(\Big((F_1(\overline{u}|Z^t,\theta_1^\dagger) - F_1(\underline{u}|Z^t,\theta_1^\dagger)) - \big(F_0(\overline{u}|Z^t,\theta_0) - F_0(\underline{u}|Z^t,\theta_0)\big)\Big)^2\right) \\ &= \quad E\left(\Big((F_2(\overline{u}|Z^t,\theta_2^\dagger) - F_2(\underline{u}|Z^t,\theta_2^\dagger)) - \big(F_0(\overline{u}|Z^t,\theta_0) - F_0(\underline{u}|Z^t,\theta_0)\big)\Big)^2\right) \end{aligned}$$

versus

$$\begin{aligned} H_A' \quad &: \quad E\left(\Big((F_1(\overline{u}|Z^t,\theta_1^\dagger) - F_1(\underline{u}|Z^t,\theta_1^\dagger)) - \big(F_0(\overline{u}|Z^t,\theta_0) - F_0(\underline{u}|Z^t,\theta_0)\big)\Big)^2\right) \\ &\neq \quad E\left(\Big((F_2(\overline{u}|Z^t,\theta_2^\dagger) - F_2(\underline{u}|Z^t,\theta_2^\dagger)) - \big(F_0(\overline{u}|Z^t,\theta_0) - F_0(\underline{u}|Z^t,\theta_0)\big)\Big)^2\right) \end{aligned}$$

Needless to say, if the benchmark model is correctly specified, we do not reject the null. Related tests that instead focus on dynamic correct specification of a conditional interval models (as opposed to allowing for misspecification under both hypotheses, as is done with all of our tests) are discussed in Christoffersen (1998).

If the objective is to test for the correct specification of a single conditional interval model, say model 1, for a given information set, then we can define the hypotheses as:

$$H_0'' : \Pr\left(\underline{u} \leq Y_t \leq \overline{u}|Z^t\right) = F_1(\overline{u}|Z^t,\theta_1^\dagger) - F_1(\underline{u}|Z^t,\theta_1^\dagger) \text{ a.s. for some } \theta_1^\dagger \in \Theta$$

versus[5]

$$H_A'' : \text{the negation of } H_0''.$$

[5]In the definition of H_0'', $\theta_1^\dagger$ should be replaced by θ_0, if Z^t is meant as the information set including all the relevant history.

Tests of this sort that consider the correct specification of the conditional distribution for given information set (i.e. conditional distribution tests that allow for the possibility of dynamic misspecification under both hypotheses) are discussed in Corradi and Swanson (2005a).

In order to test H_0 versus H_A, form the following statistic:

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

where

$$\begin{aligned} Z_T(1,k) &= \frac{1}{\sqrt{T}} \sum_{t=s}^{T} \left(\left(1\{\underline{u} \leq Y_t \leq \overline{u}\} - \left(F_1(\overline{u}|Z^t, \widehat{\theta}_{1,T}) - F_1(\underline{u}|Z^t, \widehat{\theta}_{1,T}) \right) \right)^2 \right. \\ &\quad \left. - \left(1\{\underline{u} \leq Y_t \leq \overline{u}\} - \left(F_k(\overline{u}|Z^t, \widehat{\theta}_{k,T}) - F_k(\underline{u}|Z^t, \widehat{\theta}_{k,T}) \right) \right)^2 \right) \end{aligned} \tag{2}$$

with $s = \max\{s_1, s_2\}$,

$$\widehat{\theta}_{i,T} = \arg\max_{\theta_i \in \Theta_i} \frac{1}{T} \sum_{t=s}^{T} \ln f_i(Y_t|Z^t, \theta_i), \ i = 1, ..., m, \tag{3}$$

and

$$\theta_i^{\dagger} = \arg\max_{\theta_i \in \Theta_i} E(\ln f_i(Y_t|Z^t, \theta_i)), \ i = 1, ..., m,$$

where $f_i(Y_t|Z^t, \theta_i)$ is the conditional density under model i. As $f_i(\cdot|\cdot)$ does not in general coincide with the true conditional density, $\widehat{\theta}_{i,T}$ is the QMLE, and $\theta_i^{\dagger} \neq \theta_0$,in general. More broadly speaking, the results discussed below hold for any estimator for which $\sqrt{T}\left(\widehat{\theta}_{i,T} - \theta_i^{\dagger}\right)$ is asymptotically normal. This is the case for several extremum estimators, for example, such as (nonlinear) least squares, (Q)MLE, etc. However, it is not advisable to use overidentified GMM estimators because $\sqrt{T}\left(\widehat{\theta}_{i,T} - \theta_i^{\dagger}\right)$ is not asymptotically normal, in general, when model i is not correctly specified (see e.g. Hall and Inoue (2003)). Needless to say, if interest focuses on testing H_0' versus H_A', one should use the statistic $Z_T(1,2)$, and if interest focuses on testing H_0'' versus H_A'', the appropriate test statistic is:

$$\sup_{v \in V} Z_T(v) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left(1\{\underline{u} \leq Y_t \leq \overline{u}\} - \left(F_1(\overline{u}|Z^t, \widehat{\theta}_{1,T}) - F_1(\underline{u}|Z^t, \widehat{\theta}_{1,T}) \right) \right) 1\{Z^t \leq v\}, \tag{4}$$

which is a special case of the statistic considered in Theorem 2 of Corradi and Swanson (2005a), in the context of testing for the correct specification of the "entire" conditional distribution, for a given information set. The limiting distribution of (4), and the construction of valid critical values

via the bootstrap follow from Theorem 2 and Theorem 4 in Corradi and Swanson (2005a), who also provide some Monte Carlo evidence. Discussion of the test statistic in (4) in relation with the existing literature on testing for the correct conditional distribution is given in the paper just mentioned.

The intuition behind equation (2) is very simple. First, note that $E(1\{\underline{u} \leq Y_t \leq \overline{u}\}|Z^t) = \Pr(\underline{u} \leq Y_t \leq \overline{u} \leq u|Z^t) = F_0(\overline{u}|Z^t,\theta_0) - F_0(\underline{u}|Z^t,\theta_0)$. Thus, $1\{\underline{u} \leq Y_t \leq \overline{u}\} - \left(F_i(\overline{u}|Z^t,\theta_i^\dagger) - F_i(\underline{u}|Z^t,\theta_i^\dagger)\right)$ can be interpreted as an "error" term associated with computation of the conditional expectation, under F_i. Now, write the statistic in equation (2) as:

$$\frac{1}{\sqrt{T}}\sum_{t=s}^{T}\left(\left(\left(1\{\underline{u} \leq Y_t \leq \overline{u}\} - \left(F_1(\overline{u}|Z^t,\widehat{\theta}_{1,T}) - F_1(\underline{u}|Z^t,\widehat{\theta}_{1,T})\right)\right)^2 - \mu_1^2\right)\right.$$
$$\left. - \left(\left(1\{\underline{u} \leq Y_t \leq \overline{u}\} - \left(F_k(\overline{u}|Z^t,\widehat{\theta}_{k,T}) - F_k(\underline{u}|Z^t,\widehat{\theta}_{k,T})\right)\right)^2 - \mu_k^2\right)\right) + \frac{T-s}{\sqrt{T}}(\mu_1^2 - \mu_k^2), \quad (5)$$

where $\mu_j^2 = E\left(\left(1\{\underline{u} \leq Y_t \leq \overline{u}\} - \left(F_j(\overline{u}|Z^t,\theta_j^\dagger) - F_j(\underline{u}|Z^t,\theta_j^\dagger)\right)\right)^2\right)$, $j = 1,...,m$. In the appendix, it is shown that the first term in equation (5) weakly converges to a gaussian process. Also, for $j = 1,...,m$:

$$\begin{aligned}
\mu_j^2 &= E\left(\left(1\{\underline{u} \leq Y_t \leq \overline{u}\} - \left(F_j(\overline{u}|Z^t,\theta_j^\dagger) - F_j(\underline{u}|Z^t,\theta_j^\dagger)\right)\right)^2\right) \\
&= E\left(\left(1\{\underline{u} \leq Y_t \leq \overline{u}\} - \left(F_0(\overline{u}|Z^t,\theta_0) - F_0(\underline{u}|Z^t,\theta_0)\right)\right)\right. \\
&\quad \left. - \left(\left(F_j(\overline{u}|Z^t,\theta_j^\dagger) - F_j(\underline{u}|Z^t,\theta_j^\dagger)\right) - \left(F_0(\overline{u}|Z^t,\theta_0) - F_0(\underline{u}|Z^t,\theta_0)\right)\right)\right)^2 \\
&= E\left(\left(1\{\underline{u} \leq Y_t \leq \overline{u}\} - \left(F_0(\overline{u}|Z^t,\theta_0) - F_0(\underline{u}|Z^t,\theta_0)\right)\right)^2\right) \\
&\quad + E\left(\left(\left(F_j(\overline{u}|Z^t,\theta_j^\dagger) - F_j(\underline{u}|Z^t,\theta_j^\dagger)\right) - \left(F_0(\overline{u}|Z^t,\theta_0) - F_0(\underline{u}|Z^t,\theta_0)\right)\right)^2\right),
\end{aligned}$$

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

$$\begin{aligned}
\mu_1^2 - \mu_k^2 &= E\left(\left(\left(F_1(\overline{u}|Z^t,\theta_1^\dagger) - F_1(\underline{u}|Z^t,\theta_1^\dagger)\right) - \left(F_0(\overline{u}|Z^t,\theta_0) - F_0(\underline{u}|Z^t,\theta_0)\right)\right)^2\right) \\
&\quad - E\left(\left(\left(F_k(\overline{u}|Z^t,\theta_k^\dagger) - F_k(\underline{u}|Z^t,\theta_k^\dagger)\right) - \left(F_0(\underline{u}|Z^t,\theta_0) - F_0(\underline{u}|Z^t,\theta_0)\right)\right)^2\right). \quad (6)
\end{aligned}$$

Before outlining the asymptotic properties of the statistic in equation (1) two comments are worth making.

First, following the reality check approach of White (2000), the problem of testing multiple hypotheses has been reduced to a single test by applying the (single valued) max function to

multiple hypotheses. This approach has the advantage that it avoids sequential testing bias and also captures the correlation across the various models. On the other hand, if we reject the null, we can conclude that there is at least one model that outperforms the benchmark, but we do not have available to us a complete picture concerning which model(s) contribute to the rejection of the null. Of course, some information can be obtained by looking at the distributional analog of mean square error associated with the various models, and forming a crude ranking of the models, although the usual cautions associated with using a MSE type measure to rank models should be taken. Alternatively, our approach can be complemented by a multiple comparison approach, such as the false discovery rate (FDR) approach of Benjamini and Hochberg (1995), which 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 the 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. Overall, we think that a sound practical strategy could be to first implement our 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 we simply choose the benchmark model. Nevertheless, even in this case, it may not hurt to see whether some of the individual hypotheses in the joint null are rejected via a multiple test comparison approach.

Second, it perhaps worth pointing out that simulation based versions of the tests discussed here are given in Corradi and Swanson (2005b), in the context of the evaluation of dynamic stochastic general equilibrium models.

3 Asymptotic Results

The results stated below require the following assumption.

Assumption A: (i) (Y_t, X_t) is a strictly stationary and absolutely regular $\beta-$mixing process with size -4, for $i = 1, ..., m$; (ii) $F_i(u|Z^t, \theta_i)$ is continuously differentiable on the interior of Θ_i,

where Θ_i is a compact set in $\Re^{p_i}$, and $\nabla_{\theta_i} F_i(u|Z^t, \theta_i^\dagger)$ is $2r$-dominated on Θ_i, for all u, $r > 2$;[6] (iii) $\theta_i^\dagger$ is uniquely identified (i.e. $E(\ln f_i(Y_t|Z^t, \theta_i^\dagger)) > E(\ln f_i(Y_t|Z^t, \theta_i))$, for any $\theta_i \neq \theta_i^\dagger$), where f_i is the density associated with F_i; (iv) f_i is twice continuously differentiable on the interior of Θ_i, and $\nabla_{\theta_i} \ln f_i(Y_t|Z^t, \theta_i)$ and $\nabla^2_{\theta_i} \ln f_i(Y_t|Z^t, \theta_i)$ are $2r-$dominated on Θ_i, with $r > 2$; (v) $E\left(-\nabla^2_{\theta_i} \ln f_i(Y_t|Z^t, \theta_i)\right)$ is positive definite, uniformly on Θ_i, and $\lim_{T\to\infty} Var\left(\frac{1}{\sqrt{T}} \sum_{t=s}^{T} \nabla_{\theta_i} \ln f_i(Y_t|Z^t, \theta_i^\dagger)\right)$ is positive definite; and (vi) let

$$
\begin{aligned}
v_{kk} \;=\; & \lim_{T\to\infty} Var\left(\frac{1}{\sqrt{T}} \sum_{t=s}^{T} \left(\left(1\{\underline{u} \leq Y_t \leq \overline{u}\} - \left(F_1(\overline{u}|Z^t, \theta_1^\dagger) - F_1(\underline{u}|Z^t, \theta_1^\dagger)\right)\right)^2\right.\right. \\
& \left.\left. - \left(1\{\underline{u} \leq Y_t \leq \overline{u}\} - \left(F_k(\overline{u}|Z^t, \theta_k^\dagger) - F_k(\underline{u}|Z^t, \theta_k^\dagger)\right)\right)^2\right)\right),
\end{aligned}
$$

for $k = 2, ..., m$. Define analogous covariance terms, v_{jk}, $j, k = 2, ..., m$, and assume that $COV = [v_{jk}]$ is positive semi-positive definite.

Recalling that $Z^t = (Y_{t-1}, ..., Y_{t-s_1}, X_t, ..., X_{t-s_2+1})$, A1(i) ensures that Z^t is strictly stationary mixing with size -4. Note that A(vi) requires at least one of the competing models to be neither nested in nor nesting the benchmark model. The nonnestedness of at least one competitor ensures that the long-run covariance matrix is positive definite even in the absence of parameter estimation error. However assumption A(vi) can be relaxed, in which case the limiting distribution of the test statistic takes exactly the same form as given in Theorem 1 below, except that the covariance kernel contains only terms which reflect parameter estimation error.[7]

3.1 Limiting Distributions

Theorem 1: Let Assumption A hold. Then:

$$
\max_{k=2,...,m} \left(Z_T(1,k) - \sqrt{T}\left(\mu_1^2 - \mu_k^2\right)\right) \xrightarrow{d} \max_{k=2,...,m} Z_{1,k},
$$

[6] We say that $\nabla_{\theta_i} F(u|Z^t, \theta_i)$ is $2r-$dominated on Θ_i uniformly in u, if its $k^{th}-$element, $k = 1, ...p_i$, is such that $\left|\nabla_{\theta_i} F_i(u|Z^t, \theta_i)\right|_k \leq D_t(u)$, and $\sup_{u\in R} E(|D_t(u)|^{2r}) < \infty$. For more details on domination conditions, see Gallant and White (1988, pp. 33).

[7] Note that in White (2000), the nonnestedness of at least one competitor is a necessary condition, given that in his context parameter estimation error vanishes asymptotically, while in the present context it does not. More precisely, White (2000) considers out of sample comparison, using the first R observations for model estimation and the last P observations for model validation, where $T = P + R$. Parameter estimation error vanishes in his setup either because $P/R \to 0$ or because the same loss function is used for estimation and model validation.

where $Z_{1,k}$ is a zero mean Gaussian process with covariance $c_{kk} = v_{kk} + p_{kk} + pc_{kk}$, v_{kk} denotes the component of the long-run covariance matrix that would obtain in the absence of parameter estimation error, p_{kk} denotes the contribution of parameter estimation error, and pc_{kk} denotes the covariance across the two components. In particular:[8]

$$\begin{aligned} v_{kk} &= E\sum_{j=-\infty}^{\infty}\left(\left(\left(1\{\underline{u}\leq Y_s\leq\overline{u}\}-\left(F_1(\overline{u}|Z^s,\theta_1^\dagger)-F_1(\underline{u}|Z^s,\theta_1^\dagger)\right)\right)^2-\mu_1^2\right)\right. \\ &\quad \left.\left(\left(1\{\underline{u}\leq Y_{s+j}\leq\overline{u}\}-\left(F_1(\overline{u}|Z^{s+j},\theta_1^\dagger)-F_1(\underline{u}|Z^{s+j},\theta_1^\dagger)\right)\right)^2-\mu_1^2\right)\right) \end{aligned} \tag{7}$$

$$\begin{aligned} &+E\sum_{j=-\infty}^{\infty}\left(\left(\left(1\{\underline{u}\leq Y_s\leq\overline{u}\}-\left(F_k(\overline{u}|Z^s,\theta_k^\dagger)-F_k(\underline{u}|Z^s,\theta_k^\dagger)\right)\right)^2-\mu_k^2\right)\right. \\ &\left.\left(\left(1\{\underline{u}\leq Y_{s+j}\leq\overline{u}\}-\left(F_k(\overline{u}|Z^{s+j},\theta_k^\dagger)-F_k(\underline{u}|Z^{s+j},\theta_k^\dagger)\right)\right)^2-\mu_k^2\right)\right) \end{aligned} \tag{8}$$

$$\begin{aligned} &-2E\sum_{j=-\infty}^{\infty}\left(\left(\left(1\{\underline{u}\leq Y_s\leq\overline{u}\}-\left(F_1(\overline{u}|Z^s,\theta_1^\dagger)-F_1(\underline{u}|Z^s,\theta_1^\dagger)\right)\right)^2-\mu_1^2\right)\right. \\ &\left.\left(\left(1\{\underline{u}\leq Y_{s+j}\leq\overline{u}\}-\left(F_k(\overline{u}|Z^{s+j},\theta_k^\dagger)-F_k(\underline{u}|Z^{s+j},\theta_k^\dagger)\right)\right)^2-\mu_k^2\right)\right) \end{aligned} \tag{9}$$

$$p_{kk} = 4m_{\theta_1^\dagger}{}'A(\theta_1^\dagger)E\left(\sum_{j=-\infty}^{\infty}\nabla_{\theta_1}\ln f_1(Y_s|Z^s,\theta_1^\dagger)\nabla_{\theta_1}\ln f_1(Y_{s+j}|Z^{s+j},\theta_1^\dagger)'\right)A(\theta_1^\dagger)m_{\theta_1^\dagger} \tag{10}$$

$$+4m'_{\theta_k^\dagger}A(\theta_k^\dagger)E\left(\sum_{j=-\infty}^{\infty}\nabla_{\theta_k}\ln f_k(Y_s|Z^s,\theta_k^\dagger)\nabla_{\theta_k}\ln f_k(Y_{s+j}|Z^{s+j},\theta_k^\dagger)'\right)A(\theta_k^\dagger)m_{\theta_k^\dagger} \tag{11}$$

$$-8m'_{\theta_1^\dagger}A(\theta_1^\dagger)E\left(\sum_{j=-\infty}^{\infty}\nabla_{\theta_1}\ln f_1(Y_s|Z^s,\theta_1^\dagger)\nabla_{\theta_k}\ln f_k(Y_{s+j}|Z^{s+j},\theta_k^\dagger)'\right)A(\theta_k^\dagger)m_{\theta_k^\dagger} \tag{12}$$

[8]Note that the recentered statistic is actually

$$\max_{k=2,\ldots,m}\left(Z_T(1,k)-\frac{T-s}{\sqrt{T}}\left(\mu_1^2-\mu_k^2\right)\right).$$

However, for notational simplicity, and given that the two are asymptotically equivalent, we "approximate" $\frac{T-s}{\sqrt{T}}$ with $\sqrt{T}$, both in the text and in the Appendix.

$$
\begin{aligned}
pc_{kk} \quad = \quad & -4m'_{\theta_1^\dagger} A(\theta_1^\dagger) E\left(\sum_{j=-\infty}^{\infty} \nabla_{\theta_1} \ln f_1(Y_s|Z^s,\theta_1^\dagger)\right. \\
& \left.\left(\left(1\{\underline{u} \le Y_{s+j} \le \overline{u}\} - \left(F_1(\overline{u}|Z^{s+j},\theta_1^\dagger) - F_1(\underline{u}|Z^{s+j},\theta_1^\dagger)\right)\right)^2 - \mu_1^2\right)\right) \\
& +8m_{\theta_1^\dagger}{}' A(\theta_1^\dagger) E\left(\sum_{j=-\infty}^{\infty} \nabla_{\theta_1} \ln f_1(Y_s|Z^s,\theta_1^\dagger)\right. \\
& \left.\left(\left(1\{\underline{u} \le Y_{s+j} \le \overline{u}\} - \left(F_k(\overline{u}|Z^{s+j},\theta_k^\dagger) - F_k(\underline{u}|Z^{s+j},\theta_k^\dagger)\right)\right)^2 - \mu_k^2\right)\right) \qquad (13)
\end{aligned}
$$

$$
-4m'_{\theta_k^\dagger} A(\theta_k^\dagger) E\left(\sum_{j=-\infty}^{\infty} \nabla_{\theta_k} \ln f_k(Y_s|Z^s,\theta_k^\dagger)\left(\left(1\{\underline{u} \le Y_{s+j} \le \overline{u}\} - \left(F_k(\overline{u}|Z^{s+j},\theta_k^\dagger) - F_k(\underline{u}|Z^{s+j},\theta_k^\dagger)\right)\right)^2 - \mu_k^2\right)\right) \qquad (14)
$$

with[9] $m_{\theta_i^\dagger}{}' = E\left(\nabla_{\theta_i}\left(F_i(\overline{u}|Z^t,\theta_i^\dagger) - F_i(\overline{u}|Z^t,\theta_i^\dagger)\right)\left(1\{\underline{u} \le Y_t \le \overline{u}\} - \left(F_i(\overline{u}|Z^t,\theta_i^\dagger) - F_i(\overline{u}|Z^t,\theta_i^\dagger)\right)\right)\right)$
and $A(\theta_i^\dagger) = \left(E\left(-\ln \nabla^2_{\theta_i} f_i(y_t|Z^t,\theta_i^\dagger)\right)\right)^{-1}$.

As an immediate corollary, note the following.

Corollary 2: Let Assumptions A(i)-A(v) hold, and suppose A(vi) is violated. Then:

$$
\max_{k=2,\ldots,m}\left(Z_T(1,k) - \sqrt{T}\left(\mu_1^2(u) - \mu_k^2(u)\right)\right) \xrightarrow{d} \max_{k=2,\ldots,m} \widetilde{Z}_{1,k},
$$

where $\widetilde{Z}_{1,k}$ is a zero mean normal random variable with covariance equal to p_{kk}, as defined in equations (10)-(12) above.

From Theorem 1 and Corollary 2, it follows that when all competing models provide an approximation to the true conditional interval model that is as (mean square) accurate as that provided by the benchmark (i.e. when $\mu_1^2 - \mu_k^2 = 0, \forall k$), then the limiting distribution corresponds to the maximum of an $m-1$ dimensional zero-mean normal random vector, with a covariance kernel that reflects both the contribution of parameter estimation error and the dependent structure of the data. Additionally, when all competitor models are worse than the benchmark, the statistic diverges to minus infinity, at rate $\sqrt{T}$. Finally, when only some competitor models are worse than the benchmark, the limiting distribution provides a conservative test, as Z_T will always be smaller than $\max_{k=2,\ldots,m}\left(Z_T(1,k) - \sqrt{T}\left(\mu_1^2 - \mu_k^2\right)\right)$, asymptotically, and therefore the critical values of $\max_{k=2,\ldots,m}\left(Z_T(1,k) - \sqrt{T}\left(\mu_1^2 - \mu_k^2\right)\right)$ provide upper bounds for the critical values of

[9] Note that $m_{\theta_i^\dagger}$ depends on chose interval $(\underline{u},\overline{u})$. Hovever, for notational simplicity we omit such dependence.

$\max_{k=2,...,m} Z_T(1,k)$. Of course, when H_A holds, the statistic diverges to plus infinity at rate $\sqrt{T}$. It is well known that the maximum of a normal random vector is not a normal random variable, and hence critical values cannot immediately be tabulated. In a related paper, White (2000) suggests obtaining critical values either via Monte Carlo simulation or via use of the bootstrap. Here, we focus on use of the bootstrap, although White's results do not apply in our case, as contribution of parameter estimation error does not vanish in our setup, and hence must be properly taken into account when forming critical values. Before turning our attention to the bootstrap, however, we briefly outline an out-of-sample version of our test statistic.

Thus far, we have compared conditional interval models via a distributional generalization of in-sample mean square error. Needless to say, an out-of-sample version of the statistic may also be constructed. Let $T = R + P$, let $\widehat{\theta}_{i,t}$ $i = 1, ..., m$ be a recursive estimator computed using $t = R, R+1, ..., R+P-1$ observations, and let $\widetilde{Z}^t = (Y_t, ..Y_{t-s_1}, X_t, ..., X_{t-s_2})$. A 1-step ahead out-of-sample version of the statistic in equations (1) and (2) is given by:

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

where

$$\begin{aligned} OZ_P(1,k) &= \frac{1}{\sqrt{P}} \sum_{t=R+s}^{T-1} \left(\left(1\{\underline{u} \leq Y_{t+1} \leq \overline{u}\} - \left(F_1(\overline{u}|Z^t, \widehat{\theta}_{1,t}) - F_1(\underline{u}|Z^t, \widehat{\theta}_{1,t}) \right) \right)^2 \right. \\ &\quad \left. - \left(1\{\underline{u} \leq Y_{t+1} \leq \overline{u}\} - \left(F_k(\overline{u}|Z^t, \widehat{\theta}_{k,t}) - F_k(\underline{u}|Z^t, \widehat{\theta}_{k,t}) \right) \right)^2 \right). \end{aligned}$$

Now, Theorem 1 and Corollary 2 still apply (Corollary 2 requires $P/R \to \pi > 0$), although the covariance matrices will be slightly different. However, Theorem 3 (see below) no longer applies, as the block bootstrap is no longer valid, and is indeed characterized by a bias term whose sign varies across samples. This is because of the use of recursive estimation. This issue is studied in Corradi and Swanson (2004), who propose a proper recentering of the quasi likelihood function.[10]

3.2 Bootstrap Critical Values

In this subsection we outline how to obtain valid critical values for the asymptotic distribution of $\max_{k=2,...,m} \left(Z_T(1,k) - \sqrt{T} \left(\mu_1^2 - \mu_k^2 \right) \right)$, via use of a version of the block bootstrap that properly

[10] Corradi and Swanson (2004) study the case of rolling estimators.

captures the contribution of parameter estimation error to the covariance kernel associated with the limiting distribution of the test statistic.[11]

In order to show the first order validity of the bootstrap, we shall obtain the limiting distribution of the bootstrap statistic and show that it coincides with the limiting distribution given in Theorem 1. As all candidate models are potentially misspecified under both hypotheses, the parametric bootstrap is not generally applicable in our context. In fact, if observations are resampled from one of the candidate models, then we cannot ensure that the resampled statistic has the appropriate limiting distribution. Our approach is thus to establish the first order validity of the block bootstrap in the presence of parameter estimation error, by drawing in part upon results of Goncalves and White (2002, 2004).[12]

Assume that bootstrap samples are formed as follows. Let $W_t = (Y_t, Z^t)$. Draw b overlapping blocks of length l from $W_s, ..., W_T$, where $s = \max\{s_1, s_2\}$, so that $bl = T - s$. Thus, $W_s^*, ..., W_{s+l}^*, ..., W_{T-l+1}^*, ..., W_T^*$ is equal to $W_{I_1+1}, ..., W_{I_1+l}, ..., W_{I_b+1}, ..., W_{I_b+l}$, where $I_i, i = 1, ..., b$ are identically and independently distributed discrete uniform random variates on $s-1, s, ..., T-l$. It follows that, conditional on the sample, the pseudo time series W_t^*, $t = s, ..., T$, consists of b independent and identically distributed blocks of length l.

Now, consider the bootstrap analog of Z_T. Define the block bootstrap QMLE as,

$$\widehat{\theta}_{i,T}^* = \arg \max_{\theta_i \in \Theta_i} \frac{1}{T} \sum_{t=s}^{T} \ln f_i(Y_t^* | Z^{*t}, \theta_i), \ i = 1, ...m,$$

[11] In principle, we could have obtained an estimator for $C = [c_{kj}]$, as defined in the statement of Theorem 1, which takes into account the contribution of parameter estimation error, call it $\widehat{C}$. Then, we could draw N $m-1$-dimensional standard normal random vectors, say $\eta^{(i)}$, $i = 1, ..., N$, and for each i : form $\widehat{C}^{1/2}\eta^{(i)}$, take the maximum of the $m-1$ elements, and finally compute the empirical distribution of the N maxima. However, as pointed out by White (2000), when the sample size is moderate and the number of models is large, $\widehat{C}$ is a rather poor estimator for C.

[12] Goncalves and White (2002,2004) consider the more general case of heterogeneous and near epoch dependent observations.

and define the bootstrap statistic as[13]:

$$Z_T^* = \max_{k=2,\ldots,m} Z_T^*(1,k),$$

where

$$\begin{aligned} Z_{T,u}^*(1,k) &= \frac{1}{\sqrt{T}} \sum_{t=s}^{T} \Bigg(\Bigg(\Big(1\{\underline{u} \leq Y_t^* \leq \overline{u}\} - \Big(F_1(\overline{u}|Z^{*t}, \widehat{\theta}_{1,T}^*) - F_1(\underline{u}|Z^{*t}, \widehat{\theta}_{1,T}^*) \Big) \Big)^2 \\ &\quad - \Big(1\{\underline{u} \leq Y_t \leq \overline{u}\} - \Big(F_1(\overline{u}|Z^t, \widehat{\theta}_{1,T}) - F_1(\underline{u}|Z^t, \widehat{\theta}_{1,T}) \Big) \Big)^2 \Bigg) \\ &\quad - \Bigg(\Big(1\{\underline{u} \leq Y_t^* \leq \overline{u}\} - \Big(F_k(\overline{u}|Z^{*t}, \widehat{\theta}_{k,T}^*) - F_k(\underline{u}|Z^{*t}, \widehat{\theta}_{k,T}^*) \Big) \Big)^2 \\ &\quad - \Big(1\{\underline{u} \leq Y_t \leq \overline{u}\} - \Big(F_k(\overline{u}|Z^t, \widehat{\theta}_{k,T}) - F_k(\underline{u}|Z^t, \widehat{\theta}_{k,T}) \Big) \Big)^2 \Bigg) \Bigg). \end{aligned}$$

Theorem 3: Let Assumption A hold. If $l \to \infty$ and $l/T^{1/2} \to 0$, as $T \to \infty$, then,

$$P\Bigg(\omega : \sup_{v \in \Re} \Bigg| P^* \Bigg(\max_{k=2,\ldots,m} Z_T^*(1,k) \leq v \Bigg)$$

$$- P\Bigg(\max_{k=2,\ldots,m} \Big(Z_T(1,k) - \sqrt{T}\left(\mu_1^2 - \mu_k^2\right) \Big) \leq v \Bigg) \Bigg| > \varepsilon \Bigg) \to 0,$$

where P^* denotes the probability law of the resampled series, conditional on the sample, and $\mu_1^2 - \mu_k^2$ is defined as in equation (6).

The above result suggests proceeding in the following manner. For any bootstrap replication, compute the bootstrap statistic, Z_T^*. Perform B bootstrap replications (B large) and compute the quantiles of the empirical distribution of the B bootstrap statistics. Reject H_0 if Z_T is greater than the $(1-\alpha)th$-quantile. Otherwise, do not reject. Now, for all samples except a set with probability measure approaching zero, Z_T has the same limiting distribution as the corresponding bootstrap statistic, when $\mu_1^2 - \mu_k^2 = 0, \forall k$, which is the least favorable case under the null hypothesis. Thus, the above approach ensures that the test has asymptotic size α. On the other hand, when one or more, but not all of the competing models are strictly dominated by the benchmark, the above approach

[13]It should be pointed out that $\ln f_i(Y_t|Z^t, \theta_i)$ and $\ln f_i(Y_t^*|Z^{*t}, \theta_i)$ can be replaced by generic functions $m_i(Y_t, Z^t, \theta_i)$ and $m_i(Y_t^*, Z^{*t}, \theta_i)$, provided they satisfy assumptions A and $A2.1$ in Goncalves and White (2004), and provided $E^*\left(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} m_i(Y_t^*, Z^{*t}, \theta_i)\right) = o(1)$ Pr $-P$. Thus, the results for QMLE straightforwardly extend to generic $m-$estimators, such as NLS or exactly identified GMM. On the other hand, they do not apply to overidentified GMM, as $E^*\left(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} m_i(Y_t^*, Z^{*t}, \theta_i)\right) = O(1)$ Pr $-P$. In that case, even for first order validity, one has to properly recenter $m_i(Y_t^*, Z^{*t}, \theta_i)$ (see e.g. Hall and Horowitz (1996), Andrews (2002) or Inoue and Shintani (2004)).

ensures that the test has asymptotic size between 0 and α. When all models are dominated by the benchmark, the statistic vanishes to minus infinity, so that the rule above implies zero asymptotic size. Finally, under the alternative, Z_T diverges to (plus) infinity, while the corresponding bootstrap statistic has a well defined limiting distribution. This ensures 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 $\max_{k=2,\ldots,m}\left(\mu_1^2-\mu_k^2\right)=0$, but $\left(\mu_1^2-\mu_k^2\right)<0$ for some k, then the bootstrap critical values lead to conservative inference. An alternative to our bootstrap critical values in this case is to construct critical values using subsampling (see e.g. Politis, Romano and Wolf (1999), Ch.3). Heuristically, construct $T-2b_T$ statistics using subsamples of length b_T, where $b_T/T\rightarrow 0$. 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 $\max_{k=2,\ldots,m}\left(\mu_1^2-\mu_k^2\right)=0$, but $\left(\mu_1^2-\mu_k^2\right)<0$, 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.

Hansen (2005) points out that the conservative nature of the reality check of White (2000), leads to reduced power, and that it should be feasible to improve the power and reduce the sensitivity of the reality check test to poor and irrelevant alternatives via use of the modified reality check test outlined in his paper. Given the similarity between the approach taken in our paper, and that taken by White (2000), it may also be possible to improve our test performance using the approach of Hansen (2005) to modify our test.

4 Monte Carlo Findings

The experimental setup used in this section is as follows. We begin by generating $(y_t, y_{t-1}, w_t, x_t, q_t)'$ as,

$$\begin{pmatrix} y_t \\ y_{t-1} \\ x_t \\ w_t \\ q_t \end{pmatrix} \sim St\left(0,\Sigma,v\right)$$

where $St(0, \Sigma, v)$ denotes a Student's t distribution with mean zero, variance Σ, and v degrees of freedom; with

$$\Sigma = \begin{pmatrix} \sigma_y^2 & \sigma_{12} & 0 & 0 & 0 \\ \sigma_{12} & \sigma_y^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma_X^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma_W^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma_Q^2 \end{pmatrix}.$$

The DGP of interest is assumed to be (see e.g. Spanos (1999))

$$y_t | y_{t-1} \sim St\left(\alpha y_{t-1}, \left(\frac{v}{v-1}\left(1 + \frac{y_{t-1}^2}{\sigma_y^2}\right)(\sigma_y^2 - \sigma_y^2 \alpha)\right); v\right). \tag{15}$$

where $\alpha = \frac{\sigma_{12}}{\sigma^2}$, so that the conditional mean is a linear function of y_{t-1} and the conditional variance is a linear function of y_{t-1}^2.

In our experiments, we impose misspecification upon all estimated models by assuming normality (i.e. we assume that F_i, $i = 1, ..., m$, is the normal CDF). Our objective is to ascertain whether a given benchmark model is "better", in the sense of having lower squared approximation error, than two given alternative models. Thus, $m = 3$. Level and power experiments are defined by adjusting the conditioning information sets used to estimate (via QMLE) the parameters of each conditional model, and subsequently to form $F_i(u|Z^t, \widehat{\theta}_{i,T})$, $F_i(u|Z^{*t}, \widehat{\theta}^*_{i,T})$, Z_T, and Z^*_T. In all experiments, values of $\alpha = \{0.4, 0.6, 0.8, 0.9\}$ are used, samples of $T = 60$ and 120 are tried, $v = 5$, $\sigma^2 = 1$, and $\sigma_X^2 = \sigma_W^2 = \sigma_Q^2 = \{0.1, 1.0, 10.0\}$. Throughout, the conditional confidence interval version of the test is constructed, and the upper and lower bounds of the interval are fixed at $\mu_Y + \gamma\sigma_Y$ and $\mu_Y - \gamma\sigma_Y$, respectively, where μ_Y and σ_Y are the mean and variance of y_t, and where $\gamma = 1/2$.[14] Additionally, 5% and 10% nominal level bootstrap critical values are constructed using 100 bootstrap replications, block lengths of $l = \{2, 3, 5, 6\}$ are tried, and all reported rejection frequencies are based on 5000 Monte Carlo simulations.[15] Given $Z^t = (y_{t-1}, x_t, w_t, q_t)$, the experiments reported on are organized as follows:

Empirical Level Experiments: In these experiments, we define the conditioning variable sets as follows: For the benchmark model (F_1), use $\widetilde{Z}^t = (y_{t-1}, x_t)$, where $\widetilde{Z}^t$ is a proper subset of Z^t. For the two alternative models (F_2 and F_3) we set $\widetilde{Z}^t = (y_{t-1}, w_t)$ and $\widetilde{Z}^t = (y_{t-1}, q_t)$, respectively.

[14] Findings corresponding to $\gamma = \{\frac{1}{16}, \frac{1}{8}\}$ are very similar and are available from the authors upon request.

[15] Additional results for cases where $x = \{\frac{1}{4}, 1\}$, $l = \{10, 12\}$, and where critical values are constructed using 250 bootstrap replications are available upon request, and yield qualitatively similar results to those reported in Tables 1-2.

In this case, the estimated coefficients associated with $x_t, w_t,$ and q_t have probability limits equal to zero, as none of these variables enters into the true conditional mean function. In addition, all models are misspecified, as conditional normality is assumed throughout. Therefore, the benchmark and the two competitors are equally misspecified. Finally, the limiting distribution of the test statistic in this case is driven by parameter estimation error, as assumption A(vi) does not hold (see Corollary 2 for this case).

Empirical Power Experiments: In these experiments, we set the conditioning variable sets as follows: For the benchmark model (F_1), $\widetilde{Z}^t = (w_t)$. For the two alternative models (F_2 and F_3) we set $\widetilde{Z}^t = (y_{t-1})$ and $\widetilde{Z}^t = (q_t)$, respectively. In this manner, it is ensured that the first of the two alternative models has smaller squared approximation error than the benchmark model. In fact, all three models are incorrect for both the marginal distribution (normal instead of Student-t) and for the conditional variance, which is set equal to the unconditional value, instead of being a linear function of y_{t-1}^2. However, one of the competitors, model 2, is correctly specified for the conditional mean, while the other two are not. Therefore, model 2 is characterized by a smaller squared approximation error.

Our findings are summarized in Table 1 (empirical level experiments) and Table 2 (empirical power experiments). In these tables, the first column reports the value of α used in a particular experiment, while the remaining entries are rejection frequencies of the null hypothesis that the benchmark model is not outperformed by any of the alternative models. A number of conclusions emerge upon inspection of the tables. Turning first to the empirical level results given in Table 1, note, for example, that empirical level varies from values grossly above nominal levels (when block lengths and values of α are large), to values below or close to nominal levels (when values of α are smaller). However, note that it is often the case that moving from 60 to 120 observations results in rejection frequencies being closer to the nominal level of the test, as expected (with the exception that the test becomes even more conservative when l is 5 or 6, in many cases). Notice also that when $\alpha = 0.4$ (low persistence) a block length of 2 usually suffices to capture the dependence structure of the series, while for $\alpha = 0.9$ (high persistence) a larger block length is necessary. Finally, it is worth noting that, overall, the empirical rejection frequencies are not too distant from nominal levels, a result which is somewhat surprising given the small sample sizes used in our experiments. However, the test could clearly be expected to exhibit improved behavior were larger samples of data used.

With regard to empirical power (see Table 2), note that rejection frequencies increase as α increases. This is not surprising, as the contribution of y_{t-1} to the conditional mean, which is neglected by models 1 and 3, becomes more substantial as α increases. Overall, for $\alpha \geq 0.6$ and for a nominal level of 10%, rejection frequencies are above 0.5 in many cases, again suggesting the need for larger samples.[16]

As noted above, rejection frequencies are sensitive to the choice of the blocksize parameter. This suggests that it should be useful to choose the block length in a data-driven manner. One way in which this may be accomplished is by use of a 2-step procedure as follows. First, one defines the optimal rate at which the block length should grow, as the sample grows. This rate usually depends on what one is interested in (for example, the focus is confidence intervals in our setup – see chapter 6 in Lahiri (2003) for further details). Second, one computes the optimal blocksize for a smaller sample via subsampling techniques, as proposed by Hall, Horowitz and Jing (HHJ: 1995), and then obtains the optimal block length for the full sample, using the optimal rate in the first step.[17] However, it is not clear whether application of the HHJ approach leads to an optimal choice (i.e. to the blocksize which minimizes the appropriate mean squared error, say). The reason for this is that the theoretical optimal blocksize is obtained by comparing the first (or second) term of the Edgeworth expansion of the actual and bootstrap statistics. However, in our case the statistic is not pivotal, as Z_T and Z_T^* are not scaled by a proper variance estimator, and consequently we cannot obtain an Edgeworth expansion with a standard normal variate as the leading term in the expansion. In principle, we could begin by scaling the test statistic by an autocorrelation and heteroskedasticity robust (HAC) variance estimator, but in such a case the statistic could no longer be written as a smooth function of the sample mean, and it is not clear whether data-driven blocksize selection of the variety outlined above would actually be optimal.[18] Although these issues remain unresolved, and are the focus of ongoing research, we nevertheless suggest using a data driven approach, such as the HHJ approach, with the caveat that the method should at this stage only be thought of as providing a rough guide for blocksize selection.

[16]Note that our Monte Carlo findings are not directly comparable with those of Christoffersen (1998), as his null corresponds to correct dynamic specification of the conditional interval model.

[17]Further data driven methods for computing the blocksize are reported in Lahiri (Ch.6, 2003).

[18]For higher order properties for statistics studentized with HAC estimators (see e.g. Götze and Künsch (1996) for the sample mean, and Inoue and Shintani (2004) for linear IV estimators).

5 Concluding Remarks

We have provided a test that allows for the joint comparison of multiple misspecified conditional interval models, for the case of dependent observations, and for the case where accuracy is measured using a distributional analog of mean square error. We also outlined the construction of valid asymptotic critical values based on a version of the block bootstrap, which properly takes into account the contribution of parameter estimation error. A small number of Monte Carlo experiments were also run in order to assess the finite sample properties of the test, and results indicate that the test does not have unreasonable finite sample properties given very small samples of 60 and 120 observations, although the results do suggest that larger samples should likely be used in empirical application of the test.

6 Appendix

Proof of Theorem 1: Recall that

$$\begin{aligned}\mu_i^2 &= E\left(\left(1\{\underline{u} \le Y_t \le \overline{u}\} - \left(F_i(\overline{u}|Z^t,\theta_i^\dagger) - F_i(\underline{u}|Z^t,\theta_i^\dagger)\right)\right)^2\right) \\ &= E\left(\left(1\{\underline{u} \le Y_t \le \overline{u}\} - \left(F_0(\overline{u}|Z^t,\theta_0) - F_0(\underline{u}|Z^t,\theta_0)\right)\right)^2\right) \\ &\quad + E\left(\left(\left(F_0(\overline{u}|Z^t,\theta_0) - F_0(\underline{u}|Z^t,\theta_0)\right) - \left(F_i(\overline{u}|Z^t,\theta_i^\dagger) - F_i(\underline{u}|Z^t,\theta_i^\dagger)\right)\right)^2\right).\end{aligned}$$

Thus, from (5),

$$\begin{aligned}Z_T(1,k) &= \frac{1}{\sqrt{T}}\sum_{t=s}^{T}\left(\left(\left(1\{\underline{u} \le Y_t \le \overline{u}\} - \left(F_1(\overline{u}|Z^t,\widehat{\theta}_{1,T}) - F_1(\underline{u}|Z^t,\widehat{\theta}_{1,T})\right)\right)^2 - \mu_1^2\right)\right. \\ &\quad \left. - \left(\left(1\{\underline{u} \le Y_t \le \overline{u}\} - \left(F_k(\overline{u}|Z^t,\widehat{\theta}_{k,T}) - F_k(\underline{u}|Z^t,\widehat{\theta}_{k,T})\right)\right)^2 - \mu_k^2\right)\right) + \frac{T-s}{\sqrt{T}}(\mu_1^2 - \mu_k^2) \\ &= \frac{1}{\sqrt{T}}\sum_{t=s}^{T}\left(\left(\left(1\{\underline{u} \le Y_t \le \overline{u}\} - \left(F_1(\overline{u}|Z^t,\theta_1^\dagger) - F_1(\underline{u}|Z^t,\theta_1^\dagger)\right)\right)^2 - \mu_1^2\right)\right. \\ &\quad \left. - \left(\left(1\{\underline{u} \le Y_t \le \overline{u}\} - \left(F_k(\overline{u}|Z^t,\theta_k^\dagger) - F_k(\underline{u}|Z^t,\theta_k^\dagger)\right)\right)^2 - \mu_k^2\right)\right) \\ &\quad - \frac{2}{T}\sum_{t=s}^{T}\nabla_{\theta_1}\left(F_1(\overline{u}|Z^t,\overline{\theta}_{1,T}) - F_1(\underline{u}|Z^t,\overline{\theta}_{1,T})\right)' \\ &\quad \times\left(1\{\underline{u} \le Y_t \le \overline{u}\} - \left(F_1(\overline{u}|Z^t,\theta_1^\dagger) - F_1(\underline{u}|Z^t,\theta_1^\dagger)\right)\right)\sqrt{T}\left(\widehat{\theta}_{1,T} - \theta_1^\dagger\right) \\ &\quad + \frac{2}{T}\sum_{t=s}^{T}\nabla_{\theta_k}\left(F_k(\overline{u}|Z^t,\overline{\theta}_{k,T}) - F_k(\underline{u}|Z^t,\overline{\theta}_{k,T})\right)' \\ &\quad \times\left(1\{\underline{u} \le Y_t \le \overline{u}\} - \left(F_k(\overline{u}|Z^t,\theta_k^\dagger) - F_k(\underline{u}|Z^t,\theta_k^\dagger)\right)\right)\sqrt{T}\left(\widehat{\theta}_{k,T} - \theta_k^\dagger\right) \\ &\quad + \frac{T-s}{\sqrt{T}}(\mu_1^2 - \mu_k^2) + o_P(1),\end{aligned}$$

where $\overline{\theta}_{i,T} \in (\widehat{\theta}_{i,T}, \theta_i^\dagger)$. Note that, given Assumption A(i) and A(iii), for $i = 1, ..., m$,

$$\sqrt{T}\left(\widehat{\theta}_{i,T} - \theta_i^\dagger\right) = A(\theta_i^\dagger)\frac{1}{\sqrt{T}}\sum_{t=s}^{T}\nabla_{\theta_i}\ln f_i(Y_t|Z^t,\theta_i^\dagger) + o_P(1),$$

where $A(\theta_i^\dagger) = \left(E\left(-\nabla^2_{\theta_i} f_i(y_t|Z^t,\theta_i^\dagger)\right)\right)^{-1}$. Thus, $Z_T(1,k)$ converges in distribution to a normal random variable with variance equal to c_{kk}. The statement in Theorem 1 then follows as a straightforward application of the Cramer Wold device and the continuous mapping theorem.

Proof of Corollary 2: Immediate from the proof of Theorem 1.

Proof of Theorem 3: In the sequel, P^*, E^*, and Var^* denote the probability law of the resampled series, conditional on the sample, the expectation, and the variance operators associated with P^*, respectively. With the notation $o_{P^*}(1)$ $\Pr - P$, and $O_{P^*}(1)$ $\Pr - P$, we mean a term approaching zero in P^*-probability and a term bounded in P^*-probability, conditional on the sample and for all samples except a set with probability measure approaching zero, respectively. Write $Z^*_{T,u}(1,k)$ as

$$\begin{aligned}
Z^*_{T,u}(1,k) &= \frac{1}{\sqrt{T}}\sum_{t=s}^{T}\Big(\Big(\Big(1\{\underline{u}\le Y_t^*\le\overline{u}\}-\Big(F_1(\overline{u}|Z^{*t},\theta_1^\dagger)-F_1(\underline{u}|Z^{*t},\theta_1^\dagger)\Big)\Big)\\
&\quad -\nabla_{\theta_1}\Big(F_1(\overline{u}|Z^{*t},\overline{\theta}^*_{1,T})-F_1(\underline{u}|Z^{*t},\overline{\theta}^*_{1,T})\Big)\Big(\widehat{\theta}^*_{1,T}-\theta_1^\dagger\Big)\Big)^2\\
&\quad -((1\{\underline{u}\le Y_t\le\overline{u}\}-\Big(F_1(\overline{u}|Z^{t},\theta_1^\dagger)-F_1(\underline{u}|Z^{t},\theta_1^\dagger)\Big))\\
&\quad -\nabla_{\theta_1}\left(F_1(\overline{u}|Z^{t},\overline{\theta}_{1,T})-F_1(\underline{u}|Z^{t},\overline{\theta}_{1,T})\right)\Big(\widehat{\theta}_{1,T}-\theta_1^\dagger\Big)\Big)^2\Big)\\
&\quad -\Big(\Big(1\{\underline{u}\le Y_t^*\le\overline{u}\}-\Big(F_k(\overline{u}|Z^{*t},\theta_k^\dagger)-F_k(\underline{u}|Z^{*t},\theta_k^\dagger)\Big)\Big)\\
&\quad -\nabla_{\theta_k}\Big(F_k(\overline{u}|Z^{*t},\overline{\theta}^*_{k,T})-F_k(\underline{u}|Z^{*t},\overline{\theta}^*_{k,T})\Big)\Big(\widehat{\theta}^*_{k,T}-\theta_k^\dagger\Big)\Big)^2\\
&\quad -((1\{\underline{u}\le Y_t\le\overline{u}\}-\Big(F_k(\overline{u}|Z^{t},\theta_k^\dagger)-F_k(\underline{u}|Z^{t},\theta_k^\dagger)\Big)\Big)\\
&\quad -\nabla_{\theta_k}\left(F_k(\overline{u}|Z^{t},\overline{\theta}_{k,T})-F_k(\underline{u}|Z^{t},\overline{\theta}_{k,T})\right)\Big(\widehat{\theta}_{k,T}-\theta_k^\dagger\Big)\Big)^2\Big)\Big)
\end{aligned}$$

where $\overline{\theta}^*_{i,T}\in\left(\widehat{\theta}^*_{i,T},\theta_i^\dagger\right)$, $\overline{\theta}_{i,T}\in\left(\widehat{\theta}_{i,T},\theta_i^\dagger\right)$. Now,

$$\begin{aligned}
&Vec\left(\frac{1}{\sqrt{T}}\sum_{t=s}^{T}\nabla_{\theta_i}\Big(F_i(\overline{u}|Z^{*t},\overline{\theta}^*_{i,T})-F_i(\underline{u}|Z^{*t},\overline{\theta}^*_{i,T})\Big)'\Big(\widehat{\theta}^*_{i,T}-\theta_i^\dagger\Big)\right.\\
&\left.\Big(\widehat{\theta}^*_{i,T}-\theta_i^\dagger\Big)'\nabla_{\theta_i}\Big(F_i(\overline{u}|Z^{*t},\overline{\theta}^*_{i,T})-F_i(\underline{u}|Z^{*t},\overline{\theta}^*_{i,T})\Big)\right)\\
=\;&\left[\frac{1}{T}\sum_{t=s}^{T}\nabla_{\theta_i}\Big(F_i(\overline{u}|Z^{*t},\overline{\theta}^*_{i,T})-F_i(\underline{u}|Z^{*t},\overline{\theta}^*_{i,T})\Big)'\otimes\nabla_{\theta_i}\Big(F_i(\overline{u}|Z^{*t},\overline{\theta}^*_{i,T})-F_i(\underline{u}|Z^{*t},\overline{\theta}^*_{i,T})\Big)\right]\\
&\times\sqrt{T}vec\Big(\widehat{\theta}^*_{i,T}-\theta_i^\dagger\Big)\Big(\widehat{\theta}^*_{i,T}-\theta_i^\dagger\Big)'\\
=\;&o_{P^*}(1),\Pr-P, \qquad (16)
\end{aligned}$$

as $\sqrt{T}\left(\widehat{\theta}^*_{i,T} - \theta^\dagger_i\right) = \sqrt{T}\left(\widehat{\theta}^*_{i,T} - \widehat{\theta}_{i,T}\right) + \sqrt{T}\left(\widehat{\theta}_{i,T} - \theta^\dagger_i\right) = O_{P^*}(1) + O(1) = O_{P^*}(1)$ $\Pr - P$, by Theorem 2.2 in Goncalves and White (GW: 2004), and $\sqrt{T}\left(\widehat{\theta}^*_{i,T} - \widehat{\theta}_{i,T}\right) = O_{P^*}(1)$ $\Pr - P$, as it converges in P^*-distribution, and because the term in square brackets is $O_{P^*}(1)$, $\Pr - P$. Thus, $Z^*_T(1,k)$ can be written as,

$$
\begin{aligned}
&\frac{1}{\sqrt{T}}\sum_{t=s}^{T}\left(\left(1\{\underline{u} \leq Y^*_t \leq \overline{u}\} - \left(F_1(\overline{u}|Z^{*t},\theta^\dagger_1) - F_1(\underline{u}|Z^{*t},\theta^\dagger_1)\right)\right)^2\right. \\
&\left. - \left(1\{\underline{u} \leq Y_t \leq \overline{u}\} - \left(F_1(\overline{u}|Z^{t},\theta^\dagger_1) - F_1(\underline{u}|Z^{t},\theta^\dagger_1)\right)\right)^2\right) \\
&-\frac{2}{T}\sum_{t=s}^{T}\left(\left(1\{\underline{u} \leq Y^*_t \leq \overline{u}\} - \left(F_1(\overline{u}|Z^{*t},\theta^\dagger_1) - F_1(\underline{u}|Z^{*t},\theta^\dagger_1)\right)\right)\right. \\
&\left.\times\nabla_{\theta_1}\left(F_1(\overline{u}|Z^{*t},\overline{\theta}^*_{1,T}) - F_1(\underline{u}|Z^{*t},\overline{\theta}^*_{1,T})\right)'\sqrt{T}\left(\widehat{\theta}^*_{1,T} - \theta^\dagger_1\right)\right) \\
&+\frac{2}{T}\sum_{t=s}^{T}\left(\left(1\{\underline{u} \leq Y_t \leq \overline{u}\} - \left(F_1(\overline{u}|Z^{t},\theta^\dagger_1) - F_1(\underline{u}|Z^{t},\theta^\dagger_1)\right)\right)\right. \\
&\left.\times\nabla_{\theta_1}\left(F_1(\overline{u}|Z^{t},\overline{\theta}_{1,T}) - F_1(\underline{u}|Z^{t},\overline{\theta}_{1,T})\right)'\sqrt{T}\left(\widehat{\theta}_{1,T} - \theta^\dagger_1\right)\right)
\end{aligned} \tag{17}
$$

$$
\begin{aligned}
&\frac{1}{\sqrt{T}}\sum_{t=s}^{T}\left(\left(1\{\underline{u} \leq Y^*_t \leq \overline{u}\} - \left(F_k(\overline{u}|Z^{*t},\theta^\dagger_k) - F_k(\underline{u}|Z^{*t},\theta^\dagger_k)\right)\right)^2\right. \\
&\left. - \left(1\{\underline{u} \leq Y_t \leq \overline{u}\} - \left(F_k(\overline{u}|Z^{t},\theta^\dagger_k) - F_k(\underline{u}|Z^{t},\theta^\dagger_k)\right)\right)^2\right) \\
&-\frac{2}{T}\sum_{t=s}^{T}\left(\left(1\{\underline{u} \leq Y^*_t \leq \overline{u}\} - \left(F_k(\overline{u}|Z^{*t},\theta^\dagger_k) - F_k(\underline{u}|Z^{*t},\theta^\dagger_k)\right)\right)\right. \\
&\left.\times\nabla_{\theta_k}\left(F_k(\overline{u}|Z^{*t},\overline{\theta}^*_{k,T}) - F_k(\underline{u}|Z^{*t},\overline{\theta}^*_{k,T})\right)'\sqrt{T}\left(\widehat{\theta}^*_{k,T} - \theta^\dagger_k\right)\right) \\
&+\frac{2}{T}\sum_{t=s}^{T}\left(\left(1\{\underline{u} \leq Y_t \leq \overline{u}\} - \left(F_k(\overline{u}|Z^{t},\theta^\dagger_k) - F_k(\underline{u}|Z^{t},\theta^\dagger_k)\right)\right)\right. \\
&\left.\times\nabla_{\theta_1}\left(F_1(\overline{u}|Z^{t},\overline{\theta}_{k,T}) - F_k(\underline{u}|Z^{t},\overline{\theta}_{k,T})\right)'\sqrt{T}\left(\widehat{\theta}_{k,T} - \theta^\dagger_k\right)\right) + o^*_P(1)\ Pr - P
\end{aligned}
$$

We begin by showing that for $i = 1, ..., m$, conditional on the sample and for all samples except a set of probability measure approaching zero:

(a) The term in first two lines in (17) has the same limiting distribution ($\Pr - P$) as:

$$
\frac{1}{\sqrt{T}}\sum_{t=s}^{T}\left(\left(1\{\underline{u} \leq Y_t \leq \overline{u}\} - \left(F_1(\overline{u}|Z^{t},\theta^\dagger_1) - F_1(\underline{u}|Z^{t},\theta^\dagger_1)\right)\right)^2 - \mu_1^2\right).
$$

(b) The term in the last four lines of (17) has the same limiting distribution ($\Pr - P$) as:

$$-\frac{2}{T}\sum_{t=s}^{T}\nabla_{\theta_1}\left(F_1(\overline{u}|Z^t,\theta_1^\dagger)-F_1(\underline{u}|Z^t,\theta_1^\dagger)\right)'$$
$$\times\left(1\{\underline{u}\leq Y_t\leq\overline{u}\}-\left(F_1(\overline{u}|Z^t,\theta_1^\dagger)-F_1(\underline{u}|Z^t,\theta_1^\dagger)\right)\right)\sqrt{T}\left(\widehat{\theta}_{1,T}-\theta_1^\dagger\right),\ \Pr-P$$

We begin by showing **(a)**. Given the block resampling scheme described in Section 3.2, it is easy to see that,

$$\overline{F}=E^*\left(\frac{1}{\sqrt{T}}\sum_{t=s}^{T}\left(1\{\underline{u}\leq Y_t^*\leq\overline{u}\}-\left(F_1(\overline{u}|Z^{*t},\theta_1^\dagger)-F_1(\underline{u}|Z^{*t},\theta_1^\dagger)\right)\right)^2\right)$$
$$=\frac{1}{\sqrt{T}}\sum_{t=s}^{T}\left(1\{\underline{u}\leq Y_t\leq\overline{u}\}-\left(F_1(\overline{u}|Z^t,\theta_1^\dagger)-F_1(\underline{u}|Z^t,\theta_1^\dagger)\right)\right)^2+O\left(\frac{l}{\sqrt{T}}\right),\ \Pr-P.$$

For notational simplicity, just set $\underline{u}=-\infty$. Needless to say, the same argument applies to any generic $\underline{u}<\overline{u}$. Recalling that each block, conditional on the sample, is identically and independently distributed,

$$Var^*\left(\frac{1}{\sqrt{T}}\sum_{t=s}^{T}\left(\left(1\{Y_t^*\leq\overline{u}\}-F_1(\overline{u}|Z^{*t},\theta_1^\dagger)\right)^2\right)\right)$$
$$=E^*\left(\left(\frac{1}{\sqrt{T}}\sum_{t=s}^{T}\left(\left(1\{Y_t^*\leq\overline{u}\}-F_1(\overline{u}|Z^{*t},\theta_1^\dagger)\right)^2-\overline{F}\right)\right)^2\right)+O\left(\frac{l}{\sqrt{T}}\right)$$
$$=\frac{1}{bl}E^*\left(\left(\sum_{k=1}^{b}\sum_{i=1}^{l}\left(\left(1\{Y_{I_k+i}^*\leq\overline{u}\}-F_1(\overline{u}|Z^{*I_{k+i}},\theta_1^\dagger)\right)^2-\overline{F}\right)\right)^2\right)+O\left(\frac{l}{\sqrt{T}}\right)$$

$$=\frac{1}{l}E^*\left(\left(\sum_{i=1}^{l}\left(\left(1\{Y_{I_1+i}^*\leq\overline{u}\}-F_1(\overline{u}|Z^{*I_{1+i}},\theta_1^\dagger)\right)^2-\overline{F}\right)\right)^2\right)+O\left(\frac{l}{\sqrt{T}}\right)$$
$$=\frac{1}{T}\sum_{t=l}^{T-l}\sum_{i=-l}^{l}\left(\left(1\{Y_t\leq\overline{u}\}-F_1(\overline{u}|Z^t,\theta_1^\dagger)\right)^2-\overline{F}\right)\left(\left(1\{Y_{t+i}\leq\overline{u}\}-F_1(\overline{u}|Z^{t+i},\theta_1^\dagger)\right)^2-\overline{F}\right)$$
$$+O\left(\frac{l}{\sqrt{T}}\right)$$
$$=\lim_{T\to\infty}Var\left(\frac{1}{\sqrt{T}}\sum_{t=s}^{T}\left(\left(1\{Y_t\leq\overline{u}\}-F_1(\overline{u}|Z^t,\theta_1^\dagger)\right)^2\right)\right)+O\left(\frac{l}{\sqrt{T}}\right)\ \Pr-P,\tag{18}$$

where the last equality follows from Theorem 1 in Andrews (1991), given Assumption A, and given the growth rate conditions on l. Therefore, given Assumption A, by Theorem 3.5 in Künsch (1989), **(a)** holds.

We now need to establish (b). First, note that given the mixing and domination conditions in Assumption A, from Lemmas 4 and 5 in GW, it follows that,

$$\begin{aligned} & \frac{2}{T}\sum_{t=s}^{T}\left(\left(1\{Y_t^* \leq \overline{u}\} - F_1(\overline{u}|Z^{*t},\theta_1^\dagger)\right)\nabla_{\theta_1}F_1(\overline{u}|Z^{*t},\overline{\theta}_{1,T}^*)' \right. \\ & \left. - \left(1\{Y_t \leq \overline{u}\} - F_1(\overline{u}|Z^{t},\theta_1^\dagger)\right)\nabla_{\theta_1}F_1(\overline{u}|Z^{t},\overline{\theta}_{1,T})'\right) \\ = \;& o_P^*(1) \;\Pr - P \end{aligned}$$

Thus, we can write the sum of the last two terms in equation (17) as,

$$\begin{aligned} & -\frac{2}{T}\sum_{t=s}^{T}\left(\left(1\{Y_t^* \leq \overline{u}\} - F_1(\overline{u}|Z^{*t},\theta_1^\dagger)\right)\nabla_{\theta_1}F_1(\overline{u}|Z^{*t},\overline{\theta}_{1,T}^*)'\right) \\ & \sqrt{T}\left(\widehat{\theta}_{1,T}^* - \widehat{\theta}_{1,T}\right) + o_{P^*}(1), \;\; \Pr - P. \end{aligned}$$

Also, by Theorem 2.2 in GW, there exists an $\varepsilon > 0$ such that,

$$\Pr\left(\sup_{x\in\Re^{p_1}}\left|P^*\left(\sqrt{T}\left(\widehat{\theta}_{1,T}^* - \widehat{\theta}_{1,T}\right) \leq x\right) - P\left(\sqrt{T}\left(\widehat{\theta}_{1,T} - \theta_1^\dagger\right) \leq x\right)\right| > \varepsilon\right) \to 0.$$

Thus, $\sqrt{T}\left(\widehat{\theta}_{1,T}^* - \widehat{\theta}_{1,T}\right)$ has the same asymptotic normal distribution as $\sqrt{T}\left(\widehat{\theta}_{1,T} - \theta_1^\dagger\right)$, conditional on the sample and for all samples except a set with probability measure approaching zero. Finally, again by the same argument used in Lemmas A4 and A5 in GW,

$$\begin{aligned} & \frac{1}{T}\sum_{t=s}^{T}\left(\nabla_{\theta_1}F_1(\overline{u}|Z^{*t},\overline{\theta}_{1,T}^*)'\left(1\{Y_t^* \leq \overline{u}\} - F_1(\overline{u}|X_t^*,\theta_1^\dagger)\right)\right) \\ = \;& m_{\theta_1^\dagger}{}' + o_{P^*}(1), \;\; \Pr - P, \end{aligned}$$

where $m_{\theta_i^\dagger}{}' = E\left(\nabla_{\theta_i}F_i(\overline{u}|Z^t,\theta_i^\dagger)\left(Y_t \leq \overline{u}\} - F_i(\overline{u}|Z^t,\theta_i^\dagger)\right)\right)$. Needless to say, the corresponding terms for model k can be treated in the same manner. Thus, $Z_T(1,k)^*$ has the same limiting distribution as $Z_T(1,k)$, conditional on the sample and for all samples except a set with probability measure approaching zero.

7 References

Altissimo, F. and A. Mele, (2002), Testing the Closeness of Conditional Densities by Simulated Nonparametric Methods, Working Paper, LSE.

Andrews, D.W.K., (1991), Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation, *Econometrica*, 59, 817-858.

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.

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

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.

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

Chatfield, C., (1993), Calculating Interval Forecasts, *Journal of Business and Economic Statistics,* 11, 121-135.

Christoffersen, P.F., (1998), Evaluating Interval Forecasts, *International Economic Review*, 39, 841-862.

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

Clements, M.P. and N. Taylor, (2001), Bootstrapping Prediction Intervals for Autoregressive Models, *International Journal of Forecasting,* 17, 247-276.

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

Corradi, V. and N.R. Swanson, (2004), Predictive Density Accuracy Tests, Working Paper, Rutgers University.

Corradi, V. and N.R. Swanson, (2005a), Bootstrap Conditional Distribution Tests in the Presence of Dynamic Misspecification, *Journal of Econometrics,* forthcoming.

Corradi, Valentina and Norman R. Swanson, (2005b), Evaluation of Dynamic Stochastic General Equilibrium Models Based on Distributional Comparison of Simulated and Historical Data, *Journal of Econometrics*, forthcoming.

Davidson, J., (1994), *Stochastic Limit Theory*, Oxford University Press, Oxford.

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., 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.

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, 153-180.

Gallant, A.R. and H. White, (1988), *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models,* Blackwell, Oxford.

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 I. Komunjer, (2003), Evaluation and Combination of Conditional Quantile Forecasts, Working Paper, Boston College.

Goncalves, S., and H. White, (2002), The Bootstrap of the Mean for Dependent and 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.*

Götze, F., and H.R. Künsch, (1996), Second-Order Correctness of the Blockwise Bootstrap for Stationary Observations, *Annals of Statistics*, 24, 1914-1933.

Granger, C.W.J., H. White, and M. Kamstra, (1989), Interval Forecasting - An Analysis Based Upon ARCH-Quantile Estimators, *Journal of Econometrics,* 40, 87-96.

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

Hall, P., J.K. Horowitz, and N.J. Jing, (1995), On Blocking Rules for the Bootstrap with Dependent Data, *Biometrika*, 82, 561-574.

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.

Hansen, P.R., (2005), An Unbiased Test for Superior Predictive Ability, Working Paper, Stanford University.

Hochberg, Y., (1988), A Sharper Bonferroni Procedure for Multiple Significance Tests, *Biometrika,*

75, 800-803.

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., and H. Li, (2005), Out of Sample Performance of Spot Interest Rate Models, *Review of Financial Studies,* 18, 37-84.

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

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

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

Lahiri, S.N., (2003), *Resampling Methods for Dependent Data,* Springer and Verlag, New York.

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, (2003), Consistent Testing for Stochastic Dominance Under General Sampling Schemes, forthcoming *Review of Economic Studies.*

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

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

Spanos, A., (1999), *Probability Theory and Statistical Inference: Econometric Modelling with Observational Data,* Cambridge University Press.

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

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.

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

Table 1: Empirical Level Experiments: Interval = $\mu_Y + \frac{1}{2}\sigma_Y$

α	Sample Size = 60 Observations				Sample Size = 120 Observations			
	$l=2$	$l=3$	$l=5$	$l=6$	$l=2$	$l=3$	$l=5$	$l=6$
Panel A: 5% Nominal Level – Exogenous Variate Variance = 0.1								
0.4	0.024	0.013	0.014	0.033	0.034	0.021	0.008	0.014
0.6	0.044	0.035	0.042	0.050	0.051	0.033	0.029	0.025
0.8	0.082	0.084	0.126	0.143	0.071	0.076	0.088	0.088
0.9	0.132	0.131	0.198	0.252	0.096	0.114	0.147	0.157
Panel B: 5% Nominal Level – Exogenous Variate Variance = 1.0								
0.4	0.066	0.022	0.017	0.005	0.106	0.037	0.009	0.008
0.6	0.131	0.059	0.027	0.009	0.166	0.076	0.021	0.017
0.8	0.174	0.081	0.035	0.028	0.177	0.079	0.024	0.015
0.9	0.154	0.061	0.030	0.032	0.145	0.052	0.023	0.023
Panel C: 5% Nominal Level – Exogenous Variate Variance = 10.0								
0.4	0.060	0.023	0.005	0.008	0.136	0.033	0.009	0.007
0.6	0.119	0.066	0.019	0.022	0.182	0.084	0.027	0.016
0.8	0.170	0.085	0.034	0.024	0.183	0.095	0.036	0.018
0.9	0.161	0.073	0.043	0.036	0.153	0.062	0.031	0.020
Panel D: 10% Nominal Level – Exogenous Variate Variance = 0.1								
0.4	0.049	0.029	0.029	0.053	0.073	0.047	0.018	0.024
0.6	0.066	0.059	0.059	0.074	0.083	0.060	0.046	0.046
0.8	0.118	0.105	0.153	0.173	0.094	0.101	0.113	0.104
0.9	0.157	0.159	0.229	0.278	0.110	0.134	0.172	0.179
Panel E: 10% Nominal Level – Exogenous Variate Variance = 1.0								
0.4	0.128	0.067	0.046	0.030	0.207	0.106	0.038	0.034
0.6	0.212	0.121	0.069	0.042	0.242	0.149	0.062	0.064
0.8	0.249	0.136	0.085	0.058	0.258	0.142	0.069	0.058
0.9	0.235	0.119	0.080	0.071	0.211	0.106	0.057	0.060
Panel F: 10% Nominal Level – Exogenous Variate Variance = 10.0								
0.4	0.124	0.070	0.034	0.027	0.218	0.085	0.044	0.042
0.6	0.200	0.146	0.061	0.074	0.245	0.153	0.067	0.057
0.8	0.258	0.145	0.085	0.067	0.259	0.155	0.086	0.051
0.9	0.235	0.131	0.094	0.069	0.213	0.117	0.066	0.042

Notes: Empirical rejection frequencies are reported in the 2nd through 9th columns of entries in the table. The first column reports the value of α, the autoregressive parameter in the DGP. In all experiments, $v = 5$, $\sigma^2 = 1$, and $\sigma_X^2 = \sigma_W^2 = \sigma_Q^2 = \{0.1, 1.0, 10.0\}$ The upper and lower bounds of the interval are fixed at $\mu_Y + \gamma\sigma_Y$ and $\mu_Y - \gamma\sigma_Y$, respectively, where $\gamma = \frac{1}{2}$. The 5% and 10% nominal level bootstrap critical values used in the experiments are constructed using 100 bootstrap replications, block lengths of $l = \{2, 3, 5, 6\}$ are tried, and all reported rejection frequencies are based on 5000 Monte Carlo simulations. See Section 4 for further details.

Table 2: Empirical Power Experiments: Interval = $\mu_Y + \frac{1}{2}\sigma_Y$

α	Sample Size = 60 Observations				Sample Size = 120 Observations			
	$l = 2$	$l = 3$	$l = 5$	$l = 6$	$l = 2$	$l = 3$	$l = 5$	$l = 6$
Panel A: 5% Nominal Level – Exogenous Variate Variance = 0.1								
0.4	0.177	0.075	0.037	0.023	0.250	0.142	0.058	0.046
0.6	0.616	0.525	0.453	0.442	0.626	0.516	0.423	0.401
0.8	0.496	0.365	0.217	0.170	0.591	0.441	0.213	0.154
0.9	0.635	0.473	0.263	0.227	0.651	0.514	0.309	0.224
Panel B: 5% Nominal Level – Exogenous Variate Variance = 1.0								
0.4	0.168	0.081	0.035	0.038	0.317	0.163	0.050	0.035
0.6	0.614	0.498	0.420	0.414	0.671	0.550	0.421	0.415
0.8	0.577	0.395	0.219	0.177	0.614	0.426	0.230	0.148
0.9	0.630	0.479	0.280	0.206	0.663	0.521	0.298	0.194
Panel C: 5% Nominal Level – Exogenous Variate Variance = 10.0								
0.4	0.171	0.083	0.035	0.028	0.349	0.178	0.053	0.048
0.6	0.639	0.480	0.380	0.402	0.662	0.530	0.401	0.374
0.8	0.571	0.398	0.208	0.169	0.608	0.454	0.217	0.162
0.9	0.639	0.487	0.279	0.238	0.652	0.505	0.290	0.232
Panel D: 10% Nominal Level – Exogenous Variate Variance = 0.1								
0.4	0.263	0.169	0.105	0.074	0.345	0.271	0.150	0.132
0.6	0.666	0.605	0.510	0.501	0.673	0.597	0.495	0.468
0.8	0.557	0.461	0.327	0.279	0.635	0.527	0.327	0.264
0.9	0.676	0.541	0.375	0.330	0.687	0.577	0.409	0.338
Panel E: 10% Nominal Level – Exogenous Variate Variance = 1.0								
0.4	0.272	0.187	0.093	0.090	0.415	0.285	0.143	0.104
0.6	0.667	0.574	0.499	0.494	0.706	0.616	0.511	0.507
0.8	0.624	0.503	0.325	0.290	0.656	0.505	0.344	0.260
0.9	0.670	0.565	0.374	0.312	0.694	0.599	0.395	0.299
Panel F: 10% Nominal Level – Exogenous Variate Variance = 10.0								
0.4	0.278	0.171	0.101	0.090	0.437	0.310	0.157	0.121
0.6	0.691	0.558	0.469	0.472	0.707	0.596	0.490	0.460
0.8	0.638	0.503	0.345	0.289	0.648	0.540	0.349	0.267
0.9	0.682	0.591	0.416	0.351	0.699	0.585	0.400	0.329

Notes: See notes to Table 1.