RESEARCH REPORT

Serial No. 484

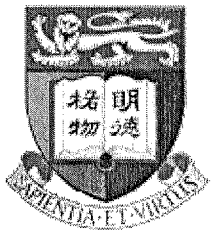
June 2012

(Revised)

ON CONDITIONALLY HETEROSCEDASTIC AR MODELS WITH THRESHOLDS

by

Kung-Sik Chan, Dong Li, Shiqing Ling and Howell Tong



THE UNIVERSITY OF HONG KONG
DEPARTMENT OF STATISTICS AND ACTUARIAL SCIENCE

Pokfulam Road, Hong Kong. E-Mail: saas@hku.hk
Tel.: (852) 2859 2466/7 Fax: (852) 2858 9041

On Conditionally Heteroscedastic AR Models with Thresholds

Kung-Sik Chan, Dong Li, Shiqing Ling and Howell Tong

University of Iowa, Hong Kong University of Science & Technology and London School of Economics & Political Science

Abstract: Conditional heteroscedasticity is often observed in many real time series data in diverse fields, including ecology, economics, finance and others. Its modelling has attracted considerable attention in the literature for more than 30 years. Probably the most popular is the autoregressive conditional heteroscedasticity (ARCH) model introduced by Engle (1982), which was later generalized by many others. By viewing the ARCH approach as a dynamic mixture of independent random variables, we explore the potential provided by visiting the most basic form of a dynamic mixing function. The exploration has enabled us to develop systematically a simple yet versatile observable mixing function, leading to the conditionally heteroscedastic AR model with thresholds, or a T-CHARM for short. Besides its capability, among others, of capturing the so-called volatility clustering and asymmetric conditional heteroscedasticity, and having heavier tail than the innovations, the T-CHARM can preserve non-negativity of the conditional variance as well as stationarity *without* placing restrictive constraints on its parameters. Indeed, the mixing function of a T-CHARM need not even be an increasing function of the magnitude of the state variable. We further provide fairly comprehensive theoretical underpinnings for statistical practice, which are supported by efficient computation procedures and algorithms. We report some of our experiences with T-CHARM of real time series from several disciplines, namely economics, biology and environmental science, which highlight the general utility of T-CHARM, especially beyond economic time series. Finally, we indicate potentials of the approach to multivariate time series as well as random fields.

Key words and phrases: Compound Poisson process; conditional variance; heavy tail; heteroscedasticity; limiting distribution; quasi-maximum likelihood estimation; random field; score test; T-CHARM; threshold model; volatility.

1 Introduction

A time series can often be modelled as the sum of a conditional mean function (i.e. the drift or trend) and a conditional variance function (i.e. the diffusion).

See, e.g., Tong (1990, p. 98). In the time series literature, the former attracted attention from very early days. Although the importance of the latter did not go entirely unnoticed, with an early example in ecological population dynamics, e.g. Moran (1953), a systematic modelling of the latter did not seem to attract serious attention before the 1980s.

For discrete-time cases, on which this paper focuses, and as far as we are aware, it is in the econometric and finance literature that the modelling of the conditional variance has been treated really seriously. Indeed Engle (1982) proposed the autoregressive conditional heteroscedasticity (ARCH) model, which has attracted a wide following in the above literature. (For a simple introduction see, e.g., Cryer and Chan (2008).)

Now, given the popularity of the ARCH model and its many variants, we might be asked as to why we should wish to introduce an alternative approach. A short answer is the fact that there is more than one way to skin a cat. We only need to cite the area of dimension reduction as a recent example, in which we have methods based on projection pursuit, sliced inverse regression, principal Hessian direction, minimum average variance estimation, and others. (See, e.g., Xia *et al.* (2002).) A longer answer is this. We view the ARCH model as a dynamic mixture of independent random variables. Specifically, it uses a mixing function that is the square-root of a linear combination of the squares of relevant past observations. Now, in order to preserve non-negativity of the conditional variance as well as stationarity, constraints on the parameters of the ARCH model are necessary. It turns out that these constraints can be rather restrictive for practical applications. Moreover, for different applications, it has been found necessary to amend the ARCH model in different directions. In fact, a wikipedia search reveals a vastly extended family with members carrying acronyms like GARCH, EGARCH, IGARCH, NGARCH, GARCH-M, QGARCH, GJR-GARCH, TGARCH, fGARCH. (The list is probably incomplete.) Each acronym corresponds to an attempt to amend the basic ARCH model to cope with different practical demands, by extending the basic ARCH mixing function to ones of increasing complexity and carrying with them associated parameter constraints. Therefore against the above background, we would argue that there is no *a priori* reason why the mixing function should be wedded to the form ini-

tiated in Eagle (1982) and generalized by its followers. Therefore, it is pertinent to go back to basics and seek afresh an alternative approach based on a most basic form of the mixing function.

Note that although unobservable mixing functions have been studied they are not the focus of this paper.

Now, the simple idea of a piecewise constant mixing function was already sown in Moran (1953) and made its debut in Tong and Lim (1980). Let us not forget that piecewise constant functions are basic to many core statistical areas, e.g. histograms for density estimation and regressograms for regression analysis. Moreover, Tong (1982) discussed how a discontinuous decision process (e.g. as that often incurred in investment strategies) could result in a smooth dynamical system becoming a piecewise smooth one. We pursue the idea systematically in this paper. The aims are multi-fold: (i) to demonstrate the existence of a viable and simple alternative to model conditional heteroscedasticity which we shall call T-CHARM; (ii) to show that *without* imposing restrictive conditions on its defining parameters, a T-CHARM can preserve non-negativity of the conditional variance as well as stationarity, and can have a distribution with a heavier tail than that of the innovation; (iii) to show that the mixing function need not be an increasing function of the magnitude of the state variable; (iv) to show the applicability of T-CHARM to areas much wider than economics; (iv) to outline scopes of extension into multivariate time series, discrete-valued processes, spatio-temporal processes, image processing and others.

Clearly, thresholds will be the key to the piecewise-constant approach to modelling the conditional variance function, just as they are to the piecewise-linear approach to modelling the conditional mean function. Recent references of the latter since its introduction by Tong (1978) include Chan (2009), Hansen (2011), Tong (2011) and a special issue edited by Chan and Li (2007). Briefly, they include, for example, ecological dynamics (Stenseth, 2009), threshold random coefficient AR models (Brockwell *et al.*, 1992), threshold moving-average models (Gooijer, 1998; Ling and Tong, 2005; Li and Li, 2008; Li, Ling and Tong, 2012; Li, Ling and Li, 2012), threshold ARMA models (Tong, 1990; Li and Li, 2011; Li *et al.*, 2011), threshold cointegration (Balke and Fomby, 1997; Hansen and Seo, 2002), threshold unit root (Enders and Granger, 1998) and many others.

Although these experiences are valuable, many new challenges remain before we can achieve our stated goal.

This paper considers a particularly simple threshold model that focuses on the conditional variance (sometimes also called volatility in econometrics and finance); we call it the threshold model for conditional heteroscedasticity, or T-CHARM for short. An appealing feature of the T-CHARM is that it is always strictly stationary and ergodic *essentially without any restriction on the parameters.* Its mixing function is so flexible that asymmetric heteroscedasticity poses no particular problem. Strong consistency is enjoyed by the quasi-maximum likelihood estimators of the parameters and convergence rates are available. In fact, the limiting distribution of the estimated threshold is the same as that of the smallest minimizer of a two-sided compound Poisson process. When the re-scaled error is normal, it is noteworthy that the percentiles of the limiting distribution can be tabulated readily. Other parameter estimates are shown to be asymptotically normal.

Our proposed approach bears some resemblance to the QGARCH model (Gourieroux and Monfort, 1992) and the Threshold ARCH model (Rabemananjara and Zakoïan, 1993; Zakoïan, 1994), in that they all carry with them some non-parametric flavour. However, the T-CHARM differs fundamentally from these earlier developments in that while the QGARCH model and the Threshold ARCH model assume *fixed and known* thresholds over which the conditional variance function jumps, the T-CHARM is a systematic development that infers from data the *unknown* number and *unknown* locations of the thresholds. Unless there is prior substantive knowledge, a fixed-threshold approach, being rather like a spline approach, generally requires stipulating a large number of thresholds for an adequate piecewise-constant approximation to the true conditional variance function. Thus, for finite samples, some of the estimates of jumps may have excessive standard errors. On the other hand, our data oriented approach to estimating the number and locations of the thresholds tends to result, in practice, in a much parsimonious parametrization involving only a minimal number of thresholds; see the real applications below.

This paper is organized as follows. Section 2 presents the model and its structure. Section 3 studies the estimation of the model parameters. Section 4

gives the limiting distribution of the estimated threshold including a useful approximation. Section 5 considers various statistical tools pertinent to T-CHARM modelling. Section 6 illustrates the methodology by reference to three sets of real data, drawn respectively from finance, the biological science and the environmental science. We conclude the paper with some discussion in Section 7. Proofs of key results are relegated to the appendix.

2 T-CHARM and its probabilistic structure

A T-CHARM is a simple(st) threshold autoregressive (TAR) model. In its simplest form (-a more general form will be given in the next section):

$$X_t = \sigma(X_{t-1})\eta_t, \tag{1}$$

where $\{\eta_t\}$ are independent and identically (but not necessarily normally) distributed (i.i.d.) random variables each with zero mean and unit variance, η_t is independent of $\{X_s : s < t\}$, and $\sigma(x)$ is a piecewise constant function of x. Assume that σ takes m distinct values. Specifically, let $\sigma(x) = \sigma_i$ for $x \in R_i$, where σ_i's are distinct positive numbers, and $\{R_i, i = 1, \ldots, m\}$ define a partition of the real line R, i.e., $R = \cup_{i=1}^{m} R_i$ and the R_i's are pairwise disjoint. The R_i's will be referred to as regimes below. Thus, the conditional variance of X_t given current and past X's depends only on the regime into which X_{t-1} falls. It is clear that $\mathbb{E}(X_t|X_s, s < t) \equiv 0$, so $\{X_t\}$ is a martingale difference sequence of random variables. Consequently, the X's are uncorrelated forming a sequence of white noise. Clearly, a T-CHARM can be generalized in many different ways, just as a TAR model can; for a survey, see Tong (2011). For example, the argument of the function σ can be replaced by a more general 'state' variable, which can be a function of either an observable or a hidden time series or both. Another direction is to soften the thresholding by smoothing the piecewise constant function σ, which can be accomplished obviously in a multitude of ways, e.g. similar to those described in Tong (2011).

Recall that the ARCH approach builds on assuming, in its simplest form, $\sigma(x) = \sqrt{\alpha_0 + \alpha_1 x^2}$, α_0 and α_1 being positive constants. Loosely speaking, the ARCH approach adopts a continuous mixture of independent random variables,

with a mixing function that lacks flexibility, e.g. it is always an increasing function of $|x|$ for positive α_1.

Now, future X's can be predicted based on the information contained in the current and past X's. As linear prediction is valueless given the white noise structure of the X's, we must turn to nonlinear prediction. The result stated below indicates that the dependence structure of the X's is generally revealed by the autocorrelation structure of some instantaneously nonlinear transformation of the X's. In fact, we shall show that for any instantaneous transformation $Y_t = h(X_t)$ with finite second moments and $\mathbb{E}(h(\eta_t)) \neq 0$, the autocorrelation function (ACF) of the transformed process is generally the same as that of some stationary ARMA$(m-1, m-1)$ process. Note that both the AR and the MA orders are generally equal to one less than the number of regimes.

Define the regime process $\{S_t\}$ for which $S_t = i$ if and only if $X_t \in R_i, i = 1, \ldots, m$. In words, the regime process signifies the regimes into which the observations fall. oNow, the regime process is a Markov chain. Let $P = (p_{ij})$ be the 1-step transition probability matrix of the regime process, i.e., $p_{ij} = \mathbb{P}(X_{t+1} \in R_j | X_t \in R_i)$. Being a stochastic matrix, P admits 1 as its maximum eigenvalue. Let 1 denote an m-dimensional vector of unit elements. Then $P1 = 1$. So, the vector 1 is a right eigenvector of P. Consequently, there exists an m-dimensional non-zero left eigenvector u corresponding to the unit eigenvalue so that

$$u^\tau P = u^\tau, \tag{2}$$

where τ denotes the transpose of a vector or matrix. If u is non-negative, it can be normalized to sum to 1. Then if the regime process has u as its initial probability distribution, the regime process becomes stationary. The existence and uniqueness of a positive stationary distribution for the regime process can be guaranteed if the transition probability matrix is irreducible, under which assumption the transition probability matrix has a simple unit eigenvalue, i.e., of unit multiplicity, and all its other eigenvalues are less than 1 in magnitude. Indeed, the irreducibility of P is a necessary and sufficient condition for the existence of a unique stationary distribution for the Markov chain. The transition matrix P is irreducible if and only if $\sum_{t=1}^{\infty} P^t$ is a positive matrix. P is irreducible under very mild conditions, for example if η_t has a positive probability density function. Henceforth, we assume that P is an irreducible matrix.

The vector of stationary probabilities u admits a closed-form solution in terms of the $m \times m$ transition probability matrix P. Partition P as follows:

$$P = \begin{pmatrix} P_{AA} & P_{AB} \\ P_{BA} & P_{BB} \end{pmatrix},$$

where P_{AA} is $(m-1) \times (m-1)$ and P_{BA} is $1 \times (m-1)$. Similarly, partition u as $(u_A^\tau, u_B)^\tau$, where u_A is of dimension $m-1$. Let 1 denote the $(m-1)-$dimensional column vector all of whose elements equal 1. Since u sums to 1, $u_B = 1 - u_A^\tau 1$. Now, (2) is equivalent to

$$u_A^\tau P_{AA} + (1 - u_A^\tau 1) P_{BA} = u_A^\tau,$$

implying $u_A^\tau = P_{BA}(I + 1P_{BA} - P_{AA})^{-1}$, where I is the $(m-1) \times (m-1)$ identity matrix. For the case of two regimes, i.e. $m = 2$, let

$$P = \begin{pmatrix} 1 - \nu_2 & \nu_2 \\ \nu_1 & 1 - \nu_1 \end{pmatrix}.$$

Then we have $u^\tau = (\frac{\nu_1}{\nu_1+\nu_2}, \frac{\nu_2}{\nu_1+\nu_2})$. Note that the irreducibility of P ensures that $\nu_1 + \nu_2 > 0$.

We now show that the stationary distribution of X_t is a mixture of distributions of $\sigma_j \eta$, where η has the same distribution as the common distribution of the innovations $\{\eta_t\}$, with u_j as the probability weights, where $u^\tau = (u_1, u_2, \ldots, u_m)$. Conditional on $X_0 = x_0 \in R_i$, $S_{t-1} = j$ with probability P_{ij}^{t-1}, in which case X_t is distributed as $\sigma_j \eta_t$. Because $P_{ij}^{t-1} \to u_j$, the conditional distribution of X_t converges in distribution to a mixture of distributions of $\sigma_j \eta$ with probability u_j. It is readily seen that this limiting mixture distribution is the stationary distribution of $\{X_t\}$. A careful examination of the above arguments show that the ℓ-step ahead predictive distribution of $X_{t+\ell}$ given X_t in the ith regime is a mixture of distributions of $\sigma_j \eta$ with probability weight $P_{ij}^{\ell-1}$. In practice, if the X's are returns, it is of interest to predict the uncertainty in the ℓ-step-ahead cumulative returns. That is to say the interest is to explore the predictive distribution of $\sum_{k=1}^{\ell} X_{t+k}$. It can be seen that the latter distribution is identical to that of a mixture of distributions corresponding to $\sum_{k=1}^{\ell} \sigma_{j_k} \eta_{t+k}$, with probability weights $P_{i,j_1} \prod_{k=2}^{\ell} P_{j_{k-1},j_k}$, for $j_k \in \{1, \ldots, m\}, k = 1, ..., \ell$. These distributional results will be useful in value-at-risk calculations in finance.

It follows from the Cayley-Hamilton theorem that if $d(x) = \det(xI - P)$ is the characteristic polynomial of P, then $d(P) = 0$. Simply let $d(x) = x^m - \sum_{j=1}^m d_j x^{m-j}$, $d(P) = P^m - \sum_{j=1}^m d_j P^{m-j}$, where the superscript denotes a matrix power and the zeroth power is I, the identity matrix. Since 1 is the unique eigenvalue of P that is of unit magnitude, $d(x) = (x-1)c(x)$, where $c(x) = x^{m-1} - \sum_{j=1}^{m-1} c_j x^{m-j-1}$ has all its root of magnitude strictly less than 1.

Theorem 2.1 *Let $\{X_t\}$ be defined by Eqn. (1) and the transition probability matrix P of the associated regime process $\{S_t\}$ be an irreducible $m \times m$ matrix. Let $Y_t = h(X_t)$, where h is a continuous function. Assume that $\{Y_t\}$ admits finite second moments and $\mathbb{E}(h(\eta_t)) \neq 0$. Let $\gamma_k = \gamma_{k,Y}$ be the kth lag auto-covariance of $\{Y_t\}$. Then $\{\gamma_k\}$ satisfies the Yule-Walker equation*

$$\gamma_k = c_1\gamma_{k-1} + \ldots + c_{m-1}\gamma_{k-m+1} \quad \text{for } k \geq m. \tag{3}$$

The fact that $\{\gamma_k\}$ satisfies the Yule-Walker equation means that the ACF of $\{Y_t\}$ is exactly the same as that of some ARMA$(m-1, m-1)$ process. To see this, let B be the backshift operator defined by $BY_t = Y_{t-1}$. Observe that because of the Yule-Walker equation, $W_t = c(B)Y_t$ is a process of memory not more than $m-1$ lags, and so $\{W_t\}$ must be an MA$(m-1)$ process by Proposition 3.2.1 of Brockwell and Davis (1991). Thus, in terms of the second order structure, $\{Y_t\}$ is an ARMA$(m-1, m-1)$ process. However, it can be seen from the proof of Theorem 2.1 that if the vector $\nu^\tau = (\mathbb{E}(h(\sigma_1\eta_t)), \ldots, \mathbb{E}(h(\sigma_m\eta_t)))$ is orthogonal to some eigenvectors of P whose corresponding eigenvalues are less than 1 in magnitude, then the ARMA orders may be lowered. However, the latter happens only if ν lies in a set of zero Lebesgue measure. For example, this exceptional case occurs if h is a linear function, in which case $\nu = 0$ and is orthogonal to all eigenvectors of P. Actually, in this case, $\{Y_t\}$ is a sequence of white noise process, an ARMA(0, 0) process! However, for a nonlinear transformation h, it is unlikely that ν is orthogonal to any eigenvector of P whose corresponding eigenvalue is less than 1 in magnitude. Thus, we have the generic result that any instantaneous nonlinear transformation of $\{X_t\}$ is an ARMA$(m-1, m-1)$ process. This result forms a basis for tentatively identifying the number of regimes of the T-CHARM. For example, we can consider the square of the X process and tentatively identify its ARMA orders by using existing methodologies such as the Extended ACF

(EACF, see Tsay and Tiao, 1984), the AIC and others; see, e.g., Cryer and Chan (2008). The number of regimes m may be estimated by adding 1 to the estimated AR order. The identification may be verified by repeating the procedure after taking the absolute value the original process. Note that the above discussion subsumes an earlier result of Gourieroux and Monfort (1992), who derived the aforementioned ARMA representation for the special case $h(x) = x^2$.

Another issue concerns the definition of regimes. The solution is likely to be application-specific. Empirically, a simple partition scheme consists of defining $R_j = \{x \in R : r_j < g(x) \le r_{j+1}\}$, where $-\infty = r_0 < r_1 < r_2 < \cdots < r_{m-1} < r_m = \infty$, and g is some function, for example, the identity function or the absolute value function. We shall illustrate the above techniques with real data later.

Denote $I_t = (I_{2t}, ..., I_{mt})^\tau$, where $I_{it} = I\{X_t \in R_i\}$, $i = 2, ..., m$, and $I\{\cdot\}$ is an indictor function. Let $a_t = (I\{\sigma_1\eta_t \in R_2\}, ..., I\{\sigma_1\eta_t \in R_m\})^\tau$ be an $(m-1)$-dimensional vector and A_t be an $(m-1) \times (m-1)$ matrix whose (i, j) entry equals $I\{\sigma_{j+1}\eta_t \in R_{i+1}\} - I\{\sigma_1\eta_t \in R_{i+1}\}$. Then, we have

$$I_t = a_t + A_t I_{t-1}.$$

After iterating k times, it follows that

$$I_t = \sum_{j=0}^{k} \Big(\prod_{i=0}^{j-1} A_{t-i} \Big) a_{t-j} + \Big(\prod_{i=0}^{k} A_{t-i} \Big) I_{t-k-1},$$

with the convention $\prod_{i=0}^{-1} = I$ and $\prod_{i=0}^{j} A_{t-i} = A_t A_{t-1} \cdots A_{t-j}$. Assume that η_t has a positive density on the real line R and the Lebesgue measure of each R_i is positive. Since the matrices in the sequence $\{A_t\}$ are i.i.d. and $\mathbb{E}|I\{\sigma_j\eta_t \in R_i\} - I\{\sigma_1\eta_t \in R_i\}| < 1$, using the technique in the proof of Theorem 2.1 in Li, Ling and Tong (2012), we can see that the first term converges to $\sum_{j=1}^{\infty} \Big(\prod_{i=0}^{j-1} A_{t-i} \Big) a_{t-j}$ almost surely (a.s.), while the second term converges to zero a.s. Clearly, the infinite series above is strictly stationary and ergodic since it is a function of i.i.d. $\{\eta_t\}$. Summarizing the above discussion, we have the following result.

Theorem 2.2 *If the density of η_t, denoted by $f(\cdot)$, is positive on R and the Lebesgue measure of each R_i is positive, then, (i) I_t has the following expansion*

$$I_t = \sum_{j=0}^{\infty} \Big(\prod_{i=0}^{j-1} A_{t-i} \Big) a_{t-j}$$

and $\sigma(X_t) = \sigma_1 + (\sigma_2 - \sigma_1, ..., \sigma_m - \sigma_1)I_{t-1}$*; (ii) model (1) is always strictly stationary and uniformly ergodic.*

It remains to verify the uniform ergodicity of the process, which follows from Corollary 6.12 of Nummelin (1984) and the fact that $p(x, y)$, the transition probability density function of $\{X_t\}$ (w.r.t. the Lebesgue measure), is bounded below by $K \times f(y)$ for all $x, y \in R$, where $0 < K < \infty$ is the minimum of $\sigma(\cdot)$, i.e. the state space R is a small set. We should mention that we do not impose any condition on $\sigma_i > 0$ in Theorem 2.2 and the results hold even when $\mathbb{E}|\eta_t| = \infty$. Note that Theorem 2.2 significantly improves Proposition 1 of Gourieroux and Monfort (1992), who obtained the strict stationarity of the process under similar regularity conditions.

We further study the autocorrelation structure of the conditional variances $\{\sigma^2(X_t)\}$ in model (1). For $k \geq 0$, simple calculations give that

$$\mathrm{cov}(\sigma^2(X_t), \sigma^2(X_{t-k})) = \sum_{i=1}^{m}\sum_{j=1}^{m} \sigma_i^2\sigma_j^2\delta_{ij}^{(k)},$$

where $\delta_{ij}^{(k)}$ satisfies the following iterative equations

$$\begin{aligned} \delta_{ij}^{(k)} &= \sum_{s=1}^{m} \mathbb{P}(\sigma_s\eta_t \in R_i)\delta_{sj}^{(k-1)} \\ \delta_{ij}^{(0)} &= \mathbb{P}(X_t \in R_i \cap R_j) - \mathbb{P}(X_t \in R_i)\mathbb{P}(X_t \in R_j). \end{aligned}$$

For $k = 0$, we can get the variance of $\sigma^2(X_t)$, namely

$$\mathrm{var}(\sigma^2(X_t)) = \sum_{1\leq i<j\leq m} (\sigma_j^2 - \sigma_i^2)^2\mathbb{P}(X_t \in R_j)\mathbb{P}(X_t \in R_i).$$

Here, $\mathbb{P}(X_t \in R_i)$'s can be uniquely determined by the following equations

$$\mathbb{P}(X_t \in R_i) = \sum_{j=1}^{m} \mathbb{P}(\sigma_j\eta_t \in R_i)\mathbb{P}(X_t \in R_j) \quad \text{and} \quad \sum_{i=1}^{m} \mathbb{P}(X_t \in R_i) = 1.$$

Thus, it is not hard to obtain the ACF $\{\rho_k\}$ of $\sigma^2(X_t)$ in principle, although the general expression can be complicated. However, for the case $m = 2$, we have simple expressions.

Theorem 2.3 *Suppose that* $m = 2$ *and the assumptions in Theorem 2.2 hold. Then, for* $k \geq 0$

1. $\mathrm{cov}(\sigma^2(X_t), \sigma^2(X_{t-k})) = (\sigma_2^2 - \sigma_1^2)^2\delta(1-\delta)\{\mathbb{P}(\sigma_2\eta_t \in R_2) - \mathbb{P}(\sigma_1\eta_t \in R_2)\}^k$, *where* $\delta = \mathbb{P}(\sigma_1\eta_t \in R_2)/\{1 - \mathbb{P}(\sigma_2\eta_t \in R_2) + \mathbb{P}(\sigma_1\eta_t \in R_2)\}$;

2. $\rho_k = \{\mathbb{P}(\sigma_2\eta_t \in R_2) - \mathbb{P}(\sigma_1\eta_t \in R_2)\}^k$.

This result shows that the ACF of $\sigma^2(X_t)$ decays to zero at an exponential rate and that clustering (due to positive covariances) may be obtained.

It is interesting to compare the ACFs of T-CHARM with those of GARCH models. For simplicity of discussion, let us consider the popular GARCH (1,1) model, namely $\sigma_t^2 = \alpha_0 + (\alpha\eta_t^2 + \beta)\sigma_{t-1}^2$, where $\alpha > 0$ and $\beta > 0$. The corresponding ACF $\{\rho_{gk}\}$ is

$$\rho_{gk} = (\alpha + \beta)^k,$$

where $\alpha + \beta < 1$. It is well-known that the estimation of ρ_{gk} requires a finite fourth moment condition, i.e., $2\alpha^2 + (\alpha+\beta)^2 < 1$, and this very strong condition is rarely satisfied by the estimated parameters in many real applications. As such the ρ_{gk} or its estimator would be hard put to provide a reliable assessment of the estimated autocorrelations of conditional variances. In contrast, we do not need any restriction on the parameters for the ρ_k of T-CHARM. We suggest that model (1) can offer a reliable alternative to modelling conditional variances. Additionally, model (1) may also capture the heavy-tailed property of financial time series because

$$\frac{\mathbb{E}X_t^4}{(\mathbb{E}X_t^2)^2} = (\mathbb{E}\eta_t^4)\frac{\sum_{i=1}^m \sigma_i^4\mathbb{P}(X_{t-1} \in R_i)}{\{\sum_{i=1}^m \sigma_i^2\mathbb{P}(X_{t-1} \in R_i)\}^2} \geq \mathbb{E}\eta_t^4$$

by Jensen's inequality.

Note that the inequality is strict for $m \geq 2$ if the innovations have infinite support and $\{\sigma_i, i = 1, 2, \cdots, m\}$ is not a singleton, implying that the T-CHARM generally has a heavier tail than the innovations. To illustrate, we consider the two following T-CHARMs:

$$y_t = [2I\{y_{t-1} \leq 0\} + \sigma I\{y_{t-1} > 0\}]\eta_t \tag{4}$$

and

$$y_t = [I\{y_{t-1} \leq r\} + 2I\{y_{t-1} > r\}]\eta_t \tag{5}$$

with η_t being standard normal. The left diagram in Fig. 1 plots the theoretical kurtosis of the stationary T-CHARM as a function of σ for the model defined by (4) while the right figure corresponds to model (5). For both models, the kurtosis exceeds that of the innovations, as it must be the case.

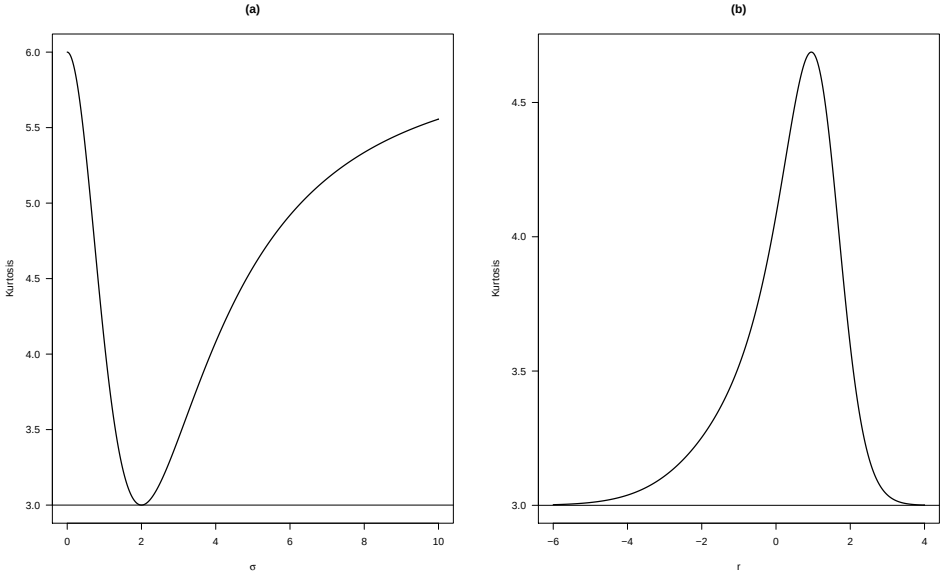


Figure 1: The kurtosis of models (4) and (5).

3 Quasi-maximum likelihood estimation

In practice, a simple way to implement model (1) is to adopt the form $R_i = (r_{i-1}, r_i]$. Here, we consider a slightly more general model than (1). The threshold variable is allowed to be a functional of the past information. That is, the model is defined as

$$\begin{aligned} X_t &= \sigma(W_{t-1})\eta_t, \\ \sigma(W_{t-1}) &= \sum_{i=1}^{m} \sigma_i I\{r_{i-1} < W_{t-1} \le r_i\}, \end{aligned} \tag{6}$$

where W_{t-1} is a known functional of $\{X_{t-1}, ..., X_{t-p}\}$, m is the number of regimes, σ_i's are positive numbers and $-\infty = r_0 < r_1 < \cdots < r_{m-1} < r_m = \infty$,

$\{r_1, ..., r_{m-1}\}$ being the threshold parameters. In this and next sections, m is assumed to be known. Denote $\theta = (\sigma_1^2, ..., \sigma_m^2)^\tau$ and $\mathbf{r} = (r_1, ..., r_{m-1})^\tau$. Let $\theta_0 = (\sigma_{10}^2, ..., \sigma_{m0}^2)^\tau$ and $\mathbf{r}_0 = (r_{10}, ..., r_{m-1,0})^\tau$ be the true values of θ and $\mathbf{r}$, respectively. Note that we do not assume that η_t is normally distributed. To estimate the parameters from data $\{X_1, ..., X_n\}$ given the initial values $\{X_{-p}, ..., X_0\}$, we adopt the following objective function:

$$L_n(\theta, \mathbf{r}) = -\frac{1}{2} \sum_{t=1}^{n} \sum_{i=1}^{m} \Big(\log \sigma_i^2 + \frac{X_t^2}{\sigma_i^2} \Big) I_{it}, \tag{7}$$

where $I_{it} = I\{r_{i-1} < W_{t-1} \le r_i\}$. For each $\mathbf{r}$, it is easy to maximize $L_n(\theta, \mathbf{r})$ with respect to θ, which we denote by $\hat{\theta}_n(\mathbf{r}) \equiv (\hat{\sigma}_{1n}^2(\mathbf{r}), ..., \hat{\sigma}_{mn}^2(\mathbf{r}))^\tau$ with

$$\hat{\sigma}_{in}^2(\mathbf{r}) = \frac{\sum_{t=1}^{n} X_t^2 I_{it}}{\sum_{t=1}^{n} I_{it}}, \quad i = 1, ..., m.$$

Note that there are at most finitely many different values of $L_n(\hat{\theta}_n(\mathbf{r}), \mathbf{r})$. By the enumeration approach, we estimate $\mathbf{r}_0$ by

$$\hat{\mathbf{r}}_n = \arg\max_{\mathbf{r}} L_n(\hat{\theta}_n(\mathbf{r}), \mathbf{r}).$$

Using the plug-in method, the estimator of θ_0 is given by $\hat{\theta}_n = \hat{\theta}_n(\hat{\mathbf{r}}_n)$. We call $(\hat{\theta}_n, \hat{\mathbf{r}}_n)$ the quasi-maximum likelihood estimator (QMLE) of $(\theta_0, \mathbf{r}_0)$.

Generally, $\widehat{\mathbf{r}}_n$ takes the form $(W_{(i_1)}, ..., W_{(i_{m-1})})^\tau$, where $i_1 < \cdots < i_{m-1}$ and $\{W_{(1)}, ..., W_{(n)}\}$ are the order statistics of $\{W_1, ..., W_n\}$. If $(W_{(j_1)}, ..., W_{(j_{m-1})})^\tau$ is an estimator of $\mathbf{r}_0$, then $L_n(\hat{\theta}_n(\mathbf{r}), \mathbf{r})$ is a constant over the $(m-1)$-dimensional cube $\widetilde{R}$, where

$$\widetilde{R} = \{\mathbf{r} = (r_1, ..., r_{m-1})^\tau : \ r_i \in [W_{(j_i)}, W_{(j_i+1)}), \ i = 1, ..., m-1\}.$$

Thus, there exist infinitely many $\mathbf{r}$ such that $L_n(\cdot)$ can achieve its global maximum and each $\mathbf{r} \in \widetilde{R}$ can be taken as an estimator of $\mathbf{r}_0$. In this case, we choose $(W_{(j_1)}, ..., W_{(j_{m-1})})^\tau$ as a representative of $\widetilde{R}$ and denote it as the estimator of $\mathbf{r}_0$. With the above procedure, it is not hard to show that $(\hat{\theta}_n, \hat{\mathbf{r}}_n)$ is the QMLE of $(\theta_n, \mathbf{r}_n)$, so

$$(\hat{\theta}_n, \hat{\mathbf{r}}_n) = \arg\max_{\Theta \times \mathcal{R}} L_n(\theta, \mathbf{r}),$$

where $\Theta \times \mathcal{R}$ is the parameter space and $\Theta = R_+^m$ with $R_+ \equiv (0, \infty)$ and $\mathcal{R} = \{\mathbf{r} : -\infty < r_1 < \cdots < r_{m-1} < \infty\}$. To state our results, we need the following assumptions.

Assumption 3.1 *The density $f(x)$ of η_t is continuous and positive on R, $\mathbb{E}\eta_t = 0$ and $\mathbb{E}\eta_t^2 = 1$.*

Assumption 3.2 *The density $f_w(\cdot)$ of W_t is continuous and $f_w(r_{j0}) > 0$ for $j = 1, ..., m-1$.*

The following theorem establishes the strong consistency of $(\hat{\theta}_n, \hat{\mathbf{r}}_n)$. The proof is similar to that of Theorem 1 in Chan (1993) and is therefore omitted.

Theorem 3.1 *If (i) Assumptions 3.1 and 3.2 hold, and (ii) $\sigma_{i0}^2 \neq \sigma_{i+1,0}^2$ for $i = 1, ..., m-1$, then, $(\hat{\theta}_n, \hat{\mathbf{r}}_n) \to (\theta_0, \mathbf{r}_0)$ a.s. as $n \to \infty$.*

Condition (ii) in Theorem 3.1 is required to ensure the identifiability of $\mathbf{r}_0$. By a technique similar to that used in the proof of Proposition 1 in Chan (1993), we have the following theorem which establishes the convergence rate of $\hat{\mathbf{r}}_n$ and the asymptotic normality of $\hat{\theta}_n$.

Theorem 3.2 *Under the conditions of Theorem 3.1, if $\sup_{x \in R}\{(1+|x|)f(x)\} < \infty$ and $\kappa_4 \equiv E\eta_t^4 < \infty$, then*

(a) $n(\hat{\mathbf{r}}_n - \mathbf{r}_0) = O_p(1)$;

(b) $\sqrt{n} \sup_{\|\mathbf{r}-\mathbf{r}_0\| \le B/n} |\hat{\sigma}_{in}^2(\mathbf{r}) - \hat{\sigma}_{in}^2(\mathbf{r}_0)| = o_p(1)$ *for any fixed $B \in (0, \infty)$.*

Furthermore,

$$\sqrt{n}(\hat{\sigma}_{in}^2(\mathbf{r}_0) - \sigma_{i0}^2) \Longrightarrow \mathcal{N}\Big(0, \ \frac{(\kappa_4 - 1)\sigma_{i0}^4}{F_w(r_{i0}) - F_w(r_{i-1,0})}\Big), \ i = 1, ..., m,$$

and all the normalized estimators are asymptotically independent, where $F_w(x)$ is the cumulative distribution function of W_t, and henceforth the symbol $\Longrightarrow$ indicates weak convergence.

The above results are similar to the corresponding results in Chan (1993), who showed that the estimated threshold is n-consistent and asymptotically independent of other estimated parameters. See also Li and Ling (2012).

4 Inference of the threshold parameter $\mathbf{r}_0$

From Theorem 3.2 (a), we know that the convergence rate of $\hat{\mathbf{r}}_n$ is n. To study the limiting distribution of $n(\widehat{\mathbf{r}}_n - \mathbf{r}_0)$, we consider the following profile log-likelihood process:

$$\widetilde{L}_n(\mathbf{s}) = -2\Big\{ L_n\big(\hat{\theta}_n(\mathbf{r}_0 + \frac{\mathbf{s}}{n}),\, \mathbf{r}_0 + \frac{\mathbf{s}}{n}\big) - L_n\big(\hat{\theta}_n(\mathbf{r}_0),\, \mathbf{r}_0\big)\Big\}, \tag{8}$$

where $\mathbf{s} = (s_1, ..., s_{m-1})^\tau \in R^{m-1}$.

Let $\mathbb{D}(R^{m-1})$ denote the function space consisting of uniform limits of sequences of simple functions defined on R^{m-1} that is equipped with the Skorokhod metric (see Seijo and Sen, 2011; Li and Ling, 2012). By Theorem 3.2 and Taylor's expansion, $\widetilde{L}_n(\mathbf{s})$ can be approximated in $\mathbb{D}(R^{m-1})$ by

$$\begin{aligned}
\wp_n(\mathbf{s}) &= L_n(\theta_0, \mathbf{r}_0 + \frac{\mathbf{s}}{n}) - L_n(\theta_0, \mathbf{r}_0) \\
&= \sum_{i=1}^{m-1} \sum_{t=1}^{n} \Big[\xi_t^{(i+1,i)} I\big\{ r_{i0} + \frac{s_i}{n} < W_{t-1} \le r_{i0} \big\} I\{s_i < 0\} \\
&\qquad\qquad + \xi_t^{(i,i+1)} I\big\{ r_{i0} < W_{t-1} \le r_{i0} + \frac{s_i}{n} \big\} I\{s_i \ge 0\} \Big],
\end{aligned}$$

where

$$\xi_t^{(i,j)} = \log \frac{\sigma_{i0}^2}{\sigma_{j0}^2} + \Big(\frac{\sigma_{j0}^2}{\sigma_{i0}^2} - 1 \Big) \eta_t^2, \quad i, j = 1, ..., m. \tag{9}$$

Next, we define $m-1$ independent one-dimensional two-sided compound Poisson processes $\{\mathcal{P}_j(z), z \in R\}$ as

$$\mathcal{P}_j(z) = I\{z < 0\} \sum_{k=1}^{N_1^{(j)}(|z|)} U_k^{(j+1,j)} + I\{z \ge 0\} \sum_{k=1}^{N_2^{(j)}(z)} V_k^{(j,j+1)}, \tag{10}$$

for $j = 1, ..., m-1$, where $\{N_1^{(j)}(z), z \ge 0\}$ and $\{N_2^{(j)}(z), z \ge 0\}$ are two independent Poisson processes with $N_1^{(j)}(0) = N_2^{(j)}(0) = 0$ a.s. and with the same jump rate $f_w(r_{j0})$. Both $\{U_k^{(i,j)}\}_{k=1}^{\infty}$ and $\{V_k^{(i,j)}\}_{k=1}^{\infty}$ are mutually independent copies of $\xi_1^{(i,j)}$. Here, we work with the left continuous version of $N_1^{(j)}(z)$ and the right continuous version of $N_2^{(j)}(z)$.

We further define a spatial compound Poisson process $\wp(\mathbf{s})$ as follows.

$$\wp(\mathbf{s}) = \sum_{j=1}^{m-1} \mathcal{P}_j(s_j), \quad \mathbf{s} = (s_1, ..., s_{m-1})^\tau \in R^{m-1}. \tag{11}$$

Clearly, $\wp(\mathbf{s}) \to \infty$ a.s. as $\|\mathbf{s}\| \to \infty$ since $\mathbb{E}U_t^{(i+1,i)} > 0$ and $\mathbb{E}V_t^{(i,i+1)} > 0$. Therefore, there exists a unique random $(m-1)$-dimensional cube $[\mathbf{M}_-, \mathbf{M}_+) \equiv [M_-^{(1)}, M_+^{(1)}) \times \cdots \times [M_-^{(m-1)}, M_+^{(m-1)})$ at which the process $\{\wp(\mathbf{s}), \mathbf{s} \in R^{m-1}\}$ attains its global minimum a.s.,i.e.,

$$[\mathbf{M}_-, \mathbf{M}_+) = \arg \min_{\mathbf{s} \in R^{m-1}} \wp(\mathbf{s}).$$

From (11), the minimization is equivalent to

$$[M_-^{(j)}, M_+^{(j)}) = \arg \min_{z \in R} \mathcal{P}_j(z), \quad j = 1, ..., m-1.$$

Note that the processes $\{\mathcal{P}_j(z)\}$ are independent, and so are $\{M_-^{(j)}\}$, $j = 1, ..., m-1$. Modifying slightly the proof of Theorem 3.3 in Li and Ling (2012), we can prove the following theorem.

Theorem 4.1 *If the conditions in Theorem 3.2 hold, then $n(\hat{\mathbf{r}}_n - \mathbf{r}_0)$ converges weakly to $\mathbf{M}_-$ and its components are asymptotically independent as $n \to \infty$. Furthermore, $n(\widehat{\mathbf{r}}_n - \mathbf{r}_0)$ is asymptotically independent of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ which is always asymptotically normal.*

Now, we describe how to implement $M_-^{(j)}$ or $\mathbf{M}_-$. From (9) and (10), we know that two factors determine the density of $M_-^{(j)}$, namely the jump rate and the jump distributions. We can simulate $M_-^{(j)}$ via simulating the compound Poisson process $\mathcal{P}_j(z)$ in (10) on the interval $[-T, T]$ for any given $T > 0$ large enough since the expectations of the jumps $U_k^{(j+1,j)}$ and $V_k^{(j,j+1)}$ are positive. Modifying Algorithm 6.2 of Cont and Tankov (2004, p.174) for a one-sided compound Poisson process, we have an algorithm for a two-sided compound Poisson process:

Algorithm

Step 1. Sample $N_1^{(j)}$ and $N_2^{(j)}$ from Poisson distribution with the same parameter $f_w(r_{j0})T$ as the total number of jumps on the intervals $[-T, 0]$ and $[0, T]$, respectively.

Step 2. Sample two independent jump time sequences $\{U_1, \cdots, U_{N_1^{(j)}}\}$ and $\{V_1, \cdots, V_{N_2^{(j)}}\}$, where $\{U_i\} \overset{i.i.d.}{\sim} U[-T, 0]$ and $\{V_i\} \overset{i.i.d.}{\sim} U[0, T]$. Here, $U[a, b]$ denotes the uniform distribution on the interval $[a, b]$.

Step 3. Sample two mutually independent sequences $\{\eta_1, \cdots, \eta_{N_1^{(j)}}\}$ and $\{\eta_1, \cdots, \eta_{N_2^{(j)}}\}$ from $f(x)$. Use them to produce two mutually independent jump-size sequences $\{Y_1, \cdots, Y_{N_1^{(j)}}\}$ and $\{Z_1, \cdots, Z_{N_2^{(j)}}\}$ by (9), respectively.

For $z \in [-T, T]$, the trajectory of (10) is given by

$$\mathcal{P}_j(z) = I\{z < 0\} \sum_{i=1}^{N_1^{(j)}} I\{U_i > z\} Y_i + I\{z \geq 0\} \sum_{j=1}^{N_2^{(j)}} I\{V_j < z\} Z_j. \tag{12}$$

Then, we take the smallest minimizer of $\mathcal{P}_j(z)$ in (12) on $[-T, T]$ as an observation of $M_-^{(j)}$. By repeating the above algorithm, we can get a sequence of observations of $M_-^{(j)}$, from which we can infer the distribution of $n(\hat{r}_{jn} - r_{j0})$.

In practice, however, since only one sample $\mathcal{X}_n = \{X_1, ..., X_n\}$ is available, we can use it to estimate θ_0 and $f_w(r_{j0})$, denoting the estimators as $\hat{\theta}_n$ and $\widehat{f}_w(\widehat{r}_{jn})$, respectively, where $\widehat{f}_w(\cdot)$ is the kernel density estimator of $f_w(\cdot)$. Then we can calculate the residuals $\{\hat{\eta}_1, ..., \hat{\eta}_n\}$ and use them to construct kernel density estimator $\widehat{f}(\cdot)$ of $f(\cdot)$.

When θ_0, $f_w(r_{j0})$ and $f(x)$ are all unknown, we substitute their consistent estimators $\hat{\theta}_n$, $\widehat{f}_w(\widehat{r}_{jn})$ and $\widehat{f}(x)$ in the Algorithm and denote the corresponding compound Poisson processes as $\{\widehat{\mathcal{P}}_j(z)\}$. Then we can get an approximation $\widehat{M}_-^{(j)}$ of $M_-^{(j)}$. Actually, by Theorem 16 in Pollard (1984, p.134), we have that $\widehat{\mathcal{P}}_j(z) \Longrightarrow \mathcal{P}_j(z)$ conditionally on $\mathcal{X}_n$ in $\mathbb{D}(R)$, in probability. By Theorem 3.1 (on the continuity of the smallest argmax functional) in Seijo and Sen (2011), $\widehat{M}_-^{(j)} \Longrightarrow M_-$ conditionally on $\mathcal{X}_n$, in probability. Thus, we have proved the following theorem.

Theorem 4.2 *If the conditions in Theorem 3.2 hold, then, in probability,*

$$\lim_{n\to\infty} |\mathbb{P}(\widehat{M}_-^{(j)} \leq x | \mathcal{X}_n) - \mathbb{P}(M_-^{(j)} \leq x)| = 0$$

at each x for which $\mathbb{P}(M_-^{(j)} = x) = 0$. That is, $\widehat{M}_-^{(j)}|_{\mathcal{X}_n} \Longrightarrow M_-^{(j)}$, in probability.

To illustrate the efficacy of the Algorithm, we consider a simple example

$$X_t = [2I\{X_{t-1} \leq 0\} + 0.5I\{X_{t-1} > 0\}]\eta_t, \tag{13}$$

where η_t is i.i.d. standard normal. Here, the sample size is 400. In Figure 2, (a) gives the density of $n(\widehat{r}_n - r_0)$, which is obtained by 10,000 replications; (b) shows the density of M_- when θ_0, $f_w(r_0)$ and $f(x)$ are all known. When a sample $\mathcal{X} = \{x_1, ..., x_{400}\}$ is given and fixed, (c) and (d) display the density of $\widehat{M}_-$. Here, 1,000 replications are used for (c) and 10,000 replications for (d). From the figure, we can see that they all match well. Comparing (c) with (d), we can find that the more the number of the replications, the more precise the density of $\widehat{M}_-$.

5 Testing for T-CHARM (m) against T-CHARM ($m+1$)

Let us denote a T-CHARM with m regimes by T-CHARM (m). In applications of threshold models, it is relevant to determine the number of regimes. Uncritical applications of conventional techniques such as AIC or BIC can be misleading because, among other issues, we have a non-standard inferential problem, so much so that it is not obvious how to count the number of *independently* adjusted parameters. In this section, we consider testing the T-CHARM(m) against the T-CHARM($m+1$). A testing approach is appropriate because our experience suggests that m is unlikely to be much larger than 2 or 3 in many practical applications. Specifically, under the null H_0, the T-CHARM is defined as in (6). Under the alternative H_1, there is an additional threshold, denoted by r, which lies in the k-regime $(r_{k-1}, r_k]$, such that $\sigma(W_{t-1})$ in the T-CHARM($m+1$) can be written as

$$\sigma(W_{t-1}) = \sum_{\substack{i=1 \\ i\neq k}}^{m} \sigma_i^2 I_{it} + \sigma_{1k}^2 I\{r_{k-1} < W_{t-1} \le r\} + \sigma_{2k}^2 I\{r < W_{t-1} \le r_k\}.$$

Under H_1, the log quasi-likelihood function (ignoring a constant) is

$$L_n(\tilde{\theta}, \mathbf{r}, \sigma_{1k}, \sigma_{2k}, r) = -\frac{1}{2}\sum_{t=1}^{n}\left\{\sum_{\substack{i=1 \\ i\neq k}}^{m}\Big(\log\sigma_i^2 + \frac{X_t^2}{\sigma_i^2}\Big)I_{it} + \Big(\log\sigma_{1k}^2 + \frac{X_t^2}{\sigma_{1k}^2}\Big)I_{1kt} + \Big(\log\sigma_{2k}^2 + \frac{X_t^2}{\sigma_{2k}^2}\Big)I_{2kt}\right\},$$

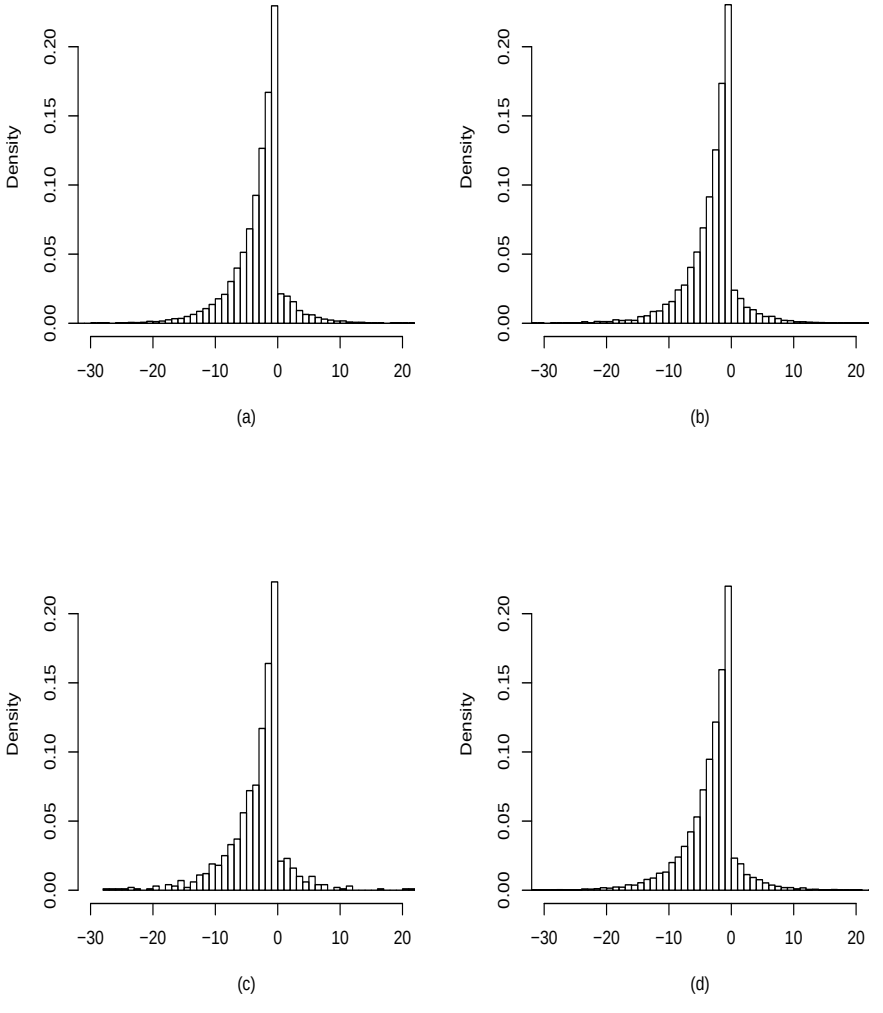


Figure 2: (a) The density of $n(\widehat{r}_n - r_0)$ when $n = 400$ and $\eta_t \sim \mathcal{N}(0,1)$; (b) The density of M_- when θ_0, $f_w(r_0)$ and $f(x)$ are all known and 10,000 replications are used. (c) The density of $\widehat{M}_-$ when $\hat{\theta}_n$, $\widehat{f}_w(\widehat{r}_n)$ and $\widehat{f}(x)$ and 1,000 replications are used. (d) The density of $\widehat{M}_-$ when $\hat{\theta}_n$, $\widehat{f}_w(\widehat{r}_n)$ and $\widehat{f}(x)$ and 10,000 replications are used.

where $\tilde{\theta} = (\sigma_1^2, \cdots \sigma_{k-1}^2, \sigma_{k+1}^2, \cdots, \sigma_m^2)^\tau$, $I_{1kt} = I\{r_{k-1} < W_{t-1} \le r\}$ and $I_{2kt} = I\{r < W_{t-1} \le r_k\}$. For each r, the profile log-likelihood ratio test statistic is

$$\begin{aligned} L_{kn}(r) &= L_n(\hat{\tilde{\theta}}, \hat{\mathbf{r}}, \hat{\sigma}_{1k}(r), \hat{\sigma}_{2k}(r), r) - L_n(\hat{\theta}, \hat{\mathbf{r}}) \\ &= -\frac{1}{2}\sum_{t=1}^{n}\Big[\Big\{\log\hat{\sigma}_k^2(\hat{\mathbf{r}}) + \frac{X_t^2}{\hat{\sigma}_k^2(\hat{\mathbf{r}})}\Big\}I_{kt}(\hat{\mathbf{r}}) + \Big\{\log\hat{\sigma}_{1k}^2(r) + \frac{X_t^2}{\hat{\sigma}_{1k}^2(r)}\Big\}I_{1kt}(\hat{r}_{k-1}, r) \\ &\qquad + \Big\{\log\hat{\sigma}_{2k}^2(r) + \frac{X_t^2}{\hat{\sigma}_{2k}^2(r)}\Big\}I_{2kt}(r, \hat{r}_k)\Big], \end{aligned}$$

where

$$\hat{\sigma}_{1k}^2(r) = \frac{\sum_{t=1}^n X_t^2 I_{1kt}}{\sum_{t=1}^n I_{1kt}} \quad \text{and} \quad \hat{\sigma}_{2k}^2(r) = \frac{\sum_{t=1}^n X_t^2 I_{2kt}}{\sum_{t=1}^n I_{2kt}}.$$

Routine analysis then yields

$$\begin{aligned} 2L_{kn}(r) &= A_k(r)\Big\{\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(\eta_t^2 - 1)I_{1kt} - \frac{F_w(r) - F_w(r_k)}{F_w(r_{k+1}) - F_w(r_k)}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(\eta_t^2 - 1)I_{kt}\Big\}^2 + o_p(1) \\ &= A_k(r)\{D_{kn}(r)\}^2 + o_p(1), \end{aligned}$$

where

$$A_k(r) = \frac{F_w(r_k) - F_w(r_{k-1})}{\{F_w(r) - F_w(r_{k-1})\}\{F_w(r_k) - F_w(r)\}},$$

and $o_p(1)$ holds uniformly in $r \in [r_{i-1}, r_i]$. By Theorem 3.1 in Ling and Tong (2011), $\{D_{kn}(r) : r \in [r_{i-1}, r_i]\}$ converges weakly to a centered Gaussian process $G_k(r) : r \in [r_{i-1}, r_i]\}$ with covariance kernel $K(r, s) = \mathrm{cov}(G_k(r), G_k(s))$ given by

$$\begin{aligned} (\kappa_4 - 1)\Big[&\min\{F_w(r) - F_w(r_{k-1}), F_w(s) - F_w(r_{k-1})\} \\ &- \frac{\{F_w(r) - F_w(r_{k-1})\}\{F_w(s) - F_w(r_{k-1})\}}{F_w(r_k) - F_w(r_{k-1})}\Big]. \end{aligned}$$

Thus, $\{F_w(r_k) - F_w(r_{k-1})\}^{-1/2} D_{kn}(r)$ converges weakly to a scalar multiple of the standard Brownian bridge $B_k(s) - sB_k(1)$, where $s = \frac{F_w(r) - F_w(r_k)}{F_w(r_{k+1}) - F_w(r_k)} \in [0, 1]$. In principle, we should use $2\max_{r\in(r_{k-1}, r_k]} L_{kn}(r)$ to construct a test. But since $2\max_{r\in(r_{k-1}, r_k]} L_{kn}(r) = \infty$, we have to restrict the range of r to $[c_{1,k}, c_{2,k}]$ so that the corresponding s lies in $[a_{1,k}, a_{2,k}]$, a proper subset of $(0, 1)$, say $(0.05, 0.95)$, in each regime. In this way, we can obtain a useful limiting distribution, similar

to those in Bai and Perron (1998) for testing multiple change-point problems and Chan (1990) for testing a threshold AR model. Since we have m regimes under the null hypothesis, the LR test can be applied regime by regime, with an adjustment for multiple testing via the Bonferroni inequality. Specifically, the LR test statistic for an additional threshold in the k-th regime equals

$$T_{k,n} = 2(\kappa_4 - 1)^{-1} \sup_{r \in [c_{1,k},\, c_{2,k}]} L_{kn}(r),$$

where $\kappa_4 \equiv E\eta_t^4$ assumed finite. Summarizing we have the following theorem.

Theorem 5.1 *Under the H_0, if the density function of η_t is bounded and positive, then it follows that*

$$T_{k,n} \Longrightarrow \sup_{s \in [a_{1,k},\, a_{2,k}]} \frac{(B_k(s) - sB_k(1))^2}{s(1-s)},$$

where $B_k(s)$, $k = 1, \cdots, m$, are independent standard Brownian motions.

Since the test is applied regime by regime, we consider the case of $m = 1$ and the range of s equals $[1, 1-a]$ for some $0 < a < 1/2$. Asymptotically, the null distribution of the square root of likelihood ratio test (after some further normalization) is then equivalent to the distribution of the maximum of the normalized absolute Brownian bridge over the interval $[a, 1-a]$. Dirkse (1975) showed that the latter is equal to the distribution of the maximum of the absolute value of a stationary Ornstein-Uhlenbeck process over the interval $[-\alpha, \alpha]$, where $\alpha = \log(1/a - 1)/2$. Specifically, the normalized Brownian Bridge $\{B_s/\sqrt{s - s^2}, 0 < s < 1\}$ is transformed to an Ornstein-Uhlenbeck process $\{U(t), -\alpha < t < \alpha\}$, where s is transformed to $t = \log(s/(1-s))/2$. Moreover, Dirkse (1975) gave the following asymptotic formula:

$$\mathbb{P}(\sup_{-\alpha \le t \le \alpha} |U(t)| > c) \sim \sqrt{2/\pi}\exp(-c^2/2)(\alpha c - \alpha/c + 2/c), \tag{14}$$

for c large. (Here, we replace the typographical error of $1/c$ in Dirkse (1975) by $2/c$, thanks to Professor D. O. Siegmund.) Thus, we can compute the approximate p-value of the LRT by the formula

$$p_0(\tilde{c}) = \sqrt{2/\pi}\exp(-\tilde{c}^2/2)(\alpha\tilde{c} - \alpha/\tilde{c} + 2/\tilde{c}), \tag{15}$$

where $\tilde{c}$ is the square root of the (normalized) LRT statistic. The asymptotic formula derived by Dirkse (1975) then implies that the p-value p_0 is asymptotically uniformly distributed over $[0, 1]$ in the sense that $\mathbb{P}(p_0 < p|H_0) \sim p$ for $p \to 0$.

Next, we consider two modified LRT that make use of the location of the quasi-likelihood estimator of the threshold. Our first modification is based on the observation that were the threshold value under the alternative known to be r_0, then we could use a more powerful likelihood ratio test with the threshold fixed at r_0, whose asymptotic null distribution is χ^2_1. In practice, the threshold is unknown and we consider the likelihood ratio test with a wide range of possible threshold values, which reduces the power of the test. For the case of a Brownian Bridge, it is known that its global maximum value is independent of the location when the maximum is attained, with the latter having a uniform distribution. This independence property also holds, at least asymptotically, for the Ornstein-Uhlenbeck process restricted to a fixed finite interval, say $[h_1, h_2]$, and that the marginal distribution of the location of the global maximum has a uniform distribution over $[h_1, h_2]$; we sketch a proof in the Appendix. Hence, we may modify the calibration of the LRT as follows. Consider the LRT implemented with the threshold searched over the $a \times 100$ to $(1 - a) \times 100$ percentiles of the threshold variable, and that the square root of the LRT attains its maximum value, say, $\tilde{c}$ at the β percentile. We then compute the p-value by formula (15) with α there replaced by $\beta_M = \log(1/\min(\beta, 1 - \beta) - 1)$, i.e. we compute the p-value as if the range of the threshold is from $[\beta, 1 - \beta]$ if $\beta < 0.5$, and then double the corresponding transformed time range in the Ornstein-Uhlenbeck process. Specifically, denote by p_1 the p-value so computed; it is then given by

$$p_1(\tilde{c}, \beta) = \sqrt{2/\pi} \exp(-\tilde{c}^2/2)(\beta_M \tilde{c} - \beta_M/\tilde{c} + 2/\tilde{c})$$

We claim that p_1 is asymptotically uniformly distributed over [0,1], so it provides a valid calibration of the LRT under the null hypothesis. In order to see this, recall that under our conjecture, β is asymptotically independent of $\tilde{c}$, and $\log((1 - \beta)/\beta)/2$ is uniformly distributed over $[-\alpha, \alpha]$. Thus, β_M is uniformly distributed over $[0, 2\alpha]$. The independence of $\tilde{c}$ and β then implies that $\mathbb{E}(p_1|\tilde{c}) = \sqrt{2/\pi} \exp(-\tilde{c}^2/2)(\alpha\tilde{c} - \alpha/\tilde{c} + 2/\tilde{c})$, and hence p_1 shares the same asymptotic behaviour of p_0. For alternatives with the threshold parameter close to the 50 percentile, β is close to one half, rendering the calculation of p_1 to be

based on a narrow interval around the the 50 percentile, which then increases the power of the LRT in discerning threshold structure. On the other hand, if the threshold parameter is close to the data extremes, then p_1 is now calibrated on a very wide interval, thereby decreasing the power of detecting the alternative. This motivates us to develop a third approach for calibrating the p-value of the LRT by searching from the end-points. The idea is now to compute the p-value as if the search was over the union of the intervals from $a \times 100$ to $\min(\beta, 1-\beta) \times 100$ percentiles and from $\max(\beta, 1-\beta) \times 100$ to $(1-a) \times 100$ percentiles. Effectively, the third approach computes the p-value as follows:

$$p_2(\tilde{c}, \beta) = \sqrt{2/\pi} \exp(-\tilde{c}^2/2)(\beta_T \tilde{c} - \beta_T/\tilde{c} + 2/\tilde{c}),$$

where $\beta_T = \log(\min(\beta, 1-\beta)/(1-\min(\beta, 1-\beta)) - \log(a/(1-a))$. We can similarly show that $\mathbb{E}(p_2|\tilde{c}) = p_0$.

We now use simulation to examine the empirical properties of the three approaches for calibrating the p-value of the LRT. We simulated the following threshold martingale difference process:

$$X_t = \sigma(X_{t-1})\eta_t,$$

where $\sigma(X_{t-1}) = \{1 + \gamma \times I(X_{t-1} > r_0)\}\sigma^2$ and $\{\eta_t\}$ are independent standard normal random variables. The noise variance ratio ranges from 0.5 to 2, with increment 0.1. We tried two different threshold values, namely, $r_0 = -0.8$ and 0. The threshold value -0.8 corresponds to the 6.6 percentile with $\gamma = 0.5$ but increases to the 30.7 percentile with γ increasing to 2. On the other hand, the threshold value 0 ranges within the 49–50 percentiles as the noise variance ratio increases from 0.5 to 2. We considered four sample sizes, namely, 100, 200, 400 and 1,000. All experiments were replicated 10,000 times. Table 1 displays the empirical sizes of the LRT with the p-values computed by the three proposed methods. For sample size 100, all result in higher rejection rates than the nominal 5% level, especially for p_1 and p_2. However, for sample sizes $\geq$ 200, the empirical sizes of the tests are increasingly closer to the nominal 5%. Figures 3 and 4 display the empirical power curves of the LRT corresponding to the three methods of calibrating the p-values; in comparing the power of the three methods, we have corrected for slight size differences by shifting the p-values so that they all have 5% empirical size. All methods enjoy good power for

Table 1: Empirical size of the nominal 5% LRT with p-values computed by the three proposed methods.

sample size	p_0	p_1	p_2
100	0.075	0.089	0.096
200	0.057	0.070	0.077
500	0.048	0.061	0.067
1,000	0.052	0.064	0.068

detecting threshold martingale differences, but as expected, p_1 has the highest power when the threshold is close to the median of the threshold variable, whereas p_2 dominates the other two approaches when the threshold is close to the extremes of the data.

6 T-CHARM of some real time series

We now illustrate the application of the T-CHARM with some real time series.

The first example is a financial time series – the daily values of a unit of the CREF stock fund over the period from August 26, 2004 to August 15, 2006. The CREF stock fund is a fund of several thousand stocks. Since stocks are traded only on the so-called trading days, which exclude weekends and holidays, the CREF data do not change over the non-trading days. For simplicity, we shall analyze the returns of the CREF data, namely the first differences of the logarithmic transformed daily values, as if they were equally spaced. The CREF returns, denoted by $\{X_t\}$, were earlier studied by Cryer and Chan (2008), who fitted a GARCH(1,1) model in order to capture the conditional variance structure of the data. See Figure 5. Changes in the conditional variance of the innovations may be signified by substantial fluctuations in past returns. This suggests the potential of using a more complex threshold variable, for example a function of finitely many past returns, than simply some lag of X for financial data. We therefore consider threshold variables of the form $W_{t-1} = \sum_{j=1}^{k} |X_{t-j} - X_{t-j-1}|$, i.e. the cumulative sum of the past k absolute changes in the daily returns. (The choice of $k = 3$ will be justified as being sufficient for the CREF example.)

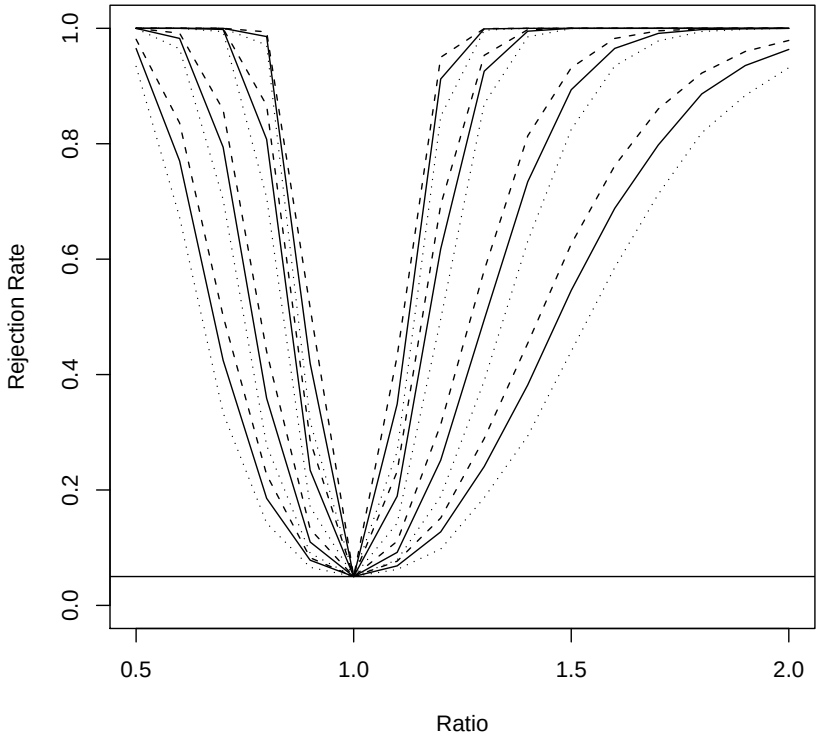


Figure 3: Empirical rejection rate of the LRT ; all three approaches for calculating the p-values are corrected to have exactly 5% empirical size by appropriately shifting the p-values. The threshold equals 0. Solid lines are the power curves for the LRT with the p-value computed by p_0; dashed lines are those for p_1 and dotted lines those for p_2. The four lines of each type, from bottom to top, correspond to sample size $n = 100, 200, 500$ and 1,000, respectively.

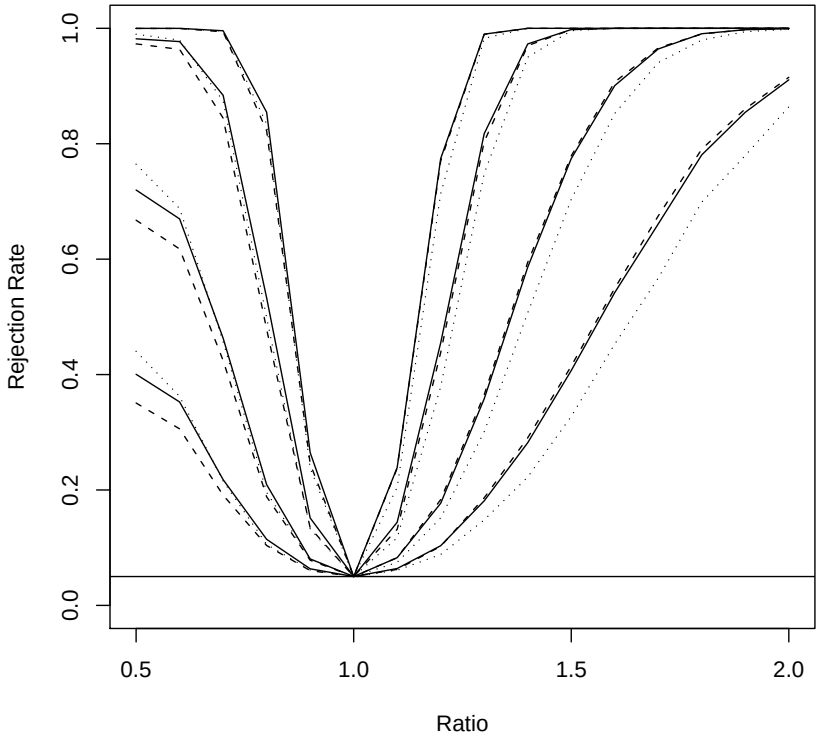


Figure 4: Empirical rejection rate of the LRT ; all three approaches for calculating the p-values are corrected to have exactly 5% empirical size by appropriately shifting the p-values. The threshold equals -0.8. Solid lines are the power curves for the LRT with the p-value computed by p_0; dashed lines are those for p_1 and dotted lines those for p_2. The four lines of each type, from bottom to top, correspond to sample size $n = 100, 200, 500$ and 1,000, respectively.

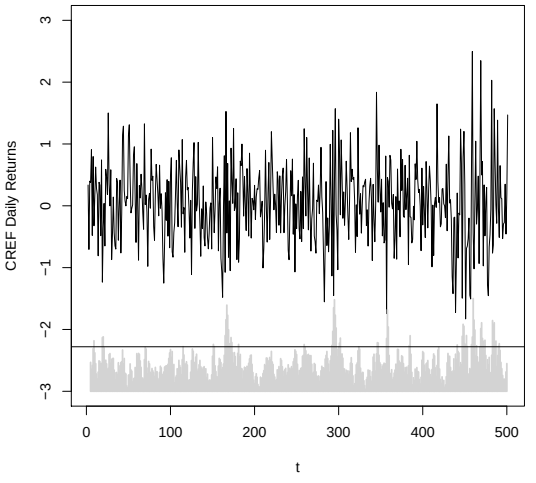


Figure 5: CREF daily returns: black solid line; the threshold variables are plotted as vertical gray lines on the bottom with the height of the lines proportional to the W's. The horizontal line indicates the estimated threshold.

Specifically, we consider the following T-CHARM:

$$X_t = \sum_{i=1}^{m} \sigma_i I\{r_{i-1} < W_{t-1} \le r_i\}\eta_t \tag{16}$$

The theories for quasi-likelihood estimation established in earlier sections can be readily adapted to this model. We fit a two-regime T-CHARM with the threshold searched between the 5 and 95 percentiles of the threshold variable W_t where $k = 3$, by quasi-likelihood estimation: $\hat{\sigma}_1^2 = 0.3765(0.0272)$, $\hat{\sigma}_2^2 = 0.7420(0.147)$, $\hat{r} = 3.333$, where the standard errors are enclosed in parentheses; see Figure 6. Based on the method detailed in Section 4 and using the empirical standardized residual distribution, we obtain (2.256, 4.024) as a 95% confidence interval of the threshold parameter, which is asymmetric about the threshold estimate. Below, model diagnostics suggests the normality of the residuals. Assuming normality, the 95% confidence interval of the threshold parameter becomes (2.321, 4.144), which is quite close to the preceding confidence interval. The presence of a threshold is justified by the LR test whose p-value equals $p_0 = 0.018$. The other

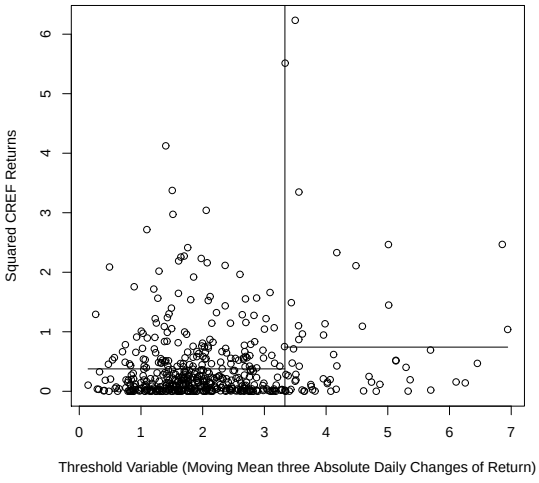


Figure 6: Scatter diagram of the squared CREF daily returns versus the threshold variable. Vertical line separates the the two regimes and the horizontal lines indicate the estimated variances of the two regimes.

two approaches for calculating the conditional p-values yield $p_1 = 0.025$ and $p_2 = 0.012$. The threshold is approximately the 88 percentile of the threshold variable with the number of data falling in the two regimes being 438 and 58. No further thresholds are needed based on LR tests for the presence of further thresholds in each of the two regimes. The choice of $k = 3$ is justified by treating k as a parameter and estimating it by profile quasi log-likelihood; upon fitting the model with k ranging from 1 to 5 and identical effective sample size yields the profile likelihood -25.54, -29.32, -25.00, -28.01 and -26.29, respectively, which is maximized at $k = 3$.

Model diagnostics (Figure 7) shows that the model provides a good fit to the data. In particular, the standardized residuals from the fitted T-CHARM appear not to show any conditional heteroscedasticity as judged by the sample ACF of their absolute values. This conclusion is corroborated by the McLeod-Li test (see Li and Li, 1996) as well as a (unreported) Lagrange multiplier test for residual ARCH effects; see Li (2004) and Ling and Tong (2011) for surveys on goodness-of-fit tests in time series modelling. The standardized residuals also

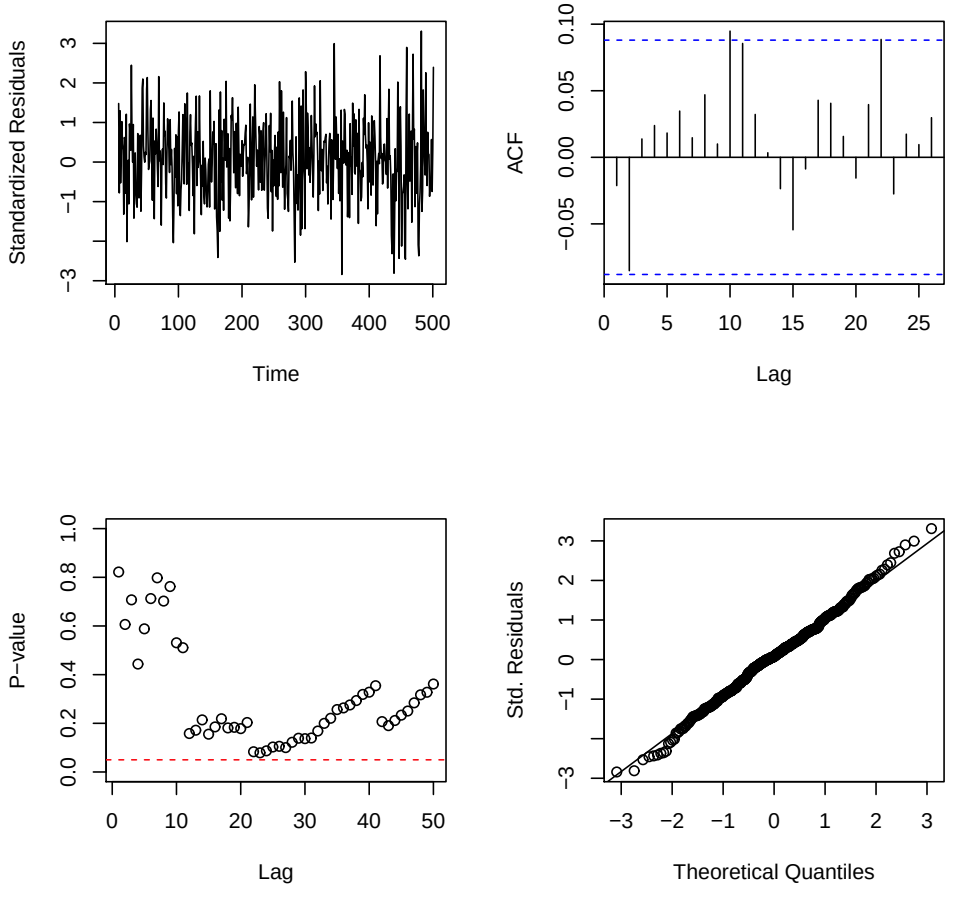


Figure 7: Upper left diagram shows the time plot of the standardized residuals from the fitted T-CHARM. Upper right diagram is the sample ACF of the standardized residuals. Lower left diagram plots the p-values of the McLeod-Li test for residual ARCH effects in the residuals, based on the first k lags of the autocorrelations of the squared standardized residuals, where $k = 1, \cdots, 50$; the dotted horizontal line shows the 5% level. Lower right diagram is the quantile-quantile normal score plot for the standardized residuals.

appear to be normally distributed, as its quantile-quantile normal score plot is quite straight.

Figure 8 shows the conditional variance processes from the fitted T-CHARM

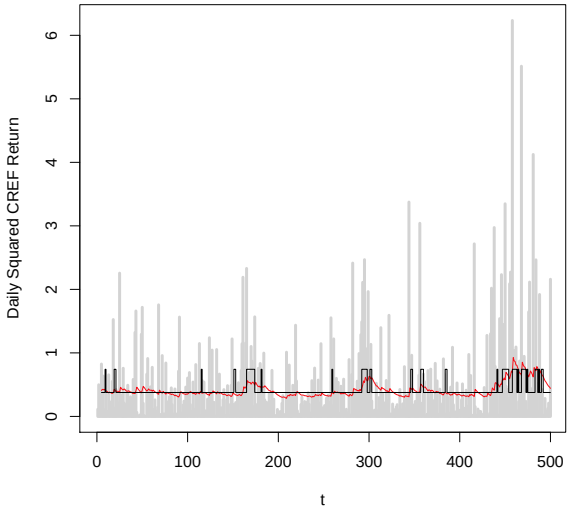


Figure 8: Fitted conditional variance process from T-CHARM (solid line) and that of the GARCH(1,1) model (red line), with the squared CREF returns plotted as gray background.

and the GARCH(1,1) model reported by Cryer and Chan (2008). They appear to complement each other on a global scale. On a finer scale, it seems that the T-CHARM captures some of the troughs during periods of high conditional variance whereas the GARCH(1,1) model tends to smooth them away. On the other hand, as three-parameter models, both models have log-likelihoods of similar order of magnitude, bearing in mind their approximate nature. Figure 9 shows the out-of-sample 1-step-ahead predictive performance of both models with fifty new observations collected from August 16 to October 24 in the year 2006. It plots the average cumulative predictive log-likelihood of the new observations; their performances seem reasonably comparable with the T-CHARM performing perhaps slightly better. Interestingly, all the fifty new observations fall into the first regime. If we extend the out-of-sample comparison further, the GARCH(1,1)

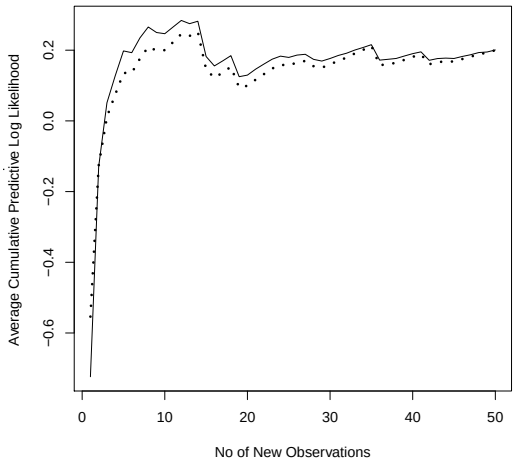


Figure 9: Average cumulative 1-step ahead predictive log-like likelihood, from the T-CHARM (solid line) and the GARCH model (dotted line).

model begins to outperform the T-CHARM, but then the dependence structure of the CREF series also experienced an unprecedented change as the market was heading into the financial crisis in 2007–2010 (Figure 10). Since financial markets are almost invariably nonstationary, it is wise to note the limitations of *stationary* models for financial time series. At best, stationary models such as T-CHARM and GARCH models merely serve to capture the approximate dynamics of the conditional variance over a relatively stationary (therefore limited) period. A really challenging research problem is to model the nonstationarity in the conditional variance based on past data; this is a daunting task since market collapses are often triggered by extraneous circumstances, e.g. the credit crisis from 2007 to 2009.

In the next two examples, which lie outside economics and finance, each time series has a non-trivial conditional mean structure that can be modelled by an ARIMA model with possibly additive outliers, but the errors are conditionally heteroscedastic white noise that we shall model by some T-CHARM. The parameters in the mean function are generally distinct from those parametrizing the conditional variance function. Consequently, the mean structure may be es-

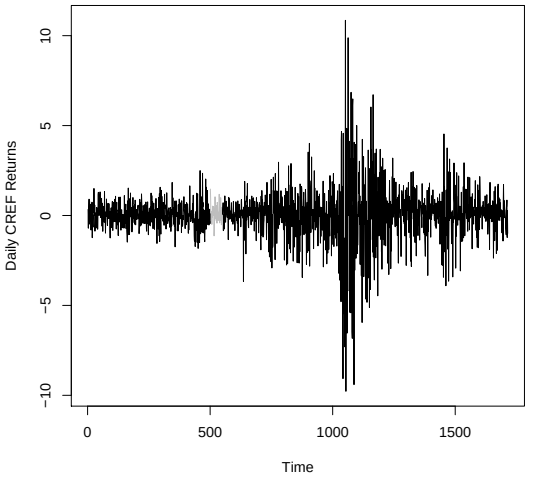


Figure 10: Time plot of an extended daily CREF returns. The initial black solid line shows the data initially analysed, the middle gray line corresponds to the period of out-of-sample forecast and final black solid line draws the more recent data, which shows the dramatic changes in the dependence structure of the CREF series as the market went through the 2007-2010 financial crisis.

timated first, by making use of the white-noise nature of the errors and ignoring the conditional variance structure. Then the parameters of the conditional variance function can be subsequently estimated through a quasi-likelihood with the ARIMA residuals treated as if they were the true innovations. It can be readily checked that, under some mild regularity conditions, the mean parameters and the variance parameters are asymptotically independent of each other, so that the their standard errors can be obtained separately from each of the two steps.

The second example is a long time series of annual tree ring width (Figure 11), with the measurements taken from a tree in a location at high altitude in

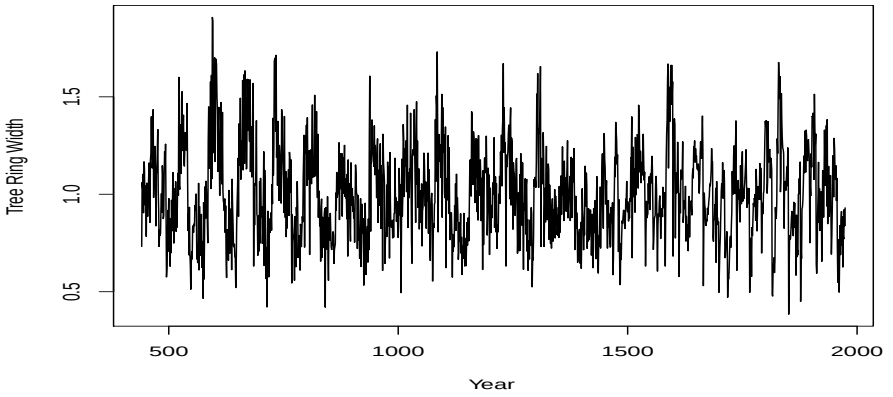


Figure 11: Time plot of the annual tree ring width.

Argentina. The time series spans over the period from year 441 to 1974 and it was contributed by J. Boninsegna to the NOAA Paleoclimatology database

`http://www.ncdc.noaa.gov/paleo/metadata/noaa-tree-2782.html`.

An IMA(1,1) model is initially identified and fitted to the data with the MA coefficient given by -0.6110 with standard error 0.0216. The residuals of the fitted IMA(1,1) model appear to be white noise in the sense that the residual ACF is only marginally significant at lag 11 and five higher lags out of the 100 lags examined. This observation is corroborated by the Box-Ljung test that is based on the first k lags of the residual ACF with k ranging from 3 to 100. On the other hand, it is unlikely that the residuals are independent as the absolute residuals appear to be correlated; see Figure 12. We fit a two-regime T-CHARM

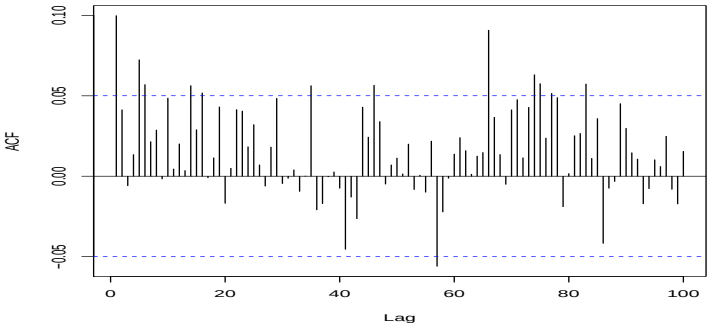


Figure 12: ACF of the absolute residuals from the IMA(1,1) model fitted to the tree ring data.

to the residuals to account for the conditional heteroscedasticity, with lag 1 of the IMA(1,1) error as the threshold variable. The following parameter estimates are obtained: $\hat{\sigma}_1^2 = 0.03205(0.00153)$, $\hat{\sigma}_2^2 = 0.05729(0.00893)$, $\hat{r} = 0.2366$ (95% confidence interval: (0.1619, 0.3049)), which is approximately the 91 percentile; see Figure 13.

The threshold structure is supported by the LR test for T-CHARM with p-value $p_0 = 0.005$. The other two methods of calculating the conditional p-values yield $p_1 = 0.008$ and $p_2 = 0.002$. The first regime contains 1402 observations while the second regime 131 observations. No further thresholds are needed by reference to the LR test for the presence of further thresholds in each of the two regimes. The fitted T-CHARM has successfully captured the conditional heteroscedasticity in the data as there are no residual ARCH effects in the standardized residuals from the fitted T-CHARM, by reference to the McLeod-Li test up to 100 lags. The fitted T-CHARM suggests that during fast-growing years, tree growth is much more variable, with a variance that almost doubles that during non-fast-growing years. What caused the observed variations is unclear, but it may be related to the fact that over non-fast-growing regime the tree ring width is bounded whereas this is not so in the fast-growing regime.

Next, we try to fit a GARCH model. The sample EACF of the absolute residuals of the IMA(1,1) model tentatively suggests GARCH(1,2) and GARCH(2,2) models, but only the ARCH(1) coefficient is found to be significant, at 5% level.

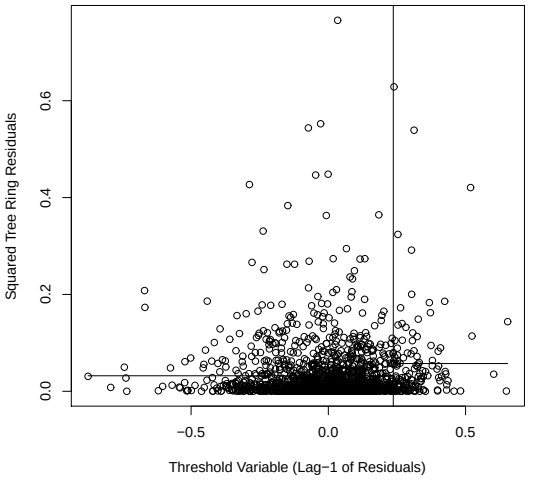


Figure 13: Scatter diagram of the squared tree ring residuals versus the lag 1 of the residuals, threshold variable. Vertical line separates the the two regimes and the horizontal lines indicate the estimated variances of the two regimes.

Eventually we choose the GARCH(1,1) model whose conditional variance equals $h_t = \beta h_{t-1} + \alpha_0 + \alpha_1 X_{t-1}^2$ where X_t stands for the IMA(1,1) errors, as this model passes the McLeod-Li test but the simpler ARCH(1) model does not. The GARCH estimates, with their standard errors enclosed in parentheses, are $\hat{\alpha}_0 = 0.0284(0.00656)$, $\hat{\alpha}_1 = 0.0987(0.0288)$ and $\hat{\beta} = 0.0747(0.198)$. While the T-CHARM has shed some light on the tree-growing process as we have seen, it is unclear to us as to how to interpret the fitted GARCH(1,1) model. Finally, both fitted models involve three parameters each with comparable quasi-likelihoods, again bearing in mind their approximate nature.

The third example concerns the time series of waiting times between the starts of two consecutive eruptions of the old faithful geyser. The data have been extensively studied in the literature. We use the version of the data collected from August 1–15, 1985; see Azzalini and Bowman (1990), Härdle (1991), the recent review by Zucchini and MacDonald (2009, Chapter 10) and the references therein. Besides the series of waiting time, the corresponding series of eruption duration is also available, but, for simplicity, we shall confine our analysis to

the waiting times. Figure 14 plots the scatter diagrams of the log waiting time

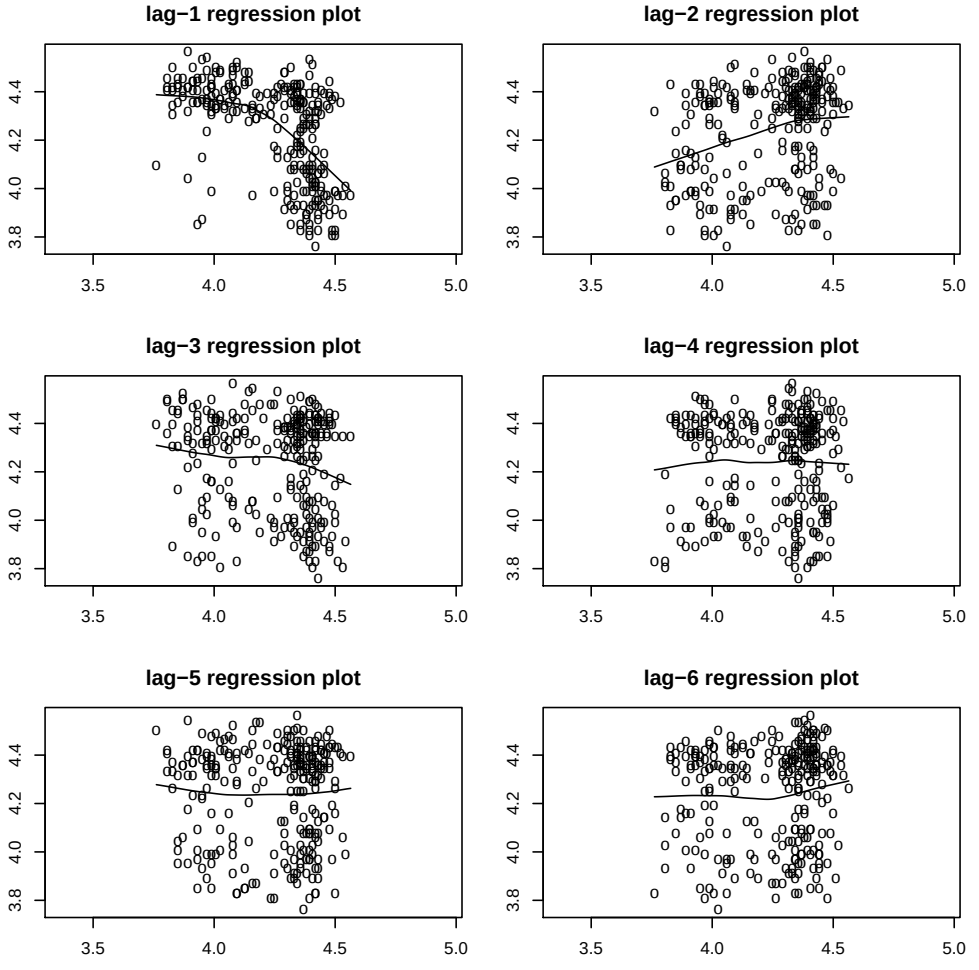


Figure 14: Lagged regression plot of the log waiting time against its lag $k, k = 1, ..., 6$. Open circles are data and solid lines are nonparametric curve fits.

against its lag k, for $k = 1, \cdots, 6$. The plot highlights three main features of the data: (i) the waiting time is strongly associated with lag 1 values but much less so with values at higher lags, (ii) the presence of conditional heteroscedasticity and (iii) a number of outliers. The last two features seem to be largely ignored in the literature, which generally focuses on nonparametric techniques and Markov chain analysis.

Preliminary statistical analysis suggests the possibility of an AR(1) model plus additive outliers for the mean structure, which we then fit to the data yielding the AR(1) coefficient estimate $-0.571(0.0497)$ and mean 4.248 (0.00650) and adjusted for five outliers at epochs 22, 37, 172, 237 and 266. The residuals of the preceding model appear to be white, based on the Box-Ljung test up to lag 50. However, the residuals are highly conditionally heteroscedastic: if the previous waiting time is under-predicted, then the subsequent waiting time is subject to much larger uncertainty; see Figure 15. We then fit a two-regime T-

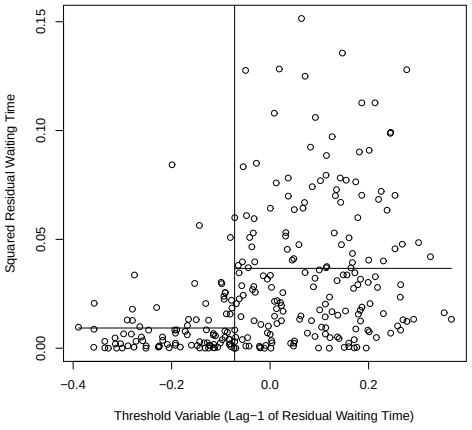


Figure 15: Squared AR(1) residual waiting time plotted against the lag-1 of the log waiting time; vertical line separates the two regimes and horizontal lines indicate the estimated variances of the two regimes.

CHARM with the lag 1 of the AR(1) errors as the threshold variable, giving the following parameter estimates: $\hat{\sigma}_1^2 = 0.0093(0.0011), \hat{\sigma}_2^2 = 0.037(0.0031)$, which is almost four times larger than the variance of the lower regime, and $\hat{r} = -0.072$ (95% confidence interval: $(-0.088, -0.058)$), which is about the 34 percentile. The existence of the threshold is supported by the LR test with p-value $< 10^{-5}$, for all three methods of calculation. And there is no need for more thresholds, based on the LR test. The standardized residuals from the T-CHARM are no longer conditionally heteroscedastic, based on the McLeod-Li test up to lag 50. However, the standardized residuals seems to be somewhat heavy tailed based

on the quantile-quantile normal score plot (not shown). Next, it seems that the GARCH formulation is unsuited for the AR(1) residuals. We tried to fit a GARCH(1,1) model to the AR(1) errors as suggested by the sample EACF of the absolute residuals. Unfortunately, the quasi-likelihood estimation has failed to converge, this remaining so even after simplifying the model to an ARCH(1) model.

7 Discussion

We have presented the theory and practice of a viable alternative to the ARCH-type models by going back to basics. It gives a simple and yet quite versatile alternative mixing function in the approach to conditional heteroscedasity based on an observable mixture of independent random variables. The model requires minimal conditions for statistical inference and often offers interpretable results. Computation is quite straightforward too.

The asymptotic results derived in Theorems 3.2 and 4.1 may provide a basis for calculating the approximate confidence intervals of the threshold parameters. Preliminary investigation of the two-regime case suggests that tabulation of the relevant quantiles, based on Monte Carlo, is precise if the variances of the two regimes are quite different, e.g. if the ratio of the smaller variance to the larger variance is less than 0.5. However, tabulation loses its precision as the variance-ratio approaches 1 because then the T-CHARM behaves more like a white noise process, making the threshold parameter nearly non-identifiable. An interesting research problem is to augment this approach with the asymptotic framework for the case of nearly equal variances (Yao, 1987), with the ultimate goal of providing confidence statements for the threshold parameters, see also Hansen (1997, 2000, 2011).

There are many different and interesting ways to generalize the above threshold approach. Tong (2011) has described fruitful experiences. There are increasing interests in modelling conditional variance clustering exhibited by random functions with a multi-dimensional index set, due to the many applications, e.g. image processing. Yan (2007) generalized the stochastic volatility approach in time series to Markov random fields by modelling the logarithmic variance process

as a conditional autoregressive (CAR) model (Besag *et al.* 1991). The T-CHARM may be generalized to modelling nonlinear spatial conditional heteroscedascity. We illustrate the basic idea with a simple example of modelling a random, noisy picture punctuated by spatial clustering. Consider a random field $\{X_s, s \in S\}$ consisting of integer-valued random variables and $S = \{(i,j), i,j = 1,2,\cdots,n\}$ is a finite lattice of pixels. We adopt a hierarchical approach with the X's specified as independent and marginally Poisson distributed with mean μ_s given the latent process $\{\mu_s, s \in S\}$; on the logarithmic scale, the latter is decomposed into the sum $\log(\mu_s) = m_s + \epsilon_s$, where m_s concerns the "mean" structure and ϵ_s the "conditional variance" structure. The process $\{m_s, s \in S\}$ may be modelled by a CAR model or some nonlinear generalization, so we set it to zero henceforth for simplicity. Spatial conditional heteroscedasticity may be captured by the following spatial T-CHARM:

$$\epsilon_s = \sigma(W_s)\eta_s, \tag{17}$$

where $\sigma(W_s) = \sum_{i=1}^{m} \sigma_i I(r_{i-1} < W_s \le r_i)$,

$$W_s = \frac{\sum_{s' \in N(s)} |X_{s'} - \exp(m_s)|}{\sum_{s' \in N(s)} 1}, \tag{18}$$

$N(s)$ denotes the set of neighbouring pixels of s, e.g. s' is a neighbour of s if its L^1-distance from s equals 1, η's are i.i.d. standard normal random variables and r's the threshold parameters. The functional W attempts to measure the local variation of the process, and the proposed T-CHARM the situation that high local variation triggers a jump in the variance. The existence and uniqueness of a joint probability distribution of $\mathbb{X} = (X_s, s \in S)^T$ satisfying the given conditional distributions can be established by considering an n^2-dimensional Markov chain $\mathbb{X}_t = (X_{s,t}, s \in S)^T, t = 0,1,\cdots$, whose transition probability kernel is the composition of the conditional distributions specified by the spatial T-CHARM; details will be reported elsewhere. Left side of Figure 16 shows a re-scaled realization of a two-regime spatial T-CHARM, where $r = 1.0, \sigma_1 = 0.4, \sigma_2 = 2\sigma_1$ and $n = 50$; the realization is re-scaled to have minimum 0 and maximum 1. The realization is obtained by iterating the aforementioned Markov chain 1,000 times and treating the last iterate as from the stationary distribution for the spatial T-CHARM. Spatial clustering is apparent in the diagram, especially when com-

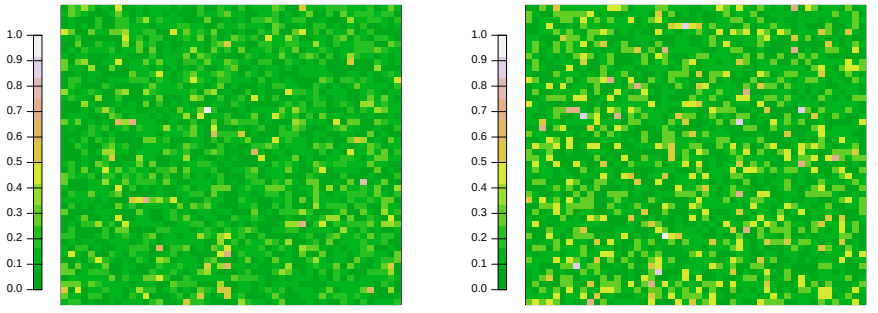


Figure 16: Left diagram is a standardized realization of a two-regime spatial T-CHARM with $\sigma_2 = 2\sigma_1$ and $\sigma_1 = 0.4$. The values are standardized to make them lie between 0 and 1, and plotted in terrain colours; see the legend. About 24% of the data fall in the regime corresponding to $\sigma_2 = 0.8$. Right diagram is a standardized realization for the case of $\sigma_1 = \sigma_2 = 0.4$.

pared with the right figure, where $\sigma_1 = \sigma_2$ so that that process is homoscedastic. We remark that it is possible to extend some of the theoretical results obtained earlier for model (6) to the spatial setting with discrete response variables, but a thorough study awaits future investigation.

Appendix: Proofs of Theorems

A.1. Proof of Theorem 2.1

Because P is irreducible, it admits a unique left eigenvector of 1, say u, such that u is a positive vector that sums to 1, and satisfies the condition

$$u^\tau P = u^\tau.$$

Hence, P can be decomposed as

$$P = uu^\tau + Q, \tag{19}$$

where for all $v \in R^m$, $Qv = Pv - u^\tau vu$ and hence $Qu = 0$ so that 0 is an eigenvalue of Q. Note that any right eigenvector w of eigenvalue less than 1

in magnitude is orthogonal to u. Consequently, Eqn. (19) implies that, for any positive integer k,

$$P^k = uu^\tau + Q^k. \tag{20}$$

Moreover, it is readily seen that the non-zero eigenvalues of Q are identical to those of P. Consequently,

$$c(Q) = Q^{m-1} - c_1 Q^{m-2} - \ldots - c_{m-1} I = 0.$$

Without loss of generality, assume that $\mathbb{E}(h(X_t)) = 0$, otherwise we shall subtract the nonzero mean from h and argue as below. Let $u^\tau = (u_1, \ldots, u_m)$ be the stationary probability vector of P, i.e. $u^\tau P = u^\tau$. Further, let $\nu = (\nu_1, \ldots, \nu_m)$ and $\eta^\tau = (\eta_1, \ldots, \eta_m)$, where, for $1 \le j \le m$, $\nu_j = \mathbb{E}(h(\sigma_j e))$ and $\eta_j = \mathbb{E}(h(X_0) I\{X_0 \in R_j\})$. It can be seen that the vector of stationary probabilities u is orthogonal to the vector ν, by the following arguments.

$$0 = \mathbb{E}(h(X_t)) = \mathbb{E}(\mathbb{E}(h(X_t)|X_{t-1})) = \sum_{j=1}^{m} u_j \mathbb{E}(h(\sigma_j e_t)) = u^\tau \nu.$$

It can be checked that the lag-k autocovariance of $\{Y_t\}$ equals

$$\gamma_k = \mathbb{E}(Y_0 Y_k) = \mathbb{E}(Y_0 \mathbb{E}(Y_k|X_0, X_{k-1})) = \eta^\tau P^{k-1} \nu = \eta^\tau Q^{k-1} \nu,$$

where the last equality follows from Eqn. (20) and the orthogonality of u and ν. The validity of the Yule-Walker equation defined by Eqn. (3) for the $\{Y_t\}$ process then follows from the equation $c(Q) = 0$. This completes the proof of Theorem 2.1. ■

A.2. Sketch of Proof on the Asymptotic Independence of the Maximum and its Location of an OU process over a Sufficiently Large Interval.

We adapt some arguments in Berman (1971) who derived the asymptotic distribution of the maximum of a stationary Gaussian process. It is readily seen from the proof of Theorem 3.1 in Berman (1971) that, under some regularity conditions including a certain asymptotic rate of the autocorrelation function with lag approaching 0 for which the OU process holds, with a fixed $h > 0$ and

T tending to infinity, the maximum of such a process over the interval $[0, T]$ is asymptotically equivalent to the maximum of the maxima of the process over $[ih, (i+1)h], i = 0, .., T-1$, with the maxima over the unit intervals being asymptotically independent. (With no loss of generality, T may be assumed to be a positive integer.) Hence, the maximum of the process over $[0, T]$ and the location of the maximum up to an uncertainty of h are asymptotically independent. By passing to the limit with $h \to 0$, the maximum of the process over $[0, T]$ can be shown to be asymptotically independent of the location of the maximum, with the latter uniformly distributed, for large T. The stationarity of the OU process then allows the result to hold for any sufficiently large, finite interval in lieu of $[0, T]$. ■

Acknowledgements

The research was partially supported by the US National Science Foundation (NSF-0934617) (to Chan), Hong Kong RGC Grants numbered HKUST601607 and HKUST602609 (to Ling), the Distinguished Visiting Professorship at the University of Hong Kong, the Saw Swee Hock Professorship and the Institute of Mathematical Science, the last two at the National University of Singapore (to Tong).

References

Azzalini, A. and Bowman, A.W. (1990). A look at some data on the Old Faithful geyser. *Applied Statistics* **39**, 357–365.

Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica* **66**, 47–78.

Balke, N.S. and Fomby, T.B. (1997). Threshold cointegration. *Int. Econ. Rev.* **38**, 627–645.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics (Disc: p.21–59). *Ann. Inst. Statist. Math.* **43**, 1–59.

Berman, S.M. (1971). Maxima and high level excursions of stationary Gaussian processes. *Transactions of the American mathematical Society* **160**, 65–85.

Brockwell, P. and Davis, R.A. (1991). *Time series: theory and methods.* (2nd) Springer-Verlag, New York.

Brockwell, P., Liu, J. and Tweedie, R.L. (1992). On the existence of stationary threshold autoregressive moving-average processes. *J. Time Ser. Anal.* **13**, 95–107.

Chan, K.S. (1990). Testing for threshold autoregression. *Ann. Statist.* **18**, 1886–1894.

Chan, K.S. (1993). Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model. *Ann. Statist.* **21**, 520–533.

Chan, K.S. (2009). *Exploration of a Nonlinear World.* World Scientific, Singapore.

Chan, K.S. and Li, W.K. (2007). Threshold models and new developments in time series. *Stat. Sinica* **17**, issue number 1, 1–287.

Cont, R. and Tankov, P. (2004). *Financial modelling with jump processes.* Chapman & Hall.

Cryer, J.D. and Chan, K.S. (2008). *Time series analysis: with applications in R.* Springer-Verlag, New York.

Dirkse, J.P. (1975). An absorption probability for the Ornstein-Uhlenbeck process. *J. Appl. Probability* **12**, 595–599.

Enders, W. and Granger, C.W.J. (1998). Unit root tests and asymmetric adjustment with an example using the term structure of interest rates. *J. Bus. & Econ. Stat.* **16**, 304–311.

Engle, R.F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of the United Kingdom inflation. *Econometrica* **50**, 987–1007.

Gooijer, J. G. (1998). On threshold moving-average models. *J. Time Ser. Anal.* **19**, 1–18.

Gourieroux, C. and Monfort, A. (1992). Qualitative threshold ARCH models. *J. Econometrics* **52**, 159–199.

Hansen, B.E. (1997). Inference in TAR models. *Stud. Nonlinear Dyn. Econom.* **2**, 1–14.

Hansen, B.E. (2000). Sample splitting and threshold estimation. *Econometrica* **68** 575–603.

Hansen, B.E. (2011). Threshold autoregression in economics. *Stat. & Its Interface.* **4**, 123–128.

Hansen, B.E. and Seo, B. (2002). Testing for two-regime threshold cointegration in vector error-correction models. *J. Econometrics* **110**, 292–318.

Härdle, W. (1991). *Smoothing Techniques with Implementation in S*. New York: Springer.

Li, C.W. and Li, W.K. (1996). On a double threshold autoregressive heteroscedastic time series model. *J. Appl. Econometrics* **11**, 253–74.

Li, D., Li, W.K. and Ling, S. (2011). On the least squares estimation of threshold autoregressive and moving-average models. *Stat. & Its Interface.* **4**, 183–196.

Li, D. and Ling, S. (2012). On the least squares estimation of multiple-regime threshold autoregressive models. *J. Econometrics* **167**, 240–253.

Li, D., Ling, S. and Li, W.K. (2012). Asymptotic theory on the least squares estimation of threshold moving-average models. *Econometric Theory* (to appear).

Li, D., Ling, S. and Tong, H. (2012). On moving-average models with feedback. *Bernoulli* **18**, 735–745.

Li, G.D. and Li, W. K. (2008). Testing for threshold moving average with conditional heteroscedasticity. *Statist. Sinica* **18**, 647–665.

Li, G. D. and Li, W. K. (2011). Testing for linear time series models against its threshold extension. *Biometrika* **98**, 243–250.

Li, W.K. (2004). *Diagnostic checks in time series.* London: Chapman & Hall/CRC.

Ling, S. (2007). Self-weighted and local quasi-maximum likelihood estimators for ARMA-GARCH/ IGARCH models. *J. Econometrics* **140**, 847–873.

Ling, S. and Tong, H. (2005). Testing for a linear MA model against threshold MA models. *Ann. Statist.* **33**, 2529–2552.

Ling, S. and Tong, H. (2011). Score based goodness-of-fit tests for time series. *Statist. Sinica* **21**, 1807–1829.

Moran, P.A.P. (1953). The statistical analysis of the Canadian lynx cycle. *Aust. J. Zoo.* **1**, 163–173.

Nummelin, E. (1984). *General irreducible Markov chains and non-negative operators.* Cambridge: Cambridge University Press.

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

Rabemananjara, R. and Zakoïan, J.M. (1993). Threshold ARCH models and asymmetries in volatility. *J. Appl. Econometrics* **8**, 31–49.

Seijo, E. and Sen, B. (2011). A continuous mapping theorem for the smallest argmax functional. *Electron. J. Stat.* **5**, 421–439.

Stenseth, N.C. (2009). The importance of TAR-modelling for understanding the structure of ecological dynamics: the hare-lynx population cycles as an example. *Exploration of a nonlinear world.* ed. by K.S. Chan, Singapore: World Scientific.

Tong, H. and Lim, K.S. (1980). Threshold autoregression, limit cycles and cyclical data (with discussions). *J. R. Statist. Soc.* B. **42**, 245–292.

Tsay, R. S. and Tiao, G. (1984). Consistent estimates of autoregressive parameters and extended sample autocorrelation function for stationary and nonstationary ARMA Models. *J. Amer. Statist. Assoc.* **79**, 84–96.

Tong, H. (1978). On a threshold model. *Pattern recognition and signal processing*. (ed. C.H.Chen). Amsterdam: Sijthoff and Noordhoff.

Tong, H. (1982). Discontinuous decision processes and threshold autoregressive time series modelling. *Biometrika* **69**, 274-276.

Tong, H. (1990). *Nonlinear time series. A dynamical system approach*. Oxford: Oxford University Press.

Tong, H. (2011). Threshold models in time series analysis – 30 years on. *Stat. & Its Interface.* **4**, 107–118.

Xia, Y., Tong, H., Li, W.K. and Zhu, L-X. (2002). An adaptive estimation of dimension reduction space. *J. R. Statist. Soc.* B. **64**, 363–410.

Yan, J. (2007). Spatial stochastic volatility for lattice data. *J. Agric. Biol. Env. Stat.* **12**, 25–40.

Yao, Y. C. (1987). Approximating the distribution of the maximum likelihood estimate of the change-point in a sequence of independent random variables. *Ann. Statist.* **15**, 1321–1328.

Zakoïan, J.M. (1994). Threshold heteroskedastic models. *J. Econom. Dynam. Control* **18**, 931–955.

Zucchini, W. and MacDonald, I. (2009) *Hidden Markov models for Time Series: an Introduction Using R*, London: Chapman & Hall/CRC.

Department of Statistics & Actuarial Science, University of Iowa, IA, 52242, USA.

E-mail: kung-sik-chan@uiowa.edu *and* li20072010@gmail.com

Department of Mathematics, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong.

E-mail: maling@ust.hk

Department of Statistics, London School of Economics and Political Science.

E-mail: howell.tong@gmail.com

Current Departmental Research Reports

Serial No.	Date	Research Report Title	Author(s)
484	Jul-11	Threshold modelling of martingale differences	Kung-Sik Chan, Dong Li, Shiqing Ling and Howell Tong
485	Jul-11	On mixture memory GARCH models	Muyi Li, Wai Keung Li and Guodong Li
486	Jul-11	On the quasi-likelihood estimation for random coefficient autoregressions	L. Truquet and J. Yao
487	Aug-11	Moment-based tests for random effects in panel data models	Jianhong Wu and Guodong Li
488	Sep-11	Least absolute deviation estimation for nonstationary vector autoregressive time series models with pure unit roots	Guodong Li, Jianhong Wu and Wai Keung Li
489	Sep-11	On the detection of the number of signals with possibly equal strengths in the high-dimensional case	Damien Passemier and Jian-Feng Yao
490	Sep-11	Estimation of the population spectral distribution from a large dimensional sample covariance matrix	W.M. Li, J.Q. Chen, Y.L. Qin, J.F. Yao and Z.D. Bai
491	Nov-11	On uniform correctness of bootstrap confidence intervals under M-estimation	Zhuqing Yu and Stephen M.S. Lee
492	Dec-11	Threshold Poisson autoregression	Chao Wang, Jian-Feng Yao and W.K. Li
493	Feb-12	A hybrid procedure for density estimation amid model uncertainties	Mehdi Soleymani and Stephen M.S. Lee
494	Feb-12	Discussion of 'An analysis of global warming in the Alpine region based on nonlinear nonstationary time series models' by Battaglia and Protopapa	Howell Tong
495	May-12	Sequential combination of weighted and nonparametric bagging for classification	Mehdi Soleymani and Stephen M.S. Lee
484 (Revised)	Jun-12	On conditionally heteroscedastic AR Models with thresholds	Kung-Sik Chan, Dong Li, Shiqing Ling and Howell Tong

❖❖❖ *The complete listing can be found at* ❖❖❖

http://www.saasweb.hku.hk/research/staff_research_report.php

Requests for off prints may be sent to saas@hku.hk by e-mail