Predictive Density and Conditional Confidence Interval Accuracy Tests*

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

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

September 2004

Abstract

This paper outlines testing procedures for assessing the relative out-of-sample predictive accuracy of multiple conditional distribution models. The tests that are discussed are based on either the comparison of entire conditional distributions or the comparison of predictive confidence intervals. We also briefly survey existing related methods in the area of predictive density evaluation, including methods based on the probability integral transform and the Kullback-Leibler Information Criterion. The procedures proposed in this paper are similar in many ways to Andrews' (1997) conditional Kolmogorov test and to White's (2000) reality check. In particular, a predictive density test is outlined that involves comparing square (approximation) errors associated with models i, $i = 1, ..., n$, by constructing weighted averages 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 $F_0(\cdot|\cdot)$ and $F_i(\cdot|\cdot)$ are true and model-i distributions, $u \in U$, and U is a possibly unbounded set on the real line. A conditional confidence interval version of this test is also outlined, and appropriate bootstrap procedures for obtaining critical values when predictions used in the formation of the test statistics are obtained via rolling and recursive estimation schemes are developed. An empirical illustration comparing alternative predictive models for U.S. inflation is given for the predictive confidence interval test.

JEL classification: C22, C51.
Keywords: block bootstrap, rolling and recursive estimation scheme, parameter estimation error, predictive density.

----

* 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. The authors owe great thanks to Clive W.J. Granger, whose discussions on the matter provided much of the impetus for the research reported in this paper. The authors also wish to thank participants of the conference in honor of Clive W.J. Granger on Predictive Methodology and Application in Economics and Finance held at the University of California, San Diego in January 2004 for many useful comments. Finally, a great many thanks are owed to the editor, Graham Elliott, and an anonymous referee for many useful comments. Corradi gratefully acknowledges ESRC grant RES-000-23-0006, and Swanson acknowledges financial support from a Rutgers University Research Council grant.

1 Introduction

In the management of financial risk in the insurance and banking industries, there is often a need for examining confidence intervals or entire 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). These considerations in part account for the development over the last few years of a new strand of literature addressing the issue of predictive density evaluation. Some of the important recent papers in this area include Diebold, Gunther and Tay (DGT: 1998), Christoffersen (1998), Bai (2003), Diebold, Hahn and Tay (1999), Hong (2001) and Christoffersen, Hahn and Inoue (2001), and Giacomini (2002).[1] This paper has two primary objectives. First, we build on the results of Corradi and Swanson (2004a,b) by outlining a procedure for assessing the relative out-of-sample predictive accuracy of multiple conditional distribution models that can be used with rolling and recursive estimation schemes. Second, we provide a brief survey of related techniques, such as those based on the use of the probability integral transform and the Kullback-Leibler Information Criterion (KLIC).

The literature on the evaluation of predictive densities is largely concerned with testing the null of correct dynamic specification of an individual conditional distribution model. However, in the literature on the evaluation of point forecast models it is acknowledged that all models in a group that is being evaluated may be misspecified (see e.g. White (2000) and Corradi and Swanson (2002)). In this paper, we draw on elements of these two literatures in order to provide tests for choosing amongst a group of misspecified out-of-sample predictive density (or confidence interval) models. Reiterating our above point, the focus of most of the papers cited above is that the density associated with the true conditional distribution is clearly the best predictive density. Therefore, evaluation of predictive densities is usually performed via tests for the correct (dynamic) specification of the conditional distribution. Along these lines, by making use of the probability integral transform, DGT suggest a simple and effective means by which predictive densities can be evaluated. Using the DGT terminology, if $p_{t+1}(y_{t+1}|\Omega_t)$ is the "true" conditional distribution

[1]Ten years ago, when Clive Granger was asked by one of the authors of this paper in an interview what he thought the most important future areas in time series analysis were, he replied that predictive density construction and evaluation was one of the most critical areas which needed to be developed.

of $y_{t+1}|\Omega_t$, then $p_{t+1}(y_{t+1}|\Omega_t)$ is an identically and independently distributed uniform random variable on $[0,1]$; so that the difference between an empirical version of $p_{t+1}(y_{t+1}|\Omega_t)$ constructed using estimated parameters and the 45 degree line can be used as measure of goodness of fit.[2]

A feature common to the papers cited above is that the null hypothesis is that of (dynamic) correct specification. Our approach differs from these as we do not assume that any of the competing models (including the benchmark) are correctly specified.[3] Thus, we posit that *all* models should be viewed as approximations of some true unknown underlying data generating process. For this reason, it is our objective in this paper to provide a conditional Kolmogorov test, in the spirit of Andrews (1997), that allows for the joint comparison of multiple misspecified conditional distribution models, for the case of dependent observations. In particular, assume that the object of interest is the conditional distribution of a scalar, y_{t+1}, given a (possibly vector valued) conditioning set, Z^t, where Z^t contains lags of y_{t+1} and/or lags other variables. Now, given a group of (possibly) misspecified conditional distributions, $F_1(u|Z^t,\theta_1^\dagger),...,F_n(u|Z^t,\theta_n^\dagger)$, assume that the objective is to compare these models in terms of their closeness to the true conditional distribution, $F_0(u|Z^t,\theta_0)=\Pr(y_{t+1}\leq u|Z^t)$. If $n>2$, we follow White (2000), in the sense that we choose a particular conditional distribution model as the "benchmark" and test the null hypothesis that no competing model can provide a more accurate approximation of the "true" conditional distribution, against the alternative that at least one competitor outperforms the benchmark model. However, unlike White, we evaluate predictive densities rather than point forecasts. Needless to say, pairwise

[2]Using the same approach, Bai (2003) proposes a Kolmogorov type test based on the comparison of $p_{t+1}(y_{t+1}|\Omega_t,\widehat{\theta}_T)$ 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 device based on a martingalization argument to construct a modified Kolmogorov test which has a nuisance parameter free limiting distribution. His test has power against violations of uniformity but not against violations of independence. Hong (2001) proposes an 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. For the case where the null is rejected, Hong (2001) also proposes a test for uniformity that is based on a comparison between a kernel density estimator and the uniform density, and that is robust to non independence (see also Hong and Li (2003)). Diebold, Hahn and Tay (1999) propose a nonparametric correction for improving the density forecast when the uniform (but not the independence) assumption is violated. Finally, Bontemps and Meddahi (2003, 2004) suggest a GMM type approach for testing normality and various distributional assumptions, which is robust to parameter estimation error.

[3]Corradi and Swanson (2003) allow for dynamic misspecification under both hypotheses.

comparison of alternative models, in which no benchmark need be specified, follows from our results 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, $i = 1, ..., n$, is measured 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. Additionally, integration over u in the formation of the actual test statistic is governed by $\phi(u) \geq 0$, where $\int_U \phi(u) = 1$. Thus, one can control not only the range of u, but also the weights attached to different values of u, so that more weight can be attached to important tail events, for example.[4]

We also consider tests based on an analogous conditional confidence interval version of the above measure. Namely, $E(\left(\left(F_1(\overline{u}|Z^t, \theta_1^\dagger) - F_1(\underline{u}|Z^t, \theta_1^\dagger)\right) - (F_0(\overline{u}|Z^t, \theta_0) - F_0(\underline{u}|Z^t, \theta_0))\right)^2)$, where $\underline{u}$ and $\overline{u}$ are "lower" and "upper" bounds on the confidence interval to be evaluated.[5]

Our test that is based on the conditional confidence interval loss measure as similar to the White (2000) reality check in the sense that both tests involve fixing $u \in U$, so that distributional asymptotics hinge upon showing pointwise convergence. However, our density accuracy test requires convergence as a process, as it is based upon a functional over $u \in U$, and hence stochastic equicontinuity must be established, unlike in the case of the reality check. Additionally, the bootstrap procedures outlined in this paper account for parameter estimation error rather than assuming that it vanishes, asymptotically. Finally, it should be noted that standard bootstrap statistics in contexts similar to that considered in this paper (such as in White (2000)) are usually formed by subtracting the actual statistic from an analogous statistic formed using bootstrap data and estimators. We show, however, that the use of recursive and rolling estimation schemes in-

[4] Berkowitz (2001) also considers evaluation of particular regions of a distribution. In particular, he proposes a likelihood ratio test for the null of normality against autoregressive alternatives, when the region of interest is the left tail of a density forecast. The advantage of his test is that is easy to implement and has a standard limiting distribution, although it is designed to have power against fixed alternatives, and does not account for parameter estimation error.

[5] An in-sample test based on the conditional confidence interval loss measure is outlined in Corradi and Swanson (2004b). It should perhaps be stressed, though, that the out-of-sample version of the test discussed here is substantively different, due in large part to the rolling and recursive estimation schemes used in the construction of predictions (see below for further discussion). On the other hand, the predictive density test discussed below which is based on $E\left(\left(F_i(u|Z^t, \theta_i^\dagger) - F_0(u|Z^t, \theta_0)\right)^2\right)$ has not been previously proposed in an in-sample context. However, the in-sample version of the test is technically much less complicated than the out-of-sample version, and follows directly from the results presented below.

stead requires bootstrap terms, say $1\{y^*_{t+1} \leq u\} - F_i(u|Z^{*,t}, \widetilde{\theta}^*_{i,t,\tau})$, $t \geq R$, to be recentered around the (full) sample mean, namely $\frac{1}{T}\sum_{j=s+1}^{T-1}\left(1\{y_{j+1} \leq u\} - F_i(u|Z^i, \widehat{\theta}_{i,t,\tau})\right)^2$, where $*$ denotes bootstrapped data, $\widehat{\theta}$ and $\widetilde{\theta}$ are discussed below, and i, j, τ are indexes defined in Section 2. This is necessary as the bootstrap statistic is constructed using the last P resampled observations (where P is the number of predictions), which in turn have been resampled from the full sample.

One well known measure of distributional accuracy is the Kullback-Leibler Information Criterion (KLIC). This measure is useful because the "most accurate" model can be shown to be that which minimizes the KLIC (see Section 2 for a more precise discussion). 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.[6] 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 our measure of accuracy is more "natural" than the KLIC, or vice-versa. However, in the next section we outline how certain problems (such as comparing conditional confidence intervals) that are difficult to address using the KLIC can be handled quite easily using our measure of distributional accuracy.

The limiting distribution of the suggested predictive density evaluation statistic turns out to be a functional of a Gaussian process with a covariance kernel reflecting both (dynamic) misspecification and parameter estimation error (PEE). The limiting distribution is not nuisance parameter free and critical values cannot be directly tabulated. Valid asymptotic critical values can be obtained via an empirical version of the block bootstrap which properly takes into account PEE, however. The PEE contribution is summarized by the limiting distribution of $P^{-1/2}\sum_{t=R}^{T-1}\left(\widehat{\theta}_t - \theta^\dagger\right)$, where R denotes the length of the initial estimation period, P the number of predictions (as mentioned above), $\widehat{\theta}_t$ is either a recursive $m-$estimator constructed using the first t observations or a rolling $m-$estimator constructed using observations from $t - R + 1$ to t, and $\theta^\dagger$ is its probability limit. In this context, it follows intuitively that in the recursive case, earlier observations are used more

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

frequently than temporally subsequent observations, while in the rolling case, observations in the center of the sample are used more frequently than observations either at the beginning or at the end of the sample. This introduces a location bias to the usual block bootstrap, as under standard resampling with replacement schemes, any block from the original sample has the same probability of being selected.[7] In order to circumvent this problem, we suggest a re-centering of the bootstrap score, resulting in a new bootstrap estimator, say $\widetilde{\theta}_t$, which is no longer the direct analog of $\widehat{\theta}_t$, but which is asymptotically unbiased, as required. 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 QMLE estimators (and $m-$estimators in general) 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)). In the case of rolling and recursive QMLE estimators, though, 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 such estimation schemes, as discussed in our section on the bootstrap. In addition to outlining appropriate re-centering methods, we discuss the case in which all parameters are jointly estimated as well as the case where the conditional mean parameters are first estimated via OLS or NLS, and the error variance is subsequently estimated using the residuals from the conditional mean model.[8]

The rest of the paper is organized as follows. Section 2 outlines the setup, presents the predictive density accuracy test, and states the asymptotic properties of the test statistic for both the case of recursive and rolling parameter estimation schemes. Section 3 is broken into three subsections. The

[7]Note that in the fixed sampling scheme, we just need to take into account the contribution of $\sqrt{R}\left(\widehat{\theta}_R - \theta^{\dagger}\right)$, whose limiting distribution is properly captured by "standard" block bootstrap techniques, using for example the results of Goncalves and White (2003). This case has been considered by Corradi and Swanson (2004b), within the context of in sample evaluation of conditional misspecified distribution models.

[8]From a theoretical perspective, it should be noted that all of our rolling estimation scheme results are new to this paper. Additionally, our recursive estimation scheme results for the case where parameters are estimated sequentially are new, while those for the joint estimation case summarize previous results reported in Corradi and Swanson (2004a).

first subsection outlines bootstrap procedures for mimicking the limiting distribution of parameter estimation error in rolling estimation schemes, while the second subsection summarizes results from Corradi and Swanson (2004a) for recursive estimation schemes. Finally, the third subsection applies the results of the previous two subsections in order to obtain asymptotically valid critical values for the predictive density accuracy tests. In Section 4, an empirical example based on predicting U.S. inflation is presented. Finally, concluding remarks are gathered in Section 6. All proofs are in an appendix. Hereafter, P^* denotes the probability law governing the resampled series, conditional on the sample, E^* and Var^* 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 except a subset of probability measure approaching zero, and finally $O_P^*(1)$ $\Pr-P$ denotes a term which is bounded in P^*-probability, conditional on the sample except a subset of probability measure approaching zero.

2 Predictive Density Evaluation

Our objective is to "choose" a conditional distribution model that provides the most accurate out-of-sample approximation of the true conditional distribution, given multiple predictive densities, and allowing for misspecification under both the null and the alternative hypothesis. One strategy that yields tests of the null of correct specification that are equally as useful as those discussed above is the conditional Kolmogorov test approach of Andrews (1997), which is based on a direct comparison of empirical joint distributions with the product of parametric conditional and non-parametric marginal distributions. Corradi and Swanson (2004b) extend Andrews (1997) in order to allow for the in-sample comparison of multiple misspecified models. As discussed above, one of our main objectives in this paper is the extension of those results to out-of-sample predictive density evaluation in the context of different estimation schemes. From the perspective of prediction, we assume that the objective is to form parametric conditional distributions for a scalar random variable, y_{t+1}, given Z^t, and to select among these, where $Z^t = (y_t, ..., y_{t-s_1+1}, X_t, ..., X_{t-s_2+1})$, $t = s, ..., \widetilde{T}, ...\widetilde{T} + s$, with $s = \max\{s_1, s_2\}$, and $\widetilde{T} + s = T$, with $T = (s + R) + P$. Assume that $i = 1, ..., n$ different models are estimated. In order to examine *rolling estimation schemes*, define the *rolling* m-estimator for the parameter vector associated with model i as:

$$\widehat{\theta}_{i,t,rol} = \arg\max_{\theta_i \in \Theta_i} \frac{1}{R} \sum_{j=t-R+1}^{t} \ln f_i(y_j, Z^{j-1}, \theta_i), \quad R+s \leq t \leq T-1, \; i = 1, ..., n \tag{1}$$

and

$$\theta_i^{\dagger} = \arg\max_{\theta_i \in \Theta_i} E(\ln f_i(y_j, Z^{j-1}, \theta_i)), \tag{2}$$

where $f_i(\cdot|\cdot, \theta_i)$ is the conditional density associated with $F_i(\cdot|\cdot)$, $i = 1, ..., n$, so that $\theta_i^{\dagger}$ is the probability limit of a quasi maximum likelihood estimator (QMLE). If model i is correctly specified, then $\theta_i^{\dagger} = \theta_0$. We compute a sequence a P estimators, first using observations from $s+1$ to $R+s$, then from to $s+2$ to $R+s+1$, and so on until we use the last R observations, that is from $P+s$ to $T-1$. These estimators are then used to construct sequences of P 1-step ahead forecasts and associated forecast errors, for example. In the context of such rolling estimators, it is necessary to distinguish between the cases of $P \leq R$ and $P > R$, as we shall see below. The rolling and recursive estimation schemes defined above are commonly used in out of sample forecast evaluation (see e.g. West (1996), West and McCracken (1998), Clark and McCracken (2001 and 2003)). Notable exceptions are Giacomini and White (2003), who propose the use of a rolling scheme with a fixed window, not increasing with the sample size, so that estimated parameters are treated as mixing variables, and Pesaran and Timmerman (2004), who, in order to take into account possible structure breaks, suggest an adaptive manner for choosing the window of observations.

We also consider *recursive estimation schemes*, for which we define the *recursive* m-estimator for the parameter vector associated with model i as:

$$\widehat{\theta}_{i,t,rec} = \arg\max_{\theta_i \in \Theta_i} \frac{1}{t} \sum_{j=s}^{t} \ln f_i(y_j, Z^{j-1}, \theta_i), \quad R+s \leq t \leq T-1, \; i = 1, ..., n \tag{3}$$

and $\theta_i^{\dagger}$ defined as in (2). Again following standard practice, this estimator is first computed using observations from $s+1$ to $R+s$ observations, and then from $s+1$ to $R+s+1$ observations, and so on until the last estimator is constructed using $T-1-s$ observations. As previously, these estimators are then used to construct sequences of P 1-step ahead forecasts and associated forecast errors.

Now, define the group of conditional distribution models from which we want to make a selection as $F_1(u|Z^t, \theta_1^{\dagger}), ..., F_n(u|Z^t, \theta_n^{\dagger})$, and define the true conditional distribution as $F_0(u|Z^t, \theta_0) =$

$\Pr(y_{t+1} \leq u|Z^t)$. In the sequel, $F_1(\cdot|\cdot, \theta_1^\dagger)$ 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.[9]

Following Corradi and Swanson (2004b), we begin by assuming that accuracy is measured using a distributional analog of mean square error. More precisely, the squared (approximation) error associated with model i, $i = 1, ..., n$, is measured 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.

In particular, we say that 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 $u \in U \subset \Re$. For any given evaluation point, this measure defines a norm and it implies a standard goodness of fit measure. Notice that this measure is intuitively appealing as it is related closely to standard mean square error loss measures. Additionally, the measure has the added appeal that the weighting function, $\phi(u)$, can be used to essentially "focus attention" on certain regions of U that are of interest (i.e. larger weights can be placed on certain regions). Finally, U can be defined so that certain ranges of the data can be examined. This sort of flexibility should be useful in many financial applications, for example.

As mentioned above, another measure of distributional accuracy available in the literature is the KLIC (see e.g. White (1982), Vuong (1989), Giacomini (2002), and Kitamura (2004)), according to which we should choose Model 1 over Model 2 if

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

The KLIC is a sensible measure of accuracy, as it chooses the model which on average gives higher probability to events which have actually occurred. 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. Although our approach and the KLIC approach should perhaps be viewed as alternatives, and as such one might want to implement both tests in some contexts, it should be noted that if we are interested

[9] In our framework, the competing models are known. Thus, our approach is different than the probability integral transform approach, where only the null model is explicitly stated.

in measuring accuracy over a specific region, or in measuring accuracy for a given conditional confidence interval, say, this cannot be done in a straightforward manner using the KLIC, while it can easily be done using our measure. 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.

The hypotheses of interest are:

$$H_0 : \max_{k=2,\ldots,n} \int_U E\left(\left(F_1(u|Z^t,\theta_1^\dagger) - F_0(u|Z^t,\theta_0)\right)^2 - \left(F_k(u|Z^t,\theta_k^\dagger) - F_0(u|Z^t,\theta_0)\right)^2\right)\phi(u)du \leq 0 \tag{4}$$

versus

$$H_A : \max_{k=2,\ldots,n} \int_U E\left(\left(F_1(u|Z^t,\theta_1^\dagger) - F_0(u|Z^t,\theta_0)\right)^2 - \left(F_k(u|Z^t,\theta_k^\dagger) - F_0(u|Z^t,\theta_0)\right)^2\right)\phi(u)du > 0, \tag{5}$$

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.[10] The statistic is:

$$Z_{P,\tau} = \max_{k=2,\ldots,n} \int_U Z_{P,u,\tau}(1,k)\phi(u)du, \quad \tau = 1,2 \tag{6}$$

where for $\tau = 1$ (rolling estimation scheme),

$$Z_{P,u,1}(1,k) = \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(\left(1\{y_{t+1} \leq u\} - F_1(u|Z^t,\widehat{\theta}_{1,t,rol})\right)^2 - \left(1\{y_{t+1} \leq u\} - F_k(u|Z^t,\widehat{\theta}_{k,t,rol})\right)^2\right) \tag{7}$$

and for $\tau = 2$ (recursive estimation scheme),

$$Z_{P,u,2}(1,k) = \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(\left(1\{y_{t+1} \leq u\} - F_1(u|Z^t,\widehat{\theta}_{1,t,rec})\right)^2 - \left(1\{y_{t+1} \leq u\} - F_k(u|Z^t,\widehat{\theta}_{k,t,rec})\right)^2\right), \tag{8}$$

[10] If interest focuses on testing the null of equal accuracy of only two predictive conditional distribution models, say F_1 and F_k, we can use an extension of the Diebold-Mariano (1995) test where conditional distributions instead of conditional means are evaluated. In this case, we can restate the above 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

$$H_A : \int_U E\left(\left(F_1(u|Z^t,\theta_1^\dagger) - F_0(u|Z^t,\theta_0)\right)^2 - \left(F_k(u|Z^t,\theta_k^\dagger) - F_0(u|Z^t,\theta_0)\right)^2\right)\phi(u)du \neq 0.$$

where $\widehat{\theta}_{i,t,rol}$ and $\widehat{\theta}_{i,t,rec}$ are defined as in (1) and in (3).

Now, note the hypotheses above can be restated in the following convenient form:

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

versus

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

where $\mu_i^2(u) = E\left(\left(1\{y_{t+1} \leq u\} - F_i(u|Z^t, \theta_i^\dagger)\right)^2\right)$, $i = 1, ..., n$. The intuition underlying this restatement of hypotheses 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, $1\{y_{t+1} \leq u\} - F_i(u|Z^t, \theta_i^\dagger)$ can be interpreted as an "error" term associated with computation of the conditional expectation under F_i. Now, write the statistic in equation (7) as:

$$\frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(\left(\left(1\{y_{t+1} \leq u\} - F_1(u|Z^t, \widehat{\theta}_{1,t,rol})\right)^2 - \mu_1^2(u) \right) \right.$$
$$\left. - \left(\left(1\{y_{t+1} \leq u\} - F_k(u|Z^t, \widehat{\theta}_{k,t,rol})\right)^2 - \mu_k^2(u) \right) \right) + \frac{T-R}{\sqrt{P}} (\mu_1^2(u) - \mu_k^2(u)), \tag{9}$$

In the appendix, it is shown that the first term in equation (9) weakly converges as a process on U. Also,

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

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,

$$\mu_1^2(u) - \mu_k^2(u) = E\left(\left(F_1(u|Z^t, \theta_1^\dagger) - F_0(u|Z^t, \theta_0)\right)^2\right) - E\left(\left(F_k(u|Z^t, \theta_k^\dagger) - F_0(u|Z^t, \theta_0)\right)^2\right). \tag{10}$$

In the sequel, we require the following assumptions.

Assumption A1: (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 $-4(4 + \psi)/\psi$, $\psi > 0$.

Assumption A2: (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 $i = 1, ..., n$, 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.[11]

Assumption A3: $T = R + P$, and as $T \to \infty$, $P/R \to \pi$, with $0 < \pi < \infty$.

Assumption A4: (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, $r > 2$, $i = 1, ..., n$;[12] and (ii) let $v_{kk}(u) = \text{plim}_{T\to\infty}$ $Var\left(\frac{1}{\sqrt{T}}\sum_{t=s}^{T}\left(\left(\left(1\{y_{t+1} \leq u\} - F_1(u|Z^t, \theta_1^\dagger)\right)^2 - \mu_1^2(u)\right) - \left(\left(1\{y_{t+1} \leq u\} - F_k(u|Z^t, \theta_k^\dagger)\right)^2 - \mu_k^2(u)\right)\right)\right)$, $k = 2, ..., n$, define analogous covariance terms, $v_{k,k'}(u)$, $k, k' = 2, ..., n$, and assume that $[v_{k,k'}(u)]$ is positive semi-definite, uniformly in u.

Assumptions A1 and A2 are standard memory, moment, smoothness and identifiability conditions. A1 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 A3 requires that R and P grow at the same rate. Of course, if R grows faster than P, then there is no need to capture the contribution of parameter estimation error when constructing bootstrap critical values for the tests discussed in the sequel. Assumption A4(i) states standard smoothness and domination conditions imposed on the conditional distributions of the models, and assumption A4(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 1a: Let Assumptions A1-A4 hold. Then,

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

where $Z_{P,k,\tau}(u)$ is a zero mean Gaussian process with covariance $C_{k,k'}(u,u')$. Here, $\tau = 1$ corresponds to the rolling estimation scheme, $\tau = 2$ corresponds to the recursive estimation scheme, and $C_{k,k}(u,u')$ equals:

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

[11] We say that $\nabla_{\theta_i} \ln f_i(y_t, Z^{t-1}, \theta_i)$ is $2r-$dominated on Θ_i if its $v-th$ element, $v = 1, ..., \varrho(i)$, is such that $\left|\nabla_{\theta_i} \ln f_i(y_t, Z^{t-1}, \theta_i)\right|_v \leq D_t$, and $E(|D_t|^{2r}) < \infty$. For more details on domination conditions, see Gallant and White (1988, pp. 33).

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

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

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

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

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

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

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

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

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

$$+4C\Pi_j m_{\theta_k^\dagger}(u)'A(\theta_k^\dagger)E\left(\sum_{j=-\infty}^{\infty}\nabla_{\theta_k}\ln f_k(y_{s+1}|Z^s,\theta_k^\dagger)'\left(\left(1\{y_{s+j+1}\leq u\}-F_1(u|Z^{s+j},\theta_1^\dagger)\right)^2-\mu_1^2(u)\right)\right) \quad (11)$$

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 $A(\theta_i^\dagger)=A_i^\dagger=\left(E\left(-\nabla^2_{\theta_i}\ln f_i(y_{t+1}|Z^t,\theta_i^\dagger)\right)\right)^{-1}$. For $\tau=1$: when $P\leq R$, $\Pi_1=\left(\pi-\frac{\pi^2}{3}\right)$, $C\Pi_1=\frac{\pi}{2}$, and when $P>R$, $\Pi_1=\left(1-\frac{1}{3\pi}\right)$ and $C\Pi_1=\left(1-\frac{1}{2\pi}\right)$. Finally, for $\tau=2$, $\Pi_2=2\left(1-\pi^{-1}\ln(1+\pi)\right)$ and $C\Pi_2=0.5\Pi_2$.

From this proposition, we see 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 parameters 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 $\max_{k=2,...,n} \int_U \left(Z_{P,u}(1,k) - \sqrt{P}\left(\mu_1^2(u) - \mu_k^2(u)\right)\right) \phi(u)du$, asymptotically. Of course, when H_A holds, the statistic diverges to plus infinity at rate $\sqrt{P}$.

Rather than using statistics such as ours, which are based on taking the maximum of a statistic formed by pairwise comparison of multiple models, as a way to avoid sequential testing bias associated with comparison of multiple models, one might instead rely on bounds, such as (modified) Bonferroni bounds. However, a well known drawback of such approaches is that they are 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).[13] 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 our test. The test 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. Of course, another way to obtain a arguably very "crude" ranking of models is to simply order the models according to their individual values (e.g.

[13] 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.

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

As mentioned above, one situation in which our approach offers an interesting alternative to the KLIC approach is when interests lies in the comparison of conditional confidence interval models. Following Corradi and Swanson (2004b), define the hypotheses of interest as:

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

versus

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

and consider the following statistic

$$V_{P,\tau}=\max_{k=2,\dots,n} V_{P,\underline{u},\overline{u},\tau}(1,k) \tag{12}$$

where

$$\begin{aligned} V_{P,\underline{u},\overline{u},\tau}(1,k) &= \frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\left(1\{\underline{u}\leq y_{t+1}\leq\overline{u}\}-\left(F_1(\overline{u}|Z^t,\widehat{\theta}_{1,t,\tau})-F_1(\underline{u}|Z^t,\widehat{\theta}_{1,t,\tau})\right)\right)^2\right. \\ &\quad \left.-\left(1\{\underline{u}\leq y_{t+1}\leq\overline{u}\}-\left(F_k(\overline{u}|Z^t,\widehat{\theta}_{k,t,\tau})-F_k(\underline{u}|Z^t,\widehat{\theta}_{k,t,\tau})\right)\right)^2\right) \end{aligned} \tag{13}$$

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 $\tau=2$. Further, note that the conditional interval version of $\mu_1^2-\mu_k^2$ above can be analogously written as:[14]

$$\begin{aligned} \mu_1^2-\mu_k^2 &= E\left(\left((F_1(\overline{u}|Z^t,\theta_1^\dagger)-F_1(\underline{u}|Z^t,\theta_1^\dagger))-\left(F_0(\overline{u}|Z^t,\theta_0)-F_0(\underline{u}|Z^t,\theta_0)\right)\right)^2\right) \\ &\quad -E\left(\left((F_k(\overline{u}|Z^t,\theta_k^\dagger)-F_k(\underline{u}|Z^t,\theta_k^\dagger))-\left(F_0(\underline{u}|Z^t,\theta_0)-F_0(\underline{u}|Z^t,\theta_0)\right)\right)^2\right). \end{aligned} \tag{14}$$

Consider the rolling estimation case. We then have the following result.

Proposition 1b: Let Assumptions A1-A4 hold. Then for $\tau=1$,

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

[14]Note that μ_1^2,μ_2^2, defined below, depend on the specific interval, i.e. depend on $\underline{u},\overline{u}$. However, for notational brevity we omit the dependence on $\underline{u},\overline{u}$.

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:

$$\begin{aligned} v_{kk} &= E\sum_{j=-\infty}^{\infty}\left(\left(\left(1\{\underline{u} \le y_{s+1} \le \overline{u}\} - \left(F_1(\overline{u}|Z^s,\theta_1^\dagger) - F_1(\underline{u}|Z^s,\theta_1^\dagger)\right)\right)^2 - \mu_1^2\right)\right. \\ &\left.\left(\left(1\{\underline{u} \le y_{s+1+j} \le \overline{u}\} - \left(F_1(\overline{u}|Z^{s+j},\theta_1^\dagger) - F_1(\underline{u}|Z^{s+j},\theta_1^\dagger)\right)\right)^2 - \mu_1^2\right)\right) \end{aligned} \tag{15}$$

$$\begin{aligned} &+E\sum_{j=-\infty}^{\infty}\left(\left(\left(1\{\underline{u} \le y_{s+1} \le \overline{u}\} - \left(F_k(\overline{u}|Z^s,\theta_k^\dagger) - F_k(\underline{u}|Z^s,\theta_k^\dagger)\right)\right)^2 - \mu_k^2\right)\right. \\ &\left.\left(\left(1\{\underline{u} \le y_{s+1+j} \le \overline{u}\} - \left(F_k(\overline{u}|Z^{s+j},\theta_k^\dagger) - F_k(\underline{u}|Z^{s+j},\theta_k^\dagger)\right)\right)^2 - \mu_k^2\right)\right) \end{aligned} \tag{16}$$

$$\begin{aligned} &-2E\sum_{j=-\infty}^{\infty}\left(\left(\left(1\{\underline{u} \le y_{s+1} \le \overline{u}\} - \left(F_1(\overline{u}|Z^s,\theta_1^\dagger) - F_1(\underline{u}|Z^s,\theta_1^\dagger)\right)\right)^2 - \mu_1^2\right)\right. \\ &\left.\left(\left(1\{\underline{u} \le y_{s+1+j} \le \overline{u}\} - \left(F_k(\overline{u}|Z^{s+j},\theta_k^\dagger) - F_k(\underline{u}|Z^{s+j},\theta_k^\dagger)\right)\right)^2 - \mu_k^2\right)\right) \end{aligned} \tag{17}$$

$$p_{kk} = 4m_{\theta_1^\dagger}{}' A(\theta_1^\dagger) E\left(\sum_{j=-\infty}^{\infty} \nabla_{\theta_1} \ln f_1(y_{s+1}|Z^s,\theta_1^\dagger) \nabla_{\theta_1} \ln f_1(y_{s+1+j}|Z^{s+j},\theta_1^\dagger)'\right) A(\theta_1^\dagger) m_{\theta_1^\dagger} \tag{18}$$

$$+4m'_{\theta_k^\dagger} A(\theta_k^\dagger) E\left(\sum_{j=-\infty}^{\infty} \nabla_{\theta_k} \ln f_k(y_s|Z^s,\theta_k^\dagger) \nabla_{\theta_k} \ln f_k(y_{s+j}|Z^{s+j},\theta_k^\dagger)'\right) A(\theta_k^\dagger) m_{\theta_k^\dagger} \tag{19}$$

$$-8m'_{\theta_1^\dagger} A(\theta_1^\dagger) E\left(\sum_{j=-\infty}^{\infty} \nabla_{\theta_1} \ln f_1(y_s|Z^s,\theta_1^\dagger) \nabla_{\theta_k} \ln f_k(y_{s+j}|Z^{s+j},\theta_k^\dagger)'\right) A(\theta_k^\dagger) m_{\theta_k^\dagger} \tag{20}$$

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

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

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.

In the next section, we discuss the construction of bootstrap critical values for the above tests.

3 Bootstrap Critical Values

We begin by outlining bootstrap methods for mimicking the limiting distribution of $\frac{1}{\sqrt{P}}\sum_{t=R+s}^{T-1}\left(\widehat{\theta}_{i,t,rol}-\theta^\dagger\right)$ and $\frac{1}{\sqrt{P}}\sum_{t=R+s}^{T-1}\left(\widehat{\theta}_{i,t,rec}-\theta^\dagger\right)$ where $\widehat{\theta}_{i,t,rol}$ and $\widehat{\theta}_{i,t,rec}$ are the rolling and recursive estimators as defined in (1) and (3). For fixed sampling schemes, the properties of the block bootstrap for $m-$estimators and/or GMM estimators with dependent observations have been studied by several authors. For example, Hall and Horowitz (1996) and Andrews (2002a,b) show that the block bootstrap provides improved critical values, in the sense of asymptotic refinements, 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. In addition, Inoue and Shintani (2003) show that the block bootstrap provides asymptotic refinements for linear overidentified GMM estimators for general mixing processes. A recent contribution which is useful in our context is that of Goncalves and White (2004), who show that for $m-$estimators, the limiting distribution of $\sqrt{T}(\widehat{\theta}^*_{i,T}-\widehat{\theta}_{i,T})$ provides a valid first order approximation to that of $\sqrt{T}(\widehat{\theta}_{i,T}-\theta_i^\dagger)$ for heterogeneous and near epoch dependent series, where $\widehat{\theta}^*_{i,T}$ is a resampled estimator, and T denotes the length of the entire sample. Based on the results mentioned above, one might expect $\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widehat{\theta}^*_{i,t,\tau}-\widehat{\theta}_{i,t,\tau}\right)$ to have the same limiting distribution as $\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widehat{\theta}_{i,t,\tau}-\theta_i^\dagger\right)$, $\tau = 1,2$ for the rolling and recursive estimation cases, respectively. However, in the rolling case, observations in the middle of the sample are used more frequently than observation at either the beginning or the end of the sample, while in the recursive case, earlier observations are used more frequently than temporally subsequent observations. 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 next three subsections, we address these issues by first developing bootstrap estimators that capture parameter estimation error, and thereafter outlining appropriate bootstrap test statistics for use in critical values construction.

3.1 The Block Bootstrap for Parameter Estimation Error: Rolling Estimation Scheme

Consider the overlapping block resampling scheme of Künsch (1989), which can be applied in our context as follows:[15] At each replication, draw b blocks (with replacement) of length l from the sample $W_t = (y_t, Z^{t-1})$, where $bl = T - s$. Thus, the each block is equal to $W_{i+1}, ..., W_{i+l}$, for some $i = s-1, ..., T-l+1$, with probability $1/(T-s-l+1)$. More formally, let I_k, $k = 1, ..., b$ be iid discrete uniform random variables on $[s-1, s, ..., T-l+1]$. 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.

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

$$\widehat{\theta}^*_{i,t,rol} = \arg\max_{\theta_i \in \Theta_i} \frac{1}{R} \sum_{j=t-R+1}^{t} \ln f_i(y_j^*, Z^{*,j-1}, \theta_i), \;\; R + s \leq t \leq T - 1, \;\; i = 1, ..., n. \tag{21}$$

By the first order conditions, $\frac{1}{R}\sum_{j=t-R+1}^{t} \nabla_\theta \ln f_i(y_j^*, Z^{*,j-1}, \widehat{\theta}^*_{i,t,rol}) = 0$, and via a mean value expansion of $\frac{1}{R}\sum_{j=t-R+1}^{t} \nabla_\theta \ln f_i(y_j^*, Z^{*,j-1}, \widehat{\theta}^*_{i,t,rol})$ around $\widehat{\theta}_{i,t,rol}$, after a few simple manipulations,

[15] 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.

for the case of $R \geq P$, we have that

$$\begin{aligned}
&\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widehat{\theta}^*_{i,t,rol}-\widehat{\theta}_{i,t,rol}\right)\\
=\ &\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\left(\frac{1}{R}\sum_{j=t-R+1}^{t}\nabla^2_\theta \ln f_i(y^*_j, Z^{*,j-1}, \overline{\theta}^*_{i,t,rol})\right)^{-1}\frac{1}{R}\sum_{j=t-R+1}^{t}\nabla_\theta \ln f_i(y^*_j, Z^{*,j-1}, \widehat{\theta}_{i,t,rol})\right)\\
=\ &A^\dagger_i\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\frac{1}{R}\sum_{j=t-R+1}^{t}\nabla_\theta \ln f_i(y^*_j, Z^{*,j-1}, \widehat{\theta}_{i,t,rol})\right)+o_{P^*}(1)\ \Pr-P\\
=\ &A^\dagger_i\frac{1}{\sqrt{P}R}\left(\sum_{j=s+1}^{P+s}(j-s)\nabla_\theta \ln f_i(y^*_j, Z^{*,j-1}, \widehat{\theta}_{i,t,rol})+P\sum_{j=P+s+1}^{R+s}\nabla_\theta \ln f_i(y^*_j, Z^{*,j-1}, \widehat{\theta}_{i,t,rol})\right.\\
&\left.+\sum_{j=R+s+1}^{T-1}(P+s-(j-R))\nabla_\theta \ln f_i(y^*_j, Z^{*,j-1}, \widehat{\theta}_{i,t,rol})\right)+o_{P^*}(1)\ \Pr-P, \qquad (22)
\end{aligned}$$

where $A^\dagger_i = A(\theta^\dagger_i)$ is as defined in the statement of Proposition 1a, and $\overline{\theta}^*_{i,t,rol} \in \left(\widehat{\theta}_{i,t,rol}, \widehat{\theta}^*_{i,t,rol}\right)$. Analogously,

$$\begin{aligned}
&\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widehat{\theta}_{i,t,rol}-\theta^\dagger_i\right)\\
=\ &A^\dagger_i\frac{1}{\sqrt{P}R}\left(\sum_{j=s+1}^{P+s}(j-s)\nabla_\theta \ln f_i(y_j, Z^{j-1}, \theta^\dagger_i)+P\sum_{j=P+s+1}^{R+s}\nabla_\theta \ln f_i(y_j, Z^{j-1}, \theta^\dagger_i)\right.\\
&\left.+\sum_{j=R+s+1}^{T-1}(P+s-(j-R))\nabla_\theta \ln f_i(y_j, Z^{j-1}, \theta^\dagger_i)\right)+o_{P^*}(1)\ \Pr-P. \qquad (23)
\end{aligned}$$

Now, given (2), $E\left(\nabla_\theta \ln f_i(y_j, Z^{j-1}, \theta^\dagger_i)\right)=0$ for all j, and $\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widehat{\theta}_{i,t,rol}-\theta^\dagger_i\right)$ 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,

$$E^*\left(\nabla_\theta \ln f_i(y^*_j, Z^{*,j-1}, \widehat{\theta}_{i,t,rol})\right)=\frac{1}{T-s}\sum_{j=s}^{T-1}\nabla_\theta \ln f_i(y_j, Z^{j-1}, \widehat{\theta}_{i,t,rol})+O\left(\frac{l}{T}\right)\ \Pr-P, \qquad (24)$$

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)).[16] Now, $\frac{1}{T-s}\sum_{j=s}^{T-1}\nabla_\theta \ln f_i(y_j, Z^{j-1}, \widehat{\theta}_{i,t,rol}) \neq 0$, and is instead of order $O_P\left(T^{-1/2}\right)$. Thus, $\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\frac{1}{T-s}\sum_{j=s}^{T-1}\nabla_\theta \ln f_i(y_j, Z^{j-1}, \widehat{\theta}_{i,t,rol}) = O_P(1)$,

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

and does not vanish in probability. This clearly contrasts with the full sample case, in which $\frac{1}{T-s}\sum_{j=s}^{T-1}\nabla_\theta \ln f_i(y_j, Z^{j-1}, \widehat{\theta}_{i,T}) = 0$, because of the first order conditions.[17] Thus, $\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widehat{\theta}^*_{i,t,rol} - \widehat{\theta}_{i,t,rol}\right)$ 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 (24), our objective is thus to have the bootstrap score centered around $\frac{1}{T-s}\sum_{j=s}^{T-1}\nabla_\theta f_i(y_j, Z^{j-1}, \widehat{\theta}_{i,t,rol})$. Hence, define a new bootstrap estimator, $\widetilde{\theta}^*_{i,t,rol}$, as:

$$\widetilde{\theta}^*_{i,t,rol} = \arg\max_{\theta_i\in\Theta_i}\frac{1}{R}\sum_{j=t-R+1}^{t}\left(\ln f_i(y^*_j, Z^{*,j-1}, \theta_i) - \theta_i'\frac{1}{T}\sum_{j'=s+1}^{T-1}\nabla_\theta \ln f_i(y_{j'}, Z^{j'-1}, \widehat{\theta}_{i,t,R})\right), \tag{25}$$

where $R+s\leq t\leq T-1$ and $i=1,...,n$. Given first order conditions, $\frac{1}{R}\sum_{j=t-R+1}^{t}\left(\nabla_\theta f_i(y^*_j, Z^{*,j-1}, \widetilde{\theta}^*_{i,t,rol}) - \left(\frac{1}{T}\sum_{j'=s}^{T-1}\nabla_{\theta_i} f_i(y_{j'}, Z^{j'-1}, \widehat{\theta}_{i,t,rol})\right)\right) = 0$, and via a mean value expansion of $\frac{1}{R}\sum_{j=t-R+1}^{t}\nabla_\theta f_i(y^*_j, Z^{*,j-1}, \widetilde{\theta}^*_{i,t,rol})$ around $\widehat{\theta}_{i,t,rol}$, after a few simple manipulations, we have that:

$$\begin{aligned}&\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widetilde{\theta}^*_{i,t,rol} - \widehat{\theta}_{i,t,rol}\right)\\ = \;& B^\dagger\frac{1}{\sqrt{P}}\sum_{t=R}^{T}\left(\frac{1}{R}\sum_{j=t-R+1}^{t}\left(\nabla_\theta f_i(y^*_j, Z^{*,j-1}, \widehat{\theta}_{i,t,R}) - \left(\frac{1}{T}\sum_{j'=s}^{T-1}\nabla_\theta f_i(y_{j'}, Z^{j'-1}, \widehat{\theta}_{i,t,R})\right)\right)\right)\\ &+o_{P^*}(1)\ \Pr-P.\end{aligned}$$

Given (24), it is immediate to see that the bias associated with $\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widetilde{\theta}^*_{i,t,rol} - \widehat{\theta}_{i,t,rol}\right)$ 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 then holds.

Proposition 2: Let Assumptions A1-A3 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 $R\to\infty$,

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

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

Broadly speaking, Proposition 2 states $\frac{1}{\sqrt{P}}\sum_{t=R}^{T}\left(\widetilde{\theta}^*_{i,t,rol} - \widehat{\theta}_{i,t,rol}\right)$ has the same limiting distribution as $\frac{1}{\sqrt{P}}\sum_{t=R+s}^{T-1}\left(\widehat{\theta}_{i,t,rol} - \theta^\dagger_i\right)$, conditional on the sample, and for all samples except a set

[17] Note that if $P/R\to 0$ and so $P/T\to 0$, then $\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\frac{1}{T-s}\sum_{j=s}^{T-1}\nabla_\theta \ln f_i(y_j, Z^{j-1}, \widehat{\theta}_{i,t,rol}) = o_P(1)$, and thus the location bias is negligible. However, this possibility is ruled out by Assumption A3.

with probability measure approaching zero. Note that given A3, both R and P grow with the sample size at the same rate as T.[18]

Thus far, we have considered the case in which all parameters are jointly estimated. However, it is quite customary to first estimate conditional mean parameters via OLS or NLS and subsequently estimate the error variance using residuals. Along these lines, let $\theta_i = (\beta_i, \sigma^2)$, where β_i is $\Re^{\varrho(i)-1}$ valued and σ^2 is a scalar. Additionally, let $\ln f_i(y_j, Z^{j-1}, \beta_i) = -(y_j - g_i(Z^{j-1}, \beta_i))^2$,

$$\widehat{\beta}_{i,t,rol} = \arg\min_{\beta_i \in B_i} \frac{1}{R} \sum_{j=t-R+1}^{t} (y_j - g_i(Z^{j-1}, \beta_i))^2 =, \quad R + s \leq t \leq T-1, \ i = 1, ..., n$$

where g is twice differentiable and $2r-$dominated on B, and $\widehat{\sigma}^2_{i,t,rol} = \frac{1}{R}\sum_{j=t-R+1}^{t}(y_j - g_i(Z^{j-1}, \widehat{\beta}_{i,t,rol}))^2$. The bootstrap estimator is instead

$$\widetilde{\beta}^*_{i,t,rol} = \arg\min_{\beta_i \in B_i} \frac{1}{R} \sum_{j=t-R+1}^{t} \left((y^*_j - g_i(Z^{*,j-1}, \beta_i))^2 - \beta' \frac{1}{T} \sum_{j'=s+1}^{T-1} \widehat{e}_{j',t,rol} \nabla_\beta g_i(Z^{j'-1}, \widehat{\beta}_{i,t,rol}) \right)$$

where $\widehat{e}_{j',t,rol} = (y_{j'} - g_i(Z^{j'-1}, \widehat{\beta}_{i,t,rol}))$ and
$\widetilde{\sigma}^{2,*}_{i,t,rol} = \widehat{\sigma}^2_{i,t,rol} + \frac{1}{R}\sum_{j=t-R+1}^{t}\left((y^*_j - g_i(Z^{*,j-1}, \widehat{\beta}_{i,t,rol}))^2 - \frac{1}{T}\sum_{j'=s+1}^{T-1}(y_{j'} - g_i(Z^{j'-1}, \widehat{\beta}_{i,t,rol}))^2\right)$.

Note that $\widetilde{\sigma}^{2,*}_{i,t,rol}$ is computed using residuals evaluated at $\widehat{\beta}_{i,t,rol}$ and not at $\widetilde{\beta}^*_{i,t,rol}$.

Proposition 3: Let A1-A3 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 $R \to \infty$,

$$\begin{aligned} & P\left(\omega : \sup_{v \in \Re^{\varrho(i)}} \left| P^*_T\left(\frac{1}{\sqrt{P}} \sum_{t=R}^{T} \left(\begin{pmatrix} \widetilde{\beta}^*_{i,t,rol} \\ \widetilde{\sigma}^{2,*}_{i,t,rol} \end{pmatrix} - \begin{pmatrix} \widehat{\beta}_{i,t,rol} \\ \widehat{\sigma}^2_{i,t,rol} \end{pmatrix} \right) \leq v \right) \right.\right. \\ & \left.\left. -P_T\left(\frac{1}{\sqrt{P}} \sum_{t=R}^{T} \left(\begin{pmatrix} \widehat{\beta}_{i,t,rol} \\ \widehat{\sigma}^2_{i,t,rol} \end{pmatrix} - \begin{pmatrix} \beta^\dagger_i \\ \sigma^{2\dagger}_i \end{pmatrix} \right) \leq v \right) \right| > \varepsilon \right) \\ \to \ & 0 \end{aligned}$$

3.2 The Block Bootstrap for Parameter Estimation Error: Recursive Estimation Scheme

Let $W^*_t = (y^*_t, Z^{*,t-1})$ be defined as in the previous subsection. Also, define a new recursive bootstrap m-estimator as,

$$\widetilde{\theta}^*_{i,t,rec} = \arg\max_{\theta_i \in \Theta_i} \frac{1}{t} \sum_{j=s}^{t} \left(\ln f_i(y^*_j, Z^{*,j-1}, \theta_i) - \theta'_i \frac{1}{T} \sum_{j'=s+1}^{T-1} \nabla_\theta \ln f_i(y_{j'}, Z^{j'-1}, \widehat{\theta}_{i,t,rec}) \right),$$

[18]If $P/R \to 0$, then both $\frac{1}{\sqrt{P}}\sum_{t=R}^{T}\left(\widehat{\theta}_{i,t,rol} - \theta^\dagger_i\right) = o_P(1)$ and $\frac{1}{\sqrt{P}}\sum_{t=R}^{T}\left(\widetilde{\theta}^*_{i,t,rol} - \widehat{\theta}_{i,t,rol}\right) = o_{P^*}(1)$ Pr $-P$.

where $R+s \leq t \leq T-1$, and $i=1,...,n$. The following result then holds.

Proposition 4: Let Assumptions A1-A3 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 $R \to \infty$,

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

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

Now let $\widehat{\beta}_{i,t,rec}$, $\widetilde{\beta}^*_{i,t,rec}$, $\widehat{\sigma}^2_{i,t,rec}$, and $\widetilde{\sigma}^{2,*}_{i,t,rec}$ be defined as $\widehat{\beta}_{i,t,rol}$, $\widetilde{\beta}^*_{i,t,rol}$, $\widehat{\sigma}^2_{i,t,rol}$, and $\widetilde{\sigma}^{2,*}_{i,t,rol}$, except that a recursive rather than a rolling estimation scheme is used. The following result then holds.

Proposition 5: Let A1-A3 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 $R \to \infty$,

$$\begin{aligned}
& P\left(\omega : \sup_{v \in \Re^{\varrho(i)}} \left| P_T^*\left(\frac{1}{\sqrt{P}}\sum_{t=R}^{T}\left(\left(\begin{array}{c}\widetilde{\beta}^*_{i,t,rec}\\ \widetilde{\sigma}^{2,*}_{i,t,rec}\end{array}\right) - \left(\begin{array}{c}\widehat{\beta}_{i,t,rec}\\ \widehat{\sigma}^2_{i,t,rec}\end{array}\right)\right) \leq v\right)\right.\right. \\
& \left.\left. -P_T\left(\frac{1}{\sqrt{P}}\sum_{t=R}^{T}\left(\left(\begin{array}{c}\widehat{\beta}_{i,t,rec}\\ \widehat{\sigma}^2_{i,t,rec}\end{array}\right) - \left(\begin{array}{c}\beta_i^{\dagger}\\ \sigma_i^{2\dagger}\end{array}\right)\right) \leq v\right)\right| > \varepsilon\right) \\
\to \quad & 0.
\end{aligned}$$

Given the results of the above subsections, it is now straightforward to write down appropriate bootstrap test statistics for use in the construction of valid critical values for our tests. This is done in the next subsection.

3.3 Bootstrap Critical Values for the Predictive Density Accuracy and Predictive Confidence Interval Tests

We can now construct appropriate bootstrap statistics, from whence bootstrap critical values can be constructed. Using the bootstrap sampling procedures defined in the previous section, one first constructs appropriate bootstrap samples. Thereafter, form bootstrap statistics as follows,

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

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

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

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 limit of P/R as $T \to \infty$. However, it should be noted that if $P/R \to 0$, then there is no need to mimic parameter estimation error, and so one can simply use $\widehat{\theta}_{1,t,\tau}$ instead of $\widetilde{\theta}^*_{1,t,\tau}$, in the formulation of $Z^*_{P,u,\tau}(1,k)$. In this case, though, it is still necessary to recenter around the (full) sample mean.

For the confidence interval case, define:

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

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

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

Proposition 6: Let Assumptions A1-A3 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 $\tau = 1, 2$:

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

where $Z^\mu_{P,u,\tau}(1,k) = Z_{P,u,\tau}(1,k) - \sqrt{P}\left(\mu_1^2(u) - \mu_k^2(u)\right)$, and where $\mu_1^2(u) - \mu_k^2(u)$ is defined as in equation (10).

Proposition 7: Let Assumptions A1-A3 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 $\tau = 1, 2$:

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

where $V^\mu_{P,j}(1,k) = V_{P,j}(1,k) - \sqrt{P}\left(\mu_1^2 - \mu_k^2\right)$, and where $\mu_1^2 - \mu_k^2$ is defined as in equation (14).

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 $E\left(\mu_1^2(u) - \mu_k^2(u)\right) = 0$, $\forall\, k$, 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 $\max_{k=2,\dots,n} \int_U \left(\mu_1^2(u) - \mu_k^2(u)\right) \phi(u)du = 0$, 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 $\max_{k=2,\dots,n} \int_U \left(\mu_1^2(u) - \mu_k^2(u)\right) \phi(u)du = 0$, 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 $\max_{k=2,\dots,n} \int_U \left(\mu_1^2(u) - \mu_k^2(u)\right) \phi(u)du = 0$, 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.

4 Empirical Illustration - Forecasting Inflation

In this section we use a simple stylized macroeconomic example to illustrate how to apply the predictive density accuracy test discussed in Section 2.[19] In particular, assume that the objective is to select amongst 4 different predictive density models for inflation, including an linear AR model and an ARX model, where the ARX model differs from the AR model only through the inclusion of unemployment as an additional explanatory variable. Assume also that 2 versions of each of these models are used, one assuming normality, and one assuming that the conditional distribution being evaluated follows a Student's t distribution with 5 degrees of freedom. Further, assume that the number of lags used in these models is selected via use of either the SIC or the AIC. This example can thus be thought of as an out-of-sample evaluation of simplified Phillips curve type models of inflation.

The data used were obtained from the St. Louis Federal Reserve website. For unemployment, we use the seasonally adjusted civilian unemployment rate. For inflation, we use the 12th difference of the log of the seasonally adjusted CPI for all urban consumers, all items. Both data series were found to be I(0), based on application of standard augmented Dickey-Fuller unit root tests. All data are monthly, and the sample period is 1954:1-2003:12. This 600 observation sample was broken into two equal parts for test construction, so that $R = P = 300$. Additionally, all predictions were 1-step ahead, and were constructed using the recursive estimation scheme discussed above.[20] Bootstrap percentiles were calculated based on 100 bootstrap replications, and we set $u \in U \subset [Inf_{\min}, Inf_{\max}]$, where Inf_t is the inflation variable being examined, and 100 equally spaced values for u across this range were used (i.e. $\phi(u)$ is the uniform density). Lags were selected as follows. First, and using only the initial R sample observations, autoregressive lags were selected according to both the SIC and the AIC. Thereafter, fixing the number of autoregressive lags, the number of lags of unemployment ($Unem_t$) was chosen, again using each of the SIC and the AIC. This framework enabled us to compare various permutations of 4 different models using the $Z_{P,2}$

[19]Monte Carlo experiments examining the finite sample properties of the tests are ongoing, and will be reported in subsequent research. A preliminary indication of the finite sample performance of the tests proposed in this paper, however, is available in Corradi and Swanson (2004a), where related tests that focus on the conditional mean rather than the conditional distribution are discussed.

[20]Results based on the rolling estimation scheme have been tabulated, and are available upon request from the authors.

statistic, where

$$Z_{P,2} = \max_{k=2,\ldots,4} \int_U Z_{P,u,2}(1,k)\phi(u)du$$

and

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

as discussed in Section 2. In particular, we consider (i) a comparison of AR and ARX models, with lags selected using the SIC; (ii) a comparison of AR and ARX models , with lags selected using the AIC; (iii) a comparison of AR models, with lags selected using either the SIC or the AIC; and (iv) a comparison of ARX models, with lags selected using either the SIC or the AIC. Recalling that each model is specified with either a Gaussian or Student's t error density,we thus have 4 applications, each of which involves the comparison of 4 different predictive density models. Results are gathered in Tables 1-4. The tables contain: mean square forecast errors - MSFE (so that our density accuracy results can be compared with model rankings based on conditional mean evaluation); lags used; $\int_U \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(1\{Inf_{t+1} \leq u\} - F_1(u|Z^t, \widehat{\theta}_{1,t}) \right)^2 \phi(u)du = DMSFE$ (for "ranking" based on our density type mean square error measures), and {50,60,70,80,90} split and full sample bootstrap percentiles for block lengths of {3,5,10,15,20} observations (for conducting inference using $Z_{P,2}$).

Although this empirical application is presented only for illustrative purposes, we feel that the results presented in Tables 1-4 are indicative of the types of results that may generally be obtained upon application of the tools developed in this paper. For example, notice that lower MSFEs are uniformly associated with models that have lags selected via the AIC. This rather surprising result suggests that parsimony is not always the best "rule of thumb" for selecting models for predicting conditional mean, and is a finding in agreement with one of the main conclusions of Marcellino, Stock and Watson (2004). Interestingly, though, the density based mean square forecast error measure that we consider (i.e. $DMSFE$) is not generally lower when the AIC is used. This suggests that the choice of lag selection criterion is sensitive to whether individual moments or entire distributions are being evaluated. Of further note is that $\max_{k=2,\ldots,4} \int_U Z_{P,u,2}(1,k)\phi(u)du$ in Table 1 is -0.046, which fails to reject the null hypothesis that the benchmark AR(1)-normal density model is at least as "good" as any other SIC selected model. Furthermore, when only AR models are evaluated (see Table 3), there is nothing gained by using the AIC instead of the SIC, and the normality assumption is again not "bested" by assuming fatter predictive density tails (notice that in this case, failure to reject occurs even when 50th percentiles of either the split or full sample recursive block bootstrap distributions are used to form critical values). In contrast to the above results, when either the AIC is used for all competitor models (Table 2), or when only ARX models are considered with lags selected by either SIC or AIC (Table 4), the null hypothesis of normality is rejected using 90th percentile critical values. Further, in both of these cases, the "preferred model", based on ranking according to $DMSFE$, is (i) an ARX model with Student's t errors (when only the AIC is used to select lags) or (ii) an ARX model with Gaussian errors and lags selected via the SIC (when only ARX models are compared). This result indicates the importance of comparing a wide variety of models. If we were only to compare AR and ARX models using the AIC, as in Table 2, then we would conclude that ARX models beat AR models, and that fatter tails should replace Gaussian tails in error density specification. However, inspection of the density based MSFE measures across all models considered in the tables makes clear that the lowest $DMSFE$ values are always associated with more parsimonious models (with lags selected using the SIC) that assume Gaussianity.

5 Concluding Remarks

In this paper we have outlined predictive density and predictive conditional confidence interval accuracy tests. In addition, we briefly surveyed related predictive density evaluation methods, and stressed that our methods differ from many of these in the sense that we allow all competing models to be misspecified. We also outlined simple block bootstrap procedures applicable to a wide class of test statistics (those for which limit distributions are functionals of Gaussian processes) constructed using estimators obtained via rolling and recursive estimation schemes. Interestingly, in the context of predictive density and conditional confidence interval tests, it turns out that the standard block bootstrap is not the same as the simple bootstrap that we propose - instead, the usual bootstrap approach must be modified in a number of ways prior to use in conjunction with predictions formed using recursive and rolling estimation schemes. An empirical example based on forecasting models of inflation is used to illustrate our methodology, and it is found that evaluation based on AR models leaves nothing to choose between $AR(1)$ models under normality and models under alternative Student's t distributional assumptions and those with lags selected using the AIC instead of the SIC. On the other hand, when the lag selection device is fixed to be the AIC, then ARX predictive density models "win", and the Student's t distribution better mimics the actual distribution of the predictive density than the Gaussian distribution.

This paper is meant as a starting point. Much further research is needed, both theoretical and empirical, before the full impact of the bootstrap procedures and predictive density accuracy tests that we have outlined will become clear. For example, our bootstrap procedures need to be examined via Monte Carlo experimentation. Additionally, empirical and Monte Carlo investigation comparing and contrasting the different predictive models using tests of the conditional mean as well as the conditional distribution will shed further light on the trade-off between using alternative varieties of predictive accuracy tests.

6 Appendix

The main theoretical contributions of this paper are contained in the proofs of Propositions 2 and 3, as the other propositions follow in a fairly straightforward manner, given the results of Corradi and Swanson (2004a,b).

Proof of Proposition 1a: Let $\mu_i^2(u) = E\left(\left(1\{y_{t+1} \leq u\} - F_i(u|Z^t, \theta_i^\dagger)\right)^2\right)$ $= E\left(\left(1\{y_{t+1} \leq u\} - F_0(u|Z^t, \theta_0)\right)^2\right) + E\left(\left(F_0(u|Z^t, \theta_0) - F_i(u|Z^t, \theta_i^\dagger)\right)^2\right)$. We begin by considering the rolling case. For any given u,

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

$$\begin{aligned}
= & \frac{1}{\sqrt{P}} \sum_{t=R+s}^{T-1} \left(\left(1\{y_{t+1} \leq u\} - F_1(u|Z^t, \widehat{\theta}_{1,t,rol})\right)^2 - \mu_1^2(u)\right) \\
& - \frac{1}{\sqrt{P}} \sum_{t=R+s}^{T-1} \left(\left(1\{y_{t+1} \leq u\} - F_k(u|Z^t, \widehat{\theta}_{k,t,rol})\right)^2 - \mu_k^2(u)\right) + \sqrt{P}(\mu_1^2(u) - \mu_k^2(u)) \\
= & \frac{1}{\sqrt{P}} \sum_{t=R+s}^{T-1} \left(\left(1\{y_{t+1} \leq u\} - F_1(u|Z^t, \theta_1^\dagger)\right)^2 - \mu_1^2(u)\right) \\
& - \frac{1}{\sqrt{P}} \sum_{t=s}^{T-1} \left(\left(1\{y_{t+1} \leq u\} - F_k(u|Z^t, \theta_k^\dagger)\right)^2 - \mu_k^2(u)\right) \\
& - \frac{2}{P} \sum_{t=R+s}^{T-1} \nabla_{\theta_1} F_1(u|Z^t, \overline{\theta}_{1,t,rol})' \left(1\{y_{t+1} \leq u\} - F_1(u|Z^t, \theta_1^\dagger)\right) \sqrt{P}\left(\widehat{\theta}_{1,t,rol} - \theta_1^\dagger\right) \\
& + \frac{2}{P} \sum_{t=R+s}^{T-1} \nabla_{\theta_k} F_k(u|Z^t, \overline{\theta}_{k,t,rol})' \left(1\{y_{t+1} \leq u\} - F_k(u|Z^t, \theta_k^\dagger)\right) \sqrt{P}\left(\widehat{\theta}_{k,t,rol} - \theta_k^\dagger\right) \\
& + \sqrt{P}(\mu_1^2(u) - \mu_k^2(u)) + o_P(1)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\left(1\{y_{t+1}\leq u\}-F_1(u|Z^t,\theta_1^\dagger)\right)^2-\mu_1^2(u)\right)\\
&\quad -\frac{1}{\sqrt{P}}\sum_{t=s}^{T-1}\left(\left(1\{y_{t+1}\leq u\}-F_k(u|Z^t,\theta_k^\dagger)\right)^2-\mu_k^2(u)\right)\\
&\quad -2m_{\theta_1^\dagger}(u)'A(\theta_1^\dagger)\frac{1}{\sqrt{P}}\sum_{t=R+s}^{T-1}\frac{1}{R}\sum_{j=t-R+1}^{t}\ln f_1(y_j,Z^{j-1},\theta_1)\\
&\quad +2m_{\theta_k^\dagger}(u)'A(\theta_k^\dagger)\frac{1}{\sqrt{P}}\sum_{t=R+s}^{T-1}\frac{1}{R}\sum_{j=t-R+1}^{t}\ln f_k(y_j,Z^{j-1},\theta_k)\\
&\quad +\sqrt{P}(\mu_1^2(u)-\mu_k^2(u))+o_P(1)\\
&= I_P(u)+\sqrt{P}(\mu_1^2(u)-\mu_k^2(u))+o_P(1) \qquad (26)
\end{aligned}
$$

where $\overline{\theta}_{i,t,rol}\in(\widehat{\theta}_{i,t,rol},\theta_i^\dagger)$, $i=1,...,n$, and $m_{\theta_i^\dagger}(u)'=E\left(\nabla_{\theta_i}F_i(u|Z^t,\theta_i^\dagger)'\left(1\{y_{t+1}\leq u\}-F_i(u|Z^t,\theta_i^\dagger)\right)\right)$ and $A(\theta_i^\dagger)=\left(E\left(-\nabla^2_{\theta_i}\ln f_i(y_{t+1}|Z^t,\theta_i^\dagger)\right)\right)^{-1}$ and where the $o_P(1)$ term holds uniformly in $u\in U$. We need to distinguish between the case of $P\leq R$ and $P>R$. In the former case, by Lemma 4.1 in West and McCracken (1998, WM), $\frac{1}{\sqrt{P}}\sum_{t=R+s}^{T-1}\frac{1}{R}\sum_{j=t-R+1}^{t}\ln f_k(y_j,Z^{j-1},\theta_1^\dagger)$ is asymptotically normal with variance $\left(\pi-\frac{\pi^2}{3}\right)E\left(\sum_{j=-\infty}^{\infty}\nabla_{\theta_1}\ln f_1(y_{s+1}|Z^s,\theta_1^\dagger)\nabla_{\theta_1}\ln f_1(y_{s+j+1}|Z^{s+j},\theta_1^\dagger)'\right)$, while the long run covariance between
$\frac{1}{\sqrt{P}}\sum_{t=R+s}^{T-1}\frac{1}{R}\sum_{j=t-R+1}^{t}\ln f_k(y_j,Z^{j-1},\theta_1^\dagger)$ and
$\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\left(1\{y_{t+1}\leq u\}-F_1(u|Z^t,\theta_1^\dagger)\right)^2-\mu_1^2(u)\right)$ is given by
$\frac{\pi}{2}E\left(\sum_{j=-\infty}^{\infty}\nabla_{\theta_1}\ln f_1(y_{s+1}|Z^s,\theta_1^\dagger)\left(\left(1\{y_{s+j+1}\leq u\}-F_k(u|Z^{s+j},\theta_k^\dagger)\right)^2-\mu_k^2(u)\right)\right)$. Again from Lemma 4.1 in WM, for the case of $P>R$, $\left(\pi-\frac{\pi^2}{3}\right)$ and $\frac{\pi}{2}$ are replaced by $\left(1-\frac{1}{3\pi}\right)$ and $\left(1-\frac{1}{2\pi}\right)$.
In the recursive case, the second last line in (26) becomes,

$$
-2m_{\theta_1^\dagger}(u)'A(\theta_1^\dagger)\frac{1}{\sqrt{P}}\sum_{t=R+s}^{T-1}\frac{1}{t}\sum_{j=s+1}^{t}\ln f_1(y_j,Z^{j-1},\theta_1)+2m_{\theta_k^\dagger}(u)'A(\theta_k^\dagger)\frac{1}{\sqrt{P}}\sum_{t=R+s}^{T-1}\frac{1}{t}\sum_{j=s+1}^{t}\ln f_k(y_j,Z^{j-1},\theta_k)
$$

and the asymptotic variance of the parameter estimation error component as well as the covariance term follow from Lemma A5 in West (1996). The statement in the Proposition will follow once we have shown convergence of the finite dimensional distribution and stochastic equicontinuity in U. Now, convergence of finite dimensional distributions follows straightforwardly from the Cramer-Wold device. Now, in order to show that $I_P(u)$, as defined in (26) weakly converges as a process on U, we need to show that it is stochastic equicontinuous on U. First note that,

$$\begin{aligned}
I_P(u) &= \frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(F_1^2(u|Z^t,\theta_1^\dagger)-E\left(F_1^2(u|Z^t,\theta_1^\dagger)\right)\right) \\
&-\frac{2}{\sqrt{P}}\sum_{t=R}^{T-1}\left(F_1(u|Z^t,\theta_1^\dagger)1\{Y_{t+1}\le u\}-E\left(F_1(u|Z^t,\theta_1^\dagger)1\{Y_{t+1}\le u\}\right)\right) \\
&-\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(F_k^2(u|Z^t,\theta_k^\dagger)-E\left(F_k^2(u|Z^t,\theta_k^\dagger)\right)\right) \\
&+\frac{2}{\sqrt{P}}\sum_{t=R}^{P}\left(F_k(u|Z^t,\theta_k^\dagger)1\{Y_{t+1}\le u\}-E\left(F_k(u|Z^t,\theta_k^\dagger)1\{Y_{t+1}\le u\}\right)\right) \\
&-2m_{\theta_1^\dagger}(u)'\sqrt{P}\left(\widehat{\theta}_{1,t,R}-\theta_1^\dagger\right)+2m_{\theta_k^\dagger}(u)'\sqrt{P}\left(\widehat{\theta}_{k,t,R}-\theta_k^\dagger\right)+o_P(1), \qquad (27)
\end{aligned}$$

where $m_{\theta_i^\dagger}(u)'=E\left(\nabla_{\theta_i}F_i(u|Z^t,\theta_i^\dagger)'\left(1\{y_{t+1}\le u\}-F_i(u|Z^t,\theta_i^\dagger)\right)\right)$, and the $o_P(1)$ term holds uniformly in $u\in U$. Let $I_{i,P}(u)$ be the term in the $i-th$ line of (27). Then,

$$\begin{aligned}
I_P(u) &= \frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}I_{1,t}(u)-\frac{2}{\sqrt{P}}\sum_{t=R}^{T-1}I_{2,t}(u)-\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}I_{3,t}(u)+\frac{2}{\sqrt{T}}\sum_{t=R}^{T}I_{4,t}(u) \\
&-2m_{\theta_1^\dagger}(u)'\sqrt{P}\left(\widehat{\theta}_{1,t,rol}-\theta_1^\dagger\right)+2m_{\theta_k^\dagger}(u)'\sqrt{P}\left(\widehat{\theta}_{k,t,rol}-\theta_k^\dagger\right)+o_P(1), \qquad (28)
\end{aligned}$$

and noting that $m_{\theta_i^\dagger}(u)'$ is equicontinuous on U, it suffices to show that the first four terms on the right hand side of (28) are stochastic equicontinuous on U. Thus, it suffices to show that,

$$\limsup_{T\to\infty}\sum_{k=1}^{4}P\left(\sup_{\substack{u,u_j\in U\\ \rho_k(u,u_j)<\delta}}\left|\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}I_{k,t}(u)-\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}I_{k,t}(u_j)\right|>\varepsilon\right)=0,\ \text{as } \delta\to 0,$$

where $\rho_k(u,u_j)=\left(E((I_{k,t}(u)-I_{k,t}(u_j))^4)\right)^{1/4}$. Now, define the bracketing number, $N_{k,4}(\epsilon,U)$, to be the smallest number, $n\in N$, for which there exists $(u_1,...,u_n)\in U$, such that for any $u\in U$ there exists u_j, $j=1,...,n$ ensuring that $\left(E((I_{k,t}(u)-I_{k,t}(u_j))^4)\right)^{1/4}\le\epsilon$. Once we have shown that for $k=1,...,4$, $\int_0^1\sqrt{\log N_{k,4}(\epsilon,U)}d\epsilon<\infty$, then stochastic equicontinuity of $I_T(u)$ on U follows from Theorem 1 (Application 1) in Doukhan, Massart and Rio (DMR: 1995). In fact, given the size of the mixing coefficients in A(i), $\sum_{j=1}^{\infty}j^{-1}\beta_j$ is a convergent series, and thus condition (2.10) in DMR can be replaced with the condition that $\int_0^1\sqrt{\log N_{k,4}(\epsilon,U)}d\epsilon<\infty$, for $k=1,...,4$. Now,

$$\begin{aligned}
\left(E((I_{1,t}(u_j)-I_{1,t}(u))^4)\right)^{1/4} &\le \left(E\left(\left(F_1^2(u|Z^t,\theta_1^\dagger)-F_1^2(u_j|Z^t,\theta_1^\dagger)\right)^4\right)\right)^{1/4} \\
&+\left|\int_V\left(F_1^2(u|z,\theta_1^\dagger)-F_1^2(u_j|z,\theta_1^\dagger)\right)f_0(z)dz\right|, \qquad (29)
\end{aligned}$$

where $f_0(\cdot)$ denotes the "true" marginal density of the conditioning variable(s), and V is the support of Z^t. With regard to the first term on the RHS of (29), note that there exists a $\widetilde{z} \in V$ such that,

$$\left(E\left(\left(F_1^2(u|Z^t,\theta_1^\dagger) - F_1^2(u_j|Z^t,\theta_1^\dagger)\right)^4\right)\right)^{1/4} = C_1\left(\int_V \left(F_1^2(u|z,\theta_1^\dagger) - F_1^2(u_j|\widetilde{z},\theta_1^\dagger)\right)^4 f_0(z)dz\right)^{1/4},$$

with $C_1 = \frac{\left(\int_V \left(F_1^2(u|z,\theta_1^\dagger) - F_1^2(u_j|z,\theta_1^\dagger)\right)^4 f_0(z)dz\right)^{1/4}}{\left(\int_V \left(F_1^2(u|z,\theta_1^\dagger) - F_1^2(u_j|\widetilde{z},\theta_1^\dagger)\right)^4 f_0(z)dz\right)^{1/4}} < \infty$. We can choose $(u_1,...,u_n) \in U$, such that $F_1^2(u_j|\widetilde{z},\theta_1^\dagger) = j\delta$, $j = 1,...,1/\delta$, so that the LHS of the last inequality above is majorized by $C_1\delta$. With regard to the second term on the RHS of (29),

$$\left|\int_V \left(F_1^2(u|z,\theta_1^\dagger) - F_1^2(u_j|z,\theta_1^\dagger)\right) f_0(z)dz\right| \leq C_2\int_V \left|F_1^2(u|z,\theta_1^\dagger) - F_1^2(u_j|\widetilde{z},\theta_1^\dagger)\right| f_0(z)dz, \tag{30}$$

where $C_2 = \frac{\int_V \left|F_1^2(u|z,\theta_1^\dagger) - F_1^2(u_j|z,\theta_1^\dagger)\right| f_0(z)dz}{\int_V \left|F_1^2(u|z,\theta_1^\dagger) - F_1^2(u_j|\widetilde{z},\theta_1^\dagger)\right| f_0(z)dz} < \infty$. For the same $(u_1,...,u_n)$ as above, set $F_1^2(u_j|\widetilde{z},\theta_1^\dagger) = j\delta$, $j = 1,...,1/\delta$, so that for any u we can find an u_j ensuring that the last term in (30) is majorized by $C_2\delta$. Now, set $\delta(\epsilon) = \epsilon/(C_1+C_2)$, so that $N_{1,4}(\epsilon,U) = 1/\delta(\epsilon) = (C_1+C_2)\epsilon^{-1}$, for $0 < \epsilon < 1/2$, and $\int_0^1 \sqrt{\log((C_1+C_2)/\epsilon)}d\epsilon \leq \int_0^1 \log((C_1+C_2)/\epsilon)d\epsilon = \log(C_1+C_2)+1 < \infty$. Finally, note that $I_{2,T}(u)$, $I_{3,T}(u)$, and $I_{4,T}(u)$ can be treated in an analogous manner. It has been already shown in Section 2 that

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

Thus, by the continuous mapping theorem it follows that,

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

where $Z_{1,k}(u)$ is a zero mean Gaussian process with covariance $C_k(u,u')$ defined as in (11), for $k = 2,...,m$.

Proof of Proposition 1b: Follows directly from Theorem 1 in Corradi and Swanson (2004b).

Proof of Proposition 2: As all models can be treated at the same manner, for notational simplicity, we drop the subscript denoting the model. Given (**??**), by first order conditions,

$$\frac{1}{R}\sum_{j=t-R+1}^{t}\left(\nabla_\theta \ln f(y_j^*, Z^{*,j-1}, \widetilde{\theta}_{t,rol}^*) - \left(\frac{1}{T}\sum_{k=s}^{T-1}\nabla_\theta \ln f(y_k, Z^{k-1}, \widehat{\theta}_{t,rol})\right)\right) = 0,\ t \geq R+s$$

Thus, a Taylor expansion around $\widehat{\theta}_{t,rol}$ yields:

$$\begin{aligned}\left(\widetilde{\theta}^{*}_{t,rol}-\widehat{\theta}_{t,rol}\right) &= \left(\frac{1}{R}\sum_{j=j-R+1}^{t}\nabla^2_\theta \ln f(y^*_j, Z^{*,j-1}, \overline{\theta}^{*}_{t,rol})\right)^{-1} \\ &\quad\times\left(\frac{1}{R}\sum_{j=t-R+1}^{t}\left(\nabla_\theta \ln f(y^*_j, Z^{*,j-1}, \widehat{\theta}_{t,rol}) - \left(\frac{1}{T}\sum_{k=s}^{T-1}\nabla_\theta \ln f(y_k, Z^{k-1}, \widehat{\theta}_{t,rol})\right)\right)\right),\end{aligned}$$

where $\overline{\theta}^{*}_{t,rol} \in \left(\widetilde{\theta}^{*}_{t,rol}, \widehat{\theta}_{t,rol}\right)$. Hereafter, let $A^\dagger = \left(E\left(\nabla_\theta \ln f(y_j, Z^{j-1}, \theta^\dagger)\right)\right)^{-1}$. Recalling that we resample from the entire sample, regardless the value of t, it follows that:

$$\frac{1}{R}\sum_{j=t-R+1}^{t} E^*\left(\nabla^2_\theta \ln f(y^*_j, Z^{*,j-1}, \theta)\right) = \frac{1}{T}\sum_{k=s}^{T-1}\nabla^2_\theta \ln f(y_k, Z^{k-1}, \theta) + O_{P^*}\left(\frac{l}{T}\right),\ \ \Pr-P, \tag{31}$$

where the $O_{P^*}\left(\frac{l}{T}\right)$ term is due to the end effect (i.e. due to the contribution of the first and last l observations, as shown in Lemma A1 in Fitzenberger (1997)). Thus,

$$\begin{aligned}&\sup_{t\geq R}\sup_{\theta\in\Theta}\left|\left(\frac{1}{t}\sum_{j=s}^{t}\nabla^2_\theta \ln f(y^*_j, Z^{*,j-1}, \theta)\right)^{-1} - A^\dagger\right| \\ \leq\ &\sup_{t\geq R}\sup_{\theta\in\Theta}\left|\left(\frac{1}{t}\sum_{j=s}^{t}\nabla^2_\theta \ln f(y^*_j, Z^{*,j-1}, \theta)\right)^{-1} - \left(\frac{1}{t}\sum_{j=s}^{t}E^*\left(\nabla^2_\theta \ln f(y^*_j, Z^{*,j-1}, \theta)\right)\right)^{-1}\right| \\ &+\sup_{t\geq R}\sup_{\theta\in\Theta}\left|\left(\frac{1}{t}\sum_{j=s}^{t}E^*\left(\nabla^2_\theta \ln f(y^*_j, Z^{*,j-1}, \theta)\right)\right)^{-1} - A^\dagger\right|.\end{aligned} \tag{32}$$

Given (31), and Assumptions A1-A2, the second term on the RHS of (32) is $o_P(1)$. Recalling also that the resampled series consists of b independent and identically distributed blocks, and that $b/T^{1/2} \to \infty$, it follows that the first term on on the RHS of (32) is $o_{P^*}(1)$ $\Pr-P$, given the uniform law of large number for *iid* random variables. Thus,

$$\begin{aligned}&\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widetilde{\theta}^{*}_{t,rol}-\widehat{\theta}_{t,rol}\right) \\ =\ & B^\dagger\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\frac{1}{R}\sum_{j=t-R+1}^{t}\left(\nabla_\theta q(y^*_j, Z^{*,j-1}, \widehat{\theta}_{t,rol}) - \left(\frac{1}{T}\sum_{k=s}^{T-1}\nabla_\theta q(y_k, Z^{k-1}, \widehat{\theta}_{t,rol})\right)\right)\right) \\ &+o_{P^*}(1)\ \Pr-P,\end{aligned} \tag{33}$$

and a first order expansion of the RHS of (33) around $\theta^{\dagger}$ yields:

$$\begin{aligned}
&\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widetilde{\theta}^{*}_{t,rol}-\widehat{\theta}_{t,rol}\right)\\
=\ & A^{\dagger}\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\frac{1}{R}\sum_{j=t-R+1}^{t}\left(\nabla_{\theta}\ln f(y^{*}_{j},Z^{*,j-1},\theta^{\dagger})-\left(\frac{1}{T}\sum_{k=s}^{T-1}\nabla_{\theta}\ln f(y_{k},Z^{k-1},\theta^{\dagger})\right)\right)\right)\\
&+A^{\dagger}\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\left(\frac{1}{R}\sum_{j=t-R+1}^{t}\left(\nabla^{2}_{\theta}\ln f(y^{*}_{j},Z^{*,j-1},\overline{\theta}_{t,rol})-\left(\frac{1}{T}\sum_{k=s}^{T-1}\nabla^{2}_{\theta}\ln f(y_{k},Z^{k-1},\overline{\theta}_{t,rol})\right)\right)\right)\right.\\
&\left.\times\left(\widehat{\theta}_{t,rol}-\theta^{\dagger}\right)\right)+o_{P^{*}}(1)\ \Pr-P. \qquad (34)
\end{aligned}$$

We need to show that the second term on the RHS of (34) is $o_{P^{*}}(1)$ $\Pr-P$. Note that this term is majorized by

$$A^{\dagger}\sup_{t\geq R}\sup_{\theta\in\Theta}\frac{\sqrt{P}}{R^{1+\vartheta}}\left|\sum_{j=t-R+1}^{t}\left(\nabla^{2}_{\theta}\ln f(y^{*}_{j},Z^{*,j-1},\theta)-\left(\frac{1}{T}\sum_{k=s}^{T-1}\nabla^{2}_{\theta}\ln f(y_{k},Z^{k-1},\theta)\right)\right)\right|\sup_{t\geq R}T^{\vartheta}\left|\widehat{\theta}_{t,rol}-\theta^{\dagger}\right|,$$

with $1/3<\vartheta<1/2$. Recalling also that $bl=T$ and $l=o(T^{1/4})$, it follows that $b/T^{3/4}\to\infty$. Thus, by the same argument used in Lemma 1(i) in Altissimo and Corradi (2002), and given (31), it follows that:

$$\sup_{t\geq R}\sup_{\theta\in\Theta}\left|\frac{1}{R}\sum_{j=s}^{t}\left(\nabla^{2}_{\theta}\ln f(y^{*}_{j},Z^{*,j-1},\theta)-\left(\frac{1}{T}\sum_{k=s}^{T-1}\nabla^{2}_{\theta}\ln f(y_{k},Z^{k-1},\theta)\right)\right)\right|=O_{a.s.^{*}}\left(\sqrt{\frac{\log\log b}{b}}\right),\ a.s.-P.$$

Thus,

$$\sup_{t\geq R}\sup_{\theta\in\Theta}\frac{\sqrt{P}}{R^{1+\vartheta}}\left|\sum_{j=t-R+1}^{t}\left(\nabla^{2}_{\theta}\ln f(y^{*}_{j},Z^{*,j-1},\theta)-\left(\frac{1}{T}\sum_{j=s}^{T}\nabla^{2}_{\theta}\ln f(y_{j},Z^{j-1},\theta)\right)\right)\right|=o_{P^{*}}(1),\ \ \Pr-P,$$

for $\vartheta>1/3$. Finally, for all $\vartheta<1/2$, $\sup_{t\geq R}R^{\vartheta}\left|\widehat{\theta}_{t,rol}-\theta^{\dagger}\right|=o_{P}(1)$ by Lemma A3 in West (1996). Let $\nabla_{\theta}\ln f(y^{*}_{j},Z^{*,j-1},\theta^{\dagger})=h^{*}_{t}$ and $\nabla_{\theta}\ln f(y_{j},Z^{j-1},\theta^{\dagger})=h$ and recall that,

$$\frac{1}{R}\sum_{j=t-R+1}^{t}E^{*}\left(\nabla_{\theta}\ln f(y^{*}_{j},Z^{*,j-1},\theta^{\dagger})\right)=\frac{1}{T}\sum_{k=s}^{T-1}\nabla_{\theta}\ln f(y_{k},Z^{k-1},\theta^{\dagger})+O_{P}\left(\frac{l}{T}\right).$$

Assuming $P \leq R$, the right hand side of (34), can be written as,

$$\begin{aligned}
&\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widetilde{\theta}^{*}_{t,rol}-\widehat{\theta}_{t,rol}\right)= \\
&A^{\dagger}\frac{1}{\sqrt{P}R}\sum_{t=s+1}^{P+s}(t-s)\left(h^{*}_{t}-\overline{h}_{T}\right)+A^{\dagger}\frac{\sqrt{P}}{R}\sum_{t=P+s+1}^{R+s}\left(h^{*}_{t}-\overline{h}_{T}\right) \\
&+A^{\dagger}\frac{1}{\sqrt{P}R}\sum_{t=R+s+1}^{T-1}\left(P+s-(t-R)\right)\left(h^{*}_{t}-\overline{h}_{T}\right) \\
&+o^{*}_{P}(1),\ \ \Pr-P. \qquad (35)
\end{aligned}$$

Now, $\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widetilde{\theta}^{*}_{t,rol}-\widehat{\theta}_{t,rol}\right)$ satisfies a central limit theorem for triangular independent arrays (see e.g. White and Wooldridge (1988)), and thus, conditional on the sample, it converges in distribution to a zero mean normal random variable.

Furthermore, by Lemma 4.1 in West and McCracken (1998):

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

where $C_{00}=\sum_{j=-\infty}^{\infty}E\left(\left(\nabla_{\theta}\ln f(y_{1+s},Z^{s},\theta^{\dagger})\right)\left(\nabla_{\theta}\ln f(y_{1+s+j},Z^{s+j},\theta^{\dagger})\right)'\right)$ and $\Pi=\pi-\pi^{2}/3$ for $P\leq R$ and $\Pi=1-\pi^{2}/3$ for $P>R$. Therefore, the statement in the theorem will follow once we have shown that:

$$Var^{*}\left(\frac{1}{\sqrt{P}}\sum_{t=R}^{T}\left(\widetilde{\theta}^{*}_{t,rol}-\widehat{\theta}_{t,rol}\right)\right)=2\Pi A^{\dagger}C_{00}A^{\dagger},\ \ \Pr-P. \qquad (36)$$

Now,

$$Var^{*}\left(\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\frac{1}{R}\sum_{j=t-R+1}^{t}h^{*}_{j}\right)=Var^{*}\left(\frac{1}{\sqrt{P}R}\sum_{j=s+1}^{P+s}(j-s)h^{*}_{j}\right)$$

$$+Var^{*}\left(\frac{\sqrt{P}}{R}\sum_{j=P+s+1}^{R+s}h^{*}_{j}\right)+Var^{*}\left(\frac{1}{\sqrt{P}R}\sum_{j=R+s+1}^{T-1}\left(P-s-(j-R)\right)h^{*}_{j}\right)+o(1)\ \ \Pr-P \qquad (37)$$

where the $o(1)$ $\Pr-P$ term comes from the fact that the covariance term are $o(1)$ $\Pr-P$, given the independence of the blocks.[21]

$$Var^{*}\left(\frac{1}{\sqrt{P}R}\sum_{j=s+1}^{P}(j-s)h^{*}_{j}\right)=Var^{*}\left(\frac{1}{\sqrt{P}E}\sum_{k=1}^{b_1}\sum_{i=1}^{l}((k-1)l+i)h_{I^{1}_{k}+i}\right)$$

[21] For notational simplicity, we start summation from 1 instead than from s. As s is finite, this has no consequence on the asymptotic behavior.

$$= E^*\left(\frac{1}{\sqrt{P}R}\sum_{k=1}^{b_1}\sum_{i=1}^{l}\sum_{j=1}^{l}((k-1)l+i)((k-1)l+j)(h_{I_k+i}-\overline{h}_P)(h_{I_k+j}-\overline{h}_P)'\right)$$

$$= \frac{1}{P}\frac{1}{R^2}\sum_{k=1}^{b_1}\sum_{i=1}^{l}\sum_{j=1}^{l}((k-1)l+i)((k-1)l+j)E^*\left((h_{I_k+i}-\overline{h}_T)(h_{I_k+j}-\overline{h}_T)'\right)$$

$$= \frac{1}{P}\frac{1}{R^2}\sum_{k=1}^{b_1}\sum_{i=1}^{l}\sum_{j=1}^{l}((k-1)l+i)((k-1)l+j)\left(\frac{1}{P}\sum_{t=l}^{P-l}(h_{t+i}-\overline{h}_T)(h_{t+j}-\overline{h}_T)'\right)+O(l/P^{1/2})\ \Pr-P$$

$$\begin{aligned}
&= \frac{1}{P}\frac{1}{R^2}\sum_{k=1}^{b_1}\sum_{i=1}^{l}\sum_{j=1}^{l}((k-1)l+i)((k-1)l+j)\gamma_{|i-j|}\\
&+\frac{1}{P}\frac{1}{R^2}\sum_{k=1}^{b_1}\sum_{i=1}^{l}\sum_{j=1}^{l}((k-1)l+i)((k-1)l+j)\left(\frac{1}{P}\sum_{t=l}^{P-l}\left((h_{t+i}-\overline{h}_T)(h_{t+j}-\overline{h}_T)'-\gamma_{i-j}\right)\right)\\
&+O(l/P^{1/2})\ \Pr-P
\end{aligned}\tag{38}$$

We need to show that the last term on the last equality in (38) is $o(1)$ $\Pr-P$. First, as for all k,i,j $\frac{((k-1)l+i)((k-1)l+j)}{R^2}\leq 1$, it is majorized by

$$\begin{aligned}
&\left|\frac{b_1}{P}\sum_{i=1}^{l}\sum_{j=1}^{l}\left(\frac{1}{P}\sum_{t=l}^{P-l}\left((h_{t+i}-\overline{h}_T)(h_{t+j}-\overline{h}_T)'-\gamma_{i-j}\right)\right)\right|\\
=&\left|\frac{1}{P}\sum_{t=l}^{P-l}\sum_{j=-l}^{l}\left((h_t-\overline{h}_T)(h_{t+j}-\overline{h}_T)'-\gamma_j\right)\right|+O(l/P^{1/2})\ \Pr-P
\end{aligned}\tag{39}$$

The first term on the RHS of (39) goes to zero in probability, by the same argument as in Lemma 2 in Corradi (1999)[22]. For the first term on the RHS of the last equality in (38), note that

$$\frac{1}{P}\frac{1}{R^2}\sum_{k=1}^{b_2}\sum_{i=1}^{l}\sum_{j=1}^{l}((k-1)l+i)((k-1)l+j)\gamma_{|i-j|}=\frac{1}{P}\sum_{t=l}^{P-l}\sum_{j=-l}^{l}t(t+j)\gamma_j+O(l/P^{1/2})\ \Pr-P$$

$$=\frac{1}{P}\frac{1}{R^2}\sum_{t=l}^{P-l}t^2\sum_{j=-l}^{l}\gamma_j+\frac{1}{P}\sum_{t=l}^{P-l}\sum_{j=-l}^{l}(t(t+j)-t^2)\gamma_j+O(l/P^{1/2})\ \Pr-P$$

By the same argument as in Lemma 4.1 in West and McCracken (1998), the second term on the RHS above approaches zero, while

$$\frac{1}{P}\sum_{t=l}^{P-l}t^2\sum_{j=-l}^{l}\gamma_j\to\frac{\pi^2}{3}C_{00}.$$

[22]The domination condition here are weaker than those in Lemma 2 in Corradi (1999) as we require only convergence to zero in probability and not almot surely.

By a similar argument, and following the proof of Lemma 4.1 in West and McCracken (1998), it can be shown that

$$Var^*\left(\frac{\sqrt{P}}{R}\sum_{j=P+s+1}^{R+s} h_j^*\right) = (\pi-\pi^2)C_{00}+o_P(1)$$

$$Var^*\left(\frac{1}{\sqrt{PR}}\sum_{j=R+s+1}^{T-1}(P-s-(j-R))\,h_j^*\right) = \frac{\pi^2}{3}C_{00}+o_P(1).$$

Finally, the case of $P>R$ can be treated along the same lines.

Proof of Proposition 3: We consider only the case of $P \leq R$. The fact that $\frac{1}{P^{1/2}}\sum_{t=R}^{T-1}\left(\widetilde{\beta}^*_{t,rol}-\widehat{\beta}_{t,rol}\right)$ has the same limiting distribution as in Proposition 2, follows by exactly the same arguments in the proof of Proposition 2 above. Now,

$$\widetilde{\sigma}^{2,*}_{i,t,R}-\widehat{\sigma}^2_{i,t,R}=\frac{1}{R}\sum_{j=t-R+1}^{t}\left((y_j^*-g_i(Z^{*,j-1},\widehat{\beta}_{i,t,rol}))^2-\frac{1}{T}\sum_{k=s+1}^{T-1}(y_j-g_i(Z^{*,j-1},\widehat{\beta}_{i,t,rol}))^2\right), \tag{40}$$

and so, letting $\widehat{\epsilon}_j^{*2}=(y_j^*-g_i(Z^{*,j-1},\widehat{\beta}_{i,t,rol}))^2$, $\widehat{\overline{\epsilon}}^2_T=\frac{1}{T}\sum_{k=s+1}^{T-1}(y_j-g_i(Z^{*,j-1},\widehat{\beta}_{i,t,rol}))^2$, $\epsilon_j^{*2}=(y_j^*-g_i(Z^{*,j-1},\beta^\dagger))^2$ and $\overline{\epsilon}^2_T=\frac{1}{T}\sum_{k=s+1}^{T-1}(y_j-g_i(Z^{j-1},\beta^\dagger))^2$, for $P\leq R$,

$$\frac{1}{\sqrt{P}}\sum_{t=R}^{T}\left(\widetilde{\sigma}^{2,*}_{i,t,R}-\widehat{\sigma}^2_{i,t,R}\right)=\frac{1}{\sqrt{PR}}\sum_{t=R+s}^{T-1}\sum_{j=t-R+1}^{t}\left(\widehat{\epsilon}_j^{*2}-\widehat{\overline{\epsilon}}^2_T\right)=\frac{1}{\sqrt{PR}}\sum_{j=s+1}^{P+s}(j-s)\left(\widehat{\epsilon}_j^{*2}-\widehat{\overline{\epsilon}}^2_T\right)$$

$$+\frac{\sqrt{P}}{R}\sum_{j=P+s+1}^{R+s}\left(\widehat{\epsilon}_j^{*2}-\widehat{\overline{\epsilon}}^2_T\right)+\frac{1}{\sqrt{PR}}\sum_{j=R+s+1}^{T-1}(P+s-(j-R))\left(\widehat{\epsilon}_j^{*2}-\widehat{\overline{\epsilon}}^2_T\right).$$

$$=\frac{1}{\sqrt{PR}}\sum_{j=s+1}^{P+s}(j-s)\left(\epsilon_j^{*2}-\overline{\epsilon}^2_T\right)$$

$$+\frac{\sqrt{P}}{R}\sum_{j=P+s+1}^{R+s}\left(\epsilon_j^{*2}-\overline{\epsilon}^2_T\right)+\frac{1}{\sqrt{PR}}\sum_{j=R+s+1}^{T-1}(P+s-(j-R))\left(\epsilon_j^{*2}-\overline{\epsilon}^2_T\right)+O_P\left(\frac{l}{P^{1/2}}\right).$$

The statement then follows by the same argument used in the proof of Proposition 2.

Proof of Proposition 4: This proof follows from Theorem 1 in Corradi and Swanson (2004a).

Proof of Proposition 5: This proof follows using arguments similar to those used in the proof of Proposition 3.

Proof of Proposition 6: The proof to this proposition follows as a straightforward modification of Theorem 3 in Corradi and Swanson (2004b).

Proof of Proposition 7: The proof follows from Theorem 3 in Corradi and Swanson (2004a).

7 References

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

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

Andrews, D.W.K., (2002b), The Block-Block Bootstrap: Improved Asymptotic Refinements, Manuscript, Cowles Foundation in Economics, Yale University.

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.

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.

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

Bontemps, C., and N. Meddahi, (2004), Testing Distributional Assumptions: a GMM Approach, Working Paper, University of Montreal.

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

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

Christoffersen, P.F., J. Hahn and A. Inoue, (2001), Testing and Comparing Value-at-Risk Measures, *Journal of Empirical Finance,* 8, 325-342.

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), Forecast Based Model Selection in the Presence of Structural Breaks, *Journal of Econometrics,* forthcoming.

Corradi, V., (1999), Deciding between $I(0)$ and $I(1)$ via FLIL-Based Bounds, *Econometric Theory,* 15, 643-663.

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, (2004a), Bootstrap Procedures for Recursive Estimation Schemes With Applications to Forecast Model Selection Working Paper, Queen Mary, University of London and Rutgers University.

Corradi, V., N.R. Swanson, (2004b), A Test for Comparing Multiple Misspecified Conditional Interval Models, Working Paper, Queen Mary, University of London and Rutgers University.

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.

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

Eberlain, E., (1986), On Strong Invariance Principles under Dependence Assumptions, *Annals of Probability,* 14, 260-270.

Fernandez-Villaverde, J., and J.F. Rubio-Ramirez, (2001), Comparing Dynamic Equilibrium Models to Data, Working Paper, University of Pennsylvania.

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

Goncalves, S., and H. White, (2003), Maximum Likelihood and the Bootstrap for Nonlinear Dynamic Models, *Journal of Econometrics,* forthcoming.

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

Hansen, P.R., (2001), An Unbiased and Powerful Test for Superior Predictive Ability, Working Paper, Brown 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, (2003) Nonparametric Specification Testing for Continuous Time Models with Applications to Term Structure Interest Rates, *Review of Financial Studies,* forthcoming.

Inoue, A., and M. Shintani, (2003), Bootstrapping GMM Estimators for Time Series, Mimeo, *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.

Linton, O., E. Maasoumi and Y.J. Whang, (2003), Consistent Testing for Stochastic Dominance Under General Sampling Schemes, Manuscript, LSE, Southern Methodist University and Ewha University.

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.

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.

Pesaran M. H., and A. Timmerman, (2004), How Costly is to Ignore Breaks when Forecasting the Direction of a Time Series? *International Journal of Forecasting,* forthcoming.

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

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

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

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.

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

Wooldridge, J.M. and H. White, (1988), Some Invariance Principles and Central Limit Theorems for Dependent and Heterogeneous Processes, *Econometric Theory*, 4, 210-230.

Table 1: Comparison of Autoregressive Inflation Models with and Without Unemployment Using SIC$^{(*)}$

	Model 1 - Normal	Model 2 - Normal	Model 3 - Student's t	Model 4 - Student's t
Specification	AR	ARX	AR	ARX
lag Selection	SIC (1)	SIC (1,1)	SIC (1)	SIC (1,1)
$MSFE$	0.00083352	0.00004763	0.00083352	0.00004763
$DMSFE$	1.80129635	2.01137942	1.84758927	1.93272971
$Z_{P,u,2}(1,k)$	benchmark	-0.21008307	-0.04629293	-0.13143336

Critical Values

	Bootstrap with Adjustment					Bootstrap without Adjustment				
Percentile	3	5	10	15	20	3	5	10	15	20
50	0.094576	0.095575	0.097357	0.104290	0.105869	0.059537	0.062459	0.067246	0.073737	0.079522
60	0.114777	0.117225	0.128311	0.134509	0.140876	0.081460	0.084932	0.097435	0.105071	0.113710
70	0.142498	0.146211	0.169168	0.179724	0.200145	0.110945	0.110945	0.130786	0.145153	0.156861
80	0.178584	0.193576	0.221591	0.244199	0.260359	0.141543	0.146881	0.185892	0.192494	0.218076
90	0.216998	0.251787	0.307671	0.328763	0.383923	0.186430	0.196849	0.254943	0.271913	0.312400

$^{(*)}$ Notes: Entires in the table are given in two parts (i) summary statistics, and (ii) bootstrap percentiles. In (i): "specification" lists the model used. For each specification, lags may be chosen either with the SIC or the AIC, and the predictive density may be either Gaussian or Student's t, as denoted in the various columns of the table. The bracketed entires beside SIC and AIC denote the number of lags chosen for the autoregressive part of the model and the number of lags of unemployment used, respectively. $MSFE$ is the out-of-sample mean square forecast error based on evaluation of P=300 1-step ahead predictions using recursively estimated models, and $DMSFE = \int_U \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(1\{Inf_{t+1} \leq u\} - F_1(u|Z^t, \widehat{\theta}_{1,t})\right)^2 \phi(u)du$, where $R = 300$, corresponding to the sample period from 1954:1-1978:12, is our analogous density based square error loss measure. Finally, $Z_{P,u,2}(1,k)$ is the accuracy test statistic, for each benchmark/alternative model comparison. The density accuracy test is the maximum across the $Z_{P,u,2}(1,k)$ values. In (ii) percentiles of the bootstrap empirical distributions under different block length sampling regimes are given. The "Bootstrap with Adjustment" allows for parameter estimation error, while the "Bootstrap without Adjustment" assumes that parameter estimation error vanishes asymptotically. Testing is carried out using 90th percentiles (see above for further details).

Table 2: Comparison of Autoregressive Inflation Models with and Without Unemployment Using AIC$^{(*)}$

	Model 1 - Normal	Model 2 - Normal	Model 3 - Student's t	Model 4 - Student's t
Specification	AR	ARX	AR	ARX
lag Selection	AIC (3)	AIC (3,1)	AIC (3)	AIC (3,1)
$MSFE$	0.00000841	0.00000865	0.00000841	0.00000865
$DMSFE$	2.17718449	2.17189485	2.11242940	2.10813786
$Z_{P,u,2}(1,k)$	benchmark	0.00528965	0.06475509	0.06904664

Critical Values

	Bootstrap with Adjustment					Bootstrap without Adjustment				
Percentile	3	5	10	15	20	3	5	10	15	20
50	-0.004056	-0.003820	-0.003739	-0.003757	-0.003722	-0.004542	-0.004448	-0.004316	-0.004318	-0.004274
60	-0.003608	-0.003358	-0.003264	-0.003343	-0.003269	-0.004318	-0.003999	-0.003911	-0.003974	-0.003943
70	-0.003220	-0.002737	-0.002467	-0.002586	-0.002342	-0.003830	-0.003384	-0.003287	-0.003393	-0.003339
80	-0.002662	-0.001339	-0.001015	-0.001044	-0.000321	-0.003148	-0.001585	-0.001226	-0.001340	-0.000783
90	-0.000780	0.001526	0.002828	0.002794	0.003600	-0.000925	0.001371	0.002737	0.002631	0.003422

$^{(*)}$ Notes: See notes to Table 1.

Table 3: Comparison of Autoregressive Inflation Models Using SIC and AIC$^{(*)}$

	Model 1 - Normal	Model 2 - Normal	Model 3 - Student's t	Model 4 - Student's t
Specification	AR	AR	AR	AR
lag Selection	SIC (1)	AIC (3)	SIC (1)	AIC (3)
$MSFE$	0.00083352	0.00000841	0.00083352	0.00000841
$DMSFE$	1.80129635	2.17718449	1.84758927	2.11242940
$Z_{P,u,2}(1,k)$	benchmark	-0.37588815	-0.04629293	-0.31113305

Critical Values

	Bootstrap with Adjustment					Bootstrap without Adjustment				
Percentile	3	5	10	15	20	3	5	10	15	20
50	0.099733	0.104210	0.111312	0.114336	0.112498	0.063302	0.069143	0.078329	0.092758	0.096471
60	0.132297	0.147051	0.163309	0.169943	0.172510	0.099277	0.109922	0.121311	0.132211	0.135370
70	0.177991	0.193313	0.202000	0.217180	0.219814	0.133178	0.150112	0.162696	0.177431	0.185820
80	0.209509	0.228377	0.245762	0.279570	0.286277	0.177059	0.189317	0.210808	0.237286	0.244186
90	0.256017	0.294037	0.345221	0.380378	0.387672	0.213491	0.244186	0.280326	0.324281	0.330913

$^{(*)}$ Notes: See notes to Table 1.

Table 4: Comparison of Autoregressive Inflation Models with Unemployment Using SIC and AIC$^{(*)}$

	Model 1 - Normal	Model 2 - Normal	Model 3 - Student's t	Model 4 - Student's t
Specification	ARX	ARX	ARX	ARX
lag Selection	SIC (1,1)	AIC (3,1)	SIC (1,1)	AIC (3,1)
$MSFE$	0.00004763	0.00000865	0.00004763	0.00000865
$DMSFE$	2.01137942	2.17189485	1.93272971	2.10813786
$Z_{P,u,2}(1,k)$	benchmark	-0.16051543	0.07864972	-0.09675844

Critical Values

	Bootstrap with Adjustment					Bootstrap without Adjustment				
Percentile	3	5	10	15	20	3	5	10	15	20
50	0.013914	0.015925	0.016737	0.018229	0.020586	0.007462	0.012167	0.012627	0.014746	0.016022
60	0.019018	0.022448	0.023213	0.024824	0.027218	0.013634	0.016693	0.018245	0.019184	0.022048
70	0.026111	0.028058	0.029292	0.030620	0.033757	0.019749	0.022771	0.023878	0.025605	0.029439
80	0.031457	0.033909	0.038523	0.041290	0.043486	0.025395	0.027832	0.033134	0.034677	0.039756
90	0.039930	0.047533	0.052668	0.054634	0.060586	0.035334	0.042551	0.046784	0.049698	0.056309

$^{(*)}$ Notes: See notes to Table 1.