Testing Parametric Conditional Distributions of Dynamic Models[1]

Jushan Bai
Boston College

June 10, 2002

[1]I thank two anonymous referees and an editor for their valuable comments which help clarify many issues and lead to a much better presentation. All errors are my own responsibility. Partial financial support is acknowledged from the National Science Foundation under Grant SBR-9709508. Please address correspondence to Jushan Bai, Department of Economics, Boston College, Chestnut Hill, MA 02467, baij@bc.edu

Abstract

This paper proposes a nonparametric test for parametric conditional distributions of dynamic models. The test is of the Kolmogorov type coupled with Khmaladze's martingale transformation. The proposed test is asymptotically distribution free and has non-trivial power against root-n local alternatives. The method is applicable for various dynamic models, including autoregressive and moving average models (ARMA), generalized autoregressive conditional heteroskedasticity (GARCH), integrated GARCH (IGARCH), and general nonlinear time series regressions. The method is also applicable for cross sectional models. Finally, we apply the procedure to testing conditional normality and conditional t-distribution in a GARCH model for the NYSE equal-weighted returns.

Key Words and Phrases: Empirical process, martingale transformation, continuous-time recursive least squares, continuous-time detrending, Brownian motion.

1 Introduction

The study of probability distributions of economic variables is an important subject and has a long history, for example, the study of income distribution by Pareto (1897) and that of wealth distribution by Sargan (1957). In financial economics, the distributions of assets' returns have been extensively examined, e.g., Fama (1967). In risk management, the distribution of a portfolio's value is closely monitored by asset managers. Often undertaken in econometrics is testing distributional assumptions, with a usual focus on normality, as in Bera and Jarque (1982). This paper studies the problem of testing conditional distributions of dynamic models, where distributions evolve over time. Though not a focus of this paper, dynamic conditional distributions is related to density forecasting, which is a major concern in risk management. For further elaboration on this topic, see Diebold, Gunther and Tay (1998).

A conventional procedure for testing distributional assumptions is that of the Kolmogorov test. However the test is designed for independent and identically distributed (iid) observations with a completely specified null distribution. In the present context, the data are dependent and the null hypothesis does not completely specify the distribution of the data because of the presence of unknown parameters and unspecified distributions for the conditioning variables. As a result, the joint distribution of observations is not uniquely determined under the null. Furthermore, it is well known that when parameters are estimated, the Kolmogorov test is not asymptotically distribution free, see Durbin (1973a). This means that different critical values are needed for different distributions and for different parameter values. These problems are further compounded by the fact that the critical values are difficult to compute because the limiting distribution of the Kolmogorov test is a complicated function of the underlying true distribution and the true parameter. The objective of this paper is to develop test statistics that can overcome all these difficulties.

Suppose that a sequence of observations $(Y_1, X_1), (Y_2, X_2), ..., (Y_n, X_n)$ is given. Let $\Omega_t = \{X_t, X_{t-1}, ..., ; Y_{t-1}, Y_{t-2}, ...\}$ represent the information set at time t (not including Y_t). We are interested in the conditional distribution of Y_t conditional on the information set Ω_t. More specifically, of central interest is the following null hypothesis:

H_0 : The conditional distribution of Y_t conditional on Ω_t is in the parametric family $F_t(y|\Omega_t, \theta)$ for some $\theta \in \Theta$, where Θ is the parameter space.

Note that the conditional distribution under the null hypothesis allows for an infinite past history of information. For example, consider an MA(1) process: $Y_t = \varepsilon_t + \theta\varepsilon_{t-1}$ with $|\theta| < 1$,

where the ε_t are iid with cumulative distribution function (cdf) F_ε. Then the conditional cdf of $Y_t|\Omega_t$ is given by $F_\varepsilon(y - \sum_{k=1}^{\infty}(-1)^k\theta^k Y_{t-k})$, with $\Omega_t = \{Y_{t-1}, Y_{t-2}, ...\}$. Moreover, the conditional distribution not only varies with the information set but also evolves over time. This possible evolution is highlighted by the subscript t in $F_t(y|\Omega_t, \theta)$. The objective is to formulate test statistics for H_0 under this general setup.

The proposed test is of the Kolmogorov type, coupled with a martingale transformation of Khmaladze (1981). It will be shown that the proposed test has non-trivial local power against root-n local alternatives. In addition, the test is asymptotically distribution free and critical values are easy to obtain. Therefore, no simulation or bootstrap will be required to perform the test procedure. We apply the method to the NYSE equal-weighted real returns modelled as a GARCH process. While conditional normality is strongly rejected, conditional t-distribution cannot be rejected. One implication of this result is that the observed "heavy-tailedness" is not entirely induced by the conditional heteroskedasticity, but the conditional distribution itself has heavy tails relative to the normal distribution.

The rest of this paper is organized as follows. Section 2 provides a literature review and states our contributions. Test statistics are described in Section 3 along with martingale transformation for several concrete examples. An empirical application is also given in this section. Section 4 establishes the weak convergence results that are prerequisite for the martingale transformations. Also considered in this section are the estimation of unknown transforming functions, analysis of local power, and consistency of the test. Some limited Monte Carlo simulations are also presented in Section 4. The conclusion is provided in the last section. An introduction to martingale transformation, its computational issue, and the theoretical proofs are given in the appendix.

2 Related literature and contributions of this paper

The Kolmogorov test is formulated for iid observations and for a simple hypothesis (i.e., a completely specified distribution). In an influential paper, Durbin (1973a) considers testing for a composite hypothesis, where the distribution function depends on an unknown vector of parameters. One unpleasant feature of the K-test is that when parameters are estimated, it is no longer asymptotically distribution free. As a result, critical values change from one null hypothesis to another. Different parameter values also need different critical values, even within the same parametric family of distributions. Several approaches have been suggested in

the literature to circumvent this problem. One is the half-sample method, where parameters are estimated with a randomly chosen half sample, see Durbin (1973b). Another approach is to randomize the estimated parameters, see Loynes (1980). These approaches do not work satisfactorily. In addition, these methods do not apply to time series data.

Recently, Andrews (1997) proposes a conditional Kolmogorov test. Andrews proves the consistency of his test and justifies the use of the bootstrap method to obtain critical values. The conditional Kolmogorov test overcomes a number of difficulties associated with the Kolmogorov test. However, Andrews' test is not designed for dynamic models. In addition, the dimension of the conditioning variables is fixed and finite.

Zheng (2000) provides a test based on Kullback-Leibler information criterion together with kernel estimation of the underlying distributions. Zheng's tests are consistent against all departures from the null. The test has local power against alternatives that converge to zero slower than root-n. Fan (1994) provides a test for parametric density function using the kernel method. Stinchcombe and White (1998) also provide nonparametric tests for conditional distributions and established consistency of their test. All these papers deal with iid observations.

Inoue (1997) proposes a test statistic for testing a number of econometric problems related to conditional distributions in time series. He suggests to use an upper bound derived from the law of the iterated logarithms to obtain critical values. Diebold, Gunther, and Tay (1998) propose a framework for evaluating density forecasts, and discuss the Kolmogorov test for conditional distributions in time series, among other issues. They do not consider the effect of parameter estimation. Linton and Gozalo (1996) use the Kolmogorov type test for conditional independence of iid observations.

In this paper, we use Khmaladze's transformation approach to derive an asymptoticlly distribution-free test. In doing so, this transformation itself is also extended in some important ways.

Khmaladze's transformation has proven useful for various problems. Koul and Stute (1999) apply the transformation to marked empirical processes for AR(1) models, either linear or nonlinear. Their focus is the specification of the conditional mean, rather than the conditional distribution. Incidentally, extension of the transformation to AR(2) or multiple-regressor marked-empirical processes is non-trivial because, among other technical difficulties, Khmaladze's transformation is not unique for multivariate empirical processes, see Khmaladze (1988, 1993). A marked empirical process for high dimensional models is a multivariate process (in-

dexed by a vector). Khmaladze's approach is also used for hazard function specification test, e.g., Andersen et al. (1993). Bai and Ng (1998) construct a consistent test for conditional symmetry with the transformation method.

We make several contributions in this paper. First, we consider conditional distributions of dynamic models which, of course, include iid observations as special cases. For dynamic models, the conditioning event may depend on the entire history of the data (generally unobservable). Information truncation is required for these situations. Second, we obtain various weak convergence results for empirical processes of dynamic models under parameter estimation and information truncation. The weak convergence result for GARCH and IGARCH is particularly interesting. We also obtain weak convergence for the transformed process under the supremum norm, which forms the basis for asymptotically distribution-free tests. Third, Khmaladze's transformation requires the knowledge of a set of transforming functions (denoted by g, see Section 3). We extend this transformation to the estimated g, and under very general conditions we prove weak convergence. In particular, we do not need any rate of convergence for the estimated g. Fourth, we find that the dimension of transforming functions is not necessarily equal to the number of freely varying parameters. For example, dimension reduction can be achieved in location-scaled models (e.g., GARCH), resulting in a very simple transformation. Fifth, we explore the consistency of the test resulting from the transformation in general, and we further establish the consistency of the test for GARCH or any location-scale model in particular. Finally, an empirical application further demonstrates the usefulness of the proposed method.

For some problems, the conditional distribution $F_t(\cdot|\Omega_t, \theta)$, is not specified and instead a data generating process (DGP) is given, e.g., continuous-time finance models. Often, the implied conditional distribution is difficult to derive from the DGP. Given a set of data, one can test whether the data come from the hypothesized data generating process using the procedure of this paper. This is because one can simulate a large number of observations from the given DGP so that the underlying distribution implied by the DGP can be estimated up to any precision. The estimated distribution can be used to construct test statistics. Thompson (2000) applies a similar method to continuous-time finance models.

3 Test statistics

3.1 Description of the method. Assuming that the null hypothesis is true and that the

true parameter value θ_0 is known, then, using integral transformation, we can transform the dependent data into an iid sequence of uniformly-distributed random variables. That is, $U_t = F_t(Y_t|\Omega_t, \theta_0)$ are iid and uniform random variables.

The random variables U_t are unobservable since θ_0 is unknown. When an estimator $\hat{\theta}$ of θ_0 is available, we may use $\hat{U}_t = F_t(Y_t|\Omega_t, \hat{\theta})$ as an estimate for U_t. The random variables $\hat{U}_t$ are neither independent nor identically distributed. Furthermore, the unavailability of an infinite history of observations necessitates a truncation of the information sets. Let $\tilde{\Omega}_t = \{X_t, X_{t-1}, ..., X_1, 0, 0, ..., Y_{t-1}, ..., Y_1, 0, 0, ...\}$ represent a truncated (observable) version of Ω_t. Define

$$\hat{U}_t = F_t(Y_t|\tilde{\Omega}_t, \hat{\theta}).$$

For example, in the case of an MA(1) process, $Y_t = \varepsilon_t + \theta\varepsilon_{t-1}$ with ε_t being iid F_ϵ, $\hat{U}_t = F_\epsilon(Y_t - \sum_{k=1}^{t-1}(-\hat{\theta})^k Y_{t-k})$, whereas $U_t = F_\epsilon(Y_t - \sum_{k=1}^{\infty}(-\theta)^k Y_{t-k})$.

Let $\hat{V}_n(r)$ be the empirical process based on $\hat{U}_1, ..., \hat{U}_n$. That is,

$$\hat{V}_n(r) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} [I(\hat{U}_t \leq r) - r]. \tag{1}$$

Under some regularity conditions to be introduced later, it can be shown that $\hat{V}_n(r)$ has the following asymptotic representation

$$\hat{V}_n(r) = V_n(r) - \bar{g}(r)'\sqrt{n}(\hat{\theta} - \theta_0) + o_p(1), \tag{2}$$

where

$$V_n(r) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} [I(U_t \leq r) - r] \tag{3}$$

$$\bar{g}(r) = \text{plim} \frac{1}{n} \sum_{t=1}^{n} \frac{\partial F_t}{\partial \theta}(x|\Omega_t, \theta_0)\Big|_{x=F_t^{-1}(r|\Omega_t, \theta_0)} \tag{4}$$

Due to the presence of the second term in the right-hand-side of (2), the limiting process of $\hat{V}_n(r)$ generally depends on F_t as well as on θ_0.[1] As a result, the Kolmogorov test based on $\hat{V}_n(r)$ will not be asymptotically distribution free and critical values are difficult to obtain. However, applying the martingale transformation discussed in Appendix A, we can remove the term $\bar{g}(r)'\sqrt{n}(\theta - \theta_0)$. The transformed process will have a Brownian motion as its limit.

[1]The exact limiting process will also depend on how the parameter θ_0 is estimated. If $\hat{\theta}_n$ is asymptotically normal then $\hat{V}_n$ will generally have a limiting Gaussian process. In this paper, we only need root-n consistency of $\hat{\theta}_n$.

Specifically, let $g(r) = (r, \bar{g}(r)')'$ and $\dot{g}(r) = (1, \dot{\bar{g}}(r)')'$ so that $\dot{g}$ is the derivative of g. Let $C(r) = \int_r^1 \dot{g}(\tau)\dot{g}(\tau)'d\tau$. Consider the transformation

$$\hat{W}_n(r) = \hat{V}_n(r) - \int_0^r \left[\dot{g}(s)'C^{-1}(s)\int_s^1 \dot{g}(\tau)d\hat{V}_n(\tau)\right]ds. \tag{5}$$

It can be shown that $\hat{W}_n(r)$ converges weakly to a standard Brownian motion. Define the test statistic as

$$T_n = \sup_{0\le r\le 1} |\hat{W}_n(r)|$$

then by the continuous mapping theorem,

$$T_n \xrightarrow{d} \max_{0\le t\le 1} |W(r)|$$

where W is a standard Brownian motion process. It is easy to simulate the distributions of the right hand side variable. This only needs to be done once. The critical values at 10%, 5%, and 1% are found to be 1.94, 2.22, and 2.80, respectively, and are obtained via simulation. Each sample path of $W(r)$ is approximated by normalized partial sums of 1,000 iid N(0,1) variables. Then 100,000 sample paths are simulated and the maximum values are obtained for each path. These extreme values are ordered to obtain the quantiles.

The test statistic T_n can be easily computed, see Appendix B for details.

Martingale transformation is in effect a continuous-time detrending operation, where the trend function is $g(r) = (r, \bar{g}(r)')'$. To see this, write (2) in the differentiation form

$$d\hat{V}_n(r) = dV_n(r) - \dot{\bar{g}}(r)'dr\sqrt{n}(\hat{\theta} - \theta_0) + o_p(1)$$

Consider regressing $d\hat{V}_n(r)$ on $\dot{g}(r)$ over the interval $(s, 1]$. Then the least square estimator is given by $(\int_s^1 \dot{g}\dot{g}'dr)^{-1}\int_s^1 \dot{g}d\hat{V}_n = C(s)^{-1}\int_s^1 \dot{g}(\tau)d\hat{V}_n(\tau)$. Multiplying this estimator by $\dot{g}(s)ds$ gives the predicted value of $d\hat{V}_n(s)$. Thus the residual (detrended value) is given by

$$d\hat{V}_n(s) - [\dot{g}(s)'C^{-1}(s)\int_s^1 \dot{g}(\tau)d\hat{V}_n(\tau)]ds. \tag{6}$$

Then integrating from 0 to r gives rise to $\hat{W}_n(r)$. The above is a recursive residual. This is so because at each point s, a least squares is performed. The procedure is analogous to the discrete time recursive residuals of Brown, Durbin, and Evans (1975). As in the discrete time framework, recursive residuals are non-correlated (martingale differences) and a cumulative sum of recursive residuals leads to a Brownian motion. Here a cumulative sum of (6) (i.e., integration from [0, r]) yields a Brownian motion process.

3.2 Examples. In the following we give several concrete examples on testing distributional assumptions and on the construction of martingale transformation.

Example 1: GARCH(1, 1). A GARCH(1,1) takes the form, see Bollerslev (1986),

$$Y_t = X_t'\delta + \varepsilon_t\sigma_t,$$

$$\sigma_t^2 = \alpha + \beta\sigma_{t-1}^2 + \gamma(Y_{t-1} - X_{t-1}'\delta)^2,$$

where ε_t is iid with zero mean and unit variance. The parameters are assumed to satisfy

$$\alpha > 0, \beta \geq 0, \gamma \geq 0, \quad \text{and } \beta + \gamma \leq 1.$$

For IGARCH models, i.e., $\beta+\gamma = 1$, we assume $0 < \beta < 1$, as in Lee and Hansen (1994). The objective is to test the null hypothesis that the distribution function of ε_t is F. Commonly considered cases are the normal and t-distributions. Under the null hypothesis, the conditional distribution of Y_t conditional on Ω_t is

$$Y_t|\Omega_t \sim F\Big((y - X_t'\delta)/\sigma_t\Big).$$

Compute the conditional variance via the recursion (starting with a given initial value $\hat{\sigma}_0^2$):

$$\hat{\sigma}_t^2 = \hat{\alpha} + \hat{\beta}\hat{\sigma}_{t-1}^2 + \hat{\gamma}(Y_{t-1} - X_{t-1}'\hat{\delta})^2$$

and define

$$\hat{\varepsilon}_t = (Y_t - X_t'\hat{\delta})/\hat{\sigma}_t, \qquad \text{and } \hat{U}_t = F(\hat{\varepsilon}_t).$$

Let $\hat{V}_n(r)$ be the empirical process based on $\hat{U}_t$. It is shown in Section 4 that

$$\hat{V}_n(r) = V_n(r) + f(F^{-1}(r))\, p_n + f(F^{-1}(r))F^{-1}(r)\, q_n + o_p(1), \tag{7}$$

where f is the density and F^{-1} is the inverse of F, p_n and q_n are complicated functions of data and parameters. Therefore, the limiting process of $\hat{V}_n$ is rather complicated and a direct Kolmogorov test is difficult to implement. However, martingale transformation is easy to construct. Let

$$g(r) = (g_1, g_2, g_3)' = (r,\ f(F^{-1}(r)),\ f(F^{-1}(r))F^{-1}(r))' \tag{8}$$

Therefore, $\dot{g}_1 = 1$, $\dot{g}_2 = \dot{f}(F^{-1}(r))/f(F^{-1}(r))$ and $\dot{g}_3 = 1+\dot{g}_2(r)F^{-1}(r)$. For testing normality, then

$$\dot{g}(r) = (1,\ -\Phi^{-1}(r),\ 1 - \Phi^{-1}(r)^2)'$$

where $\Phi(r)$ is the cdf a standard normal random variable. Given $\dot{g}$, the transformation of $\hat{V}_n$ is performed using formula (5).

Remark 1: For general GARCH(p,q) processes, the transformation is identical to that of GARCH(1,1) because the g function is identical. This is because the corresponding empirical process $\hat{V}_n$ has the same asymptotic representation except that p_n and q_n have different expressions. But p_n and q_n are not functions of r and thus will not affect the transformation.

Example 2: ARMA(p,q) process. Consider a stationary and invertible ARMA(p,q) process such that

$$Y_t = \mu + \rho_1 Y_{t-1} + \cdots \rho_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots \theta_q \varepsilon_{t-q}$$

Consider testing the hypothesis that ε_t are iid $F(\cdot/\sigma)$. Let $\theta = (\mu, \rho_1, ..., \rho_p, \theta_1, ..., \theta_q, \sigma)$ and $\hat{\theta}$ be root-n consistent estimator of θ. Given $n+p$ observations $Y_{-p+1}, Y_{-p+2}, ..., Y_0, Y_1, ..., Y_n$, the residuals can be computed via the recursion

$$\hat{\varepsilon}_t = Y_t - \hat{\mu} - \hat{\rho}_1 Y_{t-1} - \cdots - \hat{\rho}_p Y_{t-p} - \hat{\theta}_1 \hat{\varepsilon}_{t-1} - \cdots - \hat{\theta}_q \hat{\varepsilon}_{t-q} \quad (t = 1, 2, ..., n)$$

the initial value of $(\hat{\varepsilon}_0, ..., \hat{\varepsilon}_{1-q})$ is set to zero. Define $\hat{U}_t = F(\hat{\varepsilon}_t/\hat{\sigma})$ and $\hat{V}_n$ as in (1). Then it can be shown that representation (7) is still valid but with different expressions for p_n and q_n. Thus the transformation takes exactly the same form as in GARCH(1,1).

Example 3: nonlinear time series regression. Consider the general nonlinear time series regressions:

$$Y_t = h(\Omega_t, \beta) + \varepsilon_t \tag{9}$$

where $\Omega_t = (X_t, X_{t-1}, ...; Y_{t-1}, Y_{t-2}, ...)$. For linear regressions, Bera and Jarque (1982) consider testing normality of ε_t based on skewness and kurtosis coefficients. It is assumed that ε_t are iid with zero mean and is independent of Ω_t. Consider testing the hypothesis that ε_t has a cdf $F(x, \lambda)$ with density function $f(x, \lambda)$, and $\lambda \in R^d$ is vector of unknown parameters. Then the conditional cdf of Y_t is $F(y - h(\Omega_t, \beta), \lambda)$. Define

$$\hat{U}_t = F(Y_t - h(\tilde{\Omega}_t, \hat{\beta}), \hat{\lambda}).$$

The estimated residuals are computed from the truncated information set. Again, let $\hat{V}_n$ be as defined in (1) with the new $\hat{U}_t$. Write $f(x)$ for $f(x, \lambda_0)$ and $F(x)$ for $F(x, \lambda_0)$, where λ_0 is the true parameter. Theorem 2 of Section 4 shows that

$$\hat{V}_n(r) = V_n(r) - f(F^{-1}(r))a_n + \frac{\partial F(F^{-1}(r))'}{\partial \lambda} b_n + o_p(1) \tag{10}$$

where a_n and b_n are random variables not depending on r. Thus the g function in this case is

$$g(r) = (r, \quad f(F^{-1}(r)), \ \frac{\partial F(F^{-1}(r))}{\partial \lambda}')'$$

and $\dot{g} = (1, \dot{f}(F^{-1}(r))/f(F^{-1}(r)), \frac{\partial f(F^{-1}(r))'}{\partial \lambda}/f(F^{-1}(r)))'$. When λ is a scale parameter such that $F(x, \lambda) = F(x/\lambda)$ (here $\lambda > 0$), simplification can be achieved. Comments concerning this are given following Theorem 2 in Section4.

Remark 2: It can be shown that the method is applicable for threshold autoregressive models (TAR) or self-exciting TAR. Consider, for example, $Y_t = \beta_1 Y_{t-1} + \varepsilon_t$ if $Y_{t-1} \leq c$ and $Y_t = \beta_2 Y_{t-1} + \varepsilon_t$ if $Y_{t-1} > c$. The model can be rewritten as $Y_t = \beta_1 Y_{t-1} I(Y_{t-1} \leq c) + \beta_2 Y_{t-1} I(Y_{t-1} > c) + \varepsilon_t$. If c is known, the TAR model is linear in $\beta = (\beta_1, \beta_2)$, thus covered by our theory. For unknown c, the conditional mean is not a smooth function of c. However, c can be estimated with a convergence rate n, see Chan (1993). The implication is that c can be treated as known. This is because n-consistency implies that only a bounded number (with large probability) of observations are misclassified from one regime to another and the rest are correctly classified (i.e., c known). A fixed number of misclassifications does not affect the limiting results. Finally, ε_t can also be regime-dependent.

3.3 An empirical application. In this section, we apply the test procedure of the previous section to the monthly NYSE equal-weighted returns fitted to a GARCH(1,1) process. The range of the data spans from January, 1926 to December, 1999, as shown in Figure 1.

Testing conditional normality. We estimate the following GARCH (1,1) process to the data

$$Y_t = \mu + \sigma_t \varepsilon_t,$$

with $\sigma_t^2 = \alpha + \beta \sigma_{t-1}^2 + \gamma (Y_{t-1} - \mu)^2$. The Gaussian maximum likelihood method is used to estimate the parameters. After obtaining the parameters, we compute the residuals according to $\hat{\varepsilon}_t = (Y_t - \hat{\mu})/\hat{\sigma}_t$ with $\hat{\sigma}_t^2 = \hat{\alpha} + \hat{\beta}\hat{\sigma}_{t-1}^2 + \hat{\gamma}(Y_{t-1} - \hat{\mu})^2$. We then compute $\hat{U}_t = \Phi(\hat{\varepsilon}_t)$ $(t = 1, ..., n)$ and $\hat{V}_n(r)$. The function $\dot{g}$ is given by

$$\dot{g}(r) = (1, -\Phi^{-1}(r), 1 - \Phi^{-1}(r)^2)'.$$

The transformation $\hat{W}_n(r)$ and the test statistic T_n are computed according to Appendix B.

Both the transformed process $\hat{W}_n(r)$ and untransformed process $\hat{V}_n(r)$ are plotted in Figure 2. The two horizontal lines give the 95% confidence band for a standard Brownian motion process on [0,1]. Since the process $\hat{W}_n(r)$ stays out of the confidence band, conditional normality is rejected. In fact, the critical values of the test procedure at significance levels 10%, 5%,

and 1% are 1.94, 2.22, and 2.80, respectively, and the value of the test statistic is $T_n = 4.08$. Thus conditional normality is rejected even at the 1% significance level.

Testing conditional t-distribution. With the same data set, we test the hypothesis that ε_t has a student-t distribution with $df = 5$, normalized to have a variance of 1. This number of degrees of freedom is close to the values usually found for asset returns fitted to GARCH models. Note that there is no need to re-estimate the model. Assuming that ε_t has a student-t distribution, quasi-Gaussian likelihood estimation still provides root-n consistent estimation for the parameters. See, for example, Lee and Hansen (1994), Lumsdaine (1996) and Newey and Steigerwald (1997).

Let t_ν be a student-t random variable with df=ν and let $q_v(x)$ and $Q_v(x)$ be the density and cdf of t_ν, respectively. Because ε_t is normalized to have a variance of 1, we have $\varepsilon_t \sim c^{-1}t_\nu$ with $c = [\nu/(\nu-2)]^{1/2}$. Thus, the cdf of ε_t under the null hypothesis is $F(x) = Q_v(cx)$ with $f(x) = q_v(cx)c$. Thus we should define $\hat{U}_t$ as $\hat{U}_t = Q_v(c\hat{\varepsilon}_t)$ $(t = 1, ..., n)$ and $\hat{V}_n$ be the empirical process based on $\hat{U}_1, ..., \hat{U}_n$. Using (8) for the given f and F, we obtain $g(r) = (r,\ q_v(Q_v^{-1}(r))c,\ q_v(Q_v^{-1}(r))Q_v^{-1}(r))'$ with a constant c in the second component. Since a constant factor will not affect the transformation (or alternatively, p_n is replaced by cp_n in Theorem 3 of Section 4), we can use the following g:

$$g(r) = (r,\ q_\nu(Q_\nu^{-1}(r)),\ q_\nu(Q_\nu^{-1}(r))Q_\nu^{-1}(r))' \tag{11}$$

This function again has the format of (8). It is easy to derive $\dot{g}$. In fact, denote $\dot{g} = (1,\ \dot{g}_2,\ \dot{g}_3)$. Then $\dot{g}_3 = 1 + \dot{g}_2 Q_\nu^{-1}(r)$. From $dq_\nu(x)/dx = -xq_{\nu+2}((\frac{\nu+2}{\nu})^{1/2}x)$ we have

$$\dot{g}_2 = \frac{-Q_\nu^{-1}(r)q_{\nu+2}([\nu+2)/\nu]^{1/2}Q_\nu^{-1}(r))}{q_\nu(Q_\nu^{-1}(r))}.$$

Given $\dot{g}$, the process $\hat{W}_n(r)$ and the test T_n can be easily obtained.

Figure 3 shows both $\hat{V}_n(r)$ and $\hat{W}_n(r)$. The process $\hat{W}_n(r)$ stays well within the 95% confidence band. In fact, the maximum value of $|\hat{W}_n(r)|$ is equal to 1.605, whereas the critical value is 2.22 at the 95% significance level. Therefore, we do not reject the hypothesis that innovations to the GARCH process have a conditional t-distribution.

4 Theoretical results

This section provides the theoretical basis for the validity of the results in Section 3. In particular, we focus on the asymptotic representations of the empirical processes of conditional

distributions. Throughout, we use " $\Rightarrow$ " to denote the weak convergence in $D[0,b]$ $(b>0)$, the space of cadlag functions endowed with the Skorohod metric, see Pollard (1984).

We start with a lemma given in Diebold, Gunther, and Tay (1998), who noted a similar idea can be traced back to Rosenblatt (1952). We provide a much simpler proof. Let $\mathcal{F}_t$ be a sequence of increasing σ-fields such that Y_t is $\mathcal{F}_t$ measurable. (Alternatively, think about $\mathcal{F}_t$ as the information set at time t, and Y_t is included in this information set.)

Lemma 1 *If the conditional distribution of Y_t conditional on $\mathcal{F}_{t-1}$ has a continuous cdf $F_t(y|\mathcal{F}_{t-1})$. Then the random variables $U_t = F_t(Y_t|\mathcal{F}_{t-1})$ are i.i.d. $U(0,1)$.*

Proof: Since the conditional cdf of Y_t is $F_t(y|\mathcal{F}_{t-1})$, the conditional distribution of $U_t = F_t(Y_t|\mathcal{F}_{t-1})$ (conditional on $\mathcal{F}_{t-1}$) is $U(0,1)$. Because the conditional distribution of U_t does not depend on $\mathcal{F}_{t-1}$, U_t is independent of $\mathcal{F}_{t-1}$. It follows that U_t is independent of U_{t-1} because U_{t-1} is $\mathcal{F}_{t-1}$ measurable (i.e., U_{t-1} is a part of $\mathcal{F}_{t-1}$). The latter is true because $U_{t-1} = F(Y_{t-1}|\mathcal{F}_{t-2})$, Y_{t-1} is $\mathcal{F}_{t-1}$ measurable and $\mathcal{F}_{t-2} \subset \mathcal{F}_{t-1}$. This implies that U_t is independent of $(U_{t-1}, U_{t-2}, ...)$ for all t. This further implies joint independence because the joint density can be written as product of marginal and conditional densities. $\square$.

4.1 General conditional distributions. Let $\Omega_t = \{X_t, X_{t-1}, ..., ; Y_{t-1}, Y_{t-2}, ...\}$ represent the information set at time t (not including Y_t). The hypothesis of interest is that the conditional distribution of Y_t conditional on Ω_t is in the parametric family $F_t(y|\Omega_t, \theta_0)$ for some θ_0 in the parameter space. By Lemma 1, $U_t = F_t(Y_t|\Omega_t, \theta_0)$ is a sequence of iid random variables. Let $\tilde{\Omega}_t = \{X_t, X_{t-1}, ..., X_1, 0, 0, ..., Y_{t-1}, ..., Y_1, 0, 0, ...\}$ represent a truncated version of Ω_t and $\hat{\theta}$ be a root-n consistent estimator of θ_0. Define $\hat{U}_t = F(Y_t|\tilde{\Omega}_t, \hat{\theta})$ and

$$\hat{V}_n = n^{-1/2} \sum_{t=1}^{n} [I(\hat{U}_t \leq r) - r]$$

To obtain the limiting process of $\hat{V}_n(r)$, we need to state the underlying assumptions. As a matter of notation, $F_t(y|\Omega_t, \theta)$ and $F_t(y|\theta)$ will be used interchangeably when no information truncation is present. Throughout, let $N(\theta_0, M) = \{\theta; |\theta - \theta_0| \leq Mn^{-1/2}\}$. We assume:

A1: The cdf $F_t(y|\Omega_t, \theta)$ and its density function $f_t(y|\Omega_t, \theta)$ are continuously differentiable with respect to θ; $F_t(y|\Omega_t, \theta)$ is continuous and strictly increasing in y, so that the inverse function F_t^{-1} is well defined; $E \sup_x \sup_u f_t(x|\Omega_t, u) \leq M_1$ and $E \sup_x \sup_u \| \frac{\partial F_t}{\partial \theta}(x, |\Omega_t, u)\|^2 \leq M_1$ for all t and for some $M_1 < \infty$, where the supremum with respect to u is taken in $N(\theta_0, M)$.

A2: There exists a continuously differentiable function $\bar{g}(r)$ such that for every $M > 0$

$$\sup_{u,v \in N(\theta_0, M)} \Big\| \frac{1}{n} \sum_{t=1}^{n} \frac{\partial F_t}{\partial \theta}(F_t^{-1}(r|u) \mid v) - \bar{g}(r) \Big\| = o_p(1),$$

where $o_p(1)$ is uniform in $r \in [0,1]$. In addition, $\int_0^1 \|\dot{g}\|^2 dr < \infty$ and $C(s) = \int_s^1 \dot{g}\dot{g}' dr$ is invertible for every $s \in [0,1)$, where $g = (r, \bar{g}')'$.

A3: The estimator $\hat{\theta}$ satisfies $\sqrt{n}(\hat{\theta} - \theta_0) = O_p(1)$.

A4: The effect of information truncation satisfies:

$$\sup_{u \in N(\theta_0, M)} n^{-1/2} \sum_{t=1}^{n} \left| F_t(F_t^{-1}(r|\tilde{\Omega}_t, u) \mid \Omega_t, \theta_0) - F_t(F_t^{-1}(r|\Omega_t, u) \mid \Omega_t, \theta_0) \right| = o_p(1)$$

Assumption A1 is concerned with the behavior of the conditional density function and the cumulative distribution function. In the iid setting, $\bar{g}(r)$ in A2 is equal to $\partial F(x, \theta_0)/\partial\theta$, evaluated at $x = F^{-1}(r, \theta_0)$. The term $\bar{g}(r)'\sqrt{n}(\hat{\theta} - \theta_0)$ reflects the effect of parameter estimation. It is equal to (up to an $o_p(1)$) the difference $F(x, \hat{\theta}) - F(x, \theta_0)$ via the Taylor expansion. A2 also assumes that $C(s)$ is a full rank matrix, which may not always be satisfied. However, all needed is that $\bar{g}(r)'\sqrt{n}(\hat{\theta} - \theta_0)$ can be written as $g^*(r)a_n$, where a_n does not depend on r and $C^*(r) = \int_s^1 g^* g'^* dr$ is invertible. This situation rises in location-scale models, such as GARCH models. In fact, this makes the transformation simpler because the dimension of $\bar{g}$ can be much smaller than the number of parameters. A3 is a standard assumption. A4 is unique to dynamic models and is associated with incomplete information sets. It says that past information becomes less relevant as time progresses. A4 is satisfied for GARCH and stationary and invertible ARMA processes. It is noted that even though the aggregation of truncation errors (the sum) is small, each summand in A4 may not be small. For example, in MA(1) process with $|\theta_0| < 1$, it can be shown that

$$\left| F_t(F_t^{-1}(r|\tilde{\Omega}_t, u) \mid \Omega_t, \theta_0) - F_t(F_t^{-1}(r|\Omega_t, u) \mid \Omega_t, \theta_0) \right| \leq B|\sum_{j=t}^{\infty} (-u)^j Y_{t-j}|$$

for some constant $B < \infty$ and $|u| < 1$. For each fixed t, the above is $O_p(1)$. But the sum of these terms is still $O_p(1)$ and becomes $O_p(n^{-1/2})$ upon dividing by $n^{-1/2}$.

Theorem 1 *Under assumptions A1-A4, the asymptotic representations (2), (3) and (4) hold.*

This result provides the basis for the martingale transformation. Let $g(r) = (r, \bar{g}(r)')'$ and $\dot{g}$ be the derivative of g. The martingale transformation is given by

$$\hat{W}_n(r) = \hat{V}_n(r) - \int_0^r \left[\dot{g}(s)' C^{-1}(s) \int_s^1 \dot{g}(\tau) d\hat{V}_n(\tau) \right] ds, \tag{12}$$

and the test statistic

$$T_n = \sup_{0 \leq r \leq 1} |\hat{W}_n(r)|.$$

The test statistic T_n can be easily computed, see Appendix B for details. We have

Corollary 1 *Under the assumptions of Theorem 1,*

$$\hat{W}_n(r) \Rightarrow W(r)$$

$$T_n \xrightarrow{d} \sup_{0\leq r\leq 1} |W(r)|,$$

where $W(r)$ is a standard Brownian motion.

4.2 Nonlinear time series regressions. This section considers an application of the general framework to nonlinear time series regressions of the form:

$$Y_t = h(\Omega_t, \beta) + \varepsilon_t \tag{13}$$

where $\Omega_t = (X_t, X_{t-1}, ...; Y_{t-1}, Y_{t-2}, ...)$. For linear regressions, Bera and Jarque (1982) consider testing normality of ε_t based on skewness and kurtosis. In what follows, let β_0 and λ_0 denote the true parameters. We write $h_t(\beta)$ for $h_t(\Omega_t, \beta)$, $f(x)$ for $f(x, \lambda_0)$ and $F(x)$ for $F(x, \lambda_0)$. We assume:

B1: ε_t are iid with mean zero, density function $f(x, \lambda)$, and cdf $F(x, \lambda)$, where $\lambda \in R^d$ are unknown parameters. The cdf F is strictly increasing and is continuously differentiable with respect to λ. Also, $f(x, \lambda)$ and $\partial F/\partial(\lambda)(x, \lambda)$ are bounded for λ in a neighborhood of λ_0 and for all x. Furthermore, ε_t is independent of Ω_t.

B2: $h_t(\beta)$ is continuously differential in β and $E\|\frac{\partial h_t(\beta_0)}{\partial \beta}\|^2 \leq M$ for some $M < \infty$.

B3: The estimators satisfy $\sqrt{n}(\hat{\beta} - \beta_0) = O_p(1)$, and $\sqrt{n}(\hat{\lambda} - \lambda_0) = O_p(1)$.

B4: The effect of information truncation satisfies:

$$n^{-1/2} \sum_{t=1}^{n} |h(\tilde{\Omega}_t, \beta_0) - h(\Omega_t, \beta_0)| = o_p(1).$$

For linear regressions: $Y_t = X_t'\beta + \varepsilon_t$, assumptions B2 is satisfied if $E\|X_t\|^2 \leq M$ for all t. B3 can be satisfied by least squares method or by some robust estimation methods. B4 is also trivially satisfied because of no information truncation.

Under assumption B1, the conditional cdf of Y_t is $F(y - h(\Omega_t, \beta), \lambda)$. Define

$$\hat{U}_t = F(Y_t - h(\tilde{\Omega}_t, \hat{\beta}), \hat{\lambda})$$

and let $\hat{V}_n(r)$ be defined as in (1).

Theorem 2 *Under assumptions B1-B4, (10) hold. That is,*

$$\hat{V}_n(r) = V_n(r) - f(F^{-1}(r))a_n + \frac{\partial F(F^{-1}(r))}{\partial \lambda}' b_n + o_p(1) \tag{14}$$

where $a_n = \frac{1}{n}\sum_{t=1}^n \frac{\partial h_t(\beta_0)}{\partial \beta}' \sqrt{n}(\hat{\beta} - \beta_0)$, $b_n = \sqrt{n}(\hat{\lambda} - \lambda_0)$, *and* $\frac{\partial F(F^{-1}(r))}{\partial \lambda}$ *is equal to* $\frac{\partial F(x,\lambda_0)}{\partial \lambda}$ *evaluated at* $x = F^{-1}(r, \lambda_0)$.

When λ is a scale parameter such that $F(x, \lambda) = F^0(x/\lambda)$ for some cdf F^0, then[2] $f(F^{-1}(r)) = f^0(F^{0-1}(r))\lambda^{-1}$ and $\frac{\partial F(F^{-1}(r))}{\partial \lambda} = -f^0(F^{0-1}(r))F^{0-1}(r)\lambda^{-1}$. Absorbing λ^{-1} into a_n and $-\lambda^{-1}$ into b_n, we obtain the following representation:

$$\hat{V}_n(r) = V_n(r) + f^0(F^{0-1}(r))a_n + f^0(F^{0-1}(r))F^{0-1}(r)b_n + o_p(1).$$

This is true for all location-scale models. For this class of models the dimension of g is at most three. When no conditional mean parameter is estimated, then $a_n = 0$ so that g has two components $g = (r, f^0(F^{0-1}(r))F^{0-1}(r))'$. When no scale parameter is estimated, that is, the distribution of ϵ_t is completely specified ($b_n = 0$), then $g = (r, f^0(F^{0-1}(r))'$. The GARCH to be considered below is a location-scale model but has a time-varying scale parameter. The corresponding $\hat{V}_n(r)$ process has a similar representation as above.

4.3 GARCH models. We consider GARCH(1,1) introduced in Section 3.2. The assumptions needed for representation (7) are the following:

C1: The ε_t are iid random variables with zero mean and unit variance. The density of ε_t is $f(x)$ and the cdf is $F(x)$. The latter is continuous and strictly increasing. In addition, $E|\epsilon_t|^{2+\tau} < \infty$ for some $\tau > 0$, and ε_t is independent of X_s for $s \leq t$.

C2: $\frac{1}{n}\sum_{t=1}^n X_t X_t'$ converges to a non-random and positive definite matrix.

C3: $\sqrt{n}(\hat{\theta} - \theta) = O_p(1)$, where $\theta = (\delta', \alpha, \beta, \gamma)'$.

We also assume the parameters satisfy the assumptions in Example 1. In particular, it is assumed that $0 \leq \beta < 1$. For $\beta = 0$, it reduces to autoregressive conditional heteroskedasticity (ARCH). For IGARCH, i.e., $\beta + \gamma = 1$, it is assumed that $\beta, \gamma \in (0, 1)$. Under C1, the conditional distribution of Y_t conditional on Ω_t is $Y_t|\Omega_t \sim F\Big((y - X_t'\delta)/\sigma_t\Big)$. Compute $\hat{\sigma}_t$ and $\hat{U}_t$ as in Example 1 and defined $\hat{V}_n$ as in (1).

Theorem 3 *Under assumptions C1-C3,*

$$\hat{V}_n(r) = V_n(r) + f(F^{-1}(r))\, p_n + f(F^{-1}(r))F^{-1}(r)\, q_n + o_p(1),$$

where p_n *and* q_n *are stochastically bounded and are given by*

$$p_n = \frac{1}{n}\sum_{t=1}^n X_t \sqrt{n}(\hat{\delta} - \delta)/\hat{\sigma}_t,$$

[2]This follows from $f(x) = f^0(x/\lambda)/\lambda$ and $F^{-1}(r, \lambda) = F^{0-1}(r)\lambda$.

$$q_n = \frac{1}{2n}\sum_{t=1}^{n}\frac{1}{\sigma_t^2}\Big[\sqrt{n}(\hat{\alpha}-\alpha)\sum_{j=0}^{t}\hat{\beta}^j + \sqrt{n}(\hat{\sigma}_0^2-\sigma_0^2)\hat{\beta}^t$$
$$+\sqrt{n}(\hat{\beta}-\beta)\sum_{j=0}^{t-1}\hat{\beta}^j\sigma_{t-1-j}^2 + \sqrt{n}(\hat{\gamma}-\gamma)\sum_{j=0}^{t-1}\hat{\beta}^j(Y_{t-1-j}-X_{t-1-j}'\hat{\delta})^2\Big].$$

For ARCH models ($\beta = 0$), there is no need to estimate β, and q_n becomes (deduced from the above with $\hat{\beta} = \beta = 0$ and $0^0 = 1$),

$$q_n = \frac{1}{2n}\sum_{t=1}^{n}\frac{1}{\sigma_t^2}\Big[\sqrt{n}(\hat{\alpha}-\alpha) + \sqrt{n}(\hat{\gamma}-\gamma)(Y_{t-1-j}-X_{t-1-j}'\hat{\delta})^2\Big].$$

It is noted that the dimension of g is at most three, regardless of the number of parameters in the conditional mean and conditional variance. As a consequence, martingale transformations for these models are straightforward.

4.4 Estimating the function $\dot{g}$. The martingale transformation requires the function $\dot{g}$, the derivative of g. For certain problems, $\dot{g}(r)$ is completely known. An example is testing conditional distributions in GARCH models (see Section 8 below). In this case, the construction of $\hat{W}_n$ is straightforward. In general, the function $\dot{g}(r)$ depends on the unknown parameter θ_0 so that $\dot{g}(r) = \dot{g}(r, \theta_0)$. A natural solution is to replace θ_0 by a root-n consistent estimator $\hat{\theta}_n$. Assume $\dot{g}$ is continuously differentiable with respect to θ, we will have a pointwise root-n consistent estimate of $\dot{g}$ because

$$\sqrt{n}(\dot{g}(r,\hat{\theta}) - \dot{g}(r,\theta_0)) = \frac{\partial \dot{g}(r,\theta^*)}{\partial\theta}\sqrt{n}(\hat{\theta}-\theta_0), \tag{15}$$

where θ^* is between $\hat{\theta}$ and θ_0 We can proceed to construct $\hat{W}_n(r)$ using $\dot{g}(r,\hat{\theta})$ in place of $\dot{g}(r)$. In veiw of (4), we can also estimate $\dot{g}$ by $\dot{g}_n(r)$ such that $\dot{g}_n(r) = (1, \dot{\tilde{g}}_n(r)')'$, where

$$\dot{\tilde{g}}_n(r) = \frac{1}{n}\sum_{t=1}^{n}\frac{\frac{\partial f_t}{\partial\theta}(x|\tilde{\Omega}_t,\hat{\theta})}{f_t(x|\tilde{\Omega}_t,\hat{\theta})},$$

evaluated at $x = F_t^{-1}(r|\tilde{\Omega}_t,\hat{\theta})$. The above is equal to the derivative (with respect to r) of the right-hand side of (4) with Ω_t replaced by $\tilde{\Omega}_t$ and θ_0 replaced by $\hat{\theta}$. The estimator is, in general, root-n consistent for $\dot{g}$.

Here we shall consider a more general framework, which allows for nonparametric estimation of $\dot{g}$. In this case, the estimated $\dot{g}$ may not be root-n consistent. For example, in testing symmetry, the functions g, F_t, and f_t are all unknown and the above estimators will not be feasible. As alluded to in the introduction, when a data generating process rather than

a conditional distribution is specified, nonparametric estimation is required. We show that root-n consistency is not necessary for the procedure to work.

D1: Let $\dot{g}_n(r)$ be an estimator of $\dot{g}(r)$, either parametric or nonparametric, such that

$$\int_0^1 \|\dot{g}_n(r) - \dot{g}(r)\|^2 dr = o_p(1) \quad and \tag{16}$$

$$\int_s^1 [\dot{g}_n(r) - \dot{g}(r)] dV_n(r) = o_p(1) \tag{17}$$

uniformly in $s \in [0,1]$.

Under D1, we show that $\dot{g}$ can be replaced by $\dot{g}_n$ without affecting the asymptotic results. Note that condition (16) is much weaker than $\sup_{0\le r\le 1} \|\dot{g}_n(r) - \dot{g}(r)\| = o_p(1)$ because the left side of (16) is bounded by the squared value of $\sup_{0\le r\le 1} \|\dot{g}_n(r) - \dot{g}(r)\|$. Consider the transformed process based on $\dot{g}_n$,

$$\tilde{W}_n(r) = \hat{V}_n(r) - \int_0^r \Big[\dot{g}_n(s)' C_n^{-1}(s) \int_s^1 \dot{g}_n(\tau) d\hat{V}_n(\tau)\Big] ds, \tag{18}$$

where $C_n(s) = \int_s^1 \dot{g}_n \dot{g}_n' dr$. The test statistic is defined as

$$T_{n,\epsilon} = \sup_{0\le r\le 1-\epsilon} |\tilde{W}_n(r)|,$$

where $\epsilon > 0$ is a small number.

Theorem 4 *Under assumptions A1-A4 and D1, we have for every $\epsilon \in (0,1)$, in the space $D[0, 1-\epsilon]$,*

$$\tilde{W}_n(r) \Rightarrow W(r)$$

$$T_{n\epsilon} \xrightarrow{d} \sup_{0\le r\le 1-\epsilon} |W(r)|.$$

It is conjectured that the theorem also holds for $\epsilon = 0$. However, the proof of Theorem 4 for $\epsilon = 0$ is extremely subtle and technically demanding. This extension will not be considered. We note that

$$T_n^* = \frac{1}{\sqrt{1-\epsilon}} T_{n\epsilon} \xrightarrow{d} \sup_{0\le s\le 1} |W(s)|$$

because $(1-\epsilon)^{-1/2} \sup_{0\le s\le 1-\epsilon} |W(s)|$ and $\sup_{0\le s\le 1} |W(s)|$ have the same distribution. Hence the same set of critical values for T_n are applicable for $T_{n\epsilon}$ after a simple rescaling.

Discussion. We now consider how to verify D1 in practice. First of all, assumption D1 does not require root-n consistency of $\dot{g}_n$ as in (15). Suppose $\dot{g}_n(r)$ has the following representation,

$$\dot{g}_n(r) - \dot{g}(r) = \kappa_n(r) a_n$$

where $\kappa_n(r)$ is a matrix of (random) functions and $a_n = o_p(1)$. For example, in (15), $\kappa_n(r) = \partial\dot{g}(r, \theta_n^*)/\partial\theta$ and $a_n = (\hat{\theta} - \theta_0)$. In this case, $a_n = O_p(n^{-1/2})$, which is more than necessary. If we assume $\int_0^1 \|k_n(r)\|^2 dr = O_p(1)$, then (16) holds because $a_n = o_p(1)$. Furthermore, if $\int_s^1 \kappa_n(r) dV_n(r)$ is stochastically bounded, i.e.

$$\int_s^1 \kappa_n(r) dV_n(r) = n^{-1/2} \sum_{t=1}^{n} \left[I(U_t > s)\kappa_n(U_t) - \int_s^1 \kappa_n(r) dr \right] = O_p(1) \tag{19}$$

then (17) holds. Equation (19) is generally a consequence of the uniform central limit theorem. For example, with $\kappa_n(r) = \kappa(r, \theta_n^*) = \partial\dot{g}(r, \theta_n^*)/\partial\theta$, the left side of (19) is bounded by

$$n^{-1/2} \sup_{\lambda \in N(\theta_0)} \left\| \sum_{i=1}^{n} \left[I(U_i > s)\kappa(U_i, \lambda) - E\{I(U_i > s)\kappa(U_i, \lambda)\} \right] \right\| \tag{20}$$

where $N(\theta_0)$ is a (shrinking) neighborhood of θ_0. The above is $O_p(1)$ by the uniform central limit theorem. When $a_n = O_p(1)n^{-1/2}$, assumption (17) can also be verified using some uniform strong law of large numbers (USLLN). In this case, we can replace $n^{-1/2}$ by n^{-1} in (19) and conclude it is $o_p(1)$ by the USLLN. Then $a_n \int_s^1 \kappa_n(r) dV_n(r) = O_p(1)n^{-1/2} \int_s^1 \kappa_n(r) dV_n(r) = O_p(1)o_p(1) = o_p(1)$.

4.5 Local Power Analysis. We shall show that the test based on martingale transformation has non-trivial power against root-n local alternatives. Consider the following local alternatives: for $\delta > 0$ and $1 > \delta/\sqrt{n}$,

$$G_{nt}(y|\Omega_t, \theta_0) = (1 - \delta/\sqrt{n})F_t(y|\Omega_t, \theta_0) + (\delta/\sqrt{n})H_t(y|\Omega_t, \theta_0) \tag{21}$$

where both F_t and H_t are conditional distribution functions. The null hypothesis states that the conditional distribution of Y_t is given by $F_t(y|\Omega_t, \theta)$, whereas, under the alternative hypothesis the conditional distribution is $G_{nt}(y|\Omega_t, \theta)$. We assume F_t and H_t are different such that

$$k(r) = \text{plim} \frac{1}{n} \sum_{t=1}^{n} H_t(F_t^{-1}(r|\Omega_t, \theta_0)|\Omega_t, \theta_0) - r \not\equiv 0 \tag{22}$$

If $H_t = F_t$, then G_{nt} is identical to F_t, and moreover, $H_t(F_t^{-1}(r)) = r$ and $k(r) = 0$. Under the alternative hypothesis, the random variables $U_t = F_t(Y_t|\Omega_t, \theta_0)$ are no longer uniform random variables and not necessarily independent. Rather,

$$U_t^* = G_{nt}(Y_t|\Omega_t, \theta_0) \quad (t = 1, 2, ..., n)$$

are i.i.d. uniformly distributed random variables.

Again let $\hat{U}_t = F_t(Y_t|\tilde{\Omega}_t, \hat{\theta})$ and let $\hat{V}_n(r)$ denote the empirical process constructed from $\hat{U}_1, ..., \hat{U}_n$. Under the local alternative, we can still assume $\sqrt{n}(\hat{\theta} - \theta_0) = O_p(1)$.

Theorem 5 *Under the local alternative hypothesis, we have*

$$\hat{V}_n(r) = V_n^*(r) - \bar{g}(r)'\sqrt{n}(\hat{\theta} - \theta_0) + \delta k(r) + o_p(1)$$

Where $k(r)$ is defined in (22), $\bar{g}$ is given in (4), and

$$V_n^*(r) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} [I(U_t^* \leq r) - r].$$

In addition,

$$\hat{W}_n(r) \Rightarrow W(r) + \delta k(r) - \delta \phi_g(k)(r),$$

where $\phi_g(k)(r) = \int_0^r [\dot{g}(s)'C(s)^{-1} \int_s^1 \dot{g} dk] ds$.

An interesting question is what kind of function k satisfies $k(r) - \phi_g(k)(r) \equiv 0$. For such a function, the test will have no local power against the corresponding departure from the null. The following lemma provides a solution to the integral equation:

$$k(r) - \phi_g(k)(r) \equiv 0. \tag{23}$$

Lemma 2 *A function $k(r)$ satisfies the integral equation (23) if and only if $k(r) = a'g(r)$ for some constant vector a.*

It is easy to verify that $a'g(r)$ satisfies (23). The "only if" part is proved in Appendix C. The lemma implies that there may exist one but only one direction (along the function $g(r)$) over which the test possibly lacks power. However, the equation $k(r) = a'g(r)$ imposes strong restrictions on possible departures from the null hypothesis. Whether there exists a genuine alternative hypothesis such that $k(r) = a'g(r)$ (for some $a \neq 0$) is an open question. For concrete problems, e.g., the distributional problem in GARCH models, it is shown below that $k = a'g$ if and only if the null hypothesis is true ($a = 0$), which implies the test has local power against all departures from the null. It should be pointed out, however, root-n consistent tests are not necessarily more powerful than those that are not root-n consistent but can adapt to unknown smoothness of the alternatives, as showed by Horowitz and Spokoiny (2001).

As an application of Lemma 2, consider the local power of the test for GARCH models. Let ε_t be iid with cdf

$$G_n(y) = (1 - \delta n^{-1/2})F(x) + \delta n^{-1/2} H(x),$$

where F and H are distribution functions. Because $k(r) = H(F^{-1}(r)) - r$, the integral equation (23) is equivalent to, by Lemma 2,

$$H(F^{-1}(r)) - r = a_1 r + a_2 f(F^{-1}(r)) + a_3 f(F^{-1}(r)) F^{-1}(r)$$

for some $a = (a_1, a_2, a_3)' \neq 0$. With a change in variable such that $x = F^{-1}(r)$, we can rewrite the above equation as

$$H(x) - F(x) = a_1 F(x) + a_2 f(x) + a_3 f(x)x. \tag{24}$$

Under the assumption that $x^3 f(x) \to 0$ for $|x| \to \infty$, we shall show that the only distribution function $H(x)$ satisfying (24) is $F(x)$ itself, and in this case, $a_i = 0$. To see this, let $x \to +\infty$, we have $a_1 = 0$ because $H(x) - F(x) \to 0$. GARCH models require the distribution function $G_n(y)$ to have zero mean and unit variance for all n. Because F is assumed to have zero mean and unit variance under the null hypothesis, this implies H has zero mean and unit variance. That is, $\int x dH(x) = \int x dF = 0$ and $\int x^2 dH(x) = \int x^2 dF = 1$. Using zero mean restriction, we have $0 = \int x dH - \int x dF = a_2 \int df(x) + a_3 \int d(f(x)x) = -a_2$ because the second integration is equal to zero. Thus $a_2 = 0$. Using unit variance restriction, we have $0 = \int x^2 dH - \int x^2 dF = a_3 \int x^2 d(f(x)x) = -2a_3 \int x^2 f(x) dx = -2a_3$. Thus $a_3 = 0$. We have used the assumption that $x^3 f(x) \to 0$ as $|x| \to \infty$. In summary, we have $H(x) = F(x)$. That is, $G_n \equiv F$. This shows the test has local power for any $H(x) \neq F(x)$. This consistency result holds for any location-scale model.

4.6 Simulations. To assess the size and power of the test statistics, we report some limited simulation results. For assessing size, random variables x_t are generated from normal and t distributions. Let $\varepsilon_t = (x_t - \mu)/\sigma$, where μ and σ^2 are, respectively, the mean and variance of the underlying distribution. Since the distribution of ε_t is invariant with μ and σ under normality, $N(0, 1)$ will be used when x_t is normal. We first estimate the mean and variance parameters and then compute the residuals as $\hat{\varepsilon}_t = (x_t - \hat{\mu})/\hat{\sigma}$, where x_t is either iid standard normal or t_ν, with $\nu = 5$, and $\hat{\mu}$ and $\hat{\sigma}$ are the sample mean and sample variance, respectively. For x_t being normal, we test ε_t as having a standard normal distribution based on residuals $\hat{\varepsilon}_t$. For x_t being t_ν, we test $\varepsilon_t(\nu/(\nu-2))^{1/2}$ as having a t_ν distribution. Because the transforming functions are known, the statistic T_n not T_n^* is used. The results are obtained from 1000 repetitions and are reported in Table 1.

Table 1. Size of the Test

	Normal distribution			t-distribution		
n	10%	5%	1%	10%	5%	1%
100	0.103	0.056	0.025	0.075	0.044	0.018
200	0.104	0.058	0.027	0.065	0.041	0.012
500	0.103	0.056	0.016	0.081	0.042	0.009

For normal distribution, the test tends to be oversized, and for t distribution, the test tends to be undersized except at the 1% level. Overall, the size appears to be acceptable.

For power, we generate data x_t from t_ν and χ^2_ν distributions (with $\nu = 5$). The residuals $\hat{\varepsilon}_t$ are calculated as before. We then test $\varepsilon_t = (x_t - \mu)/\sigma$ to have a standard normal distribution based on the residuals $\hat{\varepsilon}_t$. Note that, when the number of degrees of freedom ν is large, the standardized t or χ^2 random variable (ε_t) is approximately normal $N(0, 1)$. Thus the power of the test should decrease as ν increases. Here we only report the results for $\nu = 5$. All results are obtained from 1000 simulations.

Table 2. Power of the Test

	t-distribution			χ^2-distribution		
n	10%	5%	1%	10%	5%	1%
100	0.53	0.47	0.41	0.91	0.85	0.81
200	0.79	0.73	0.62	1.00	0.97	0.93
500	0.96	0.93	0.91	1.00	1.00	1.00

The test has better power under chi-square distribution than under t-distribution. This is expected because the former has a skewed distribution. Overall, the power is satisfactory.

5 Conclusion

This paper proposes a nonparametric test for conditional distributions of dynamic models. With Khmaladze's transformation, the test overcomes many difficulties associated with the classical Kolmogorov test. On the technical aspects, we establish some weak convergence results for empirical distribution functions under parameter estimation and information truncation. We extend Khmaladze's transformation to allow estimated transforming functions under very weak and general conditions. We also show that dimension reduction in the transformation can be achieved in conditional mean and conditional variance models. The consistency property of the test is also explored. An empirical study demonstrates the usefulness of the test procedure. It is also seen that the method is easy to implement. The result has many potential applications. For example, it is possible to test the specification of continuous-time finance models based on the framework of this paper.

Appendix A: Martingale transformation

A technique used in this paper is the martingale approach of Khmaladze (1981), which effectively transforms a non-martingale process to a martingale one. Let $V(r)$ be a standard Brownian bridge on [0,1]. Then

$$W(r) = V(r) + \int_0^r \frac{V(s)}{1-s} ds \tag{A.1}$$

is a standard Brownian motion on [0,1]. Here $W(r)$ is a martingale transformation of the Brownian bridge. Let $g(r) = (r, g_1(r), ..., g_p(r))'$ be a vector of real-valued functions on $[0,1]$ such that $C(s) = \int_s^1 \dot{g}(v)\dot{g}(v)'dv$ is invertible for each $s \in [0,1)$, where $\dot{g}(r)$ is the derivative of g. Define

$$W(r) = V(r) - \int_0^r \Big[\dot{g}(s)'C^{-1}(s) \int_s^1 \dot{g}(\tau) dV(\tau)\Big] ds. \tag{A.2}$$

It can be shown that $W(r)$ is also a standard Brownian motion. Equation (A.1) is a special case of (A.2) with $g(r) = r$.

Now suppose that $V_n(r)$ is a sequence of stochastic processes on [0,1] such that $V_n(r) \Rightarrow V(r)$, a Brownian bridge. Define

$$W_n(r) = V_n(r) - \int_0^r \Big[\dot{g}(s)'C^{-1}(s) \int_s^1 \dot{g}(\tau) dV_n(\tau)\Big] ds, \tag{A.3}$$

where $\int \dot{g} dV_n$ is defined via the integration parts assuming $\dot{g}$ has a bounded variation. Then $W_n(r) \Rightarrow W(r)$, a standard Brownian motion. The advantages of this transformation will be seen below.

Which g to choose?

Let $\hat{V}_n(r)$ be an empirical process of observations with estimated parameters. As in Theorem 1, the following asymptotic representation holds:

$$\hat{V}_n(r) = V_n(r) - \bar{g}(r)'\sqrt{n}(\hat{\theta} - \theta_0) + o_p(1) \tag{A.4}$$

where $o_p(1)$ is uniform over [0,1] and $V_n(r) \Rightarrow V(r)$, a Brownian bridge. Let $g(r) = (r, \bar{g}')'$, and $C(s)$ is defined earlier. We assume $C(s)$ is invertible for $s \in [0,1)$. Consider the transformation based on $\hat{V}_n(r)$:

$$\hat{W}_n(r) = \hat{V}_n(r) - \int_0^r \Big[\dot{g}(s)'C^{-1}(s) \int_s^1 \dot{g}(\tau) d\hat{V}_n(\tau)\Big] ds. \tag{A.5}$$

Furthermore, define the mapping $\phi_g : D[0,1] \to D[0,1]$ such that

$$\phi_g(h)(r) = \int_0^r \Big[\dot{g}(s)'C^{-1}(s) \int_s^1 \dot{g}(\tau) dh(\tau)\Big] ds. \tag{A.6}$$

Then $\hat{W}_n = \hat{V}_n - \phi_g(\hat{V}_n)$. We note that ϕ_g is a linear mapping and $\phi_g(cg) = cg$ for a constant or random variable c. For $g(r) = (r, \bar{g}')'$, then $\phi_g(c\bar{g}) = c\bar{g}$, which also holds for $c = \sqrt{n}(\hat{\theta} - \theta_0)$. Using (A.4), $\phi_g(\hat{V}_n) = \phi_g(V_n) - \bar{g}'\sqrt{n}(\hat{\theta} - \theta_0) + o_p(1)$. Using (A.4) again, we have $\hat{W}_n = \hat{V}_n - \phi_g(\hat{V}_n) = V_n - \phi_g(V_n) + o_p(1)$, cancelling out $\sqrt{n}(\hat{\theta} - \theta_0)$. Thus the transformation based on $\hat{V}_n$ is asymptotically equivalent to the transformation based on V_n. That is,

$$\hat{W}_n(r) = V_n(r) - \phi_g(V_n)(r) + o_p(1) = W_n(r) + o_p(1).$$

This implies that $\hat{W}_n(r) \Rightarrow W(r)$ because $W_n \Rightarrow W$. Thus the transformation removes the effect of parameter estimation on the limiting process.

To further appreciate this transformation, we apply it to discrete-time processes (r takes on discrete values). In this case, we use summation in place of integration. When applied to regression residuals of linear models, the transformation will transform the ordinary residuals into recursive residuals, which are white noise. Consider $y_i = x_i'\beta + e_i$ $(i = 1, 2, ..., n)$, with e_i being iid and x_i being non-random. The residuals $\hat{e}_i = e_i - x_i(\hat{\beta} - \beta)$ are dependent through $\hat{\beta}$. However, the process $\hat{e}_1, \hat{e}_2, ..., \hat{e}_n$ can be transformed into a martingale-difference sequence.

First note that the transformation (A.5) in its differential form is

$$d\hat{W}_n(r) = d\hat{V}_n(r) - \dot{g}(r)'C^{-1}(r)\int_r^1 \dot{g}(\tau)d\hat{V}_n(\tau)dr. \tag{A.7}$$

If we identify $d\hat{V}_n(r)$ with $\hat{e}_i$, $\dot{g}(r)dr$ with x_i, $C(r)$ with $X_{n-i}'X_{n-i} = \sum_{k=i+1}^n x_k x_k'$, and $\int_r^1 \dot{g}dV_n$ with $\sum_{k=i+1}^n x_k\hat{e}_k = X_{n-i}'\hat{E}_{n-i}$, where $\hat{E}_{n-i}$ is a vector of the last $n-i$ residuals, then the right hand side of (A.7) is

$$\hat{e}_i - x_i'(X_{n-i}'X_{n-i})^{-1}X_{n-i}'\hat{E}_{n-i}.$$

The above can be rewritten as $y_i - x_i'\hat{\beta}_{n-i}$, where $\hat{\beta}_{n-i}$ is the least squares estimator based on the last $n-i$ observations (follows from $\hat{E}_{n-i} = Y_{n-i} - X_{n-i}'\hat{\beta}$). Thus we obtain the ith backward recursive residual (up to the normalizing constant $1 + x_i'(X_{n-i}'X_{n-i})^{-1}x_i$). Similarly, if we use an alternative transformation formula (given in Khmaladze), we will obtain the forward recursive residuals of Brown, Durbin, and Evans (1975). It is well known that partial sum of recursive residuals leads to a Brownian motion process.

We can interpret the martingale transformation as employing a continuous-time recursive least squares method to obtain continuous-time recursive residuals. The integration of recursive residuals leads to a Brownian motion process. In the context of GMM estimation and hypothesis testing, Wooldridge (1990) proposed a transformation that can purge the effect of

parameter estimation. In the sense of projecting relevant variables on to their score functions to obtain projection residuals, Wooldridge's correction is similar in spirit to the martingale transformation. But the former is a finite dimensional correction and the latter can be viewed as an infinite dimensional correction.

Appendix B: Computing the Test Statistics

The martingale transformation involves integration. We discuss a numerical method for computing the integral.

An alternative expression for $\hat{W}_n$. Introduce $\hat{J}_n(r) = \frac{1}{n}\sum_{t=1}^n I(\hat{U}_t \le r)$. Then $\hat{V}_n(r) = \sqrt{n}(\hat{J}_n(r) - g_1(r))$, where $g_1(r) = r$, the first component of g. Recall that $\hat{W}_n = \hat{V}_n - \phi_g(\hat{V}_n)$, and ϕ_g is a linear mapping. So $\phi_g(\hat{V}_n) = \sqrt{n}\phi_g(\hat{J}_n) - \sqrt{n}\phi_g(g_1)$. Moreover, from $\phi_g(g) = g$, we have $\phi_g(g_1) = g_1$. Thus $\hat{W}_n = \sqrt{n}[\hat{J}_n - \phi_g(\hat{J}_n)]$. That is,

$$\hat{W}_n(r) = \sqrt{n}\Big(\hat{J}_n(r) - \int_0^r \dot{g}' C(s)^{-1} \int_s^1 [\dot{g}(\tau) d\hat{J}_n(\tau)] ds\Big).$$

This leads to a simpler computation.

Deriving a computable formula. Denote by $\hat{u}_1, \hat{u}_2, ..., \hat{u}_n$ the realized values of $\hat{U}_1, \hat{U}_2, ..., \hat{U}_n$. Let $\hat{u}_{(1)} < \hat{u}_{(2)} < \cdots < \hat{u}_{(n)}$ denote the ordered version of $\hat{u}_1, ..., \hat{u}_n$. In addition, $\hat{u}_{(0)} = 0$, and $\hat{u}_{(n+1)} = 1$. For notational succinctness, let $v_i = \hat{u}_{(i)}$ $(i = 0, 1, ..., n+1)$. The numbers $v_0, v_1, ..., v_{n+1}$ form a natural partition of [0, 1]. Suppose g is also given. Using

$$\int_s^1 \dot{g}(\tau) d\hat{J}_n(\tau) = \frac{1}{n} \sum_{i:\hat{u}_i \ge s} \dot{g}(\hat{u}_i)$$

and evaluating the above integral at $s = \hat{u}_{(k)}$, we have $\int_{\hat{u}_{(k)}}^1 \dot{g} d\hat{J}_n = \frac{1}{n}\sum_{i=k}^n \dot{g}(\hat{u}_{(i)})$. That is,

$$\int_{v_k}^1 \dot{g} d\hat{J}_n = \frac{1}{n} \sum_{i=k}^n \dot{g}(v_i) \tag{B.1}$$

We next approximate the following integral by

$$\int_s^1 \dot{g}\dot{g}' d\tau \doteq \sum_{i:v_i \ge s} \dot{g}(v_i)\dot{g}(v_i)'(v_{i+1} - v_i)$$

where "$\doteq$" represents an approximate equality. Evaluating the above integration at $s = v_k$ gives

$$\int_{v_k}^1 \dot{g}\dot{g}' d\tau \doteq \sum_{i=k}^n \dot{g}(v_i)\dot{g}(v_i)'(v_{i+1} - v_i). \tag{B.2}$$

We denote the right-hand-side of (B.1) by $\frac{1}{n}D_k$, and the right-hand-side of (B.2) by C_k. Then

$$\int_0^{v_j} [\dot{g}(s)'C(s)^{-1} \int_s^1 \dot{g}(\tau) d\hat{J}_n(\tau)] ds \doteq \frac{1}{n} \sum_{k=1}^{j} \dot{g}(v_k) C_k^{-1} D_k (v_k - v_{k-1}).$$

Computing the test statistic. Summarizing the above derivation and noting that $\hat{J}_n(v_j) = j/n$ (for all j), we compute T_n with

$$\sup_{1 \le j \le n} |\hat{W}_n(v_j)| \doteq \max_{1 \le j \le n} \sqrt{n} \Big| \frac{j}{n} - \frac{1}{n} \sum_{k=1}^{j} \dot{g}(v_k)' C_k^{-1} D_k (v_k - v_{k-1}) \Big|$$

where $D_k = \sum_{i=k}^{n} \dot{g}(v_i)$ and $C_k = \sum_{i=k}^{n} \dot{g}(v_i)\dot{g}(v_i)'(v_{i+1} - v_i)$, and where $v_1, ..., v_n$ are ordered values of $\hat{U}_1, ..., \hat{U}_n$.

When $\dot{g}$ is estimated by $\dot{g}_n$, simply replace $\dot{g}$ by $\dot{g}_n$ and calculate $T_{n\epsilon}$ with the same formula except that the supremum with respect to j is taken in the range $1 \le j \le n\epsilon$.

Appendix C: Proofs

In the absence of information truncation, $F_t(y|\Omega_t, \theta)$ and $F_t(y|\theta)$ will be used interchangeably.

Lemma C.1 *Under the assumptions of Theorem 1,*

$$\max_{1 \le t \le n} \sup_{u \in N(\theta_0, M)} \sup_{0 \le r \le 1} \Big| F_t(F_t^{-1}(r|u) \mid \theta_0) - r \Big| = o_p(1)$$

Proof: Let $x = F_t^{-1}(r|u)$ or $r = F_t(x|u)$. Then

$$\begin{aligned} \sup_r |F_t(F_t^{-1}(r|u) \mid \theta_0) - r| &= \sup_x |F_t(x|\theta_0) - F_t(x|u)|. \\ &= \sup_x |\partial F_t / \partial \theta(x|\theta^*)'(\theta_0 - u)| \\ &\le n^{-1/2} \sup_x \|\partial F_t / \partial \theta(x|\theta^*)\| M \end{aligned}$$

where θ^* is between θ^0 and u. By assumption A1, $E(\sup_x \sup_{\theta^* \in N(\theta_0, M)} \|\partial F_t / \partial \theta(x|\theta^*)\|^2) \le M_1$, this implies that $n^{-1/2} \sup_x \|\partial F_t / \partial \theta(x|\theta^*)\| = o_p(1)$ uniformly in $t \in [1, n]$ and $\theta^* \in N(\theta_0, M)$. $\square$.

Lemma C.2 *For every $\epsilon > 0$, there exists $\delta > 0$ such that for $u, v \in N(\theta_0, M)$ and for all large n,*

$$P\left(\sup_{\|u - v\| \le \delta n^{-1/2}} \sup_{0 \le r \le 1} n^{-1/2} \Big| \sum_{t=1}^{n} F_t(F_t^{-1}(r|u)|\theta_0) - F_t(F_t^{-1}(r|v)|\theta_0) \Big| > \epsilon \right) < \epsilon.$$

Proof: Let $x = F_t^{-1}(r|u)$ and $y = F_t^{-1}(r|v)$. Then $r = F_t(x|u) = F_t(y|v)$. Thus,

$$\begin{aligned} F_t(F_t^{-1}(r|u)|\theta_0) - F_t(F_t^{-1}(r|v)|\theta_0) &= F_t(x|\theta_0) - F_t(y|\theta_0) \\ &= F_t(x|\theta_0) - F_t(x|u) - [F_t(y|\theta_0) - F_t(y|v)] \\ &= \partial F_t/\partial\theta(x|\theta^*)'(\theta_0 - u) - \partial F_t/\partial\theta(y|\theta^\dagger)'(\theta_0 - v) \\ &= \partial F_t/\partial\theta(x|\theta^*)'(v - u) + \Big[\partial F_t/\partial\theta(x|\theta^*) - \partial F_t/\partial\theta(y|\theta^\dagger)\Big]'(\theta_0 - v) \end{aligned}$$

where θ^* is between θ^0 and u, and $\theta^\dagger$ is between θ^0 and v. From $\|u - v\| \leq \delta n^{-1/2}$ and $\|\theta^0 - v\| \leq Mn^{-1/2}$,

$$\sup_r n^{-1/2}\Big|\sum_{t=1}^n F_t(F_t^{-1}(r|u)|\theta_0) - F_t(F_t^{-1}(r|v)|\theta_0)\Big| \tag{C.1}$$

$$\leq \sup_r \frac{1}{n}\|\sum_{t=1}^n \frac{\partial F_t}{\partial\theta}(F_t^{-1}(r|u)|\theta^*)\|\delta + \sup_r \frac{1}{n}\|\sum_{t=1}^n \Big[\frac{\partial F_t}{\partial\theta}(F_t^{-1}(r|u)|\theta^*) - \frac{\partial F_t}{\partial\theta}(F_t^{-1}(r|v)|\theta^\dagger)\Big]\|M$$

The first expression is $O_p(1)\delta$ by A1 (or A2). The second expression is $o_p(1)$ because the limit is $\|\bar{g}(r) - \bar{g}(r)\|M = 0$ by A2. Thus (C.1) is bounded by $O_p(1)\delta + o_p(1)$, which implies Lemma C.2. □

Lemma C.3 *For every $\epsilon > 0$, there exists $\delta > 0$ such that for all large n,*

$$P\left(\sup_{|r_1 - r_2| \leq \delta n^{-1/2}} \sup_{u \in N(\theta_0, M)} n^{-1/2}\Big|\sum_{t=1}^n F_t(F_t^{-1}(r_1|u)|\theta_0) - F_t(F_t^{-1}(r_2|u)|\theta_0)\Big| > \epsilon\right) < \epsilon.$$

Proof. By Talyor expansion, there exists θ^* in between θ_0 and u such that

$$F_t(x|\theta_0) = F_t(x|u) + \frac{\partial F_t}{\partial\theta}(x|\theta^*)'(\theta_0 - u).$$

Evaluate the above at $x = F^{-1}(r_1|u)$ and note $F_t(F_t^{-1}(r_1|u)|u) = r_1$ for all r_1,

$$F_t(F_t^{-1}(r_1|u)|\theta_0) = r_1 + \frac{\partial F_t}{\partial\theta}(F_t^{-1}(r_1|u)|\theta^*)'(\theta_0 - u).$$

A similar identity holds for r_2. Thus,

$$n^{-1/2}\Big|\sum_{t=1}^n F_t(F_t^{-1}(r_1|u)|\theta_0) - F_t(F_t^{-1}(r_2|u)|\theta_0)\Big|$$

$$\leq \sqrt{n}|r_1 - r_2| + \frac{1}{n}\|\sum_{t=1}^n \Big[\frac{\partial F_t}{\partial\theta}(F_t^{-1}(r_1|u)|\theta^*) - \frac{\partial F_t}{\partial\theta}(F_t^{-1}(r_2|u)|\theta^\dagger)\Big]\|M$$

The above is bounded by $\delta + o_p(1)$, which implies Lemma C.3. To see this, $\sqrt{n}|r_1 - r_2| \leq \delta$ by assumption and the second expression converges to $\|\bar{g}(r_1) - \bar{g}(r_2)\|M = o(1)$ by A2 and $r_1 - r_2 \to 0$. $\square$

Clearly from the above proof, if r_1 and r_2 are such that $|r_1 - r_2| \leq n^{-1/2-d}$ $(d > 0)$, then

$$n^{-1/2}\Big|\sum_{t=1}^{n} F_t(F_t^{-1}(r_1|u)|\theta_0) - F_t(F_t^{-1}(r_2|u)|\theta_0)\Big| \leq n^{-d} + o_p(1) = o_p(1) \tag{C.2}$$

Equation (C.2) is analogous to Lemma A.3 of Bai (1996).

Proof of Theorem 1. We first consider the case of no information truncation. This occurs if the dynamic model depends only on a finite number of lagged Y_t. From $U_t = F_t(Y_t|\theta_0)$ and $\hat{U}_t = F_t(Y_t|\Omega_t, \hat{\theta}) = F_t(Y_t|\hat{\theta})$, we have $Y_t = F_t^{-1}(U_t|\theta_0)$ and $Y_t = F_t^{-1}(\hat{U}_t|\hat{\theta})$. Thus

$$F_t^{-1}(U_t|\theta_0) = F_t^{-1}(\hat{U}_t|\hat{\theta})$$

and

$$U_t = F_t(F_t^{-1}(\hat{U}_t|\hat{\theta}) \mid \theta_0). \tag{C.3}$$

This implies that $\hat{U}_t \leq r$ if and only if $U_t \leq F_t(F_t^{-1}(r|\hat{\theta}) \mid \theta_0)$. Therefore,

$$\hat{V}_n(r) = \frac{1}{\sqrt{n}}\sum_{t=1}^{n}[I(U_t \leq F_t(F_t^{-1}(r|\hat{\theta}) \mid \theta_0)) - r].$$

Now define

$$\xi_t(r, a, b) = F_t(F_t^{-1}(r|a) \mid b).$$

In particular,

$$\xi_t(r, \hat{\theta}, \theta_0) = F_t(F_t^{-1}(r|\hat{\theta}) \mid \theta_0).$$

We have

$$\hat{V}_n(r) = \frac{1}{\sqrt{n}}\sum_{t=1}^{n}[I(U_t \leq \xi_t(r, \hat{\theta}, \theta_0)) - r].$$

Adding and subtracting terms, we have

$$\hat{V}_n(r) = n^{-1/2}\sum_{t=1}^{n}[I(U_t \leq r) - r] + d_n(r) + R_n(r, \hat{\theta}),$$

where

$$d_n(r) = n^{-1/2}\sum_{t=1}^{n}[\xi_t(r, \hat{\theta}, \theta_0) - r]$$

and

$$R_n(r, \hat{\theta}) = n^{-1/2}\sum_{t=1}^{n}\Big[I(U_t \leq \xi_t(r, \hat{\theta}, \theta_0)) - \xi_t(r, \hat{\theta}, \theta_0) - I(U_t \leq r) + r\Big]. \tag{C.4}$$

Because $\xi_t(r, \hat{\theta}, \hat{\theta}) = r$, by A2 and Taylor expansion, for some θ^* between $\hat{\theta}$ and θ_0,

$$d_n(r) = \frac{1}{n}\sum_{t=1}^{n} \frac{\partial F_t}{\partial \theta}(F_t^{-1}(r|\hat{\theta}) \mid \theta^*)\sqrt{n}(\theta_0 - \hat{\theta}) = -\bar{g}(r)\sqrt{n}(\hat{\theta} - \theta_0) + o_p(1). \tag{C.5}$$

It remains to show that $R_n(r, \hat{\theta}) = o_p(1)$ uniformly in r; the proofs involves three steps. Let $K = [0, 1] \times [\theta_0 - Mn^{-1/2}, \theta_0 + Mn^{-1/2}]$. The three steps are: (i) show $R_n(r, u) = o_p(1)$ for each $(r, u) \in K$; (ii) show $\sup_r |R_n(r, u_1) - R(r, u_2)|$ is small when $\|u_1 - u_2\|$ is small; (iii) show $\sup_u |R_n(r_1, u) - R_n(r_2, u)|$ is small when $|r_1 - r_2|$ is small. These can be proved using the argument of Bai (1994, 1996). In particular, Lemma C.1 is needed in each step; Lemma C.2 is needed in proving (ii) and Lemma C.3 is needed in proving (iii). The readers are also referred to Loynes (1980) and Koul (1996). The details are omitted to save space.

Next we consider the case of information truncation. Equation (C.3) is now changed to

$$U_t = F_t(F_t^{-1}(\hat{U}_t|\tilde{\Omega}_t, \hat{\theta}) \mid \Omega_t, \theta_0) \tag{C.6}$$

Again, the above expression is understood as the function $F_t(y|\Omega_t, \theta_0)$ evaluated at $y = F_t^{-1}(Y_t|\Omega_t, \hat{\theta})$. From (C.6), we have $\hat{U}_t \le r$ if and only if

$$\begin{aligned} U_t &\le F_t(F_t^{-1}(r|\tilde{\Omega}_t, \hat{\theta}) \mid \Omega_t, \theta_0) \\ &= F_t(F_t^{-1}(r|\Omega_t, \hat{\theta}) \mid \Omega_t, \theta_0) + \\ &\quad F_t(F_t^{-1}(r|\tilde{\Omega}_t, \hat{\theta}) \mid \Omega_t, \theta_0) - F_t(F_t^{-1}(r|\Omega_t, \hat{\theta}) \mid \Omega_t, \theta_0) \\ &= \xi_t(r, \hat{\theta}, \theta_0) + \eta_t(r) \end{aligned}$$

where

$$\eta_t(r) = F_t(F_t^{-1}(r|\tilde{\Omega}_t, \hat{\theta}) \mid \Omega_t, \theta_0) - F_t(F_t^{-1}(r|\Omega_t, \hat{\theta}) \mid \Omega_t, \theta_0).$$

Thus $\hat{V}_n(r) = n^{-1/2}\sum_{t=1}^{n}[I(U_t \le \xi_t(r, \hat{\theta}, \theta_0) + \eta_t(r)) - r]$

Adding and subtracting terms, we have

$$\hat{V}_n(r) = n^{-1/2}\sum_{t=1}^{n}[I(U_t \le r) - r] + d_n^*(r) + R_n^*(r)$$

where

$$d_n^*(r) = n^{-1/2}\sum_{t=1}^{n}[\xi_t(r, \hat{\theta}, \theta_0) - r] + n^{-1/2}\sum_{t=1}^{n}\eta_t(r) \tag{C.7}$$

$$R_n^*(r) = n^{-1/2}\sum_{t=1}^{n}\Big[I(U_t \le \xi_t(r, \hat{\theta}, \theta_0) + \eta_t(r)) - \xi_t(r, \hat{\theta}, \theta_0) - \eta_t(r) - I(U_t \le r) + r\Big]. \tag{C.8}$$

The second term of the right hand side (r.h.s.) of (C.7) is $o_p(1)$ by A4. Similar to (C.5), $d_n^*(r) = -\bar{g}(r)\sqrt{n}(\hat{\theta} - \theta_0) + o_p(1)$. It remains to show $R_n^*(r) = o_p(1)$. Note that the term η_t

does not satisfy $\max_{1\le t\le T}|\eta_t| = o_p(1)$. But it does satisfy, by A4, $n^{-1/2}\sum_{t=1}^{n}|\eta_t(r)| = o_p(1)$ uniformly in r. Thus this term must be given special treatment when analyzing $R_n^*(r)$. But this can be proved using the argument of Bai (1994). The details are omitted. □

Proof of Corollary 1. Because U_t are iid $U(0,1)$, $V_n(r) \Rightarrow V(r)$, where $V(r)$ is a standard Brownian bridge. Furthermore, because $\phi_g(g) = g$, we have $\hat{W}_n(r) = W_n(r) + o_p(1)$, here $W_n(r)$ has the form of (A.3) and $W_n(r) \Rightarrow W(r)$. Thus by the continuous mapping theorem, $T_n = \sup|\hat{W}_n(r)| \xrightarrow{d} \sup|W(r)|$. □

Proof of Theorem 2. Since Assumptions B1-B4 imply A1-A4, Theorem 2 is a consequence of Theorem 1. Note that neither Theorem 1 nor Theorem 2 requires the invertibility of $C(s)$, although the martingale transformation itself does. □

Proof of Theorem 3. Given the model's concrete structure, it is easier to derive a direct proof. We use the identity

$$\hat{\varepsilon}_t = \varepsilon_t(\hat{\sigma}_t^2/\sigma_t^2)^{-1/2} - X_t'(\hat{\delta}-\delta)\hat{\sigma}_t^{-1} = \varepsilon_t(1+\eta_{nt}(\hat{\theta}))^{-1/2} - X_t'(\hat{\delta}-\delta)\hat{\sigma}_t^{-1}$$

where

$$\eta_{nt}(\hat{\theta}) = \frac{1}{\sigma_t^2}(\hat{\sigma}_t^2 - \sigma_t^2).$$

Define

$$K_n(x) = n^{-1/2}\sum_{t=1}^{n}[I(\varepsilon_t \le x) - F(x)], \quad \textit{and} \quad \hat{K}_n(x) = n^{-1/2}\sum_{t=1}^{n}[I(\hat{\varepsilon}_t \le x) - F(x)].$$

Then from the relationship between $\hat{\varepsilon}_t$ and ε_t,

$$\hat{K}_n(x) = n^{-1/2}\sum_{t=1}^{n}[I(\varepsilon_t \le x(1+\eta_{nt}(\hat{\theta}))^{1/2} + X_t'(\hat{\delta}-\delta)\hat{\sigma}_t^{-1}) - F(x)].$$

It is easy to argue that $P(\inf_t \hat{\sigma}_t > \alpha/2) \to 1$ (α is a parameter in the conditional variance). So we have $\sup_t |X_t'(\hat{\delta}-\delta)/\hat{\sigma}_t| \le \sqrt{n}|\hat{\delta}-\delta)|n^{-1/2}\max_{1\le t\le n}\|X_t\|/(\inf_t \hat{\sigma}_t) = o_p(1)$ because Assumption C2 implies that $n^{-1/2}\max_{1\le t\le n}\|X_t\| = o_p(1)$. We next argue that $\eta_{nt}(\hat{\theta})$ is also uniformly small over t. Notice that

$$\hat{\sigma}_t^2 - \sigma_t^2 = (\hat{\alpha}-\alpha) + \hat{\beta}(\hat{\sigma}_{t-1}^2 - \sigma_{t-1}^2) + (\hat{\beta}-\beta)\sigma_{t-1}^2 + (\hat{\gamma}-\gamma)(Y_{t-1} - X_{t-1}'\hat{\delta})^2. \tag{C.9}$$

The above has a recursive relationship in terms of $\hat{\sigma}_{t-1}^2 - \sigma_{t-1}^2$ [analogous to an AR(1) process for $\hat{\sigma}_{t-1}^2 - \sigma_{t-1}^2$]. By repeated substitution, we can write

$$\hat{\sigma}_t^2 - \sigma_t^2 = \hat{\beta}^t(\hat{\sigma}_0^2 - \sigma_0^2) + (\hat{\alpha}-\alpha)\sum_{j=0}^{t-1}\hat{\beta}^j + (\hat{\beta}-\beta)\sum_{j=0}^{t-1}\hat{\beta}^j\sigma_{t-1-j}^2 + (\hat{\gamma}-\gamma)\sum_{j=0}^{t-1}\hat{\beta}^j(Y_{t-1-j} - X_{t-1-j}'\hat{\delta})^2.$$

Dividing the above by σ_t^2, we have

$$\eta_{nt}(\hat{\theta}) = \hat{\beta}^t \frac{(\hat{\sigma}_0^2 - \sigma_0^2)}{\sigma_t^2} + \frac{(\hat{\alpha} - \alpha)}{\sigma_t^2} \sum_{j=0}^{t-1} \hat{\beta}^j + (\hat{\beta} - \beta) \sum_{j=0}^{t-1} \hat{\beta}^j \frac{\sigma_{t-1-j}^2}{\sigma_t^2} + \frac{(\hat{\gamma} - \gamma)}{\sigma_t^2} \sum_{j=0}^{t-1} \hat{\beta}^j (Y_{t-1-j} - X'_{t-1-j}\hat{\delta})^2. \tag{C.10}$$

Each of the last three terms on the r.h.s. of (C.10) is $o_p(1)$ uniformly in t. For example, consider the last term. From $(Y_t - X_t'\hat{\delta})^2 = (\sigma_t \varepsilon_t - X_t'(\hat{\delta} - \delta))^2$ and $(a-b)^2 \le 2(a^2 + b^2)$, the last term of (C.10) is bounded by

$$2|(\hat{\gamma} - \gamma)| \max_{1 \le t \le n} \sum_{j=0}^{t-1} |\hat{\beta}|^j \Big[\sigma_{t-1-j}^2 \varepsilon_{t-1-j}^2 / \sigma_t^2 + \|\hat{\delta} - \delta\|^2 \|X_{t-1-j}\|^2 / \sigma_t^2 \Big]$$

$$\le C_n \Big(n^{-1/2} \max_{0 \le t \le n} |\varepsilon_t^2| \Big) \Big(\max_{1 \le t \le n} \sum_{j=0}^{t-1} |\hat{\beta}|^j \sigma_{t-1-j}^2 / \sigma_t^2 \Big) + D_n n^{-1/2} \|\sqrt{n}(\hat{\delta} - \delta)\|^2 \max_{0 \le t \le n} (\|X_t\|^2 / n),$$

where $C_n = 2|\sqrt{n}(\hat{\gamma} - \gamma)| = O_p(1)$ and $D_n = C_n(\sum_{i=1}^n |\hat{\beta}|^j) = O_p(1)$ because $P(|\hat{\beta}| < 1) \to 1$. Assumption C2 implies that $\max_{0 \le t \le n}(\|X_t\|^2/n) = o_p(1)$ and thus the second term above is $o_p(n^{-1/2})$. From $E|\varepsilon_t|^{2+\tau} < \infty$, we have $n^{-1/2} \max_{0 \le t \le n} \varepsilon_t^2 = o_p(1)$. To show the first term above is $o_p(1)$, it suffices to argue $\max_{1 \le t \le n} \sum_{j=0}^{t-1} |\hat{\beta}|^j \sigma_{t-1-j}^2 / \sigma_t^2 = O_p(1)$. Now, choose $\bar{\beta} < 1$ such that $P(|\hat{\beta}| < \bar{\beta}) \to 1$, then

$$\max_{1 \le t \le n} \sum_{j=0}^{t-1} |\hat{\beta}|^j \sigma_{t-1-j}^2 / \sigma_t^2 \le \sum_{t=1}^{n} \sum_{j=0}^{t-1} |\hat{\beta}|^j \sigma_{t-1-j}^2 / \sigma_t^2 \le \sum_{t=1}^{n} \sum_{j=0}^{t-1} \bar{\beta}^j \sigma_{t-1-j}^2 / \sigma_t^2 + o_p(1). \tag{C.11}$$

For $\beta = 0$ (implying an ARCH model), it is assumed that $\gamma < 1$. Noting that $\sigma_t^2 \ge \alpha > 0$, $E(\sigma_{t-k}^2 / \sigma_t^2)$ is bounded uniformly in t and k, see Engle (1982). It follows that (C.11) is $O_p(1)$. Now consider $1 > \beta > 0$ (including IGARCH). From Lee and Hansen (1994), $E(\sigma_{t-1-j}^2 / \sigma_t^2) \le M(R/\beta)^j$ for each j, where M is bounded and $0 < R < 1$. Because $R < 1$, we can choose $\bar{\beta} > \beta$ such that $\bar{\beta} R / \beta < 1$. Thus $E \sum_{t=1}^n \sum_{j=0}^{t-1} \bar{\beta}^j \sigma_{t-1-j}^2 / \sigma_t^2 \le M \sum_{t=1}^{\infty} \sum_{j=0}^{t-1} (\bar{\beta} R / \beta)^j \le M_1$ for some finite M_1. Thus (C.11) is $O_p(1)$. Summarizing these results, the last term of (C.10) is $o_p(1)$. Similarly, the second and third terms of (C.10) are also $o_p(1)$. This implies that $\eta_{nt}(\hat{\theta}) = o_p(1) + \hat{\beta}^t(\hat{\sigma}_0^2 - \sigma_0^2)/\sigma_t^2$, where $o_p(1)$ is uniform in t. Although for each t, $\hat{\beta}^t(\hat{\sigma}_0^2 - \sigma_0^2)/\sigma_t^2$ is not $o_p(1)$, the sum of these terms divided by root-n is $o_p(1)$. That is, $n^{-1/2} \sum_{t=1}^n \hat{\beta}^t(\hat{\sigma}_0^2 - \sigma_0^2)/\sigma_t^2 = o_p(1)$. Thus these terms will not affect the limiting process of $\hat{K}_n(x)$ and can be ignored (see the proof of Theorem 1). The conditions of Theorems A.2 and A.3 of Bai (1996) are satisfied. Apply Theorems A.2 and A.3 of Bai (1996), applied with $s = 1$, $c_t = 1$, $a_t = \frac{1}{2}\sqrt{n}\eta_{nt}(\hat{\theta})$, $b_t = X_t'(\hat{\delta} - \delta)/\hat{\sigma}_t$ in the notation of Bai (note that $(1 + \eta_n(\hat{\theta}))^{1/2}$

is equal to $1 + \frac{1}{2}\eta_{nt}(\hat{\theta})$ plus a higher order term of $\eta_{nt}(\hat{\theta})$, which is negligible), we have

$$\hat{K}_n(x) = K_n(x) + f(x)\frac{1}{n}\sum_{t=1}^{n} X_t'(\hat{\delta} - \delta)/\hat{\sigma}_t + f(x)x\frac{1}{2}n^{-1/2}\sum_{t=1}^{n}\eta_{nt}(\hat{\theta}) + o_p(1).$$

But $\frac{1}{2}n^{-1/2}\sum_{t=1}^{n}\eta_{nt}(\hat{\theta})$ is equal to q_n defined in Theorem 3. Note that $q_n = O_p(1)$. This is true even if the initial estimator $\hat{\sigma}_0^2$ is not consistent for σ_0^2. In fact, $\hat{\sigma}_0^2$ can be an arbitrary random variable so that $\sqrt{n}(\hat{\sigma}_0^2 - \sigma_0^2) = O_p(\sqrt{n})$. The contribution of this term to q_n is negligible. This follows from the fact that $P(|\hat{\beta}| < 1 - \epsilon) > 1 - \epsilon$ for large n and

$$\frac{1}{n}\sum_{t=1}^{n}\frac{1}{\sigma_t^2}|\sqrt{n}(\hat{\sigma}_0^2 - \sigma_0^2)\hat{\beta}^t| \leq n^{-1/2}|(\hat{\sigma}_0^2 - \sigma_0^2)|\frac{1}{\alpha}\sum_{t=1}^{n}|\hat{\beta}|^t = O_p(n^{-1/2}) = o_p(1).$$

All other terms in q_n are $O_p(1)$. Thus $q_n = O_p(1)$.

Finally, notice that $\hat{K}_n(x) = \hat{V}_n(F(x))$. If we let $F(x) = r$, then the representation of $\hat{V}_n(r)$ follows readily. $\square$.

Remark 3. For GARCH models (or location-scale models) martingale transformations can be performed directly on $\hat{K}_n(x)$. It is well known that $K_n(x)$ converges weakly to a Brownian bridge $K(x)$ on the real line with covariance function $EK(x)K(y) = F(x)(1-F(y))$ for $x < y$. Because the limit of $K_n(x)$ is a time-stretched Brownian bridge, the martingale transformation of $\hat{K}_n(x)$ takes a new form. Let $\dot{h}(x) = (1, \dot{f}/f, 1 + x(\dot{f}/f))'$ (a vector with three components) and $C_h(x) = \int_x^\infty \dot{h}(y)\dot{h}'(y)f(y)dy$. Define

$$\hat{B}_n(x) = \hat{K}_n(x) - \int_{-\infty}^{x}\Big[\dot{h}(v)'C_h(v)^{-1}\int_v^\infty \dot{h}(\tau)d\hat{K}_n(\tau)\Big]f(v)dv$$

and the test statistic $T_n = \sup_{-\infty<x<\infty}|\hat{B}_n(x)|$. It can be shown that $\hat{B}_n(x) = \hat{W}_n(F(x))$. Thus the two transformations are equivalent. It follows that

$$\hat{B}_n(x) \Rightarrow W(F(x)), \quad \text{and} \quad T_n \xrightarrow{d} \sup_{0\leq s\leq 1}|W(s)|.$$

The transformation based on $\hat{K}_n(x)$ does not involve the quantile function $F^{-1}(r)$.

Proof of Theorem 4. It suffices to show that $\tilde{W}_n \Rightarrow W$ in the space $D[0, 1-\epsilon]$ under the sup norm. By assumption D1, $\|g_n(r) - g(r)\| \leq \int_0^r \|\dot{g}_n(s) - \dot{g}(s)\|ds \leq \int_0^1 \|\dot{g}_n(r) - \dot{g}(r)\|^2 ds = o_p(1)$. From $\hat{V}_n(r) = V_n(r) - \bar{g}(r)\sqrt{n}(\hat{\theta} - \theta_0) + o_p(1)$ and $\bar{g}_n = \bar{g} + o_p(1)$, we can write

$$\hat{V}_n(r) = V_n(r) - \bar{g}_n(r)'\sqrt{n}(\hat{\theta} - \theta_0) + o_p(1).$$

Because $\phi_{g_n}(g_n) = g_n$, the transformation $\tilde{W}_n(r)$ in (18) can be written as

$$\tilde{W}_n(r) = V_n(r) - \int_0^r \Big[\dot{g}_n(s)'C_n^{-1}(s)\int_s^1 \dot{g}_n(\tau)dV_n(\tau)\Big]ds + o_p(1). \tag{C.12}$$

That is, we can replace $\hat{V}_n$ by V_n. By comparing $\tilde{W}_n$ above with W_n in (A.3), we only need to show that, uniformly in $r \in [0, 1-\epsilon]$,

$$\int_0^r \Big[\dot{g}_n(s)' C_n^{-1}(s) \int_s^1 \dot{g}_n(\tau) dV_n(\tau)\Big] ds - \int_0^r \Big[\dot{g}(s)' C^{-1}(s) \int_s^1 \dot{g}(\tau) dV_n(\tau)\Big] ds = o_p(1). \quad \text{(C.13)}$$

Let $b_n(s) = \dot{g}_n(s)' C_n(s)^{-1} \int_s^1 \dot{g}_n dV_n$ and $b(s) = \dot{g}(s) C(s)^{-1} \int_s^1 \dot{g}' dV_n$. Then (C.13) is equivalent to $\int_0^r [b_n(s) - b(s)] ds = o_p(1)$ uniformly in $r \in [0, 1-\epsilon]$. Now

$$\begin{aligned} b_n(s) - b(s) &= \dot{g}_n' [C_n^{-1}(s) - C^{-1}(s)] \textstyle\int_s^1 \dot{g} dV_n \\ &+ (\dot{g}_n - \dot{g})' C^{-1}(s) \textstyle\int_s^1 \dot{g} dV_n \\ &+ \dot{g}_n' C_n^{-1}(s) \textstyle\int_s^1 (\dot{g}_n - \dot{g}) dV_n. \end{aligned} \quad \text{(C.14)}$$

We show each of the three terms on r.h.s. of (C.14) is small. Denote $h_n = \dot{g}_n' - \dot{g}'$. We have $\int_s^1 \dot{g}_n \dot{g}_n' = \int_s^1 \dot{g}\dot{g}' + \int_s^1 \dot{g} h_n' + \int_s^1 h_n \dot{g}' + \int_s^1 h_n h_n'$. Furthermore, $\| \int_s^1 h_n h_n' \| \le \int_0^1 \|h_n\|^2 = o_p(1)$ and $\| \int_s^1 \dot{g} h_n \|^2 \le (\int_0^1 \|\dot{g}\|^2)(\int_0^1 \|h_n\|^2) = o_p(1)$. From the matrix algebra $A^{-1} - (A+B)^{-1} = A^{-1}B(A+B)^{-1}$, applied with $A = \int \dot{g}\dot{g}'$ and $B = \int g h_n' + \int h_n \dot{g}' + \int h_n h_n'$ and noticing that $\|B\| = o_p(1)$, we have

$$\|(\int_s^1 \dot{g}_n \dot{g}_n' dr)^{-1} - (\int_s^1 \dot{g}\dot{g}' dr)^{-1}\| \le o_p(1) \|(\int_s^1 \dot{g}_n \dot{g}_n' dr)^{-1}\| \cdot \|(\int_s^1 \dot{g}\dot{g}' dr)^{-1}\|. \quad \text{(C.15)}$$

For $s \le 1-\epsilon$, we have

$$\int_s^1 \dot{g}_n \dot{g}_n' dr \ge \int_{1-\epsilon}^1 \dot{g}_n \dot{g}_n' dr \quad \text{and} \quad \int_s^1 \dot{g}\dot{g}' dr \ge \int_{1-\epsilon}^1 \dot{g}\dot{g}' dr$$

where $R \ge Q$ means that $R - Q$ is positive semi-definite. Using the fact that if $R > Q$, then $\|Q^{-1}\| \le \|R^{-1}\|$, we can rewrite (C.15) as, for all $s \le 1-\epsilon$,

$$\|C_n(s)^{-1} - C(s)^{-1}\| = o_p(1)$$

because $\|(\int_{1-\epsilon}^1 \dot{g}_n \dot{g}_n')^{-1}\| = O_p(1)$ and $\|(\int_{1-\epsilon}^1 \dot{g}\dot{g}')^{-1}\| = O(1)$.

Next, because $\int_s^1 \dot{g} dV_n = n^{-1/2} \sum_{t=1}^n [(U_t \ge s)\dot{g}(U_i) - E\{(U_t \ge s)\dot{g}(U_t)\}]$ is $O_p(1)$ uniformly in s by the functional central limit theorem, we have

$$\|\dot{g}_n' [C_n(s)^{-1} - C(s)^{-1}] \int_s^1 \dot{g} dV_n\| = \|\dot{g}_n\| o_p(1) O_p(1) = \|\dot{g}_n\| o_p(1); \quad \text{(C.16)}$$

$$\|(\dot{g}_n - \dot{g})' C(s)^{-1} \int_s^1 g dV_n\| \le \|\dot{g}_n - \dot{g}\| \cdot \|C(1-\epsilon)^{-1}\| O_p(1) = \|\dot{g}_n - \dot{g}\| O_p(1). \quad \text{(C.17)}$$

Finally, by condition (17),

$$\|\dot{g}_n' C_n(s)^{-1} \int_s^1 (\dot{g}_n - \dot{g}) dV_n\| \le \|\dot{g}_n\| \|C_n(1-\epsilon)^{-1}\| \| \int_s^1 (\dot{g}_n - \dot{g}) dV_n\| = \|\dot{g}_n\| o_p(1). \quad \text{(C.18)}$$

From (C.14), and combining (C.16)-(C.18), we see that for $s \leq 1-\epsilon$, $\|b_n(s) - b(s)\| = \|\dot{g}_n\| o_p(1) + \|\dot{g}_n - \dot{g}\| O_p(1)$. Together with (16), it follows that, for $r \leq 1-\epsilon$,

$$\| \int_0^r [b_n(s) - b(s)] ds \| \leq \int_0^r \|b_n(s) - b(s)\| ds \leq o_p(1) (\int_0^1 \|\dot{g}_n\|^2 dr)^{1/2}$$

$$+ O_p(1) (\int_0^1 \|\dot{g}_n - \dot{g}\|^2 dr)^{1/2} = o_p(1) O_p(1) + O_p(1) o_p(1) = o_p(1).$$

The proof of Theorem 4 is complete. □

Proof of Theorem 5. Assuming no information truncation for simplicity, we suppress Ω_t in the conditional distribution. We have the identity,

$$U_t^* = G_t(F_t^{-1}(\hat{U}_t|\hat{\theta}) \mid \theta_0).$$

Thus $\hat{U}_t \leq r$ if and only if

$$\begin{aligned} U_t^* &\leq G_t(F_t^{-1}(r|\hat{\theta}) \mid \theta_0) \\ &= F_t(F_t^{-1}(r|\hat{\theta}) \mid \theta_0) + e_t(r) \\ &= \xi_t(r, \hat{\theta}, \theta_0) + e_t(r) \end{aligned}$$

where $e_t(r) = \delta n^{-1/2} [H_t(F_t^{-1}(r|\hat{\theta}) \mid \theta_0) - F_t(F_t^{-1}(r|\hat{\theta})|\theta_0)]$. Thus, adding and subtracting terms, we have

$$\hat{V}_n(r) = n^{-1/2} \sum_{t=1}^n [I(U_t^* \leq r) - r] + d_n^\dagger(r) + R_n^\dagger(r)$$

where

$$d_n^\dagger(r) = n^{-1/2} \sum_{t=1}^n [\xi_t(r, \hat{\theta}, \theta_0) - r] + n^{-1/2} \sum_{t=1}^n e_t(r) \tag{C.19}$$

$$R_n^\dagger(r) = n^{-1/2} \sum_{t=1}^n \Big[I(U_t^* \leq \xi_t(r, \hat{\theta}, \theta_0) + e_t(r)) - \xi_t(r, \hat{\theta}, \theta_0) - e_t(r) - I(U_t^* \leq r) + r \Big]. \tag{C.20}$$

The first term on the r.h.s. of (C.19) is $-\bar{g}(r)' \sqrt{n}(\hat{\theta} - \theta_0) + o_p(1)$, and the second term is

$$n^{-1/2} \sum_{t=1}^n e_t(r) = \delta \frac{1}{n} \sum_{t=1}^n [H_t(F_t^{-1}(r|\hat{\theta}) \mid \theta_0) - F_t(F_t^{-1}(r|\hat{\theta}) \mid \theta_0)]$$

$$= \delta \frac{1}{n} \sum_{t=1}^n [H_t(F_t^{-1}(r|\hat{\theta})|\theta_0) - r] + \delta \frac{1}{n} \sum_{t=1}^n [F_t(F_t^{-1}(r|\hat{\theta})|\theta_0) - r].$$

The first term converges to $\delta k(r)$ in probability, and the second term is $O_p(n^{-1/2})$ by the Taylor expansion, assumptions A2 and A3. Thus

$$d_n^\dagger(r) = -\bar{g}(r)' \sqrt{n}(\hat{\theta} - \theta_0) + \delta k(r) + o_p(1).$$

Finally, the proof of $R_n^\dagger(r) = o_p(1)$ is similar to that of $R_n^*(r) = o_p(1)$. Next,

$$\begin{aligned}\hat{W}_n &= \hat{V}_n - \phi_g(\hat{V}_n) \\ &= V_n^* - \phi_g(V_n^*) + \delta k - \delta\phi_g(k) + o_p(1).\end{aligned}$$

Note that $V_n^* - \phi_g(V_n^*) \Rightarrow W$, a Brownian motion. The desired result follows. □

Proof of Lemma 2.

The proof is easy once the right approach is discovered. Differentiate the identity (23) on both sides we have

$$\dot{k}(r) = \dot{g}(r)'C(r)^{-1}\int_r^1 \dot{g}(v)\dot{k}(v)dv. \tag{C.21}$$

Let $a(x)$ be the vector function

$$a(r) = C(r)^{-1}\int_r^1 \dot{g}(v)\dot{k}(v)dv$$

we have

$$\dot{k}(r) = \dot{g}(r)'a(r).$$

We next show $a(r)$ is a constant vector by showing $\dot{a}(r) \equiv 0$. Note that the derivative of the inverse matrix $C(r)^{-1}$ is given by $C(r)^{-1}\dot{C}(r)C(r)^{-1} = C(r)^{-1}\dot{g}\dot{g}'C(r)^{-1}$. Thus

$$\begin{aligned}\dot{a}(r) &= C(r)^{-1}\dot{g}\dot{g}'C(r)^{-1}\int_r^1 \dot{g}(v)\dot{k}(v)dv \\ &- C(r)^{-1}\dot{g}(r)\dot{k}(r) \\ &= C(r)^{-1}\dot{g}(r)\dot{k}(r) \qquad \text{by } (C.21) \\ &- C(r)^{-1}\dot{g}(r)\dot{k}(r) \\ &= 0\end{aligned}$$

Thus $\dot{a}(r) \equiv 0$ and $a(r)$ is a constant vector and is denoted by a. It follows that $\dot{k}(r) = \dot{g}(r)'a$. Integrating this new identity on both sides we obtain $k(r) = g(r)'a + c$, where c is a constant. But for $c \neq 0$, $k(r)$ does not satisfy the integral equation. So $k(r) = a'g(r)$ is the only solution.

References

[1] Andersen, P. K., O. Borgan, R.D. Gill, and N. Keiding (1993): *Statistical Models Based on Counting Processes*, New York: Springer-Verlag.

[2] Andrews, D.W.K. (1997): "A conditional Kolmogorov test," *Econometrica, 65, 1097-1128.*

[3] Bai, J. (1994): "Weak convergence of sequential empirical processes of residual in ARMA models," *Annals of Statistics*, 22, 2051-2061.

[4] Bai, J. (1996): "Testing for parameter constancy in regression models: an empirical distribution function approach," *Econometrica*, 64, 597-622.

[5] Bai, J. and S. Ng (1998): "A consistent test for conditional symmetry of time series models," forthcoming *J. of Econometrics*,

[6] Bera, A. and C. Jarque (1982): "Model specification tests: a simultaneous approach." *Journal of Econometrics*, 20, 1982, 59-82.

[7] Bollerslev, T. (1986) "Generalized autoregressive conditional heteroscedasticity," *Journal of Econometrics,* 31, 307-327.

[8] Bollerslev, T. (1987): "A conditionally heteroskedastic time series model for speculative prices and rates of return," *Review of Economics and Statistics* 69, 542-547.

[9] Brown, R.L., J. Durbin, and J.M. Evans (1975): "Techniques for testing the constancy of regression relationships over time," *Journal of Royal Statistical Society, Series B*, 37, 149-192.

[10] Chan, K.S. (1993), "Consistency and limiting distribution of the least squares estimator of a continuous autoregressive model," *Annals of Statistics*, 21, 520-533.

[11] Diebold, F.X., T. Gunther, and A. Tay (1998): "Evaluating density forecasts, with applications to financial risk management," *International Economic Review*, 39, 863-883.

[12] Durbin, J. (1973a): "Weak convergence of sample distribution functions when parameters are estimated," *Annals of Statistics*, 1, 279-290.

[13] Durbin, J. (1973b): *Distribution Theory for Tests Based on Sample Distributions Function*, J.W. Arrowsmith Ltd. England.

[14] Engle, R. (1982) "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation," *Econometrica*, 50 987-1006.

[15] Fama, E. (1965). "The behavior of stock market prices," *Journal of Business*, 38, 34-105.

[16] Fan, Y.Q. (1994): "Testing the goodness-of-fit test of a parametric density function," *Econometric Theory*, 10, 316-356.

[17] Horowitz, J.L., and V. Spokoiny (2001), "An adaptive, rate optimal test of a parametric mean-regression model agaist a nonparametric alternative," *Econometrica*, 689, 599-631.

[18] Inoue, A. (1997): "A conditional goodness-of-fit test in time series," manuscript, Department of Economics, University of Pennsylvania.

[19] Khmaladze, E.V. (1981): "Martingale approach in the theory of goodness-of-tests," *Theory of probability and its applications*, XXVI, 240-257.

[20] Khmaladze, E.V. (1988): "An innovation approach to goodness-of-fit tests in R^n," *Annals of Statistics*, 16, 1503–1516.

[21] Khmaladze, E.V. (1993): "Goodness of fit problem and scanning innovation martingales," *Annals of Statistics*, 21, 798-829.

[22] Koul, H.L. (1996): "Asymptotics of some estimators and sequential residual empiricals in nonlinear time series," *Annals of Statistics*, 24, 380-404.

[23] Koul, H.L. and W. Stute (1999): "Nonparametric model checks for time series," *Annals of Statistics*, Vol 27, 204-236.

[24] Lee, S.W. and B.E. Hansen (1994): "Asymptotic theory for the GARCH(1,1) quasi-maximum likelihood estimator," *Econometric Theory*, 10, 29-52.

[25] Linton, O. and P. Gozalo (1996): "Conditional Independence restriction: testing and estimation," Manuscript, Department of Economics, Yale University.

[26] Loynes, R.M. (1980): "The empirical distribution function of residuals from generalized regression," *Annals of Statistics*, 8, 285-298.

[27] Lumsdaine, R. (1996): "Consistency and asymptotic normality of the quasi- maximum likelihood estimator in IGARCH(1,1) and covariance stationary GARCH(1,1) models," *Econometrica*, 64, 575-596.

[28] Newey, W. and D.G. Steigerwald (1997): "Asymptotic bias for quasi-maximum likelihood estimators in conditional heteroskedasticity models," *Econometrica*, 65, 587-599.

[29] Pareto, V. (1897). *Cours d'economie politique*, Vol. 2, Paris: F. Pichou.

[30] Pollard, D. (1984). *Convergence of Stochastic Processes*, Springer-Verlag, New York.

[31] Rosenbaltt, M. (1952): "Remarks on a multivariate transformation," *Annals of Mathematical Statistics*, 23, 470-472.

[32] Sargan, J.D. (1957): "The distribution of wealth," *Econometrica*, 25, 568-590.

[33] Stinchcombe, M.B. and H. White (1993): "Consistent specification testing with unidentified nuisance parameters using duality and Banach space limit theory," Discussion paper 93-14, Department of Economics, University of California, San Diego.

[34] Thompson, S. (2000): "Specification tests for continuous time models," unpublished manuscript, Department of Economics, Harvard University,.

[35] Tong, H. (1990): *Non-linear Time Series Analysis: A Dynamical System Approach*. Oxford University Press.

[36] Wooldridge, J. (1990). "A unified approach to robust, regression-based specification tests," *Econometric Theory*, 6, 17-43.

[37] Zheng, J.X. (1994): "A specification test of conditional parametric distribution using kernel estimation methods," manuscript, Department of Economics, University of Texas, Austin.

[38] Zheng, J.X. (2000): "A consistent test of conditional parametric distribution," *Econometric Theory*, Vol 16, 667-691.