Testing for Multivariate Volatility Functions Using Minimum Volume Sets and Inverse Regression

Wolfgang Polonik[1] and Qiwei Yao[2,3]

[1]Department of Statistics, University of California, Davis, CA95616-8705, USA

[2]Department of Statistics, London School of Economics, London, WC2A 2AE, UK

[3]Guanghua School of Management, Peking University, Beijing 100871, China

Email: polonik@wald.ucdavis.edu and q.yao@lse.ac.uk

Abstract

We propose two new types of nonparametric tests for investigating multivariate regression functions. The tests are based on cumulative sums coupled with either minimum volume sets or inverse regression ideas; involving no multivariate nonparametric regression estimation. The methods proposed facilitate the investigation for different features such as if a multivariate regression function is (i) constant, (ii) of a bathtub shape, and (iii) of a given parametric form. The inference based on those tests may be further enhanced through associated diagnostic plots. Although the potential use of those ideas is much wider, we focus on the inference for multivariate volatility functions in this paper, i.e. we test for (i) heteroscedasticity, (ii) the so-called "smiling effect", and (iii) some parametric volatility models. The asymptotic behavior of the proposed tests is investigated, and practical feasibility is shown via simulation studies. We further illustrate our methods with real financial data.

Keywords: Brownian bridge, empirical process, ARCH models, heteroscedasticity, integral stochastic order, level set, smiling effect.

1 Introduction

We propose and study two types of nonparametric tests for investigating if a multivariate regression function is, for instance, constant, of a bathtub shape, or of a particular parametric form. The methodology has the potential to be useful in various contexts, including regression analysis and the analysis of time series. All the procedures proposed are associated with diagnostic plots.

In terms of methodology, our approach may be seen as a generalization of the classical goodness-of-fit tests for distribution functions (such as the Kolmogorov-Smirnov test) to those for regression functions. While most the classical goodness-of-fit tests for one-dimensional distribution functions are asymptotically distribution-free under the null hypotheses, this nice property is typically lost in multivariate cases. This explains the difficulties in directly applying, for example, the Kolmogorov-Smirnov tests for multivariate distribution functions (Polonik 1999). We circumvent this problem by using either minimum volume (MV) sets or an inverse regression idea. With MV sets, we effectively test a multivariate function in term of a single-indexed empirical process. The tests based on inverse regression rely on several one-dimensional empirical processes. Therefore the asymptotic distribution-free properties may be restored. The idea of using MV sets was initially proposed by Polonik (1999) for the goodness-of-fit tests for multivariate distribution functions.

We illustrate the new methods in the context of testing various features of volatility functions, which is particularly relevant to analyzing financial time series. Let $\{Y_t\}$ be a strictly stationary and ergodic time series defined by

$$Y_t = \sigma_t \varepsilon_t, \tag{1.1}$$

where $\sigma_t \geq 0$ is $\mathcal{F}_{t-1}$-measurable, $\mathcal{F}_t$ denotes the σ-algebra generated by $\{Y_{t-k},\, k \geq 0\}$, and $\{\varepsilon_t\}$ is a sequence of independent and identically distributed random variables with mean 0, $E(\varepsilon_t^2) < \infty$, $E|\varepsilon_t| = 1$. Furthermore, we assume that ε_t is independent of $\mathcal{F}_{t-1}$. Now it is easy to see that $E(Y_t|\mathcal{F}_{t-1}) = 0$, and $E(|Y_t|\big|\mathcal{F}_{t-1}) = \sigma_t$. Hence $\sigma_t \equiv \sigma(Y_{t-1}, Y_{t-2}, \cdots)$ is a regression function of $|Y_t|$ on $Y_{t-1}, Y_{t-2}, \cdots$, and is called a volatility function. In fact (1.1) is a standard setting for modeling volatilities of financial returns (see, e.g. Morgan, 1996, p.92), although the conventional assumption is $E(\varepsilon_t^2) = 1$. This implies $E(Y_t^2|\mathcal{F}_{t-1}) = \sigma_t^2$. We adopt the parametrization implied by the condition $E|\varepsilon_t| = 1$ instead in order to relaxes the moment conditions required in the inference.

We consider three types of null hypotheses on σ_t, i.e. homoscedasticity, a "smile effect" and a specified parametric form such as ARCH models. For the first two cases, our tests are asymptotically distribution-free under the null hypotheses. However the tests for parametric models are not distribution-free due to the presence of the estimators for the parameters in the test statistics, for which two bootstrap methods are used to approximate the P-values.

There exists a large body of literature on testing regression functions. This includes, for example, tests based on analysing residuals (Eubank and Hart 1993, Härdle and Mammen 1993, and Art-Sahalia et al 2001), residual-regression based tests (Bierens 1987, Chen and An 1997, and Koul and Stude 1999), and score-type tests (Cook and Weisberg 1982, and Stute and Zhu 2005). The tests proposed in this paper may be viewed as residual-regression based methods for testing multivariate volatility functions. We restrict our attention on model (1.1) which assumes a zero-mean. Undoubtedly this simplifies our investigation significantly. Since for financial returns, the magnitudes of drift terms are typically negligible comparing to those of volatility functions. This is why (1.1) is taken as the standard volatility model in, for example, the RiskMetrics of Morgan (1996). For testing volatilities with non-zero means, we refer to Casas et al (2007), Wu and Zhu (2007) and the references within.

The rest of the paper is organized as follows. Section 2 deals with the tests for homoscedasticity. Section 3 extends the ideas for testing parametric forms of volatility functions. The tests for a bathtub shaped volatility (i.e. "smiling factor") is discussed in section 4. Numerical illustration with both simulated and real data is presented in section 5. Proofs of all the theoretical results are presented in Polonik and Yao (2007) which is effectively an extended version of this paper.

2 Tests for homoscedasticity

In this section we deal with the tests for the homoscedasticity hypothesis

$$H_0 \,:\, \sigma(\cdot) \equiv \nu_y, \qquad \nu_y > 0 \text{ is a constant.} \tag{2.1}$$

Conventional practice is to test the null hypothesis (2.1) against a specified parametric form such as ARCH models; see, e.g., section 4.2 of Fan and Yao (2003) and references therein. More recently, a nonparametric approach has been adopted for testing for conditional heteroscedasticity for univariate

volatility functions; see Chen and An (1997) and Laïb (2003). In terms of methodology, the available tests may be classified into two categories: tests solely based on analyzing residuals (Engle 1982, Lee 1991, McLeod and Li 1993, and Horváth et al. 2001), and tests based on residual-regression (Chen and An, 1997, Stute, 1997, Koul and Stute, 1999, and Laïb, 2003). The latter is based on the fact that under the null-hypothesis (2.1), it holds that

$$E\{(\,|Y_t| - E|Y_t|\,)\, I(\,\mathbf{X}_t \le \mathbf{x}\,)\,\} = 0 \qquad \text{for all } \mathbf{x} \in \mathcal{R}^p, \tag{2.2}$$

where $\mathbf{X}_t = (Y_{t-1}, \cdots, Y_{t-p})'$ $(p \ge 1)$, and $\mathbf{X}_t \le \mathbf{x}$ denotes that each component of $\mathbf{X}_t$ is not greater than the corresponding component of $\mathbf{x}$. All the work mentioned above in the second category deals with univariate regressor only. The methods proposed in this paper may be viewed as an attempt to extend these methods from the second category to multivariate cases.

Our new tests are based on the following observation: under (2.1) it holds that $F \equiv G$, where $F(\cdot)$ denotes the distribution function of $\mathbf{X}_t$, and G is a distribution function defined as

$$G(\mathbf{x}) = \nu_y^{-1} E\{|Y_t| I(\mathbf{X}_t \le \mathbf{x})\}, \tag{2.3}$$

where $\nu_y = E|Y_t|$. It is easy to see that $G(\cdot)$ is a well-defined probability measure. Thus hypothesis (2.1) may be viewed as a hypothesis on two probability distributions. Although no observations are directly available from distribution G, $F(\mathbf{x})$ and $G(\mathbf{x})$ may be estimated by, respectively,

$$F_n(\mathbf{x}) = \frac{1}{n}\sum_{t=1}^n I(\mathbf{X}_t \le \mathbf{x}), \qquad G_n(\mathbf{x}) = \frac{1}{n\,\widehat{\nu}_y}\sum_{t=1}^n |Y_t| I(\mathbf{X}_t \le \mathbf{x}),$$

where $\widehat{\nu}_y = n^{-1}\sum_{1\le t\le n}|Y_t|$. Hence, we may test hypothesis (2.1) using the statistic

$$\sup_{\mathbf{x}} \widehat{\nu}_y \left| G_n(\mathbf{x}) - F_n(\mathbf{x}) \right| \;=\; \sup_{\mathbf{x}} \left| \frac{1}{n}\sum_{t=1}^n \left\{ |Y_t| - \widehat{\nu}_y \right\} I(\mathbf{X}_t \le \mathbf{x}) \right|, \tag{2.4}$$

When $p = 1$, tests of this type have been extensively explored by, among others, Chen and An (1997), Stute (1997), Koul and Stute (1999) and Laïb (2003). For $p > 1$ the null-distribution of the test statistic (2.4) depends on the underlying distribution. Note that the lack of the (asymptotic) distribution-free property may cause non-trivial difficulties in determining the critical values of the tests since the null hypothesis $F = G$ is not simple.

We construct the tests using minimum volume (MV) sets or an inverse regression idea. A remarkable gain for these new approaches is that the null-distributions of our test statistics are asymptotically distribution-free even when $p > 1$. Note that for testing the null hypothesis (2.1), one could

simply let $p = 1$ (i.e. $\mathbf{X}_t = Y_{t-1}$) in (2.4). We argue that the tests with $p > 1$ are significantly more powerful than those with $p = 1$ when σ_t depends on several lagged values of Y_t. Numerical results in section 5 provide convincing evidence to support this argument. Note for GARCH(1,1) processes, σ_t depends on Y_{t-k} for all $k \geq 1$.

2.1 Tests based on MV sets

We first introduce the concept of minimum volume (MV) sets. An MV set under distribution F with pdf f on $\mathcal{R}^d$, indexed by $\alpha \in [0, 1]$, is defined as

$$M_f(\alpha) = \arg \min_{A \subset \mathcal{R}^p} \{\mathrm{Leb}(A) : \; F(A) \geq \alpha\}, \tag{2.5}$$

where Leb(A) denotes the Lebesgue measure of A. Obviously $M_f(\alpha)$ is a set of the minimum Lebesgue measure among the sets of F-measure not smaller than α. When f has no flat parts (see (2.9) below), MV sets exist and are essentially unique (up to F-nullsets). In fact, in this case $M_f(\alpha) = \{\, \mathbf{x} : f(\mathbf{x}) \geq \lambda_\alpha \,\}$ for an appropriate constant $\lambda_\alpha \geq 0$. The MV sets $M_g(\alpha)$ under distribution G may be defined in the similar manner. We denote by g the probability density function of G.

One of the reasons to use MV sets for constructing our tests is that they are capable of discriminating different distributions. Suppose that both f and g do not have flat parts in the sense of (2.9). Polonik (1999) showed that $F = G$ if and only if $(F - G)\{M_f(\alpha)\} = 0$ and $(F - G)\{M_g(\alpha)\} = 0$ for all $\alpha \in [0, 1]$.

Estimators for MV sets may be obtained by replacing F in (2.5) by the estimator F_n. In order to avoid "oversmoothing", we need to apply some restriction on the candidate sets A in (2.5). Let $\mathcal{C}$ be a set consisting of appropriate subsets of $\mathcal{R}^p$. An estimator for $M_f(\alpha)$ may be defined as

$$\widehat{M}_{\mathcal{C},f}(\alpha) = \arg \min_{A \in \mathcal{C}} \{\mathrm{Leb}(A) : \; F_n(A) \geq \alpha\}, \tag{2.6}$$

which is called an empirical MV set in $\mathcal{C}$. We should choose $\mathcal{C}$ such that it contains all the MV sets under both F and G. Under hypothesis (2.1) we have $E\{|Y_t| - \nu_y \big| \mathbf{X}_t \in A\} = 0$ for all $A \in \mathcal{C}$; see also (2.2). The latter is equivalent to

$$(G - F)(A) = 0 \quad \text{for all } A \in \mathcal{C}. \tag{2.7}$$

Let $\widehat{M}_{\mathcal{C},g}(\alpha)$ be the empirical MV sets in $\mathcal{C}$ under G_n; see (2.6). Put

$$\widehat{\Psi}_{\mathcal{C},f}(\alpha) = (F_n - G_n)\{\widehat{M}_{\mathcal{C},f}(\alpha)\}, \qquad \widehat{\Psi}_{\mathcal{C},g}(\alpha) = (F_n - G_n)\{\widehat{M}_{\mathcal{C},g}(\alpha)\}. \tag{2.8}$$

Relation (2.7) suggests that we may define a test statistic

$$T_1 = \sup_{\alpha \in (0,1]} \big[\, |\widehat{\Psi}_{\mathcal{C},g}(\alpha)| + |\widehat{\Psi}_{\mathcal{C},f}(\alpha)| \,\big],$$

and reject hypothesis $F = G$ for large values of T_1. We assume that $\mathcal{C}$ is chosen such that $M_f(\alpha) \in \mathcal{C}$ and $M_g(\alpha) \in \mathcal{C}$ for all α. This assumption may be interpreted as a correctly chosen model. Observe that in this case (2.7) is *equivalent* to $F = G$. For technical reasons we also assume that $\emptyset \in \mathcal{C}$.

In general the class $\mathcal{C}$ should be rich enough to distinguish F from G when H_0 does not hold. On the other hand, $\mathcal{C}$ may not be too large, so that all the MV sets may be consistently estimated by their empirical counterparts. Complexity of the class $\mathcal{C}$ may be measured by its metric entropy which leads to the concept of a covering integral; see the discussion in the beginning of section 6 of Polonik and Yao (2007). Most results below require a finite covering integral. The classes with finite covering integrals include the sets of balls, rectangles, ellipsoids, and convex sets in $\mathcal{R}^p$. In order to make the computation attainable, we typically let $\mathcal{C}$ consist of all ellipsoids in applications. This effectively imposes a shape constraint on the underlying distributions. Note that our theoretical results continue to hold (under additional assumptions) even if the true MV sets are not ellipsoids, although some power losses in the tests may occur. Ellipsoids appear to be an acceptable choice balancing out the trade-off between the computational feasibility and the practical effectiveness. Note also that elliptical distributions have featured in modern stochastic finance; see, for example, Bingham and Kiesel (2002), Hamada and Valdez (2004), and Pelagatti and Rodena (2005).

Assumption 2.1

(i) F and G have bounded and continuous Lebesgue densities f and g respectively.

(ii) The densities f and g have no flat parts, i.e.

$$\sup_{\lambda > 0} F\,\{\mathbf{x} \in \mathcal{R}^p : \, |f(\mathbf{x}) - \lambda\,| \le \eta\,\} \to 0 \qquad \text{as } \eta \to 0. \tag{2.9}$$

and the same holds with (F, f) replaced by (G, g).

Theorem 2.2 *Suppose that the assumptions of Theorem 2.1 in Polonik and Yao (2007) hold. Then as $n \to \infty$, it holds under null hypothesis $F = G$ that*

$$\sqrt{\frac{n\,\widehat{\nu}_y^2}{\widehat{\sigma}_y^2}}\,T_1 \;\overset{\mathcal{D}}{\to}\; 2\sup_{\alpha\in[0,1]}|B(\alpha)|$$

where $\widehat{\sigma}_y^2 = n^{-1}\sum_{1\le t\le n} Y_t^2 - \widehat{\nu}_y^2$, and B denotes the standard Brownian bridge process. On the other hand, if $F \neq G$, $P(T_1 > c) \to 1$ for some constant $c > 0$.

Combining the result of the above theorem and (9.39) of Billingsley (1999), we have the approximation:

$$P\Big(\sqrt{\frac{n\,\widehat{\nu}_y^2}{\widehat{\sigma}_y^2}}T_1 > z\Big) \;\approx\; 2\sum_{k=1}^{\infty}(-1)^{k-1}\exp\Big\{-\frac{k^2z^2}{2}\Big\}. \tag{2.10}$$

Diagnostic plots. The test based on statistic T_1 naturally leads to two diagnostic plots $\alpha \to \Big(G_n\{\widehat{M}_{\mathcal{C},g}(\alpha)\},\ F_n\{\widehat{M}_{\mathcal{C},g}(\alpha)\}\Big)$ and $\alpha \to \Big(G_n\{\widehat{M}_{\mathcal{C},f}(\alpha)\},\ F_n\{\widehat{M}_{\mathcal{C},f}(\alpha)\}\Big)$ which are called CC-plots (Polonik 1999). They may be viewed as a generalization of the standard QQ-plots for univariate distributions to multivariate distributions. Under null hypothesis (2.1), both plots should be approximately a 45^o straight line.

2.2 Tests based on inverse regression

Although the test T_1 is based on an empirical process indexed by a one-dimensional parameter, we still need to compute the MV sets in $\mathcal{R}^p$. In this subsection, we swap the roles of Y_t and $\mathbf{X}_t$; leading to a test based on one-dimensional sets only. The key idea is to use an inverse regression equation, i.e. under the null hypothesis (2.1),

$$E\big[(\mathbf{Z}_t - \boldsymbol{\nu}_y)\,I\{|Y_t| < y\}\big] = 0 \qquad \text{for all } y \ge 0, \tag{2.11}$$

where $\mathbf{Z}_t = (\,|Y_{t-1}|, \ldots, |Y_{t-p}|\,)'$, and $\boldsymbol{\nu}_y$ is a $p \times 1$ vector with all elements equal to $\nu_y = E|Y_t|$. Hence we may define a test statistic

$$T_2 = \frac{1}{n}\max_{1\le j\le p}\sup_x\Big|\sum_{t=1}^{n}\big(\,|Y_{t-j}| - \widehat{\nu}_{y,j}\,\big)\,I(|Y_t| \le x)\Big|,$$

and reject hypothesis $F = G$ for large values of T_2, where $\widehat{\nu}_{y,j} = n^{-1}\sum_{1\le t\le n}|Y_{t-j}|$.

Theorem 2.3 *Suppose $\sqrt{n}\,(\widehat{\nu}_y - \nu_y\,) = O_P(1)$. As $n \to \infty$, it holds under null hypothesis (2.1) that*

$$\sqrt{\frac{n\,\widehat{\nu}_y^2}{\widehat{\sigma}_y^2}}\,T_2 \xrightarrow{\mathcal{D}} \max_{1\le j\le p}\ \sup_{0\le\alpha\le 1} |B_j(\alpha)|\,,$$

where $B_1, \cdots, B_p$ denote p independent standard Brownian bridge processes. On the other hand, if (2.11) does not hold, $P(T_2 > c) \to 1$ for some constant $c > 0$.

Due to the independence of the limiting Brownian bridge processes, it holds that

$$P\{\max_{1\le j\le p}\sup_{0\le\alpha\le 1}|B_j(\alpha)| \ge z\} = 1 - \big[\,1 + 2\sum_{k=1}^{\infty}(-1)^k \exp\big\{-2\,k^2 z^2\big\}\,\big]^p,$$

see, e.g., (9.39) of Billingsley (1999).

Note that (2.11) in general does not imply $F = G$, and for such cases the test based on T_2 has no power. We argue that such a situation is rare in practice. The advantage of using one-dimensional sets in T_2 brings in considerable convenience in practice, in spite of the fact that it is not an omnibus test for the conditional heteroscedasticity. The inverse-regression idea has been employed for testing mean functions by Zhu (2003), and Khmaladze and Koul (2004).

Diagnostic plots. It is easy to see that under the null hypothesis (3.1), the plots

$$y \;\to\; \Big(\frac{1}{\sqrt{n}}\sum_{t=1}^{n}|Y_{t-j}|I(|Y_t|\le y),\ \frac{\widehat{\nu}_{y,j}}{\sqrt{n}}\sum_{t=1}^{n}I(|Y_t|\le y)\Big), \quad j = 1, \cdots, p$$

should all approximately be 45^o lines through the origin.

3 Tests for parametric heteroscedasticity

The methods presented above may be formally extended to testing for the parametric heteroscedasticity hypothesis

$$H_0:\ \sigma_t = \sigma_0(\mathbf{X}_t, \boldsymbol{\theta}_0) \quad F\text{-a.s. for some } \boldsymbol{\theta}_0 \in \boldsymbol{\Theta}, \tag{3.1}$$

where $\mathbf{X}_t = (Y_{t-1}, \cdots, Y_{t-p})'$, Y_t is defined by (1.1), the form of function σ_0 is known, and $\boldsymbol{\Theta} \subset \mathcal{R}^q$, and $p, q \ge 1$ are integers. For example, for $q = p + 1$ and $\sigma_0(\mathbf{x}, \boldsymbol{\theta})^2 = \theta_1 + \theta_2 x_1^2 + \cdots + \theta_{p+1}x_p^2$, we test the validation of ARCH(p) model.

3.1 Tests based on MV sets

Put $G_{\boldsymbol\theta}(A) = \frac{1}{\nu_\theta} E_{\boldsymbol\theta}\Big\{\frac{|Y_t|}{\sigma_0(\mathbf{X}_t,\boldsymbol\theta)} I(\mathbf{X}_t \in A)\Big\}$ where $\nu_\theta = E_\theta\{|Y_t|/\sigma_0(\mathbf{X}_t,\boldsymbol\theta)\}$ is a normalizing constant, and $E_{\boldsymbol\theta}$ denotes expectation taken with $\sigma(\cdot) = \sigma_0(\cdot,\boldsymbol\theta)$ in (1.1). It is easy to see that $G_{\boldsymbol\theta}(\cdot)$ is a well-defined probability measure on $\mathcal{R}^p$. Furthermore, the null hypothesis (3.1) holds if and only if $G_{\boldsymbol\theta_0} \equiv F$. The latter is equivalent to $(G_{\boldsymbol\theta_0} - F)(A) = 0$ for all $A \in \mathcal{C}$, if, for example, $\mathcal{C}$ contains all the MV sets under F and $G_{\boldsymbol\theta_0}$. Hence we may construct a test statistic based on a sample version of the above expression. To this end, let $\widehat{\boldsymbol\theta}$ be an estimator for $\boldsymbol\theta_0$ and define

$$e_{t,\widehat{\boldsymbol\theta}} = |Y_t|/\sigma_0(\mathbf{X}_t, \widehat{\boldsymbol\theta}), \qquad \text{and} \qquad G_{n,\widehat{\boldsymbol\theta}}(A) = \frac{1}{n\,\widehat\nu_{\widehat{\boldsymbol\theta}}} \sum_{t=1}^n e_{t,\widehat{\boldsymbol\theta}}\, I(\mathbf{X}_t \in A), \tag{3.2}$$

where $\widehat\nu_{\widehat{\boldsymbol\theta}} = \frac{1}{n}\sum_{t=1}^n e_{t,\widehat{\boldsymbol\theta}}$. Let $\widehat M_{\mathcal{C}}(\alpha)$ and $\widehat M_{\mathcal{C},\widehat{\boldsymbol\theta}}(\alpha)$ be the empirical MV set under, respectively, F_n and $G_{n,\widehat{\boldsymbol\theta}}$; see (2.6). For $\alpha \in [0,1]$, put $\widehat\Psi_{\widehat{\boldsymbol\theta}}(C) = (G_{n,\widehat{\boldsymbol\theta}} - F_n)(C)$. The test statistic is defined as

$$T_3 = \sup_{\alpha\in(0,1]} \big\{ \big|\widehat\Psi_{\widehat{\boldsymbol\theta}}(\widehat M_{\mathcal{C}}(\alpha))\big| + \big|\widehat\Psi_{\widehat{\boldsymbol\theta}}(\widehat M_{\mathcal{C},\widehat{\boldsymbol\theta}}(\alpha))\big| \big\}.$$

Due to the presence of the estimator $\widehat{\boldsymbol\theta}$, the asymptotic null-distribution of statistic T_3 depends on the underlying processes in a rather implicit manner. Therefore it cannot be used to compute P-values of the test directly. In fact, it may be proved under the hypothesis that $\sigma_t = \sigma_0(\mathbf{X}_t,\boldsymbol\theta)$, $G_{\boldsymbol\theta}$–a.s. and under appropriate regularity assumptions that

$$\sqrt{n}\,\widehat\Psi_{\widehat{\boldsymbol\theta}}(\widehat M_{\mathcal{C}}(\alpha)) = \frac{1}{\sqrt{n}} \sum_{t=1}^n (|\varepsilon_t| - 1)\big(I\{\mathbf{X}_t \in M_{\boldsymbol\theta}(\alpha)\} - \alpha \big) + \sqrt{n}(\widehat{\boldsymbol\theta} - \boldsymbol\theta)'\,\mathbf{b}_{\boldsymbol\theta}\{M_{\boldsymbol\theta}(\alpha)\} + o_{P_{\boldsymbol\theta}}(1), \tag{3.3}$$

where $M_{\boldsymbol\theta}(\alpha)$ denotes the MV-set corresponding to $G_{\boldsymbol\theta}$ and $\mathbf{b}_{\boldsymbol\theta}\{M_{\boldsymbol\theta}(\alpha)\} = E_{\boldsymbol\theta}\big[\frac{\dot\sigma_0(\mathbf{X}_t,\boldsymbol\theta)}{\sigma_0(\mathbf{X}_t,\boldsymbol\theta)}\big(I\{\mathbf{X}_t \in M_{\boldsymbol\theta}(\alpha)\} - \alpha\big)\big]$. A similar expansion holds for $\widehat\Psi_{\widehat{\boldsymbol\theta}}(\widehat M_{\mathcal{C},\widehat{\boldsymbol\theta}}(\alpha))$. The dependence on the underlying process is due to the second term on the RHS of (3.3), which would vanish if we do not need to estimate $\boldsymbol\theta$ (i.e. $\widehat{\boldsymbol\theta} = \boldsymbol\theta$).

3.2 Tests based on inverse regression

Based on the same idea as for T_2, we may test the null hypothesis (3.1) using the statistic

$$T_4 = \max_{1\le j\le p} \sup_x \frac{1}{n} \Big| \sum_{t=1}^n (|Y_{t-j}| - \widehat\nu_{y,j}) I(e_{t,\widehat{\boldsymbol\theta}} \le x) \Big|.$$

where $e_{t,\widehat{\boldsymbol\theta}}$ is defined in (3.2), and $\widehat{\nu}_{,} = n^{-1}\sum_{1\le i\le n}|Y_{i-j}|$. Similar to (3.3), it may be shown that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{Z_t-\widehat{\nu}_y\}I(e_{t,\widehat{\boldsymbol\theta}}\le y)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{Z_t-\nu_y\}\,(I(\varepsilon_t\le y)-F_\varepsilon(y))$$
$$+\ y\,f(y)\,(\sqrt{n}\,(\widehat{\boldsymbol\theta}-\boldsymbol\theta))'\,\mathbf{a}_{\boldsymbol\theta}\ +\ o_{P_{\boldsymbol\theta}}(1), \qquad (3.4)$$

where $\mathbf{a}_{\boldsymbol\theta}=(a_{\boldsymbol\theta,j},\ j=1,\ldots,p)'$, and $a_{\boldsymbol\theta,j}=\mathrm{E}_{\boldsymbol\theta}\Big(\{|Y_{t-j}|-\nu_y\}\,\frac{\dot\sigma_0(\mathbf{X}_t,\boldsymbol\theta)}{\sigma_0(\mathbf{X}_t,\boldsymbol\theta)}\Big)$. Again, the term involving $\widehat{\boldsymbol\theta}-\boldsymbol\theta$ on the RHS of the above expression makes the asymptotic null-distribution of T_4 depend on the underlying process.

Below we propose to approximate the null-distributions of both T_3 and T_4 by some bootstrap methods. One alternative would be to apply a martingale transform to the (estimated) empirical processes to make their sampling distributions independent of underlying processes (Khmaladze 1981, 1988). This method has been successfully applied in dealing with processes indexed by a one-dimensional parameter; see, for example, Stute (1997), Stute *et al.* (1998) and Koul and Stute (1999). Although the asymptotic expansions stated above indicate that the application of this method to the present context might be possible, there are several issues which have to be investigated first; including the practical implementation of the method to multivariate cases, and the potential power loss due to the transformation. We will pursue this approach in a separate study.

3.3 Bootstrap tests

For bootstrap tests for a composite null hypothesis, ideally the bootstrap sample should be drawn from the 'representative' distribution of the null hypothesis, which determines the significance levels for the tests. This may be achieved easily if the null-distribution of a test statistic is distribution-free, as then the bootstrap sample may be draw from any distribution under the null hypothesis. In general, we typically replace the 'representative' distribution by the distribution under the null hypothesis which is the 'closest' to the observations (Hinkley 1988, Hall and Wilson 1991). As we have pointed out above, the asymptotic distributions of both T_3 and T_4 are not distribution-free. We outline below the two bootstrap methods to estimate the P-values of the tests.

Parametric bootstrap test. Under some circumstances we may assume that the distribution of innovations ε_t in model (1.1) is known, say, F_1. The bootstrap sample may be drawn from the

equation

$$Y_t^* = \sigma_0(\mathbf{X}_t^*, \widehat{\boldsymbol{\theta}})\varepsilon_t^*, \tag{3.5}$$

where $\mathbf{X}_t^* = (Y_{t-1}^*, \cdots, Y_{t-p}^*)'$, $\widehat{\boldsymbol{\theta}}$ is an estimator for $\boldsymbol{\theta}$, and $\{\varepsilon_t^*\} \sim_{i.i.d.} F_1$.

Nonparametric bootstrap test. If the distribution of ε_t in model (1.1) is unknown, we may adopt a nonparametric bootstrap method as follows: define the residuals $\widehat{\varepsilon}_t = Y_t/\sigma(\mathbf{X}_t, \widehat{\boldsymbol{\theta}})$, and draw bootstrap sample from (3.5) but now $\{\varepsilon_t^*\}$ are independent drawn from the standardized residuals.

With a bootstrap sample, T_3^* (or T_4^*) is computed in the same manner as T_3 (or T_4) with $\{Y_t\}$ replaced by $\{Y_t^*\}$. The bootstrap estimate for the P-value is the relative frequency of the occurrence of the event $T_3 > T_3^*$ (or $T_4 > T_4^*$) in a repeated bootstrap sampling of B times, where $B > 0$ is a large integer. For the nonparametric bootstrap test, we may first 'standardize' the residuals $\widehat{\varepsilon}_t$ such that the sample means of $\widehat{\varepsilon}_t$ and $|\widehat{\varepsilon}_t|$ are, respectively, 0 and 1.

The nonparametric bootstrap test outlined above is more general than the parametric one. It may still apply when the innovation distribution F_1 is known. However some power-loss may be expected then, since the residuals $\widehat{\varepsilon}_t$, and therefore also the bootstrap innovations ε_t^*, will not behave like a random sample from F_1 *if* the null hypothesis (3.1) does not holds; see the numerical examples in table 5.1 in section 5.

Theorem 3.1 below shows that the parametric bootstrap method outlined above provides the correct asymptotic approximations for the significance levels. The justification for the nonparametric counterpart requires substantially more effort and deserves a separate study.

Let $Y_{t,\boldsymbol{\theta}}$ be a stationary and ergodic solution of $Y_t = \sigma_0(\mathbf{X}_t, \boldsymbol{\theta})\varepsilon_t$.

Theorem 3.1 *For any constant $c > 0$, let the statements below hold uniformly on $\mathcal{U}_{c/\sqrt{n}}(\boldsymbol{\theta}_0) \equiv \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \le c\, n^{-1/2}\}$ as $n \to \infty$.*

(i) *(3.3) holds.*

(ii) $W_n(C) \equiv \frac{1}{\sqrt{n}} \sum_{t=1}^n (|\varepsilon_t| - 1)\big[I(\mathbf{X}_{t,\boldsymbol{\theta}} \in C) - G_{\boldsymbol{\theta}}(C)\big]$, $C \in \mathcal{C}$, *is asymptotically equicontinuous with respect to the $L_1(G_{\boldsymbol{\theta}})$-norm.*

(iii) $\sup_{1 \le t \le n} |Y_{t,\boldsymbol{\theta}} - Y_{t,\boldsymbol{\theta}_0}| = o_P(1)$.

(iv) $\sqrt{n}\,(\widehat{\boldsymbol\theta}-\boldsymbol\theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \{h(\varepsilon_t)-\mathrm{E}h(\varepsilon_t)\}\,\boldsymbol\psi_{\boldsymbol\theta}(\mathbf{X}_{t,\boldsymbol\theta})+o_P(1)$, *where* h *is a scalar function for which* $\mathrm{E}h^2(\varepsilon_t)<\infty$, $\boldsymbol\psi_{\boldsymbol\theta}$ *is a Lipschitz continuous q-vector function for which* $E_{\boldsymbol\theta_0}||\boldsymbol\psi_{\boldsymbol\theta_0}(\mathbf{X}_{t,\boldsymbol\theta_0})||^2<\infty$ *and* $||\boldsymbol\psi_{\boldsymbol\theta}(\mathbf{x})-\boldsymbol\psi_{\boldsymbol\theta_0}(\mathbf{x})||\le K(\mathbf{x})||\boldsymbol\theta-\boldsymbol\theta_0||$. *Here* $K(\cdot)\ge 0$ *is a measurable function with* $E\{K(\mathbf{X}_{t,\boldsymbol\theta_0})\}^2<\infty$.

(v) The density function $g_{\boldsymbol\theta}$ *of* $G_{\boldsymbol\theta}$ *is Lipschitz continuous, and* $\dot\sigma_0(\mathbf{x},\boldsymbol\theta)$ *is continuous in both* $\mathbf{x}$ *and* $\boldsymbol\theta\in\mathcal{U}_{c/\sqrt{n}}(\boldsymbol\theta_0)$. *Further, there exist a q-vector of measurable functions* $\mathbf{k}$ *with* $|\dot\sigma_0(\mathbf{x},\boldsymbol\theta)|\le\mathbf{k}(x)$ *and* $\sup_{\boldsymbol\theta\in\mathcal{U}_{c/\sqrt{n}}(\boldsymbol\theta)}\mathrm{E}\mathbf{k}(X_{t,\boldsymbol\theta})<\infty$.

Then it holds that

$$\sup_{x\in\mathcal{R}}|P(T_3^*\le x|X_1,\ldots,X_n)-P(T_3\le x)|\to 0 \ \text{ in probability.}$$

Since we do not specify the form of $\sigma_0(\cdot)$ in (3.1), the conditions in the above theorem are generic. Nevertheless they may be established for, for example, ARCH/GARCH models. Conditions (i) and (ii) require the covering integral of $\mathcal{C}$ to be finite. Condition (iii) holds for stationary GARCH(p,q) processes (cf. (2.68) in Fan and Yao 2003). Examples for which expansion (iv) holds include qMLE for ARCH(p) models with $h(\varepsilon_t)=\varepsilon_t^2$; see Hall and Yao (2003), p.304. Note that the standard parametrization is used in Hall and Yao (2003), which entails $E(\varepsilon_t^2)=1$. We refer section 3.3 of Polonik and Yao (2007) for further discussions on those conditions.

4 Is higher volatility associated with rarer events?

In this section, we continue to explore the relationship between the two probability measures F and G in order to investigate some qualitative characteristics of the volatility function σ_t under the general framework (1.1). We will provide a statistical test to check if financial market is more volatile at occurrence of rarer events, which is reflected in GARCH models. For example, a simple ARCH(1) specifies volatility function as $\sigma_{t+1}^2=a+bY_t^2$ $(a,b\ge 0)$. Therefore large positive or large negative values of Y_t lead to large values of σ_{t+1}^2, while the chance of having excessive returns (i.e. $|Y_t|$ is excessively large) is small.

We use the same notation as in section 2. To facilitate our discussion, we assume $\sigma_t=\sigma(\mathbf{X}_t)$, where $\sigma(\cdot)$ is an unknown function. Recall $f(\cdot)$ is the density function of $\mathbf{X}_t$ and $M_f(\alpha)$ is the MV

set of F; see (2.5). The measure G admits the density function $g(\cdot) = \nu_y^{-1}\sigma(\cdot)f(\cdot)$; see the definition of $G(\cdot)$ in (2.3). We consider to test the null hypothesis

$$H_0 : \ F\{M_f(\alpha)\} = G\{M_f(\alpha)\} \quad \text{for all } \alpha \in (0,1) \tag{4.1}$$

against the one-sided alternative

$$H_1 : F\{M_f(\alpha)\} \geq G\{M_f(\alpha)\} \text{ for all } \alpha \in (0,1), \quad \text{with strict inequality for some } \alpha \in (0,1). \tag{4.2}$$

When the above H_0 is rejected, it indicates that the probability mass under G is more spreading-out than that under F. Note that $g(\cdot) \propto \sigma(\cdot)f(\cdot)$. Hence the G-measure is more spreading-out than the F-measure when $\sigma(\cdot)$ makes the probability mass thinner where $f(\cdot)$ is large, and thicker where $f(\cdot)$ is small. This phenomenon will occur when, for example, f is unimodal and decays to 0 at boundaries and $\sigma(\cdot)$ is a U-shape curve. The latter feature is termed "smiling effect" in volatility literature; see, for example, Härdle and Tsybakov (1997). It reflects the stylized feature that a financial market is more volatile when returns are large, either positively or negatively. Note that we only compare the two measures over the MV sets under F since we only look for the evidence that G is more spread-out than F in relation to the central areas of the data, which is reflected by the MV sets under F.

Hypothesis (4.2) (ignoring the strict inequalities) defines an integral stochastic order. In fact, by interpreting the probabilities in (4.2) as expected values of indicator functions of the MV sets, and taking linear combinations, one can see that (4.2) implies inequalities of the form $\int h(x)\, dG(x) \leq \int h(x) dF(x)$ for functions h whose level sets (i.e. the sets $\{\, x : h(x) \geq \lambda \,\}$, $\lambda \geq 0$) are MV sets of f. This means that (4.2) defines in integral stochastic order with generator consisting of all functions with level sets in $\{\, M_f(\alpha),\ \alpha \in [0,1] \,\}$. Recall that level sets of f are MV sets, and hence every (positive) function with each of its level set also being a level set of f is a member of the class of generators. This implies many integral relations between f and g, as for instance (by taking $h = f^k$) we obtain $\int f^k\, g \leq \int f^{k+1}$. See, e.g. Müller and Stoyan (2002) for details on integral stochastic orders.

We should reject the null hypothesis (4.2) when $(F - G)\{M_f(\alpha)\}$ takes large positive values for some $\alpha \in (0,1)$. Therefore we may define a test statistic as

$$T_5 = \int_0^1 \{\, \widehat{\Psi}_{\mathcal{C},f}(\alpha) \,\}^+ d\alpha$$

where $\widehat{\Psi}_{\mathcal{C},f}(\alpha)$ is defined as in (2.8), and we reject (4.2) for some large values of T_5.

Theorem 4.1 *Suppose that the conditions of Theorem 2.2 hold, and $F \equiv G$. Then as $n \to \infty$, it holds that*

$$\sqrt{\frac{n}{\widehat{\sigma}_y^2/\widehat{\nu}_y^2}}\; T_5 \xrightarrow{\mathcal{D}} \int_0^1 (\, B(\alpha)\,)^+ \, d\alpha,$$

where $B(\alpha)$ is the standard Brownian bridge, and $\widehat{\sigma}_y$ and $\widehat{\nu}_y$ are the same as in (2.10).

We tabulate the high quantiles of the random variable $\int_0^1 (\, B(\alpha)\,)^+ \, d\alpha$ below, which were obtained from a Monte Carlo simulation.

α	0.900	0.925	0.950	0.975	0.990	0.995
quantile	0.383	0.388	0.478	0.563	0.670	0.717

Remark 4.2 The setting (4.1) and (4.2) implicitly implies that we are concerned with the processes for which $F\{M_f(\alpha)\} \geq G\{M_f(\alpha)\}$ for all α. In practice, we may apply a pre-test with the test statistic $T_6 = \int_0^1 \{\, \widehat{\Psi}_{\mathcal{C},f}(\alpha)\,\}^- d\alpha$, where $x^- = \max(-x, 0)$. Obviously, T_6 shares the same asymptotic distribution as T_5. We should not proceed with the test T_5 if the pre-test with T_6 is significant (i.e., for example, $T_6 > 0.478$; see the table above).

Diagnostic plots. The CC-plots discussed in section 2.1 also serve as a diagnostic plot associated with the test T_5. Now, however, we are only interested in one-sided deviations.

5 Numerical illustration

We illustrate the proposed tests with numerical examples. For tests T_1 and T_5, we always choose $\widehat{M}_{\mathcal{C},f}(\alpha)$ among the sets with balls as their images under the mapping $\mathbf{X}_t \to \mathbf{S}^{-1/2}(\mathbf{X}_t - \bar{\mathbf{X}})$, where $\bar{\mathbf{X}}$ and $\mathbf{S}$ denote, respectively, the sample mean and the sample covariance matrix of $\{\mathbf{X}_t\}$; see (2.6). We choose $\widehat{M}_{\mathcal{C},g}(\alpha)$ among the sets also with balls as their images but under a different mapping $\mathbf{X}_t \to \mathbf{S}_g^{-1/2}(\mathbf{X}_t - \bar{\mathbf{X}}_g)$, where $\bar{\mathbf{X}}_g$ and $\mathbf{S}_g$ denote the mean and the covariance matrix of the distribution G_n. For test T_3, $\widehat{M}_{\mathcal{C},\widehat{\boldsymbol{\theta}}}(\alpha)$ are defined in the same manner as $\widehat{M}_{\mathcal{C},g}(\alpha)$ with G_n replaced by $G_{n,\widehat{\boldsymbol{\theta}}}$ defined in (3.2).

Before we proceed to numerical experiments, we would like to point out that the tests proposed in this paper do not facilitate a direct comparison with the existing methods. As indicated in the beginning of section 2, the existing methods based on the residual-regression idea are for univariate $\mathbf{X}_t$ for which there is no point to adopt the MV-set or inverse regression. On the other hand, those parametric tests based on the residuals obtained from fitting ARCH/GARCH models with Gaussian innovation cannot apply to heavy-tailed Models I and II in (5.1) below, and would be in disadvantage when applying to a 'wrong' model such as Model III.

5.1 Simulated examples

First we deal with tests T_1, T_2 and T_5 with the data generated from three different models:

$$\begin{aligned} &\text{I.} \qquad Y_t = e_t, \\ &\text{II.} \qquad Y_t = \sigma_t e_t, \quad \sigma_t^2 = 0.5 + 0.1Y_{t-1}^2 + 0.8\sigma_{t-1}^2, \\ &\text{III.} \qquad Y_t = \sigma_t \varepsilon_t, \quad \sigma_t = 0.5 + 0.2|Y_{t-1}| + 0.75\sigma_{t-1}, \end{aligned} \tag{5.1}$$

where ε_t be independent $N(0,1)$ random variables, and e_t be independent t_3-distributed random variables. The processes defined by model I is a sequences of i.i.d. random variables with $E|Y_t|^3 = \infty$, which rules out the use of conventional parametric testing methods; see section 4.2.6 of Fan and Yao (2003). Model II is a GARCH(1,1) with heavy tailed innovations. Model III defines a power GARCH(1,1) model with power index equal to 1.

For sample sizes 50, 100, 200 and 500, we applied the tests with $p = 1, 2, 4$ and 6. For each setting, we replicated the simulation 500 times. The relative frequencies of rejecting the null hypotheses at level 10%, 5% and 1% are listed in Tables 1 – 3.

Table 1 contains the results from applying tests to the independent t_3 observations. It indicates that the asymptotic approximations (2.10), (2.12) and Theorem 4.1 are adequate. The asymptotic approximations (2.10) and Theorem 4.1 remain the same for different values of p. With the range of sample sizes specified in the simulation, Table 1 suggests that overall those two approximations work about equally well for p between 1 and 6. The reported relative frequencies are almost always smaller than the nominal levels. This will underplay the potential power of the tests. Also the improvement in the asymptotic approximations for T_1 and T_2 when n increases is not as evident as that for T_5.

Table 1: Relative frequencies of rejecting null hypotheses; the three numbers in each entry corresponding to, respectively, the significance level 10%, 5% and 1%. The observations are drawn from Model I.

p	n	T_1			T_2			T_5		
1	50	.04	.00	.00	.05	.02	.00	.08	.03	.00
	100	.04	.01	.00	.04	.01	.00	.11	.05	.01
	200	.07	.04	.01	.07	.03	.00	.09	.05	.02
	500	.07	.03	.00	.06	.03	.00	.10	.05	.01
2	50	.03	.01	.00	.05	.02	.01	.07	.07	.00
	100	.04	.02	.00	.04	.01	.00	.09	.04	.01
	200	.04	.01	.00	.07	.03	.00	.10	.05	.01
	500	.06	.03	.00	.08	.03	.00	.10	.05	.01
4	50	.04	.01	.00	.04	.02	.01	.07	.03	.00
	100	.05	.03	.00	.04	.03	.01	.08	.03	.00
	200	.05	.02	.00	.04	.03	.00	.07	.04	.01
	500	.06	.03	.01	.07	.04	.01	.10	.05	.01
6	50	.09	.04	.01	.04	.02	.01	.05	.02	.00
	100	.06	.03	.00	.03	.02	.01	.05	.01	.00
	200	.06	.02	.00	.04	.02	.00	.07	.04	.01
	500	.07	.03	.01	.06	.04	.01	.09	.05	.01

Table 2: Legend is as for Table 1. The observations are drawn from Model II: GARCH(1,1) model with t_3-distributed innovations.

p	n	T_1			T_2			T_5		
1	50	.11	.05	.01	.23	.16	.05	.34	.27	.13
	100	.43	.33	.16	.57	.48	.33	.66	.60	.44
	200	.80	.74	.60	.89	.82	.74	.92	.89	.81
	500	1.00	1.00	.98	1.00	1.00	.99	1.00	1.00	1.00
2	50	.14	.07	.01	.25	.18	.06	.46	.35	.17
	100	.55	.45	.30	.62	.41	.35	.80	.73	.58
	200	.92	.88	.79	.92	.89	.77	.97	.96	.93
	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	50	.16	.10	.01	.20	.14	.04	.50	.40	.22
	100	.62	.56	.38	.60	.51	.34	.87	.79	.65
	200	.95	.93	.86	.92	.89	.78	.98	.97	.95
	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
6	50	.14	.07	.01	.17	.12	.06	.46	.36	.19
	100	.64	.55	.37	.58	.48	.32	.86	.81	.70
	200	.96	.94	.88	.95	.90	.79	.99	.98	.96
	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 3: Legend is as for Table 1. The observations are generated from Model III: power-GARCH(1,1) model with Gaussian innovations.

p	n	T_1			T_2			T_5		
1	50	.10	.05	.01	.14	.08	.03	.32	.24	.09
	100	.28	.20	.08	.34	.24	.11	.60	.49	.28
	200	.61	.49	.30	.61	.52	.32	.84	.74	.58
	500	.96	.93	.84	.96	.92	.80	.99	.99	.95
2	50	.10	.06	.01	.14	.08	.02	.39	.28	.12
	100	.37	.24	.10	.31	.20	.11	.71	.60	.35
	200	.73	.65	.46	.65	.55	.37	.92	.87	.73
	500	.99	.99	.95	.97	.95	.84	1.00	1.00	.99
4	50	.12	.07	.02	.14	.07	.01	.40	.29	.13
	100	.41	.29	.15	.33	.23	.10	.73	.64	.41
	200	.80	.71	.53	.61	.51	.31	.93	.88	.79
	500	1.00	0.99	.98	.97	.95	.82	1.00	1.00	1.00
6	50	.10	.06	.01	.14	.09	.03	.36	.25	.12
	100	.38	.28	.14	.29	.21	.09	.70	.60	.40
	200	.75	.69	.52	.62	.52	.32	.93	.89	.76
	500	.99	.99	.97	.96	.91	.83	1.00	1.00	0.99

With the data generated by the two heteroscedastic models, i.e. Models II and III, our tests demonstrate the power in rejecting both the null hypotheses (2.1) and (4.2); see Tables 2 and 3. In fact the power of rejection increases as the sample size n increases. Note that for both Models II and III, σ_t depends on infinite number of lagged values of Y_t, Tables 2 and 3 show that the tests with $p = 6$ and 4 are more powerful than those with $p = 1$ and 2 in most cases.

The diagnostic plots associated with the tests T_1 and T_5 are presented in Figure 1, and those associated with the test T_2 are displayed in Figure 2. To save the space we only presented the plots with $n = 200$ and $p = 2$. For each of Models I – III, five samples were randomly selected. Under the null hypothesis (2.1), those plots are closely around the diagonal line $y = x$; see the top row in both Figures 1 and 2. When there exists heteroscedasticity, some parts of the curves drifted away from the diagonal line; see Rows 2 and 3 in Figures 1 and 2. The departure is more pronounced for GARCH(1,1) model with heavy tailed innovations.

For testing the null hypothesis (4.2) against one-sided alternative (4.1) with statistic T_5, we look for the one-sided departure of the solid curves under the diagonal $y = x$ in Figure 1. This is evident in most plots in Rows 2-3 there. This indicates that "smiling effect" may well exist for all the processes

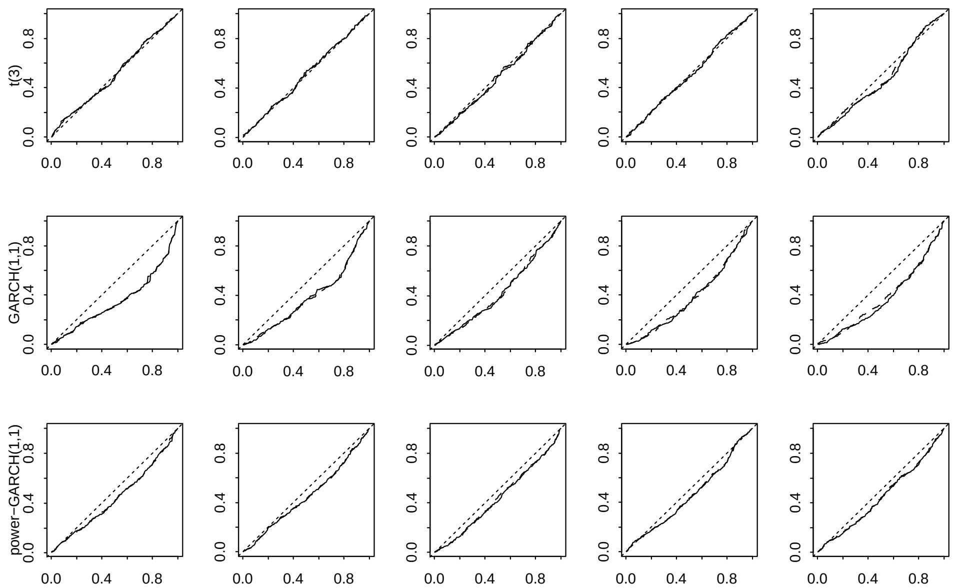

Figure 1: *Diagnostic plots associated with tests* T_1 *and* T_5 *(with* $p = 2$*): solid lines –* $G_n\{\widehat{M}_{\mathcal{C},f}(\alpha)\}$ *against* $F_n\{\widehat{M}_{\mathcal{C},f}(\alpha)\}$ *for* $\alpha \in (0,1)$*; dashed lines –* $G_n\{\widehat{M}_{\mathcal{C},g}(\alpha)\}$ *against* $F_n\{\widehat{M}_{\mathcal{C},g}(\alpha)\}$ *for* $\alpha \in (0,1)$*; dotted lines – straight line* $y = x$*. Each row represents 5 randomly selected samples (with* $n = 200$*) from, from top to bottom, each of Models I – III.*

defined by Models II and III, and again such an effect is more pronounced for the GARCH(1,1) model with t_3 innovations.

Table 4: Simulation results for parametric bootstrap (PB) and nonparametric bootstrap (nonPB) tests based on T_3 and T_4. The three numbers in each entry are the relative frequencies of rejecting the ARCH(2) null hypothesis, corresponding to, respectively, the level $\alpha = .10, .05$ and $.01$.

Model	n	T_3 (PB)			T_3 (nonPB)			T_4 (PB)			T_4 (nonPB)		
ARCH(2)	200	.11	.05	.01	.10	.05	.01	.10	.04	.01	.12	.05	.01
(with t_3-innovation)	200	.00	.00	.00	.11	.04	.01	.09	.03	.00	.11	.04	.01
ARCH(3)	200	.21	.11	.04	.12	.06	.02	.09	.04	.02	.07	.03	.01
	500	.30	.24	.12	.19	.12	.04	.13	.06	.01	.08	.04	.01
GARCH(1,1)	200	.97	.96	.87	.96	.94	.86	.45	.22	.06	.40	.21	.05
	500	1.00	1.00	1.00	1.00	1.00	1.00	.94	.80	.35	.92	.73	.26
PGARCH(1,1)	200	1.00	1.00	1.00	.99	.99	.99	.62	.12	.00	.68	.20	.00
	500	1.00	1.00	1.00	1.00	1.00	1.00	.96	.63	.00	.97	.72	.00

We also conducted a simulation study on bootstrap tests with test statistics T_3 and T_4 for testing

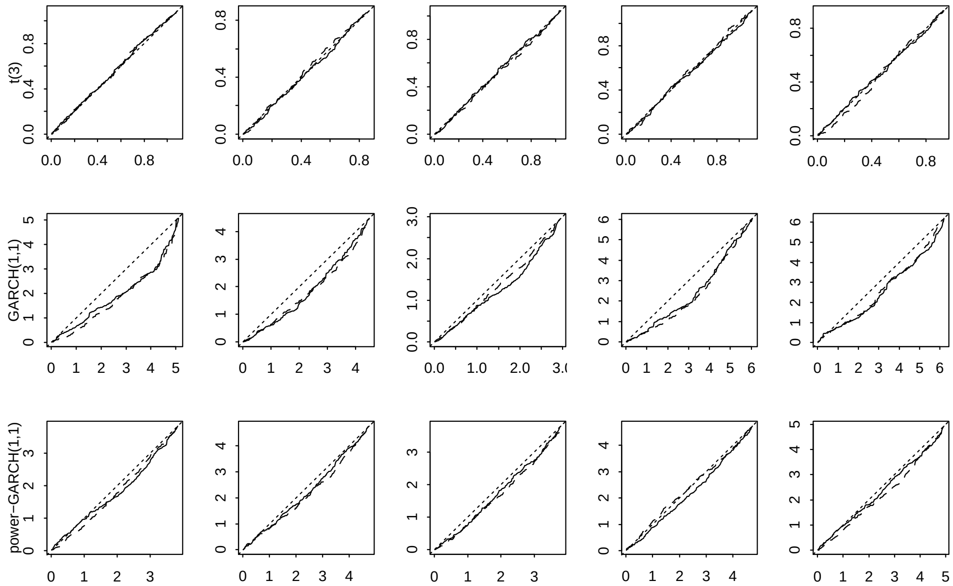

Figure 2: *Diagnostic plots associated with test* T_2 *(with* $p = 2$*):* $\frac{1}{\sqrt{n}} \sum_t |Y_{t-j}| I(|Y_t| \leq y)$ *against* $\frac{\widehat{\nu}_{y,j}}{\sqrt{n}} \sum_t I(|Y_t| \leq y)$ *for* $y > 0$*. Solid lines –* $j = 1$*, dashed lines –* $j = 2$*, dotted lines – straight line* $y = x$*. Each row represents 5 randomly selected samples (with* $n = 200$*) from, from top to bottom, each of Models I-III.*

a specified null hypothesis (3.1) with $\sigma_t^2 = c + a_1 Y_{t-1}^2 + a_2 Y_{t-2}^2$, i.e. Y_t is a ARCH(2) model. We applied both the parametric bootstrap test with normal distribution F_1, and the nonparametric bootstrap test outlined in section 4.3. Samples of size $n = 200$ or 500 were generated from model (1.1) with different forms of σ_t. For each setting, we drew 400 samples, and repeated bootstrap sampling also $B = 400$ times.

Table 4 reports the relative frequencies of rejecting the null hypothesis of ARCH(2) model in the 400 replications at the significance levels 10%, 5% and 1%. The first row contains the results for an ARCH(2) process with Gaussian innovation and the coefficients $(c, a_1, a_2) = (0.5, 0.4, 0.5)$. Now the null hypothesis H_0 is true. Both parametric and nonparametric bootstrap methods provide very accurate estimates for the significance levels. The results for the same model but with t_3-innovations are reported in the second row of the table. As expected, the nonparametric bootstrap method still provide accurate estimates for the significance levels. However the parametric bootstrap method failed with statistic T_3. This was due to the use of a wrong innovation distribution in bootstrapping.

In Table 4, Rows 3–4 list the results for an ARCH(3) model with Gaussian innovations and

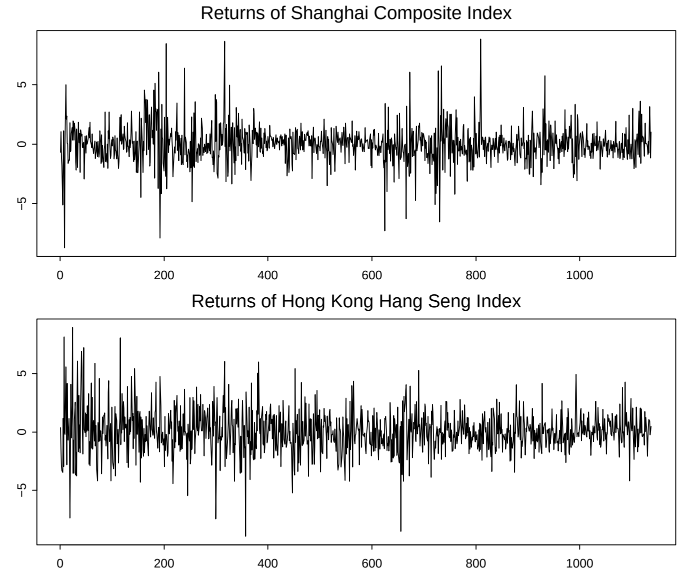

Figure 3: *Time plots of daily returns of Shanghai Composite Index and Hong Kong Hang Seng Index in 3 August 1998 – 30 December 2003.*

the coefficients $(c, a_1, a_2, a_3) = (0.5, 0.3, 0.2, 0.4)$, Rows 5-6 for for GARCH(1,1) model IV but with Gaussian innovations, and the last two rows for power-GARCH(1,1) model V. Even with sample size $n = 500$, all the tests lack the power to tell the 'subtle' difference between ARCH(3) and ARCH(2). On the other hand, the tests based on MV sets (i.e. T_3) are considerably more powerful than the tests based inverse regression (i.e. T_4). Furthermore, the nonparametric bootstrap tests are almost always less powerful than the parametric bootstrap tests which used the correct innovation distribution in bootstrap samplings.

5.2 Real data examples

Now we illustrate the tests with two real data sets: the returns of daily close prices of Shanghai Composite Index and Hong Kong Hang Seng Index in 3 August 1998 – 30 December 2003; see Figure 3. We applied the tests to the whole series ($n = 1139$), as well as to the three subseries (for each of the two data sets) in 3 August 1998 – 30 June 2000, 3 July 2000 – 31 May 2002 and 3 June 2002 – 30 Dec 2003 with sample sizes, respectively, 400, 395 and 344.

For the returns of Shanghai Composite Index, the tests T_1 and T_2 with $p = 1, 2$ and 4 are all significant at the level 1% for the whole series and the first subseries, and T_1 and T_2 with $p = 2$ and

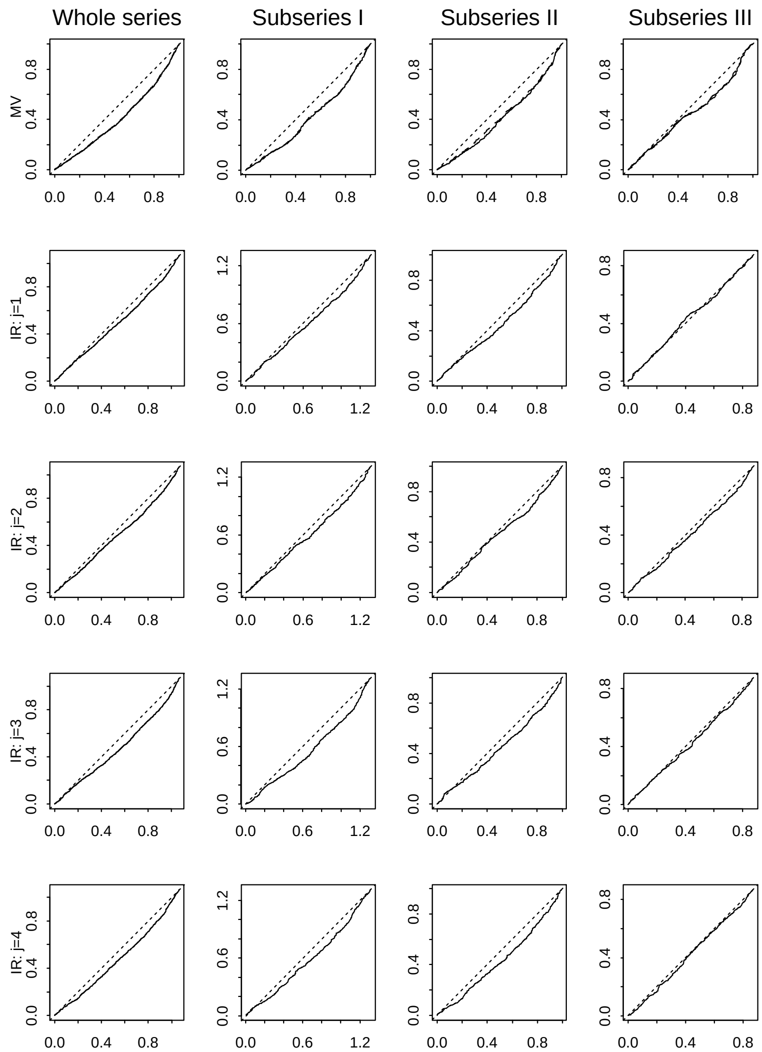

Figure 4: *Diagnostic plots associated with T_1 and T_2 (with $p = 4$) for returns of Shanghai Composite Index. Row I: $G_n\{\widehat{M}_{\mathcal{C},f}(\alpha)\}$ against $F_n\{\widehat{M}_{\mathcal{C},f}(\alpha)\}$ for $\alpha \in (0,1)$ – solid lines; $G_n\{\widehat{M}_{\mathcal{C},g}(\alpha)\}$ against $F_n\{\widehat{M}_{\mathcal{C},g}(\alpha)\}$ for $\alpha \in (0,1)$ – dashed lines. Rows II-V: $\frac{1}{\sqrt{n}}\sum_t |Y_{t-j}| I(|Y_t| \le y)$ $(y > 0)$ against $\frac{\widehat{\nu}_{y,j}}{\sqrt{n}}\sum_t I(|Y_t| \le y)$ for $j = 1,2,3,4$. Dotted lines are diagonal $y = x$.*

4 are significant at the level 1% for the second and the third subseries. The test T_1 with $p = 1$ is not significant (with P-value greater than 0.1) for both the second and the third subseries while T_2 with $p = 1$ is significant at level 5% for both the subseries. This indicates that there is overwhelming evidence to reject the null hypothesis of a constant volatility function. It also echoes the observation in the simulation study that the tests with $p > 1$ are more powerful than those with $p = 1$ for rejecting (2.1) when $\sigma(\cdot)$ depends on more than one lagged values. Figure 4 displayed the associated

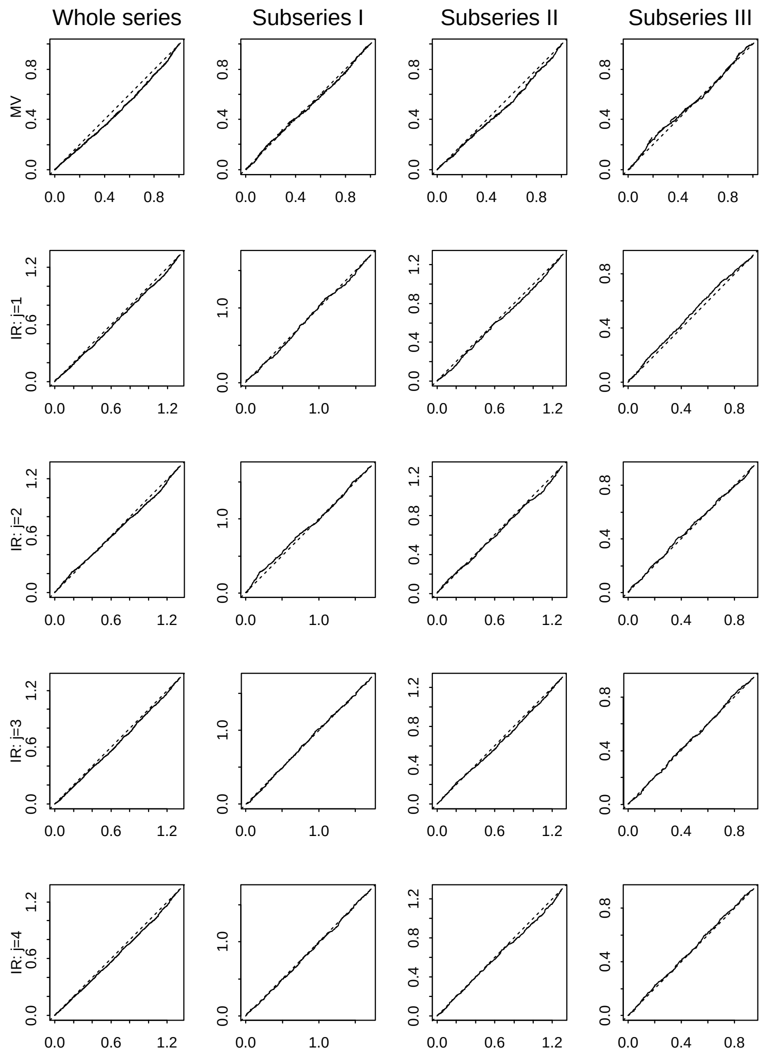

Figure 5: *Diagnostic plots associated with T_1 and T_2 (with $p = 4$) for returns of Hang Kong Hang Seng Index. Row I: $G_n\{\widehat{M}_{\mathcal{C},f}(\alpha)\}$ against $F_n\{\widehat{M}_{\mathcal{C},f}(\alpha)\}$ for $\alpha \in (0, 1)$. Rows II-V: $\frac{1}{\sqrt{n}} \sum_t |Y_{t-j}| I(|Y_t| \le y)$ against $\frac{\widehat{\nu}_{y,j}}{\sqrt{n}} \sum_t I(|Y_t| \le y)$ $(y > 0)$ for $j = 1, 2, 3, 4$. Dotted lines are diagonal $y = x$.*

diagnostic plots. It clearly indicates the evidence of the departure for the null hypothesis for the whole series as well as all the three subseries.

We also applied the tests T_3 and T_4 for testing the null hypothesis of an ARCH(2) model $\sigma_t^2 = a_0 + a_1 Y_{t-1}^2 + a_2 Y_{t-2}^2$. The quasi-MLE for (a_0, a_1, a_2) for the three subseries were, respectively (1.906, 0.405, 0.128), (1.380, 0.118, 0.325) and (0.848, 0.239, 0.335), which were very different from each other. The nonparametric bootstrap tests with both T_3 and T_4 and the parametric bootstrap test with T_4 using Gaussian innovation distribution do not reject the null hypothesis of ARCH(2) model

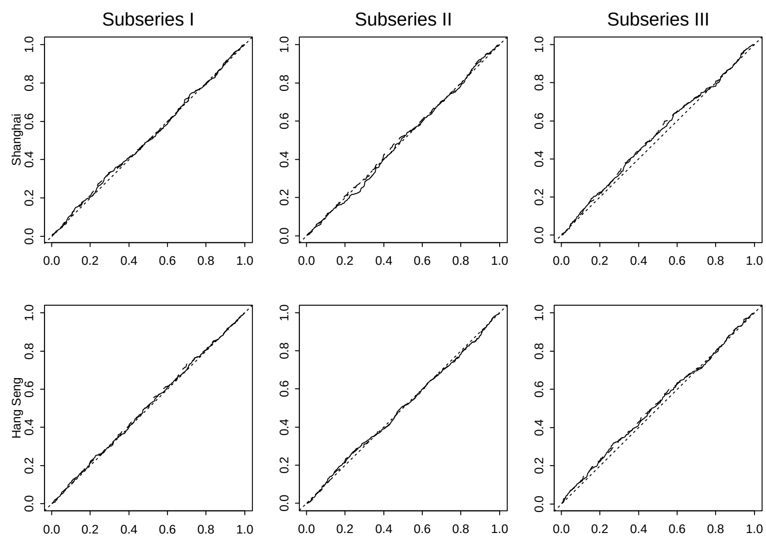

Figure 6: *Diagnostic plots associated with T_3 for subseries of both Shanghai Composite returns and Hang Kong Hang Seng returns. Solid lines – $G_{n,\widehat{\boldsymbol{\theta}}}\{\widehat{M}_{\mathcal{C},f}(\alpha)\}$ against $F_n\{\widehat{M}_{\mathcal{C},f}(\alpha)\}$ for $\alpha \in (0,1)$; dashed lines – $G_{n,\widehat{\boldsymbol{\theta}}}\{\widehat{M}_{\mathcal{C},\widehat{\boldsymbol{\theta}}}(\alpha)\}$ against $F_n\{\widehat{M}_{\mathcal{C},\widehat{\boldsymbol{\theta}}}(\alpha)\}$ for $\alpha \in (0,1)$; dotted lines – $y = x$.*

for Subseries I and II with minimum P-value 0.27. The parametric bootstrap test with T_3 yields the P-values 0.08 and 0.41 for Subseries I and II with Gaussian innovation distribution, and P-values 0.24 and 0.57 with t_7 innovation distribution. Overall there is no significant evidence to against ARCH(2) model for the first two subseries. For subseries III, the parametric bootstrap test with T_3 using either Gaussian or t_7 innovation distributions yields the P-value 0.00 while the nonparametric bootstrap test with T_3 yields the P-value 0.02; indicating the null hypothesis of ARCH(2) should be rejected. The associated diagnostic plots are displayed in Figure 6. It indeed shows a certain degree of the inadequacy of ARCH(2) model for Subseries III.

Interestingly the returns of Hong Kong Hang Seng Index showed drastically different behavior. For example, the test T_1 was not significant with $p = 1, 2$ for all three subseries, and was only significant with $p = 4$ for Subseries II at level 5% and Subseries III at level 10% only. The test T_2 was significant for all the three subseries with $p = 2, 4$ at the level 5% or 10%, was not significant with $p = 1$. For the whole series, both T_1 and T_2 were significant at level 5% with $p = 1$, and were significant at level 1% with $p = 2, 4$. One may argue that the substantial increase in significance might be due the non-stationarity rather than a genuine conditional heteroscedasticity. The associated diagnostic plots are displayed in Figure 5. In contrast to Figure 4, Figure 5 indicates little evidence

of the departure from the null hypothesis, especially for the three subseries.

We further fitted ARCH(2) model to the three subseries. The estimated coefficients for (a_0, a_1, a_2) were, respectively (5.279, 0.012, 0.000), (2.438, 0.109, 0.047) and (1.589, 0.000, 0.000). This also supports a constant conditional variance model for Subseries I and III. Not surprisingly both the parametric (with Gaussian innovation distribution) and nonparametric bootstrap tests with both T_3 and T_4 are not significant; indicating that the null hypothesis of an ARCH(2) model could not be rejected for all three subseries. The diagnostic plots in Figure 6 reinforces this argument.

Finally we applied test T_5 to explore the existence of 'smiling effect' in the subseries of those two data sets. We first applied the pre-test T_6 to examine the evidence against the prerequisite $F\{M_f(\alpha)\} \geq G\{M_f(\alpha)\}$ for any α; see Remark 4.2. The P-values of the pre-test with $p = 1, 2$ and 4 are always greater than 0.1. Hence we may proceed with the test T_5 now.

For Shanghai composite returns, T_5 is significant with $p = 2, 4$ at level 1% for all the three subseries. It is significant with $p = 1$ at level 1% for Subseries I, at level 5% for Subseries II, and not significant for Subseries III; see also the plots in Row I in Figure 4. Overall there is evidence to indicate that the 'smiling effect' exists with all the three subseries.

Again Hang Kong Hang Seng returns show different behavior. The tests T_5 with $p = 1, 2$ and 4 are all not significant even at the level 10% for Subseries I and III, are significant at level 5% for Subseries II. The diagnostic plots in Figure 5 (Row I) show that there might be some evidence for $F\{M_f(\alpha)\} > G\{M_f(\alpha)\}$ for some α for Subseries II. However the departure from the diagonal is much less pronounced than that for Shanghai composite returns; see Figure 4.

The different behaviours of the two financial markets *might* be explained as follows. Note that the data were recorded shortly after the Asian financial crisis in early 1998. As a consequence of the crisis, Hong Kong market was in a *constant* volatile status in the sense that the Hang Seng index oscillated regularly and showed hardly any tranquility spans. Therefore the volatility changed little especially in the first third and the last third periods; see the second panel in Figure 3. On the other hand, the impact of the Asian financial crisis to the mainland of China is relatively small. The spans of high volatility alternated with the spans of tranquility, which drove the dynamical structure of the Shanghai composite returns.

Acknowledgments

We thanks a Guest Editor and two referees for helpful comments and suggestions. WP was partially supported by NSF grants 0103606 and 0406431. QY was partially supported by EPSRC grants GR/R97430 and EP/C549058.

References

Ait-Sahalia, Y., Bickel, P.J. and Stoker, T.M. (2001). Goodness-of-fit tests for kernel regression with an application to option implied volatilities. *Journal of Econometrics*, **105**, 363-412.

Bierens, H.J. (1987). ARMAX model specification testing, with an application to unemployment in the Netherlands. *Journal of Econometrics*, **35**, 161-190.

Billingsley, P. (1999). *Convergence of Probability Measures* (2nd edit). Wiley, New York.

Bingham, N.H. and Kiesel, R. (2002). Semi-parametric modelling in finance: Theoretical foundation. *Quantitative Finance*, **2**, 241 - 250.

Casas, I., Gao, J. and Gijbels, I. (2007). Specification testing in semiparametric diffusion model: theory and practice. *A preprint.*

Chen, M. and An, H.Z. (1997). A Kolmogorov-Smirnov type test for conditional heteroskedasticity in time series. *Statistics and Probability Letters*, **33**, 321-331.

Cook, R.D. and Weisberg, S. (1982). *Residual and Influence in Regression.* Chapman and Hall, New York.

Eubank, R.L. and Hart, J.D. (1993). Commonalty of cusum, von Neumann and smoothing-based goodness-of-fit tests. *Biometrika*, **80**, 89-98.

Fan, J. and Yao, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods.* Springer, New York.

Härdle, W. and Mammen, E. (1993). Comparing non-parametric versus parametric regression fits. *The Annals of Statistics*, **21**, 1926-1947.

Härdle, W. and Tsybakov, A. (1997). Local polynomial estimators of the volatility function in nonparametric autoregression. *J. of Econometrics*, **81**, 223-242.

Hall, P. and Wilson, S.R. (1991). Two guide lines of bootstrap hypothesis testing. *Biometrics*, **47**, 757-762.

Hall, P. and Yao, Q. (2003). Inference in ARCH and GARCH models with heavy-tailed errors. *Econometrica* **71**, 285-317.

Hamada, M. and Valdez, E. (2004). CAPM and option pricing with elliptical distributions. *UTS Quantitative Finance Research Paper 120.*

Hinkley, D. V. (1988). Bootstrap methods (with discussion). *Journal of the Royal Statistical Society*, **B**, **50**, 321-337.

Horváth, L., Kokoszka, P. and Teyssière (2001). Empirical process of the squared residuals of an ARCH sequence. *The Annals of Statistics*, **29**, 445-469.

Khmaladze, E.V. (1981). Martingale approach in the theory of goodness-of-fit tests. *Theory. Probab. Appl.*, **26**, 240-257.

Khmaladze, E.V. (1988). An innovation approach to goodness of fit in $\mathbf{R}^m$. *The Annals of Statistics*, **16**, 1503-1516.

Khmaladze, E.V. and Koul, H.L. (2004). Martingale transforms goodness-of-fit tests in regression models. *The Annals of Statistics*, **32**, 995-1034.

Koul, H. L., Stute, W. (1999). Nonparametric model checks for time series. *The Annals of Statistics*, **27**, 204-236.

Laïb, N. (2003). Nonparametric tests for conditional variance functions in time series. *Australian & New Zealand Journal of Statistics*, **45**, 461-475.

Morgan, J.P. (1996). *RiskMetrics Technical Document*. Fourth edition, New York.

Müller, A. and Stoyan, D. (2002). Comparison methods for stochastic models and risks. Wiley, New York.

Pelagatti, M.M. and Rondena, S. (2005). Dynamic Conditional Correlation with Elliptical Distributions, *Econometrics 0503007, EconWPA*

Pollard, D. (1984). *Convergence of stochastic processes.* Springer, New York.

Polonik, W. (1995). Measuring mass concentration and estimating density contour clusters - an excess mass approach. *The Annals of Statistics*, **23**, 855-881.

Polonik, W. (1999). Concentration and goodness-of-fit in higher dimensions: (Asymptotically) distribution-free methods. *The Annals of Statistics*, **27**, 1210-1229.

Polonik, W. and Yao, Q. (2007). Testing for multivariate volatility functions. *Available at http://stats.lse.ac.uk/q.yao/qyao.links/technicalreports.html/pyLong07.pdf*

Stute, W. (1997). Nonparametric model checks for regression. *The Annals of Statistics*, **25**, 613-641.

Stute, W. and Zhu, L.X. (2005). Nonparametric checks for single-index models. *The Annals of Statistics*, **33**, 1048-1083.

van der Vaart, A. and Wellner, J.A. (1996). *Weak convergence and empirical processes.* Springer, New York.

Wu, J. and Zhu, L.X. (2007). Diagnostic checking for conditional heteroscedasticity models. *A preprint.*

Zhu, L.X. (2003). Model checking of dimension-reduction type for regression. *Statistica Sinica*, **13**, 283-296.