Quasi-likelihood Estimation of a Threshold Diffusion Process

Fei Su[a], Kung-Sik Chan[b,*]

[a] *ISO Innovative Analytics, San Francisco, CA, United States*
[b] *Department of Statistics and Actuarial Science, The University of Iowa, Iowa City, IA 52242, United States*

Abstract

The threshold diffusion process, first introduced by Tong (1990), is a continuous-time process satisfying a stochastic differential equation with a piecewise linear drift term and a piecewise smooth diffusion term, e.g., a piecewise constant function or a piecewise power function. We consider the problem of estimating the (drift) parameters indexing the drift term of a threshold diffusion process with continuous-time observations. Maximum likelihood estimation of the drift parameters requires prior knowledge of the functional form of the diffusion term, which is, however, often unavailable. We propose a quasi-likelihood approach for estimating the drift parameters of a two-regime threshold diffusion process that does not require prior knowledge about the functional form of the diffusion term. We show that, under mild regularity conditions, the quasi-likelihood estimators of the drift parameters are consistent. Moreover, the estimator of the threshold parameter is super consistent and weakly converges to some non-Gaussian continuous distribution. Also, the estimators of the autoregressive parameters in the drift term are jointly asymptotically normal with distribution the same as that when the threshold parameter is known. The empirical properties of the quasi-likelihood estimator are studied by simulation. We apply the threshold model to estimate the term structure of a long time series of US interest rates. The proposed approach and asymptotic results can be readily lifted to the case of a multi-regime threshold diffusion process.

JEL Classification:
C22
E43

[*] Corresponding author. Tel.: +1 319 335 2849 ; fax: +1 319 335 3017.
Email addresses: sophysu@gmail.com (Fei Su), kung-sik-chan@uiowa.edu (Kung-Sik Chan)

Keywords: Girsanov's theorem, Interest rates, Nonlinear time series, Stochastic differential equation, Term structure

1. Introduction

In financial and insurance markets, diffusion processes have become the standard tool for modeling returns and values for risk management purposes. For example, a number of diffusion processes have been used to model the term structure of market yields such as interest rate (Vasicek, 1977; Cox et al., 1985; Black and Karasinski, 1991), some of which include time-dependent covariates in the mean function (Hull, 2010; Black et al., 1990). While the functional form of the diffusion term differs in these models, their drift terms stay affine (or can be transformed to linear functions). Despite their relative computational convenience, linear diffusion processes fail to capture nonlinear characteristics such as multimodality, asymmetric periodic behavior, time-irreversibility, and the occurrence of occasional extreme events that are commonly found in real data.

Continuous-time nonlinear models have proved increasingly useful over the past decade for capturing the aforementioned nonlinear properties (Tong, 1990; Decamps et al., 2006). Although continuous-time nonlinear diffusion processes form a relatively large model class, the field of empirical nonlinear time series modeling is relatively under-explored, except for the first-order continuous-time threshold autoregressive (CTAR) model proposed by Tong (1990); see Section 2 for the definition of the CTAR model, and some of its properties. The first order CTAR model will be simply referred to as the threshold diffusion (TD) process below. Several approaches on the inference of TD processes with discrete-time data have been proposed in the literature, e.g., Gaussian likelihood estimation (Tong and Yeung, 1991; Brockwell and Hyndman, 1992; Brockwell, 1994; Brockwell et al., 2007), moment-based estimators (Chan et al., 1992; Coakley et al., 2003), and Bayesian approach (Pai and Pedersen, 1999). If sufficiently fine data are available, the likelihood function can be approximated by the Girsanov's formula (at least for the case of known diffusion term). An advantage of Bayesian estimation is that even when the data are not sufficiently fine, Bayesian data augmentation techniques could be used; see (Elerian et al., 2001; Eraker, 2001, 2004; Roberts and Stramer, 2001; Stramer and Roberts, 2007).

Within the under-developed literature on the inference of the TD processes with continuous-time data, maximum likelihood is preferable for efficiency consideration. Recently, Kutoyants (2012) derived the asymptotic distribution for maximum-likelihood estimation of a TD model under restrictive conditions including bounded parameter space, known ordering among some parameters, and known functional form of the diffusion term. In practice, the functional form of the diffusion term is generally unknown. Thus, it is desirable to develop an estimation method that does not require knowing the functional form of the diffusion term.

Here, we introduce a quasi-likelihood approach to estimate the drift parameters of a TD model, without requiring prior knowledge of the functional form of the diffusion term. The quasi-likelihood is obtained by applying Girsanov's theorem to the TD model with constant diffusion coefficient even though the true diffusion term may be non-constant and even nonlinear. The consistency and the limiting distributions of the quasi-likelihood drift estimators of a 2-regime TD model are derived in Section 4, under some regularity conditions. Given data over T units of time, we show that the threshold parameter is T-consistent and its limiting distribution admits a closed-form pdf. Moreover, the autoregressive parameter estimators are $\sqrt{T}$-consistent, and asymptotically independent of the threshold estimator, with a limiting normal distribution which is the same as that assuming known threshold. A simulation study is conducted in Section 5 to illustrate the asymptotic results. In Section 6, we apply the proposed method to study the term structure of the US interest rate. We conclude briefly in Section 7. All proofs are collected in Section 8.

2. Nonlinear Diffusion Processes

We begin with the general nonlinear diffusion process:

$$dX(t) = \mu(X(t), t)dt + \sigma(X(t), t)dW(t) \tag{1}$$

where the function $\mu(x,t)$ is the drift term (instantaneous mean function), $\sigma(x,t)$ is the diffusion term ($\sigma^2(x,t)$ instantaneous variance function) and $W = \{W(t)\}$ stands for the standard Brownian process. Here, we focus on the case that both the drift and diffusion terms are time-homogeneous, i.e., $\mu(x,t) \equiv \mu(x)$ and $\sigma(x,t) \equiv \sigma(x)$. The drift and the diffusion terms are generally known up to some parameters, in which case we write $\mu_{\boldsymbol{\theta}}$ for μ and $\sigma_{\boldsymbol{\gamma}}$ for σ where the drift parameter $\boldsymbol{\theta}$ and the diffusion parameter $\boldsymbol{\gamma}$ are vectors that may share some common parameters. For conciseness, these parameters are often suppressed.

Among all nonlinear diffusion processes, the first-order m-regimes threshold diffusion (TD) model, which is the first-order case of the continuous-time threshold autoregressive process (Tong, 1990; Tong and Yeung, 1991), has received much attention in the literature, and it is defined to be the solution of the following stochastic differential equation

$$dX(t) \;=\; \sum_{1}^{m}\{\boldsymbol{\beta}_i^{\top}\begin{pmatrix}1\\X(t)\end{pmatrix}dt+\sigma_i dW(t)\}I(r_{i-1}<X(t)\le r_i) \qquad (2)$$

where $-\infty = r_0 < r_1 < ... < r_m = \infty$ are the threshold parameters, $\boldsymbol{\beta}_i^{\top} = (\beta_{i0}, \beta_{i1})$ are the autoregressive parameters and σ_i's are the diffusion parameters. In other words, the drift term is piecewise linear while the diffusion term is piecewise constant, and the two functions have identical break points. Specifically, $\mu(x) = \sum_{i=1}^{m}(\beta_{i0} + \beta_{i1}x)I(r_{i-1} < x \le r_i)$ and $\sigma(x) = \sum_{i=1}^{m}\sigma_i I(r_{i-1} < x \le r_i)$. Thus, the TD process models the situation that the underlying process is governed by m Ornstein-Uhlenbeck (OU) sub-processes, with the ith OU governing mechanism in effect whenever the process $X(t)$ is in the ith regime, i.e., $X(t) \in (r_{i-1}, r_i]$. The TD process may switch regimes infinitely many times within an arbitrary small interval of time due to the properties of the Brownian motion.

Similar to Chan and Tong (1986), the hard-thresholding regime switching mechanism may be smoothed by employing a soft-thresholding rule. A smooth threshold diffusion (STD) model can be obtained by replacing $I(r_{i-1} < x \le r_i)$ by $F(x; r_i, s_i) - F(x; r_{i-1}, s_{i-1})$ where $F(\cdot; r, s)$ denotes the cumulative distribution function of some location-scale family with location parameter r and scale parameter s, for instance the family of normal or logistic distributions. The proposed estimation method and much of the theory developed below can be lifted to the STD model, with details to be reported elsewhere.

For the stationary solution of a TD model to exist, the sub-models of the two outermost regimes must be "stationary". The following theorem is due to Brockwell and Hyndman (1992) (see also Brockwell et al. (1991)).

Theorem 1. *Suppose that $\sigma_i > 0, i = 1, ..., m$. Then the process defined by (2) has a stationary distribution if and only if*

$$\lim_{x\to-\infty}\mu(x) > 0; \lim_{x\to\infty}\mu(x) < 0,$$

i.e., $\beta_{1,1} < 0$ and $\beta_{m,1} < 0$, or in the case that $\beta_{1,1} = 0$ ($\beta_{m,1} = 0$), then $\beta_{1,0} > 0$ ($\beta_{m,0} < 0$). Further, if the stationarity condition is satisfied, the stationary density is given by

$$\pi(x) = \sum_{i=1}^{m} k_i \exp\{(\beta_{i1}x^2 + 2\beta_{i0}x)/\sigma_i^2\}I(r_{i-1} < x \le r_i),$$

where the constants $\{k_i\}$ are determined by the conditions that (i) $\int_{-\infty}^{\infty}\pi(x)dx = 1$ and (ii) $\sigma_i^2\pi(r_i-) = \sigma_{i+1}^2\pi(r_i+), i = 1, \cdots, m-1$, where $\pi(r_i-)$ and $\pi(r_i+)$

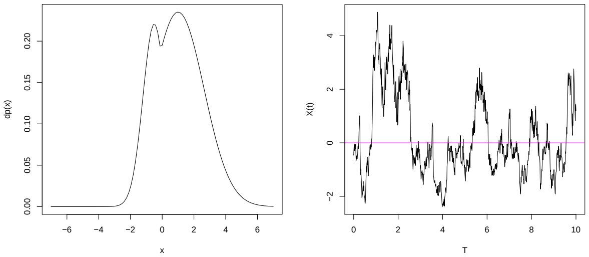

Figure 1: Left diagram: the stationary density function of $X(t)$, where $dX(t) = \{(-2 - 4X(t))I(X(t) \leq 0) + (3 - 3X(t))I(X(t) > 0)\}dt + 4dW(t)$. Right diagram: a realization of X simulated using the Euler scheme with step size equal to 0.01.

are the left and right hand limits of π *at* r_i*. That is, the function* $\sigma^2(x)\pi(x)$ *is continuous at all threshold points, and the stationary density function* $\pi(x)$ *is continuous only if the instantaneous variance function* $\sigma^2(x)$ *is continuous at the threshold points.*

Note that the stationary density is generally non-Gaussian, asymmetric and often multi-modal for a TD process. For instance, Figure 1.1 displays the stationary density function of the process $dX(t) = \{(-2 - 4X(t))I(X(t) \leq 0)+(3-3X(t))I(X(t) > 0)\}dt+4dW(t)$, which is non-Gaussian and bimodal. The form of the stationary density implies that it has finite moments of all orders. Also, a stationary TD model is geometrically ergodic (Stramer et al., 1996).

A more general TD model may be obtained by relaxing the piecewise constant diffusion term to a piecewise smooth function, for instance, a piecewise power diffusion term obtained by replacing σ_i by $\sigma_i X^{\gamma_i}(t)$ where γ_i are parameters. The preceding more general formulation enables us to model positive data without the need for data transformation. The stationarity results stated in Theorem 1 can be extended to the more general TD model. As an illustration, consider the stationarity condition for the square-root case when X is a positive process a.s., and $\sigma(x) = \sum_{i=1}^{m} \sigma_i \sqrt{x} I(r_{i-1} < x \leq r_i)$, where $0 = r_0 < r_1 < \ldots < r_m = \infty$. We shall assume that $\sigma_i > 0, \forall i$. Let $Y(t) = \sqrt{X(t)}$. Then the stationary condition for $\{X(t)\}$ and $\{Y(t)\}$ should be the same. By Ito's formula,

$$dY(t) = \sum_{i=1}^{m}\{(\frac{4\beta_{i0} - \sigma_i^2}{8Y(t)} + \frac{\beta_{i1}}{2}Y(t))dt + \frac{\sigma_i}{2}dW(t)\}I(\sqrt{r_{i-1}} < Y(t) \leq \sqrt{r_i}).$$

Thus, $\{X(t)\}$ is stationary if $4\beta_{10} - \sigma_1^2 > 0$ and $\beta_{m1} < 0$. Following an argument in Karlin and Taylor (1981, p. 221), the stationary density function can be shown to be

$$\pi(x) = \sum_{i=1}^{m} k_i x^{2\beta_{i0}/\sigma_i^2 - 1} \exp(2\beta_{i1} x/\sigma_i^2) I(r_{i-1} < x \le r_i) \tag{3}$$

where the constants k_i satisfy condition (i) of Theorem 1 and (ii') $\sigma_i^2 r_i \pi(r_i-) = \sigma_{i+1}^2 r_i \pi(r_i+), i = 1, \cdots, m-1$. Thus, the stationary density is piecewise "Gamma"-distributed. In summary, the TD model with a general diffusion term is a solution to the following stochastic differential equation.

$$dX(t) \;=\; \{\sum_{i=1}^{m} \boldsymbol{\beta}_i^\top \begin{pmatrix} 1 \\ X(t) \end{pmatrix} I(r_{i-1} < X(t) \le r_i)\}dt + \sigma(X(t))dW(t), \tag{4}$$

where $\sigma(\cdot)$ may be some piecewise smooth function.

3. Quasi-likelihood Estimation

As indicated in the preceding section, the functional form of the diffusion term of a TD model is generally unknown as it could be piecewise constant or piecewise power, or of some other piecewise smooth form. Thus, it is of interest to develop an estimation method for the drift parameters that does not requires prior knowledge on the functional form of the diffusion term. The quasi-log-likelihood function introduced below is designed to achieve this goal. As the idea is rather general and applicable to a general (possibly time-inhomogeneous) diffusion process, we motivate the quasi-likelihood in the general framework that the observations $\{X(t), 0 \le t \le T\}$ are a realization of a general diffusion process $X = \{X(t)\}$ satisfying (1) where the drift term $\mu_{\boldsymbol{\theta}}$ is parameterized by the unknown parameter $\boldsymbol{\theta}$. Suppose the diffusion term σ is known. Let dP be the probability measure induced by the standard Brownian process $W = \{W(t)\}$, and $dP_{\boldsymbol{\theta}}$ that induced by X. By the celebrated Girsanov's theorem for semimartingales, the log-likelihood function is

$$\begin{aligned} \log(\Lambda) &= \log(\frac{dP_{\boldsymbol{\theta}}}{dP}) \\ &= \int_0^T \frac{\mu_{\boldsymbol{\theta}}(X(t),t)}{\sigma(X(t),t)} dW(t) + \frac{1}{2}\int_0^T \frac{\mu_{\boldsymbol{\theta}}^2(X(t),t)}{\sigma^2(X(t),t)} dt \\ &= \int_0^T \frac{\mu_{\boldsymbol{\theta}}(X(t),t)}{\sigma^2(X(t),t)} dX(t) - \frac{1}{2}\int_0^T \frac{\mu_{\boldsymbol{\theta}}^2(X(t),t)}{\sigma^2(X(t),t)} dt. \end{aligned} \tag{5}$$

In the case of a constant diffusion term, $\log(\Lambda)$ is proportional to

$$l(\boldsymbol{\theta}) = \int_0^T \mu_{\boldsymbol{\theta}}(X(t), t) dX(t) - \frac{1}{2} \int_0^T \mu_{\boldsymbol{\theta}}^2(X(t), t) dt.$$

In the general case of a possibly non-constant diffusion term, $l(\boldsymbol{\theta})$ is no longer the log-likelihood function. However, it can be interpreted as a quasi-log-likelihood function, and the quasi-likelihood estimator $\hat{\boldsymbol{\theta}}$ is the argument maximizing $l(\cdot)$. If the drift term is a smooth function of $\boldsymbol{\theta}$ and assuming the validity of interchanging differentiation and integration and that all stochastic integrals below are well defined, the quasi-likelihood estimator satisfies the following estimating equation obtained by setting to 0 the first derivative of l w.r.t. $\boldsymbol{\theta}$:

$$0 \;=\; \int_0^T \frac{\partial \mu_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}(X(t), t)\{dX(t) - \mu_{\boldsymbol{\theta}}(X(t), t) dt\}. \tag{6}$$

Evaluated at $\boldsymbol{\theta}_0$, the true drift parameter, the right side of the preceding equation equals $\int_0^T \frac{\partial \mu_{\boldsymbol{\theta}_0}}{\partial \boldsymbol{\theta}}(X(t), t)\sigma(X(t), t) dW(t)$ where $\sigma(X(t), t)$ is the true diffusion term. Hence, it has zero mean, showing that (6) is an unbiased estimating equation. Another heuristic derivation of the quasi-log-likelihood mimics the approach of least squares. Consider

$$\begin{aligned} & \int_0^T [\{dX(t) - \mu_{\boldsymbol{\theta}}(X(t), t) dt\}^2 - \sigma^2(X(t), t) dt]/dt \\ = & \int_0^T -2\mu_{\boldsymbol{\theta}}(X(t), t) dX(t) + \mu_{\boldsymbol{\theta}}^2(X(t), t) dt. \end{aligned}$$

Hence, the quasi-log-likelihood is essentially equal to some negative "multiple" of the integrated square errors.

Here, we study the properties of the quasi-likelihood estimation of the drift term of a TD process. However, the drift term of a TD model is generally a discontinuous function, which complicates the theoretical analysis of the quasi-likelihood estimator. For simplicity, we assume a two-regime TD process and derive the large-sample properties of the quasi-likelihood estimator of the drift parameters. However, the limiting results can be readily lifted to the case of more than two regimes. The drift term of a two-regime TD model can be written as

$$\begin{aligned} \mu(X(t), t) &= \boldsymbol{\beta}_1^\top \begin{pmatrix} I(X(t) \le r) \\ X(t) I(X(t) \le r) \end{pmatrix} + \boldsymbol{\beta}_2^\top \begin{pmatrix} I(X(t) > r) \\ X(t) I(X(t) > r) \end{pmatrix} \\ &= \boldsymbol{\beta}_1^\top Y(t) I(X(t) \le r) + \boldsymbol{\beta}_2^\top Y(t) I(X(t) > r), \end{aligned}$$

where $Y(t) = (1, X(t))^\top$ and $\boldsymbol{\beta}_i^\top = (\beta_{i0}, \beta_{i1})$, $i = 1, 2$.
Then, quasi-likelihood estimation may be carried out by a two-step procedure: First, for given r, obtain the likelihood estimator for the vector of coefficients β_{ij} and denote this estimator by $\hat{\boldsymbol{\delta}}_r$, which is given by (letting $I_1(t;r) = I(X(t) \leq r), I_2(t;r) = I(X(t) > r)$)

$$\begin{aligned}\hat{\boldsymbol{\delta}}_r &= \begin{pmatrix} \int_0^T \frac{I_1(t;r)}{T} dt & \int_0^T \frac{X(t)I_1(t;r)}{T} dt & 0 & 0 \\ \int_0^T \frac{X(t)I_1(t;r)}{T} dt & \int_0^T \frac{X^2(t)I_1(t;r)}{T} dt & 0 & 0 \\ 0 & 0 & \int_0^T \frac{I_2(t;r)}{T} dt & \int_0^T \frac{X(t)I_2(t;r)}{T} dt \\ 0 & 0 & \int_0^T \frac{X(t)I_2(t;r)}{T} dt & \int_0^T \frac{X^2(t)I_2(t;r)}{T} dt \end{pmatrix}^{-1} \\ &\times \begin{pmatrix} \int_0^T \frac{I_1(t;r)}{T} dX(t) \\ \int_0^T \frac{X(t)I_1(t;r)}{T} dX(t) \\ \int_0^T \frac{I_2(t;r)}{T} dX(t) \\ \int_0^T \frac{X(t)I_2(t;r)}{T} dX(t) \end{pmatrix}.\end{aligned}$$

Substituting $\boldsymbol{\delta}$ by $\hat{\boldsymbol{\delta}}_r$ in the quasi-log-likelihood function yields the profile quasi-log-likelihood function of r

$$l(r) = l(\hat{\boldsymbol{\delta}}_r, r).$$

Then, perform a maximization of $l(r), r \in [a, b]$ to obtain the threshold estimator $\hat{r}$, where a, b are often chosen to be some percentiles, e.g. from the twenty to the eighty percentiles, of the observed data in order to guarantee data abundance for estimation in each regime.

In practice, the data are often digitized or measured by discrete-time sampling from the underlying continuous-time process, in the form of $\{X(t_i), 0 \leq i \leq m\}$, in which case the stochastic integrals can be approximated by Euler approximation. In particular, the profile function $l(r)$ becomes a piecewise constant function under the Euler scheme, and the threshold parameter can be computed via an exhaustive search. The Euler approximation scheme will be adopted in all numerical studies reported below.

It is instructive to compare the proposed estimation method with two closely related methods. Brockwell et al. (2007) derived the maximum likelihood estimator for the drift coefficients of the CTAR(p) model with constant diffusion term, which coincides with the proposed quasi-likelihood estimator, for $p = 1$. The estimation method of Tong and Yeung (1991) assumes the continuous-time sample path connecting two consecutive data to entirely lie in a single regime if both data fall in the same regime, but otherwise cross the threshold only once with the time of crossing determined by linear interpolation. Under this assumption and with the data augmented by pseudo data at the interpolated times of threshold crossing, the likelihood can then be computed by Kalman filter, the maximization of which yields the Tong-Yeung estimator. Applying first-order Taylor approximation in the Kalman filter and assuming known threshold and conditional homoscedascity, the Tong-Yeung estimator becomes the proposed estimator, except that it requires constructing pseudo data that vary with the threshold parameter. The latter requirement renders the optimization for the Tong-Yeung estimator less tractable in the general case.

The quasi-likelihood estimator of the drift parameter of a 2-regime TD process generated by (4), based on the observations $\{X(t), 0 \leq t \leq T\}$, will be denoted by $\hat{\boldsymbol{\theta}}$. The true parameter is denoted by

$$\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_{1,0}^\top, \boldsymbol{\beta}_{2,0}^\top, r_0)^\top = (\beta_{10,0}, \beta_{11,0}, \beta_{20,0}, \beta_{21,0}, r_0)^\top,$$

with the parameter space being $\Omega = \mathbb{R}^4 \times [a, b]$.

4. Large-sample Properties

The following assumptions will be used to establish the asymptotic properties of $\hat{\boldsymbol{\theta}}$, as $T \to \infty$.

(A1) $\boldsymbol{\beta}_{1,0} \neq \boldsymbol{\beta}_{2,0}$ and that $(\boldsymbol{\beta}_{1,0}^\top - \boldsymbol{\beta}_{2,0}^\top)(1, r_0)^\top \neq 0$.

(A2) The process $\{X(t)\}$ is stationary and geometrically ergodic with finite fourth moments.

(A3) The true threshold parameter r_0 lies in a pre-specified, finite interval $[a, b]$, say, $a = 20$ percentile and $b = 80$ percentile of the observed data. The process $\{X(t)\}$ admits a marginal probability density function that is a.s. continuous with possible discontinuity of first kind at the true threshold r_0. More specifically, the density function would be discontinuous only if the instantaneous variance function $\sigma^2(x)$ is discontinuous at $x = r_0$.

(A4) The variance function $\sigma^2(x)$ is a positive function, with finitely many discontinuity points (left and right limits exist for all x); Further, $\sigma(x)$ has at most linear growth, i.e., $\exists$ constants c_1, c_2 such that $|\sigma(x)| \leq c_1 + c_2|x|, x \in R$.

Though parameters in the two regimes are different, the mean function might still be continuous. (A1) excludes this possibility of continuity at the threshold point. And the assurance of discontinuity would be used in showing the super-consistency of the threshold parameter. (A2)–(A4) are useful in proving the consistency and the weak convergence of the estimators. These assumptions are essential for model identifiability. That r_0 is known to be in a fixed, finite interval $[a, b]$ guarantees adequate data for estimation in each regime, for large samples. Given the above assumptions, we are able to show the large-sample properties of the estimators. We first show that the estimators lie in a compact subset of the parameter space. The following lemma provides some technical results needed below.

Lemma 1. *Suppose (A2) is valid. Then the following averages satisfy the uniform law of large numbers*

$$\frac{1}{T}\int_0^T X^k(t)I(X(t)\le r)dt, k=0,1,2.$$

$$\frac{1}{T}\int_0^T X^k(t)I(X(t)\le r)\sigma(X(t))dW(t), k=0,1.$$

i.e. they tend to their expectations uniformly in $r\in[a,b]$, with probability approaching 1 as $T\to\infty$.

Lemma 2. *Suppose (A1)–(A4) are valid. Then it holds with probability approaching 1 as $T\to\infty$ that the quasi-likelihood estimator $\hat{\boldsymbol{\theta}}_T$ lies in a compact set, that is, there exists a finite constant $M>0$, such that $\hat{\boldsymbol{\theta}}_T$ lies in C_1 with probability approaching 1 as $T\to\infty$, where*

$$C_1=\{\boldsymbol{\theta}:|\boldsymbol{\beta}_1-\boldsymbol{\beta}_{1,0}|\le M,|\boldsymbol{\beta}_2-\boldsymbol{\beta}_{2,0}|\le M,r\in[a,b]\}.$$

Then, we show the estimators are consistent.

Theorem 2. *Assume (A1)–(A4) hold. Then the quasi-likelihood estimator $\hat{\boldsymbol{\theta}}_T=(\hat{\boldsymbol{\beta}}_1,\hat{\boldsymbol{\beta}}_2,\hat{r})$ is consistent, i.e. $\hat{\boldsymbol{\theta}}_T\to\boldsymbol{\theta}_0$ in probability.*

In particular, the estimator of the threshold parameter is T-consistent.

Theorem 3. *Assume (A1)–(A4) hold. Then the quasi-likelihood estimator of the threshold parameter is T-consistent:*

$$\hat{r}=r_0+O_p(1/T).$$

Before discussing the limiting distribution of the threshold parameter, define

$$\tilde{l}_T(\kappa)=l(\hat{\boldsymbol{\delta}}_{r_0+\kappa/T},r_0+\kappa/T)-l(\hat{\boldsymbol{\delta}}_{r_0},r_0).$$

Theorem 4. *Suppose (A1)–(A4) hold. $(\tilde{l}_T(\kappa)I(\kappa\ge 0),\tilde{l}_T(-\kappa)I(\kappa\le 0))$ converges weakly to $(\tilde{l}_1(\kappa),\tilde{l}_2(\kappa))$ in $D[0,\infty)\times D[0,\infty)$, equipped with the product topology of uniform convergence over compact sets and where $\{W(t),-\infty<t<\infty\}$ below denotes the Brownian motion with $W(0)=0$ a.s., and*
$\tilde{l}_1(\kappa)=-\frac{1}{2}f^2(r_0)\pi(r_0+)\kappa+f(r_0)\sqrt{\pi(r_0+)}\sigma(r_0+)W(\kappa)$,
$\tilde{l}_2(\kappa)=-\frac{1}{2}f^2(r_0)\pi(r_0-)\kappa+f(r_0)\sqrt{\pi(r_0-)}\sigma(r_0-)W(-\kappa)$,
where $f(r_0)=(\boldsymbol{\beta}_{1,0}-\boldsymbol{\beta}_{2,0})^\top(1,r_0)^\top$. Furthermore, $\tilde{r}_T=T(\hat{r}-r_0)$ converges

weakly to $\tilde{r}$, the unique maximizer of $\tilde{l}(\cdot)$, with the following probability density function (with $\tilde{r} \in \mathbb{R}$):

$$g(\tilde{r}) = I(\tilde{r} < 0)\frac{m}{2\sigma^2(r_0-)}[\frac{1}{\sqrt{-m\tilde{r}}}\phi(-\frac{1}{2\sigma^2(r_0-)}\sqrt{-m\tilde{r}}) - \frac{1}{2\sigma^2(r_0-)}\Phi(-\frac{1}{2\sigma^2(r_0-)}\sqrt{-m\tilde{r}})]$$
$$+I(\tilde{r} > 0)\frac{m}{2\sigma^2(r_0+)}[\frac{1}{\sqrt{m\tilde{r}}}\phi(-\frac{1}{2\sigma^2(r_0+)}\sqrt{m\tilde{r}}) - \frac{1}{2\sigma^2(r_0+)}\Phi(-\frac{1}{2\sigma^2(r_0+)}\sqrt{m\tilde{r}})],$$

where $m = \sigma^2(r_0+)f^2(r_0)\pi(r_0+) = \sigma^2(r_0-)f^2(r_0)\pi(r_0-)$, ϕ and Φ are the probability density and distribution function of a standard normal random variable, respectively.

Using Theorem 4, we can construct confidence intervals for the threshold parameter as follows. Let $m_- = m/\{2\sigma^2(r_0-)\}^2$ and $m_+ = m/\{2\sigma^2(r_0+)\}^2$. Consider the asymmetrically transformed variable $\breve{r} = m_-\tilde{r}I(\tilde{r} \leq 0) + m_+\tilde{r}I(\tilde{r} > 0)$ whose probability density function is $g(\breve{r}) = |\breve{r}|^{-1/2}\phi(-\sqrt{|\breve{r}|}) - \Phi(-\sqrt{|\breve{r}|}), \breve{r} \in \mathbb{R}$. (It is interesting to note that Hansen (1997) showed that up to scale, the preceding density function is also the limiting density of the threshold parameter estimator for a discrete-time self-exciting threshold autoregressive (SETAR) model when the autoregressive coefficients in the two regimes are asymptotically equal.) Table 1 displays several selected quantiles of $\breve{r}$. Hence, a 95% confidence interval for the threshold parameter is asymptotically equal to

$$(\hat{r} - 2.1458/(m_+T), \hat{r} + 2.1458/(m_-T)). \tag{7}$$

In practice, m_- and m_+ have to be approximated by substituting the unknown true parameter values by their estimates, which requires the specification of the diffusion term. However, the form of the limiting distribution may lend to the use of other approaches, e.g. bootstrap and subsampling, for constructing the confidence intervals.

A promising bootstrap approach is the regenerative block bootstrap (Bertail et al., 2006), which is applicable to Markov processes that can be decomposed into independent and identically distributed (IID) blocks. For instance, for a regular (irreducible) and stationary Markov process that admits an atom, it can be decomposed into IID blocks between consecutive visits to the atom. In practice, the aforementioned decomposition with data sampled over a finite interval, say, $[0, T]$, is generally marred by an incomplete initial block before the first visit to the atom and an incomplete closing block after the last visit, unless the atom is visited at $t = 0$ or T. A regenerative block bootstrap process, say $\{X^*(s), 0 \leq s \leq T\}$, can then be obtained by (i) concatenating the initial block with a number of randomly selected complete blocks with replacement, plus the closing block so that T^*, the size of the concatenated process, just exceeds or equals T, the observed sample size, (ii) subsampling the concatenated process over the interval $[t^*, t^*+T]$ where t^* is uniformly distributed over $[0, T^*-T]$, and (iii) shifting the time interval to $[0,T]$. An advantage of this approach is that it ensures that the observed process can be a realized bootstrap process. In the general case of a diffusion without an atom, a sufficiently small closed interval inside the interior of the state space, say a closed interval around the median of the process, may be utilized as an *approximate atom* for constructing an approximate regenerative block bootstrap. This bootstrap approach will be illustrated in the real application. Further investigations of the regenerative block bootstrap and other re-sampling approaches will be pursued elsewhere.

prob.	0.6000	0.7000	0.8000	0.9000	0.9500	0.9750	0.9950
quantile	0.0187	0.0923	0.2755	0.7558	1.3931	2.1454	4.1954

Table 1: Selected quantiles for $\breve{r}$.

Theorems 3 and 4 are supported by the following three lemmas.

Lemma 3. *Suppose (A1)–(A4) hold. The processes* $\{I(r_0-\Delta < X(t) \le r_0+\Delta)\}$ *and* $\{f(X(t))I(r_0-\Delta < X(t) \le r_0+\Delta)\}$ *are* $\rho-$*mixing for any function* $f(\cdot)$ *that is bounded over compact sets.*

The following two lemmas show that $\tilde{l}_T$ can be replaced by $l(\boldsymbol{\delta}_0, r_0+\kappa/T) - l(\boldsymbol{\delta}_0, r_0)$ for large samples.

Lemma 4. *Assume (A1)–(A4) hold. Then, for any positive number* K,

$$\sup_{|r-r_0|<K/T} |\hat{\boldsymbol{\beta}}_{i,r} - \hat{\boldsymbol{\beta}}_{i,r_0}| = o_p(1/\sqrt{T}), i = 1, 2.$$

Lemma 5. *Assume (A1)–(A4) hold. Then, for any fixed positive number* K,

$$\sup_{|\kappa|\le K} |\tilde{l}(\kappa) - (l(\boldsymbol{\delta}_0, r_0+\kappa/T) - l(\boldsymbol{\delta}_0, r_0))| = o_p(1).$$

Note that the $O_p(1/T)$ convergence rate of the threshold parameter implies that the threshold estimator is asymptotically independent of $\hat{\boldsymbol{\delta}}$. We can show that $\hat{\boldsymbol{\delta}}$ is $\sqrt{T}$-consistent and its limiting distribution is identical to the case of knowing the true threshold.

Theorem 5. *Suppose (A1)–(A4) hold. Then* $\hat{\boldsymbol{\delta}}_{\hat{r}} - \boldsymbol{\delta}_0 = O_p(1/\sqrt{T})$*. Moreover,* $\sqrt{T}(\hat{\boldsymbol{\delta}}_{\hat{r}} - \boldsymbol{\delta}_0)$ *is asymptotically normally distributed with the same distribution as for the case of known threshold, i.e.* $N(0, \Sigma)$ *where*

$$\Sigma = E^{-1}(\ddot{l}_{\boldsymbol{\theta}_0}) E(\dot{l}_{\boldsymbol{\theta}_0} \dot{l}_{\boldsymbol{\theta}_0}^\top) E^{-1}(\ddot{l}_{\boldsymbol{\theta}_0})^\top.$$

where $\dot{l}_{\theta_0} = \frac{\partial l(\boldsymbol{\delta},r)}{\partial \boldsymbol{\delta}}|_{\theta=\theta_0}$ *and* $\ddot{l}_{\theta_0} = \frac{\partial l^2(\boldsymbol{\delta},r)}{\partial \boldsymbol{\delta} \partial \boldsymbol{\delta}^\top}|_{\theta=\theta_0}$.

Note that

$$\ddot{l}_{\theta_0} \quad = \quad \begin{pmatrix} \int_0^T \frac{I_1(t;r)}{T} dt & \int_0^T \frac{X(t) I_1(t;r)}{T} dt & 0 & 0 \\ \int_0^T \frac{X(t) I_1(t;r)}{T} dt & \int_0^T \frac{X^2(t) I_1(t;r)}{T} dt & 0 & 0 \\ 0 & 0 & \int_0^T \frac{I_2(t;r)}{T} dt & \int_0^T \frac{X(t) I_2(t;r)}{T} dt \\ 0 & 0 & \int_0^T \frac{X(t) I_2(t;r)}{T} dt & \int_0^T \frac{X^2(t) I_2(t;r)}{T} dt \end{pmatrix}$$

and $E(\dot{l}_{\boldsymbol{\theta}_0} \dot{l}_{\boldsymbol{\theta}_0}^\top)$ equals the expectation of

$$\begin{pmatrix} \int_0^T \frac{I_1(t;r)\sigma^2(X(t))}{T} dt & \int_0^T \frac{X(t) I_1(t;r)\sigma^2(X(t))}{T} dt & 0 & 0 \\ \int_0^T \frac{X(t) I_1(t;r)\sigma^2(X(t))}{T} dt & \int_0^T \frac{X^2(t) I_1(t;r)\sigma^2(X(t))}{T} dt & 0 & 0 \\ 0 & 0 & \int_0^T \frac{I_2(t;r)\sigma^2(X(t))}{T} dt & \int_0^T \frac{X(t) I_2(t;r)\sigma^2(X(t))}{T} dt \\ 0 & 0 & \int_0^T \frac{X(t) I_2(t;r)\sigma^2(X(t))}{T} dt & \int_0^T \frac{X^2(t) I_2(t;r)\sigma^2(X(t))}{T} dt \end{pmatrix}.$$

Hence, the calculation of Σ requires knowing the diffusion term. For positive data, the diffusion term may take the form of a piecewise power function, i.e., $\sigma(x) = \sigma_1 x^\gamma I(x \le r) + \sigma_2 x^\gamma I(x > r)$. The power γ may be specified from data analysis, see the real application below. In the case of known γ, the parameters σ_i can be estimated by quadratic variation calculation. (For any continuous-time process $\{Y(t)\}$, its quadratic variation process $[Y]_t$ equals the limit $\lim \sum_{i=1}^m \{Y(t_i) - Y(t_{i-1})\}^2$ taken as the partition $0 = t_0 < t_1 < \ldots < t_m = t$ gets finer and finer, i.e., $m \to \infty$ and $\max_{i=1}^m |t_i - t_{i-1}| \to 0$.) Let $X_1(t) = X(t) I(X(t) \le r)$ and $X_2(t) = X(t) I(X(t) > r)$. Then the quadratic variations of X_i are given by

$$[X_i]_T = \int_0^T \sigma_i^2 X_i^{2\gamma}(t) dt,$$

so σ_i^2 can be estimated by

$$\hat{\sigma}_i^2 = [X_i]_T / \int_0^T X_i^{2\gamma}(t) dt. \tag{8}$$

5. Simulation

We have conducted a simulation study to illustrate the asymptotic behavior of quasi-likelihood estimation for the continuous-time threshold diffusion processes. Data were simulated from the following two models, both of which have the same mean function, but one has constant volatility function and the other has piecewise constant volatility function.

$$dX(t) = \{(-2-4X(t))I(X(t) \le 0) + (3-3X(t))I(X(t) > 0)\}dt + 4dW(t),$$

$$dY(t) = \{(-2-4Y(t))dt+4dW(t)\}I(Y(t) \le 0)+\{(3-3Y(t))dt+8dW(t)\}I(Y(t) > 0).$$

The parameter $\boldsymbol{\theta}_0^\top = (\boldsymbol{\delta}_0^\top, r_0) = (\beta_{10,0} = -2, \beta_{11,0} = -4, \beta_{20,0} = 3, \beta_{21,0} = -3, r_0 = 0)$ for both models. The diffusion processes were generated by the Euler scheme with $\Delta t = 1/100$. (The integrals in the closed-form solutions for the $\hat{\boldsymbol{\beta}}$'s are then approximated by sums.) The estimators of the $\boldsymbol{\beta}$'s and parameter r are obtained by maximizing the quasi-likelihood function, with the threshold parameter searched over the interval $[a, b]$ where a and b are chosen to be the 20 and 80 percentiles of each realization. We choose $T = 200, 500$ and 1000 respectively, and for each T, the Monte Carlo results reported below are based on 500 replications.

Table 2 lists, for each experimental setting, the sample mean, bias, and standard deviation of each drift estimate, and the corresponding empirical coverage rates of the nominally 95% confidence intervals. The confidence intervals are constructed based on Theorems 4 and 5, assuming that the functional form of the diffusion term is known, with σ_i^2 estimated by (8) and other parameters estimated by the quasi-likelihood estimates. The standard deviations and the biases of the estimators generally become smaller with larger T, confirming the derived consistency results. The normal quantile-quantile plots, in Fig. 2, for the autoregressive parameters estimated with the simulated X and $T = 500$ confirms the asymptotic normality result in Theorem 5. The plots for other cases of the X process are similar, but those for the Y process (unreported) approach straightness more slowly. A plot of the limiting density function of $\tilde{r} = T(\hat{r} - r_0)$ and its empirical counterpart for $\{X(t)\}$, with $T = 500, 1000$, are displayed in Figure 3, from which we could clearly see that the empirical density functions, obtained by kernel smoothing, are symmetric around 0, decrease quickly to 0 on both sides, and they are tracking the limiting density function closely. As the limiting density is singular at the origin, it is hard for the kernel density estimate to match the limit density at the origin. Fig 4 shows the plot for the Y process, in which case the limiting density is asymmetric around the origin. Again the empirical densities are similar to the limiting density although the match over the positive axis is poorer than that over the negative axis.

The empirical coverage probabilities of the nominal 0.95 confidence intervals are generally lower than but approach the nominal 0.95 with increasing T, but with greater disparity for the case of piecewise constant diffusion than its constant counterpart. For the threshold parameter, the confidence intervals are constructed based on (7), with the parameters estimated by the quasi-likelihood estimates and the diffusion parameter(s) estimated by (8). We also computed the confidence intervals using the true parameter values, and the corresponding empirical coverage rates are enclosed in parentheses in Table 2. For the case of constant diffusion, these two sets of confidence intervals have similar coverage rates, but their difference is larger for the case of piecewise constant diffusion.

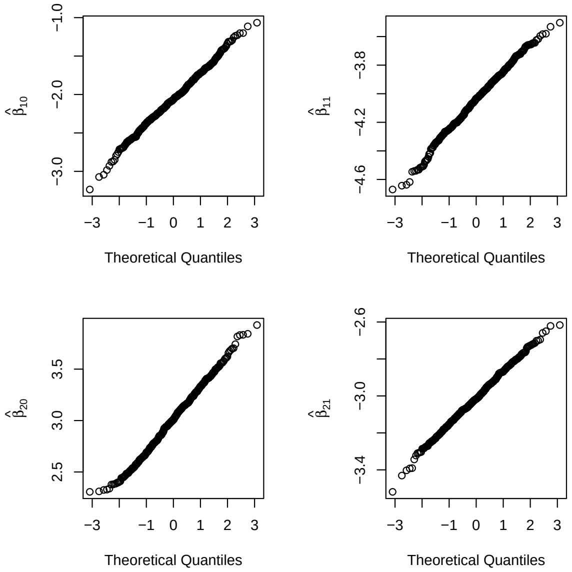

Figure 2: Normal quantile-quantile plots for $\hat{\boldsymbol{\beta}}$ estimated with data sampled from X with $T = 500$

6. Interest Rate Analysis

Consider the three-month US treasury rate based on the Federal Reserve Bank's H15 data set (Fig. 5). It is a continuous-time process with data collected on a daily basis. As the rates are only published on business days, the data are unequally spaced, but they shall be treated as equally spaced partly because it is unclear whether over a gap of, say, two non-business days, information accumulates twice as fast as over a one-day gap between

T	Parameter Estimates						Coverage Rate				
		$\hat{\beta}_{10}$	$\hat{\beta}_{11}$	$\hat{\beta}_{20}$	$\hat{\beta}_{21}$	$\hat{r}$	$\hat{\beta}_{10}$	$\hat{\beta}_{11}$	$\hat{\beta}_{20}$	$\hat{\beta}_{21}$	$\hat{r}$
	$\boldsymbol{\theta}_0$	-2	-4	3	-3	0					
X											
200	average	-2.267	-4.159	3.192	-3.087	-0.002	0.910	0.918	0.926	0.942	0.884
	sd	0.845	0.527	0.715	0.331	0.128					(0.906)
	bias	-0.267	-0.159	0.192	-0.087	-0.002					
500	average	-2.074	-4.055	3.068	-3.032	0.001	0.924	0.944	0.936	0.924	0.930
	sd	0.491	0.304	0.433	0.208	0.034					(0.930)
	bias	-0.074	-0.055	0.068	-0.032	0.001					
1000	average	-2.059	-4.047	3.014	-3.012	0.000	0.942	0.952	0.940	0.940	0.936
	sd	0.339	0.205	0.310	0.143	0.017					(0.932)
	bias	-0.059	-0.047	0.014	-0.012	0.000					
Y											
200	average	-2.098	-4.062	4.182	-3.272	0.266	0.872	0.874	0.868	0.888	0.850
	sd	0.953	0.525	2.282	0.554	0.511					(0.972)
	bias	-0.098	-0.062	1.182	-0.272	0.266					
500	average	-2.047	-4.034	3.501	-3.114	0.094	0.918	0.928	0.934	0.940	0.874
	sd	0.480	0.273	1.093	0.280	0.227					(0.956)
	bias	-0.047	-0.034	0.501	-0.114	0.094					
1000	average	-2.061	-4.039	3.223	-3.055	0.044	0.916	0.920	0.922	0.948	0.916
	sd	0.346	0.196	0.690	0.183	0.116					(0.970)
	bias	-0.061	-0.039	0.223	-0.055	0.044					

Table 2: Empirical performance of the quasi-likelihood estimators for the X and Y processes. The empirical coverage rates pertain to nominal 0.95 confidence intervals. For the threshold parameter, the confidence intervals are constructed based on the drift parameters estimated by quasi-likelihood and the diffusion parameters by quadratic variation calculations, and the empirical coverage rates for those based on the true parameter values are enclosed in parentheses.

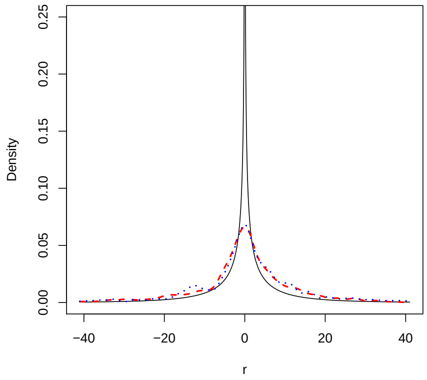

Figure 3: Empirical density function of $T(\hat{r}-r_0)$, based on the $\{X(t)\}$ process with $T = 500$ (red dashed curve) and $T = 1000$ (blue dotted curve) and the limiting density function (solid curve).

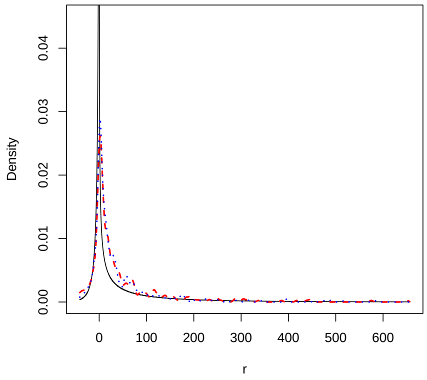

Figure 4: Empirical density function of $T(\hat{r}-r_0)$, based on the $\{Y(t)\}$ process with $T = 500$ (red dashed curve) and $T = 1000$ (blue dotted curve) and the limiting density function (solid curve).

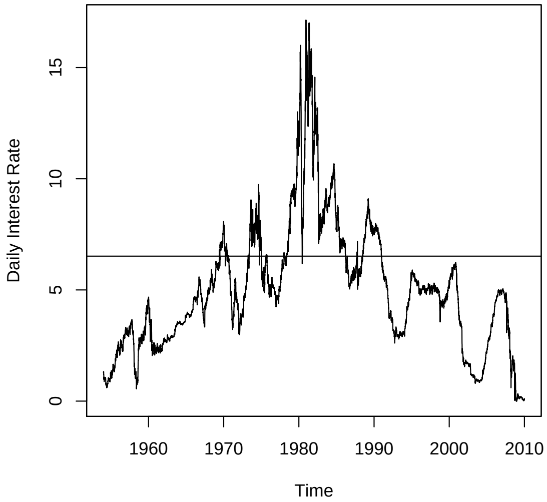

Figure 5: Daily interest rate (solid curve) with the horizontal blue line being the estimated threshold.

consecutive daily data. Moreover, the model fit changes little upon using the actual, irregular calendar sampling times; see *Supplementary Material.* We shall adopt the convention that the equal time interval for the "daily" interest rates is $\Delta t = .046$ while one unit in time represents one month.

The left diagram in Fig. 6 suggests that the short rate X may satisfy the continuous-time threshold model $dX(t) = \mu(X(t))dt + \sigma(X(t))dW(t)$. The quasi-likelihood scheme provides us the following parameter estimates for the mean function:

$$\hat{\boldsymbol{\beta}}_1 = (0.0216, -0.00498)^\top, \hat{\boldsymbol{\beta}}_2 = (0.416, -0.0480)^\top, \hat{r} = 6.52.$$

where the threshold parameter was searched from $a = 20$ percentile to $b = 80$ percentile of the data. That is,

$$\mu(X(t)) = \begin{cases} \begin{array}{lll} 0.0216 & -0.00498X(t) & \text{if } X(t) <= 6.52 \\ (0.022) & \quad (0.0066) & \\ & & \\ 0.416 & -0.0480X(t) & \text{if } X(t) > 6.52 \\ (0.28) & \quad (0.032) & \end{array} \end{cases}$$

where the standard errors of the autoregressive parameters are enclosed in parentheses and a 95% confidence interval of the threshold parameter is

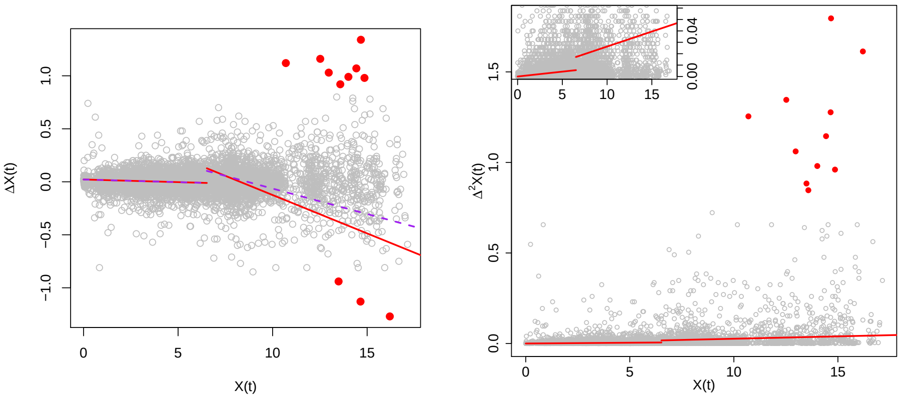

Figure 6: Left diagram: $\Delta X(t)$ versus $X(t)$, for the interest data. Fitted drift term using all data: purple dashed line. Fitted drift term using all data except the "outliers" (red solid circles): red solid line. The Right diagram: $\{\Delta X(t)\}^2$ versus $X(t)$. Fitted diffusion term based on data excluding the outliers: red solid line. Upper left insert zooms in on the lower part of the figure for better appreciation of the fitted diffusion term.

$(3.333, 6.768)$; see below on how the standard errors and the confidence interval are computed.

However, several observations appear to be outliers (red solid circles in the diagram); these observations have daily change in the interest rate being not less than 0.9 in magnitude. Excluding these outliers yields the following TD model:

$$\mu(X(t)) = \left\{ \begin{array}{lll} 0.0216 & -0.00498X(t) & \text{if } X(t) <= 6.52 \\ (0.022) & \quad (0.0066) & \\ & & \\ 0.599 & -0.0724X(t) & \text{if } X(t) > 6.52 \\ (0.26) & \quad (0.030) & \end{array} \right.$$

while a 95% confidence interval of the threshold parameter is $(4.932, 6.688)$. Hence, removing the outliers preserves the threshold but modifies the fit in the upper regime in rendering the slope more negative and the intercept more positive. The slope of the estimated drift above the threshold, however, seems greater than what would be expected from the appearance of the scatter plot, perhaps due to the influence by a few moderately outlying observations with interest rate change close to -0.9. This raises an interesting future research problem on how to robustify quasi-likelihood estimation. The right diagram in Fig. 6 plots the squared daily interest-rate change versus the daily interest rate, which suggests that the squared diffusion term is a linear function, prompting us to model $\sigma(x) = \sigma_1\sqrt{x}I(x \leq 6.52) + \sigma_2\sqrt{x}I(x > 6.52)$. Based on the quadratic variation formulas in (8), we could calculate $\hat{\sigma}_1^2 = .0186, \hat{\sigma}_2^2 = .0573$. With this diffusion specification, the standard errors for the autoregressive parameters reported in the preceding model fit are computed based on Theorem 5, while confidence intervals of the threshold parameter are constructed using Theorem 4.

The uncertainty in the parameter estimates can be alternatively assessed by the regenerative block bootstrap introduced in Section 4. The short rates are multiples of 0.01, with the median short rate being 4.83. Thus, the median may be treated as representing an approximate atom comprising all values between 4.825 and 4.835. We then carried out the regenerative block bootstrap based on the decomposition of the process between consecutive visits to 4.83. (We have also tried the regenerative bootstrap based on the atom at the threshold 6.52, which yielded similar results; see *Supplementary Material*.) The TD model was then refit to each regenerative block bootstrap process, with outliers similarly suppressed. The procedure was replicated 1000 times. Based on the percentile method, the 95% bootstrap confidence intervals of $\beta_{10}, \beta_{11}, \beta_{20}, \beta_{2,1}$ are $(-0.0466, 0.0996)$, $(-0.0201, 0.0194)$, $(0.0148, 1.39)$, $(-0.198, -0.010)$, respectively, while that of the threshold is $(4.38, 8.04)$. The bootstrap-based inference is broadly similar to the inference based on the asymptotics, although the bootstrap confidence intervals are generally wider than their theoretical counterparts. The drift term in the lower regime is then not significantly different from the zero function, showing that the short rate evolves as a martingale process, until it hits the upper regime. In the upper regime, the drift term has a significantly negative slope, effecting an autoregressive regulation to check the growth of the short rates and thereby ensuring ergodicity of the process.

Fig. 7 displays the histogram of the interest rate data, with the superimposition of the stationary density given by (3) with the coefficients determined by the model fitted with all data and that without the outliers. The stationary densities are capable of capturing the bimodality in the data, although there seems to be an excess of observations in the center than in the lower tail, as compared to the stationary densities. However, the fit excluding the outliers seems to fit the data slightly better in the upper regime than that using all data; overall, the former provides a slightly better fit to the data.

Note that the fitted TD model is consistent with the pattern displayed in Fig. 6. The slope parameter in the upper regime is larger than that in the lower regime, in magnitude; hence, there is stronger mean reversion in the

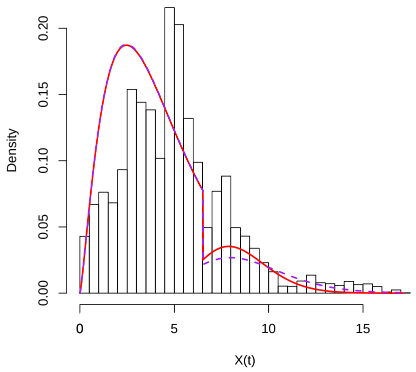

Figure 7: Histogram of the interest rate data, with the stationary density of the TD model fitted with all data (purple dashed line) and that excluding the outliers (red solid line) superimposed.

upper regime (with higher interest rates) than in the lower regime. Also, the diffusion term is proportional to $X(t)$, and the proportionality parameter σ is also larger in the upper regime.

7. Conclusion

An interesting problem is to develop the likelihood ratio test for threshold nonlinearity with continuous-time data. So far, we only considered the TD model which is the first order case of the continuous-time Autoregressive and Moving-average (CTARMA) model (Tong, 1990). An interesting problem is to extend our approach to higher-order cases of the CTARMA models. The quasi-likelihood method is only applied to estimate the piecewise linear mean function. A challenge is to explore the use of the new approach to estimate more general forms of the mean function. Another extension is weighted quasi-likelihood estimation, with weights mimicking the inverse of the diffusion term so the quasi-likelihood estimator may furnish an iterative approach to maximum likelihood estimation.

Though we assume the observations come from a continuous-time process, real data are usually observed discretely. As a result, the integrals used to calculate the estimators need to be replaced by sums. It is of interest to determine the order of convergence of these discrete sums to their integral limits and find the conditions to guarantee that the asymptotic properties of the quasi-likelihood estimator would not be affected by the discretization.

The theoretical properties of the threshold estimator is established under the condition that the threshold is known to lie in some finite interval. It would be interesting to explore whether it is possible to relax this condition, as is the case for conditional least square estimation of a discrete-time SETAR model (Chan, 1993).

Acknowledgements

We are grateful to two referees for very helpful comments. This work was partially supported by the U.S. National Science Foundation (DMS-1021896).

8. Appendix: Proofs of Lemma 1 and Theorem 4

We present the proofs of Lemma 1 and Theorem 4. The proofs of other results are similar to those in Chan (1993) and hence omitted.

8.1. Proof of Lemma 1

Below, we abuse the notation X so it now represents a random variable having the marginal stationary density of $\{X(t)\}$. First, we show that

$$\sup_{r\in R}|\frac{1}{T}\int_0^T X^k(t)I(X(t)\le r)dt - E(X^kI(X\le r))| \to 0 \text{ in probability}, k=0,1,2.$$

Let $[T]$ denote the greatest integer that is smaller than T. Then,

$$\begin{aligned} & \sup_{r\in R}|\frac{1}{T}\int_0^T X^k(t)I(X(t)\le r)dt - E(X^kI(X\le r))| \\ \le\ & \sup_{r\in R}\frac{1}{T}|\int_0^T\{X^k(t)I(X(t)\le r) - E(X^kI(X\le r))\}dt| \\ \le\ & \sup_{r\in R}|\frac{1}{[T]}\int_0^{[T]} X^k(t)I(X(t)\le r)dt - E(X^kI(X\le r))| + \frac{1}{T}\{\int_{[T]}^{[T]+1}|X^k(t)|dt + E|X^k|\}. \end{aligned}$$

Since the last term on the right side of the preceding inequality converges to 0 in probability, without loss of generality, we may and shall assume that T is an integer. The expression $\frac{1}{T}\int_0^T X^k(t)I(X(t)\le r)dt$ can be equivalently written as $\frac{1}{T}\sum_{i=1}^T\int_{i-1}^i X^k(t)I(X(t)\le r)dt, k=0,1,2$. Adapting the notion from Van der Vaart (2000, Section 19.2), given two functions l and u, denote the bracket $[l,u]$ as the set of all functions f with $l<f<u$. An ϵ-bracket in $\mathbb{L}_1(P)$ is a bracket $[l,u]$ such that $E|u-l|\le\epsilon$. The bracketing number $N_{[\,]}(\epsilon,\mathcal{F},\mathbb{L}_1)$ is the minimum number of ϵ-brackets needed to cover $\mathcal{F}$. It is known from the Glivenko-Cantelli theorem (Van der Vaart, 2000, Theorem

19.4) that for a class of measurable functions $\mathcal{F}$, if $N_{[\,]}(\epsilon, \mathcal{F}, \mathbb{L}_1) < \infty$ for every $\epsilon > 0$, then

$$\sup_{f\in\mathcal{F}} |\frac{1}{T}\sum_0^T f(X(t)) - E[f(X)]| \to 0 \text{ a.s.}$$

Without loss of generality, assume $0 \in [a, b]$, $\mathcal{F} = \mathcal{F}_1 \cup \mathcal{F}_2 = \{\int_0^1 X^k(t)I(X(t) \leq r)dt, r \in (a, 0]\} \cup \{\int_0^1 X^k(t)I(X(t) \leq r)dt, r \in (0, b]\}$, where functions in both $\mathcal{F}_1$ and $\mathcal{F}_2$ are monotone functions in r for $k = 0, 1, 2$. Hence, $\forall\epsilon$, we could find a partition $a = m_0 < m_1 < \cdots < m_n = b$ such that 0 is one of the m_i's and $E[|X^k(t)|I(m_{i-1} < X(t) \leq m_i)] < \epsilon$. So $\mathcal{F}$ can be covered by m ϵ-brackets constructed from consecutive pairs of $\int_0^1 X^k(t)I(X(t) \leq m_i)dt, i = 0, \cdots, n$. Also $\int_0^1 |X^k(t)|dt$ is an $\mathbb{L}_1$ envelope function for $\int_0^1 X^k(t)I(X(t) \leq r)dt$ $r \in [a, b]$. Thus, the class $\mathcal{F}$ belongs to the Glivenko-Cantelli class and $\frac{1}{T}\int_0^T X^k(t)I(X(t) \leq r)dt$ converges uniformly to $E[X^k I(X \leq r)]$, see Van der Vaart (2000, Theorem 19.4).

Now consider $\frac{1}{T}\int_0^T X^k(t)\sigma(X(t))I(X(t) \leq r)dW(t), k = 0, 1$. We claim that for a fixed $k = 0, 1$, the stochastic equicontinuity condition holds for the process $\frac{1}{T}\int_0^T X^k(t)\sigma(X(t))I(X(t) \leq r)dW(t), r \in [a, b]$, i.e. every $\epsilon, \eta > 0$, there exists a $\delta > 0$ such that for all T sufficiently large,

$$P(\sup_{|r_2-r_1|<\delta, r_1, r_2\in[a,b]} \frac{1}{T}|\int_0^T X^k(t)I(r_1 < X(t) \leq r_2)\sigma(X(t))dW(t)| \geq \epsilon) \leq \eta \tag{9}$$

Indeed, this can be shown by adapting the proofs of Lemmas 1 and 2 in Van Zanten (2000), by modifying the definition of the function h on p. 255 there to

$$h(u) = \int_0^u \int_0^v z^k 1_{[x,y]}(z)dzdv$$

and noticing that there exists a constant M dependent on k such that for all x, y in the interval $[a, b]$, the first derivative $|h'(u)| \leq M|x - y|$.

8.2. Proof of Theorem 4

Without loss of generality, let $r_0 = 0$. Applying Lemmas 4 and 5, we can proceed as if $\tilde{l}_T(\kappa) = l(\boldsymbol{\beta}_0, r_0 + \kappa/T) - l(\boldsymbol{\beta}_0, r_0)$. Thus, $\tilde{l}_T(\kappa) = \tilde{l}_{1,T}(\kappa) + \tilde{l}_{2,T}(\kappa)$ where $\tilde{l}_{1,T}(\kappa) = -\frac{1}{2}\int_0^T\{(\boldsymbol{\beta}_{1,0}^\top Y(t) - \boldsymbol{\beta}_{2,0}^\top Y(t))^2 I(0 < X(t) \leq \kappa/T) + (\boldsymbol{\beta}_{2,0}^\top Y(t) - \boldsymbol{\beta}_{1,0}^\top Y(t))^2 I(-\kappa/T < X(t) \leq 0)\}dt$ and $\tilde{l}_{2,T}(\kappa) = \int_0^T\{(\boldsymbol{\beta}_{1,0}^\top Y(t) - \boldsymbol{\beta}_{2,0}^\top Y(t))I(0 < X(t) \leq \kappa/T)\sigma(X(t)) + (\boldsymbol{\beta}_{2,0}^\top Y(t) - \boldsymbol{\beta}_{1,0}^\top Y(t))I(-\kappa/T < X(t) \leq 0)\sigma(X(t))\}dW(t)$.

In order to show the required weak convergence result for the process $\tilde{l}_T$, it suffices to verify the following two conditions (Van der Vaart, 2000, p. 261).

Condition (i) For every $K, \epsilon, \eta > 0$, there exists a partition of $[-K, K]$, denoted by $R_1, ..., R_q$, such that $\limsup_{T\to\infty} P(\sup_i \sup_{\kappa_1,\kappa_2 \in R_i} |\tilde{l}_T(\kappa_2) - \tilde{l}_T(\kappa_1)| \geq \epsilon) \leq \eta$.

Condition (ii) The sequence $(\tilde{l}_T(\kappa_1), \tilde{l}_T(\kappa_2), ..., \tilde{l}_T(\kappa_q))$ converges in distribution to those of the limiting Gaussian process for every finite set of real numbers $\{\kappa_i\}$.

We now verify these two conditions.

1) Let $K, \epsilon, \eta > 0$ be given. Let $-K = \delta_0 < \delta_1 < \cdots < \delta_q = K$ be a partition of $[-K, K]$, where $\delta_i - \delta_{i-1} \equiv h > 0$, for $i = 1, \cdots, q$. Define R_i as the intersection of $(\delta_{i-1}, \delta_i]$ with $\{\kappa : a \leq r_0 + \kappa/T \leq b\}$. The determination of h relies on the the following consequence of the Garsia-Rodemich-Rumsey inequality, see (3.2) in Barlow and Yor (1982), which states that for a family of real-valued random variables $\{U(a), a \in R\}$ for which there exist constants $H, \gamma > 0$ and $\alpha > 1$ such that $E(|U(a) - U(b)|^\gamma) \leq H|a - b|^\alpha$, for all a, b, then for any constant $0 < m < \alpha - 1$, there exists a constant C, dependent on α, m and γ, such that

$$E\left(\sup_{|a-b|\leq r} \frac{|U(a) - U(b)|^\gamma}{|a - b|^m}\right) \leq CHr^{\alpha-m} \tag{10}$$

We shall show below that for any constant $K > 0$, there exists a constant H such that for all $-K \leq \kappa_1, \kappa_2 \leq K$

$$E|\tilde{l}_{1,T}(\kappa_2) - \tilde{l}_{1,T}(\kappa_1)|^2 \leq H|\kappa_2 - \kappa_1|^2, \tag{11}$$

and

$$E|\tilde{l}_{2,T}(\kappa_2) - \tilde{l}_{2,T}(\kappa_1)|^4 \leq H|\kappa_2 - \kappa_1|^2. \tag{12}$$

Thus, by choosing $h > 0$ sufficiently small, Condition (i) can be readily shown to hold, by making use of (10), (11), (12) and the Markov inequality. It remains to verify (11) and (12). We can enlarge the partition to ensure that the elements in each R_i are of the same sign. Consider the case that $0 \leq \kappa_1 < \kappa_2 \leq K$ where K is fixed. Below c denotes a generic constant that may vary from occurrence to occurrence.

$$\begin{aligned}
& E(\tilde{l}_{1,T}(\kappa_2) - \tilde{l}_{1,T}(\kappa_1))^2 \\
\leq\ & cE(\int_0^T I(\kappa_1/T < X(t) \leq \kappa_2/T)dt)^2 \\
\leq\ & cE(\int_0^T I(\kappa_1/T < X(t) \leq \kappa_2/T) - E(I(\kappa_1/T < X(t) \leq \kappa_2/T))dt)^2 \\
& \qquad + cE^2(\int_0^T I(\kappa_1/T < X(t) \leq \kappa_2/T)dt) \\
\leq\ & cE^2(\int_0^T I(\kappa_1/T < X(t) \leq \kappa_2/T)dt),
\end{aligned}$$

by the ρ-mixing property of $\{I(c_1 < X(t) \leq c_2)\}$, for any two fixed constants c_1, c_2; the aforementioned ρ–mixing property can be proved by the same techniques used in proving Lemma 3. But $E^2(\int_0^T I(\kappa_1/T < X(t) \leq \kappa_2/T)dt) \leq c|\kappa_2 - \kappa_1|^2$, by (A3). The case for $-K \leq \kappa_1 < \kappa_2 < 0$ can be proved similarly. Thus, (11) holds. The verification of (12) can be proceeded similarly on noting that the Burkholder-Davis-Gundy inequality implies the existence of a universal constant C such that, for $0 \leq \kappa_1 < \kappa_2 < K$,

$$\begin{aligned}
& E(\int_0^T (\boldsymbol{\beta}_{1,0}^\top Y(t) - \boldsymbol{\beta}_{2,0}^\top Y(t)) I(\kappa_1/T < X(t) \leq \kappa_2/T)\sigma(X(t))dW(t))^4 \\
\leq\ & CE(\int_0^T (\boldsymbol{\beta}_{1,0}^\top Y(t) - \boldsymbol{\beta}_{2,0}^\top Y(t))^2 I(\kappa_1/T < X(t) \leq \kappa_2/T)\sigma^2(X(t))dt)^2
\end{aligned}$$

and the fact that $\sigma(\cdot)$ is bounded over compact sets by (A4).
2) To show the convergence of the finite-dimensional distributions, we first introduce some notations. Consider the empirical measure defined by $m_t(B) = \frac{1}{T}\int_0^T I(X(t) \in B)dt$ where B is any Borel set. Moreover, for continuous semimartingales, the empirical measure admits a (random) density function w.r.t the Lebesgue measure, known as the empirical density function and denoted by $\pi_t(\cdot)$, so that $\frac{1}{T}\int_0^T I(X(t) \in B)dt = \int_B \pi_t(x)dx$. It can be shown that for any $K > 0$,

$$\sup_{|x| \leq K} |\pi_t(x) - \pi(x)| \to 0 \text{ in probability,} \tag{13}$$

where $\pi(\cdot)$ is the stationary density function; this can be shown by adapting the proof of Theorem 7 of Van Zanten (2000) for diffusion processes with discontinuous coefficients satisfying (A2) and (A4).

Let $\kappa > 0$ be fixed. We shall first show that $\tilde{l}_{1,T}(\kappa)$ converges to a constant in probability. Recall that we let $r_0 = 0$ and $f(r_0) = \{(1, r_0)(\boldsymbol{\beta}_{1,0} - \boldsymbol{\beta}_{1,0})\}^2$

$$
\begin{aligned}
\tilde{l}_{1,T}(k) &= -\frac{1}{2}\int_0^T (\boldsymbol{\beta}_{1,0}^\top Y(t) - \boldsymbol{\beta}_{2,0}^\top Y(t))^2 I(0 < X(t) \le \kappa/T)dt \\
&= \frac{-T}{2}\int_0^{\kappa/T} \{(1, x)(\boldsymbol{\beta}_{1,0} - \boldsymbol{\beta}_{1,0})\}^2 \pi_T(x)dx \\
&\to \frac{-\kappa}{2} f^2(r_0)\pi(r_0+),
\end{aligned}
$$

in probability, owing to (13) and (A3).

Next, we show that, in terms of finite dimensional distributions, $\{\tilde{l}_{2,T}(\kappa), \kappa \ge 0\}$ converges weakly to a centred Gaussian process with covariance kernel $f^2(r_0)\pi(r_0+)\sigma^2(r_0+)(\kappa_1 \wedge \kappa_2)$ and $\{\tilde{l}_{2,T}(\kappa), \kappa \le 0\}$ converges to an independent centred Gaussian process with non-positive index and whose covariance kernel equals $K(\kappa_1, \kappa_2) = f^2(r_0)\pi(r_0-)\sigma^2(r_0-)(|\kappa_1| \wedge |\kappa_2|)$. We prove this by making use of the Central Limit Theorem for triangular array of martingale difference sequences (Durrett, 2010, Theorem 7.4) and illustrate this for the case of a fixed $\kappa > 0$. Without loss of generality, assume T is a positive integer. Then, $\tilde{l}_{2,T}(\kappa) = \sum_{i=1}^T W_{i,T}$ where $W_{i,T} = \int_i^{i+1} (\boldsymbol{\beta}_{1,0}^\top Y(t) - \boldsymbol{\beta}_{2,0}^\top Y(t)) I(0 < X(t) \le \kappa/T)\sigma(X(t))dW(t)$. We now verify that conditions (i) and (ii) of Theorem 7.4 of Durrett (2010) hold for $\tilde{l}_{2,T}(\kappa)/\sqrt{f^2(r_0)\pi(r_0+)\sigma^2(r_0+)\kappa}$.
(i) $\sum_{i \le T} E\{W_{i,T}^2 I(|W_{i,T}| > \epsilon)\} \to 0$ in probability.
(ii) $\sum_i W_{i,T}^2 \to f^2(r_0)\pi(r_0+)\sigma^2(r_0+)\kappa$ in probability.
$\{W_{i,T}\}$ is a martingale difference sequence where the underlying process $\{X(t)\}$ is stationary and ergodic,

$$
\begin{aligned}
&\sum_{i \le T} E\{W_{i,T}^2 I(|W_{i,T}| > \epsilon)\} = \frac{T}{T}\sum_{i \le T} E\{W_{i,T}^2 I(|W_{i,T}| > \epsilon)\} \\
= \quad &TE\{W_{0,T}^2 I(|W_{0,T}| > \epsilon)\} = E\{(\sqrt{T}W_{0,T})^2 I(\sqrt{T}|W_{0,T}| > \sqrt{T}\epsilon)\} \\
\to 0, \quad &\text{by the dominated convergence theorem and the boundedness of } E\{TW_{0,T}^2\}
\end{aligned}
$$

Thus Condition (i) is satisfied, and Condition (ii) directly follows from the convergence results about the empirical density functions for an ergodic stationary process (Van Zanten, 2000). The aforementioned finite-dimensional convergence result can then be verified by the Cramér-Wold device. Hence, we have shown the weak convergence of $\tilde{l}_T(.)$. More specifically,

$$
\begin{aligned}
\tilde{l}_T(\kappa)I(\kappa \ge 0) &\rightsquigarrow -\frac{1}{2}f(r_0)^2\pi(r_0+)\kappa + f(r_0)\sqrt{\pi(r_0+)}\sigma(r_0+)W(\kappa) \\
\tilde{l}_T(-\kappa)I(\kappa > 0) &\rightsquigarrow -\frac{1}{2}f(r_0)^2\pi(r_0-)\kappa + f(r_0)\sqrt{\pi(r_0-)}\sigma(r_0-)W(-\kappa),
\end{aligned}
$$

where $\{W(\kappa), \kappa \in R\}$ is the standard Brownian motion with $W(0) = 0$ almost surely. Since $\pi(x)\sigma^2(x)$ is a continuous function, $\sqrt{\pi(r_0+)}\sigma(r_0+) = \sqrt{\pi(r_0-)}\sigma(r_0-)$, after re-scaling, the limiting process is a Brownian motion with negative linear drift on the two tails. If $\sigma(x)$ is continuous at $x = r_0$, they would also be symmetric about r_0 in distribution.

Now, we show that

$$\tilde{r}_T = T(\hat{r} - r_0) = \arg\max_r \{\tilde{l}_T(r)I(r > 0) + \tilde{l}_T(-r)I(r > 0)\}$$

also converges in distribution using the continuous mapping theorem. First, note the time at which the processes $\tilde{l}_T(\kappa)I(\kappa > 0)$ and $\tilde{l}_T(-\kappa)I(\kappa > 0)$ reach their maximum is almost surely unique respectively. The maximum value of a Brownian motion $W(s), s \leq t$ is defined as $M(t) = \max_{s\leq t} W(s)$. Now denote the time that $M(t)$ is first reached as $\theta_1(t) = \inf\{s \leq t : W(s) = M(t)\}$ and the latest time that $M(t)$ is reached as $\theta_2(t) = \sup\{s \leq t : W(s) = M(t)\}$. It is known that $\theta_1(t) = \theta_2(t)$ holds almost surely (Karatzas and Shreve, 1991, p. 102); consequently, the time that each process reaches its maximum is almost surely unique. Since $\tilde{l}_T(k)I(k > 0)$ and $\tilde{l}_T(-k)I(k > 0)$ weakly converge to independent continuous processes, the probability that they reach the same maximum is 0. Then the uniqueness of the entire process is proved by the independence of the two processes.

Now, $\tilde{r}_T = \tilde{r}_T = T(\hat{r} - r_0) = \arg\max_r\{\tilde{l}_T(r)I(r > 0) + \tilde{l}_T(-r)I(r > 0)\}$ is a map from $D[0, \infty)$ to R. By the uniqueness of the maximizer, continuity and weak convergence of $\tilde{l}_T(\cdot)$, $\tilde{r}_T = T(\hat{r} - r_0)$ also converges weakly to $\tilde{r}$, (Van der Vaart, 2000, Theorem 18.11).

Finally, we drive the limiting distribution of $\tilde{r}_T$. The time that the process $\{Z(t) = \mu t + W(t), \mu < 0, t \geq 0\}$ reaches its maximum is defined as $m_Z = \arg\max_t\{Z(t)\} = \arg\max_t\{\mu t + W(t)\}$. It has the following density function:

$$g_{m_Z}(s) = -2\mu[\frac{1}{\sqrt{s}}\phi(\mu\sqrt{s}) + \mu\Phi(\mu\sqrt{s})]$$

where ϕ, Φ are standard Gaussian density and distribution function respectively (Buffet, 1900).

It can be easily seen that the density function approaches infinity when s approaches 0, and as s increases, the density converges to 0 quickly. Thus, the maximum is achieved at a small neighborhood of 0 with high probability. Note that this definition and result can be applied when t is negative by symmetry. For instance, when $\{Z(t) = -\mu t + W(-t), \mu < 0, t < 0\}$, m_Z can be defined as $m_Z = \arg\max_t\{-\mu t + W(-t)\}$. Therefore, we could readily obtain the distribution of the time when a two-sided process such as $\{Z(t) =$

$\{\mu t + W(t)\}I(t \geq 0) + \{\mu(-t) + W(-t)\}I(t < 0), \mu < 0, t \in [0, +\infty)\}$ reaches its maximum.

Now consider the density function of our objective process, $\tilde{r}$, the limiting process of $\tilde{r}_T = \arg\max_r \tilde{l}_T(r)$. Define

$$m = f^2(r_0)\sigma^2(r_0+)\pi(r_0+) = f^2(r_0)\sigma^2(r_0-)\pi(r_0-),$$

the density function of $m\tilde{r}$ is:

$$\begin{aligned} g_{m\tilde{r}}(s) &= I(s<0)\frac{1}{2\sigma^2(r_0-)}[\frac{1}{\sqrt{-s}}\phi(-\frac{1}{2\sigma^2(r_0-)}\sqrt{-s}) - \frac{1}{2\sigma(r_0-)^2}\Phi(-\frac{1}{2\sigma^2(r_0-)}\sqrt{-s})] \\ &+ I(s>0)\frac{1}{2\sigma^2(r_0+)}[\frac{1}{\sqrt{s}}\phi(-\frac{1}{2\sigma^2(r_0+)}\sqrt{s}) - \frac{1}{2\sigma^2(r_0+)}\Phi(-\frac{1}{2\sigma^2(r_0+)}\sqrt{s})] \end{aligned}$$

Then, as $m\tilde{r}$ has density function $g_{m\tilde{r}}(s)$, the density function of $\tilde{r}$ can be expressed as:

$$\begin{aligned} g_{\tilde{r}}(s) &= I(s<0)\frac{f^2(r_0)\pi(r_0-)}{2}[\frac{1}{\sqrt{-ms}}\phi(-\frac{1}{2\sigma^2(r_0-)}\sqrt{-ms}) - \frac{1}{2\sigma^2(r_0-)}\Phi(-\frac{1}{2\sigma^2(r_0-)}\sqrt{-ms})] \\ & I(s>0)\frac{f^2(r_0)\pi(r_0+)}{2}[\frac{1}{\sqrt{ms}}\phi(-\frac{1}{2\sigma^2(r_0+)}\sqrt{ms}) - \frac{1}{2\sigma^2(r_0+)}\Phi(-\frac{1}{2\sigma^2(r_0+)}\sqrt{ms})]. \end{aligned}$$

References

Barlow, M. T., Yor, M., 1982. Semi-martingale inequalities via the Garsia-Rodemich-Rumsey lemma, and applications to local times. Journal of Functional Analysis 49 (2), 198–229.

Bertail, P., Clémençon, S., et al., 2006. Regenerative block bootstrap for markov chains. Bernoulli 12 (4), 689–712.

Black, F., Derman, E., Toy, W., 1990. A one-factor model of interest rates and its application to treasury bond options. Financial Analysts Journal, 33–39.

Black, F., Karasinski, P., 1991. Bond and option pricing when short rates are lognormal. Financial Analysts Journal, 52–59.

Brockwell, P., Davis, R., Yang, Y., 2007. Continuous-time gaussian autoregression. Statistica Sinica 17 (1), 63.

Brockwell, P., Hyndman, R. J., 1992. On continuous-time threshold autoregression. International Journal of Forecasting 8 (2), 157–173.

Brockwell, P. J., 1994. On continuous-time threshold ARMA processes. Journal of Statistical Planning and Inference 39 (2), 291–303.

Brockwell, P. J., Hyndman, R. J., Grunwald, G. K., 1991. Continuous time threshold autoregressive models. Statistica Sinica 1 (2), 401–410.

Buffet, E., 1900. On the time of the maximum of Brownian motion with drift. International Journal of Stochastic Analysis 16 (3), 201–207.

Chan, K. C., Karolyi, G. A., Longstaff, F. A., Sanders, A. B., 1992. An empirical comparison of alternative models of the short-term interest rate. The Journal of Finance 47 (3), 1209–1227.

Chan, K.-S., 1993. Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model. The Annals of Statistics 21 (1), 520–533.

Chan, K. S., Tong, H., 1986. On estimating thresholds in autoregressive models. Journal of time series analysis 7 (3), 179–190.

Cline, D., Pu, H., 1999. Geometric ergodicity of nonlinear time series. Statistica Sinica 9 (4), 1103–1118.

Coakley, J., Fuertes, A.-M., Pérez, M.-T., 2003. Numerical issues in threshold autoregressive modeling of time series. Journal of Economic Dynamics and Control 27 (11), 2219–2242.

Cox, J. C., Ingersoll Jr, J. E., Ross, S. A., 1985. A theory of the term structure of interest rates. Econometrica 53, 385–407.

Decamps, M., Goovaerts, M., Schoutens, W., 2006. Self exciting threshold interest rates models. International Journal of Theoretical and Applied Finance 9 (07), 1093–1122.

Durrett, R., 2010. Probability: theory and examples, 4th Edition. Cambridge University Press, New York, NY.

Elerian, O., Chib, S., Shephard, N., 2001. Likelihood inference for discretely observed nonlinear diffusions. Econometrica 69 (4), 959–993.

Eraker, B., 2001. MCMC analysis of diffusion models with application to finance. Journal of Business and Economic Statistics 19 (2), 177–191.

Eraker, B., 2004. Do stock prices and volatility jump? Reconciling evidence from spot and option prices. The Journal of Finance 59 (3), 1367–1404.

Hansen, B. E., 1997. Inference in tar models. Studies in nonlinear dynamics and econometrics 2 (1), 1–14.

Hull, J., 2010. Options, Futures, and Other Derivatives, 7/e (With CD). Pearson Education, Upper Saddle River, New Jersey.

Karatzas, I. A., Shreve, S. E., 1991. Brownian motion and stochastic calculus. Vol. 113. Springer.

Karlin, S., Taylor, H. M., 1981. A second course in stochastic processes. Vol. 2. Academic Press.

Kutoyants, Y. A., 2012. On identification of the threshold diffusion processes. Annals of the Institute of Statistical Mathematics 64 (2), 383–413.

Pai, J., Pedersen, H. W., 1999. Threshold models of the term structure of interest rate. In: Proceedings of the 9th International AFIR Colloquium. pp. 387–400.

Roberts, G. O., Stramer, O., 2001. On inference for partially observed nonlinear diffusion models using the Metropolis–Hastings algorithm. Biometrika 88 (3), 603–621.

Stramer, O., Roberts, G. O., 2007. On Bayesian analysis of nonlinear continuous-time autoregression models. Journal of Time Series Analysis 28 (5), 744–762.

Stramer, O., Tweedie, R., Brockwell, P., 1996. Existence and stability of continuous time threshold arma processes. Statistica Sinica 6 (3), 715–732.

Tong, H., 1990. Non-linear time series: a dynamical system approach. Oxford University Press, Oxford.

Tong, H., Yeung, I., 1991. Threshold autoregressive modeling in continuous time. Statistica Sinica 1 (2), 411–430.

Van der Vaart, A. W., 2000. Asymptotic statistics. Cambridge University Press, Cambridge.

Van Zanten, J., 2000. On the uniform convergence of the empirical density of an ergodic diffusion. Statistical Inference for Stochastic Processes 3 (3), 251–262.

Vasicek, O., 1977. An equilibrium characterization of the term structure. Journal of Financial Economics 5 (2), 177–188.

Supplementary Material

Model fit with irregular sampling times

So far, the analysis proceeds under the working assumption that the short rates were regularly sampled data. This assumption may be assessed by comparing the model fit based on the working assumption with the following fit based on the actual, irregular calendar sampling times (with outliers similarly suppressed):

$$\mu(X(t)) = \begin{cases} \underset{(0.021)}{0.0207} \quad \underset{(0.0063)}{-0.00478X(t)} & \text{if } X(t) <= 6.52 \\ \\ \underset{(0.25)}{0.582} \quad \underset{(0.029)}{-0.0703X(t)} & \text{if } X(t) > 6.52 \end{cases}$$

with a 95% confidence interval of the threshold parameter being $(4.950, 6.685)$, and $\hat{\sigma}_1^2 = .0178, \hat{\sigma}_2^2 = .0551$. The two model fits are quite similar, showing that the working assumption is reasonable.

Regenerative block bootstrap for the real application

Using the median as the atom, there are 80 complete blocks with the block size distribution summarized by the following table. The longest complete block corresponds to the stretch of high interest rate period.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.0	2.0	9.0	189.7	79.5	5287.0

Table S1: Summary of complete-block sizes for the regenerative block bootstrap using 4.83 as the atom.

Using the estimated threshold value of 6.52 (the 76.9 percentile) as the atom, the number of complete blocks is reduced to 42, with the following summary of the block size distribution:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.0	4.0	11.0	192.2	133.0	2101.0

Table S2: Summary of complete-block sizes for the regenerative block bootstrap using 6.52 as the atom.

Notice that the longest block size is slightly half the maximum block size for the case of using 4.83 as the atom. The corresponding 95% bootstrap confidence intervals of $\beta_{10}, \beta_{11}, \beta_{20}, \beta_{2,1}$ are $(-0.0420, 0.0674)$, $(-0.0151, 0.0183)$, $(0.145, 1.33)$, $(-0.162, -0.0320)$, respectively, while that of the threshold is $(5.60, 8.04)$. The inference from the regenerative block bootstrap using either values as the atom is broadly similar although the bootstrap confidence intervals based on the atom at the median are slightly wider than their counterparts based on using the threshold as the atom, likely because the latter has lesser complete cycles, resulting in reduced variation in the regenerative block bootstrap.

Proofs of other results in the paper

Maximizing the quasi-likelihood ratio function $l(\boldsymbol{\theta})$ is the same as maximizing $l(\boldsymbol{\theta}) - l(\boldsymbol{\theta}_0)$ where $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_{1,0}^\top, \boldsymbol{\beta}_{2,0}^\top, r_0)^\top$ is the true parameter. For simplicity, we give the proofs for the case $r \geq r_0$ throughout this section as the proof for the case $r < r_0$ is similar. Consider the following decomposition:

$$l(\boldsymbol{\theta}) - l(\boldsymbol{\theta}_0) = R_{1,t} + R_{2,t} + R_{3,t},$$

where

$$\begin{aligned}
R_{1,t} &= \int_0^T \{(\beta_{10} + \beta_{11}X(t)) - (\beta_{10,0} + \beta_{11,0}X(t))\}I(X(t) \leq r_0))dX(t) \\
&\quad -\frac{1}{2}\int_0^T \{(\beta_{10} + \beta_{11}X(t))^2 - (\beta_{10,0} + \beta_{11,0}X(t))^2\}I(X(t) \leq r_0)dt \\
&= -\frac{1}{2}\int_0^T (\beta_{10} + \beta_{11}X(t) - \beta_{10,0} - \beta_{11,0}X(t))^2 I(X(t) \leq r_0)dt \\
&\quad + \int_0^T (\beta_{10} + \beta_{11}X(t) - \beta_{10,0} - \beta_{11,0}X(t))I(X(t) \leq r_0)\sigma(X(t))dW(t),
\end{aligned}$$

$$\begin{aligned}
R_{2,t} &= \int_0^T \{(\beta_{10} + \beta_{11}X(t)) - (\beta_{20,0} - \beta_{21,0}X(t))\}I(r_0 < X(t) \leq r))dX(t) \\
&\quad -\frac{1}{2}\int_0^T \{(\beta_{10} + \beta_{11}X(t))^2 - (\beta_{20,0} - \beta_{21,0}X(t))^2\}I(r_0 < X(t) \leq r)dt \\
&= -\frac{1}{2}\int_0^T (\beta_{10} + \beta_{11}X(t) - \beta_{20,0} - \beta_{21,0}X(t))^2 I(r_0 < X(t) \leq r)dt \\
&\quad + \int_0^T (\beta_{10} + \beta_{11}X(t) - \beta_{20,0} - \beta_{21,0}X(t))I(r_0 < X(t) \leq r)\sigma(X(t))dW(t),
\end{aligned}$$

$$
\begin{aligned}
R_{3,t} &= \int_0^T \{(\beta_{20}+\beta_{21}X(t)) - (\beta_{20,0}-\beta_{21,0}X(t))\}I(X(t)>r))dX(t) \\
&\quad -\frac{1}{2}\int_0^T \{(\beta_{20}+\beta_{21}X(t))^2 - (\beta_{20,0}-\beta_{21,0}X(t))^2\}I(X(t)>r))dt \\
&= -\frac{1}{2}\int_0^T (\beta_{20}+\beta_{21}X(t)-\beta_{20,0}-\beta_{21,0}X(t))^2 I(X(t)>r)dt \\
&\quad +\int_0^T (\beta_{20}+\beta_{21}X(t)-\beta_{20,0}-\beta_{21,0}X(t))I(X(t)>r)\sigma(X(t))dW(t).
\end{aligned}
$$

8.3. Proof of Lemma 2

Recall $dX(t) = ((\beta_{10}+\beta_{11}X(t))I(X(t)\le r) + (\beta_{20}+\beta_{21}X(t))I(X(t) > r))dt + \sigma(X(t))dW(t)$, so the quasi-likelihood ratio function for $r > r_0$ becomes:

$$
\begin{aligned}
& \frac{l(\boldsymbol{\theta}) - l(\boldsymbol{\theta}_0)}{T} \\
=\ & \frac{1}{T}(R_{1,t}+R_{2,t}+R_{3,t}) \\
=\ & -\frac{1}{2T}\int_0^T (\beta_{10}+\beta_{11}X(t)-\beta_{10,0}-\beta_{11,0}X(t))^2 I(X(t)\le r_0)dt \\
-\ & \frac{1}{2T}\int_0^T (\beta_{10}+\beta_{11}X(t)-\beta_{20,0}-\beta_{21,0}X(t))^2 I(r_0 < X(t)\le r)dt \\
-\ & \frac{1}{2T}\int_0^T (\beta_{20}+\beta_{21}X(t)-\beta_{20,0}-\beta_{21,0}X(t))^2 I(X(t)>r)dt \\
+\ & \frac{1}{T}\int_0^T (\beta_{10}+\beta_{11}X(t)-\beta_{10,0}-\beta_{11,0}X(t))I(X(t)\le r_0)\sigma(X(t))dW(t) \\
+\ & \frac{1}{T}\int_0^T (\beta_{10}+\beta_{11}X(t)-\beta_{20,0}-\beta_{21,0}X(t))I(r_0 < X(t)\le r)\sigma(X(t))dW(t) \\
+\ & \frac{1}{T}\int_0^T (\beta_{20}+\beta_{21}X(t)-\beta_{20,0}-\beta_{21,0}X(t))I(X(t)>r)\sigma(X(t))dW(t) \quad \text{(S1)}
\end{aligned}
$$

The sum of the three terms involving $dW(t)$ is uniformly bounded by an $(|\boldsymbol{\beta}_1-\boldsymbol{\beta}_{1,0}|+|\boldsymbol{\beta}_1-\boldsymbol{\beta}_{2,0}|+|\boldsymbol{\beta}_2-\boldsymbol{\beta}_{2,0}|)o_p(1)$ term and $\frac{1}{T}\int_0^T X^k(t)I(X(t)\le r)dt$ converges to its expectation uniformly for $r\in[a,b], k=0,1,2$, according to Lemma 1. Henceforth in this proof, all $o_p(1)$ terms hold uniformly for $r\in[a,b]$. Here we provide a proof that there exists an $M>0$ such that $\hat{\boldsymbol{\theta}}\in C_1$ with probability approaching 1 as $T\to\infty$, only for the case that $|\boldsymbol{\beta}_1-\boldsymbol{\beta}_{1,0}|\ge|\boldsymbol{\beta}_2-\boldsymbol{\beta}_{2,0}|$, as the other case that $|\boldsymbol{\beta}_2-\boldsymbol{\beta}_{2,0}|\ge|\boldsymbol{\beta}_1-\boldsymbol{\beta}_{1,0}|$ can be proved similarly. Let $\zeta = \min_{u^2+v^2=1} E\{(u+vX(t))^2 I(X(t) \le r_0)\}$ which is positive, as $E[(u+vX)^2 I(X \le r_0)]$ is a continuous and positive function in (u, v). Hence, for $b \ge r \ge r_0$

$$\begin{aligned}
& \frac{l(\boldsymbol{\theta}) - l(\boldsymbol{\theta}_0)}{T} \\
\le & -\frac{1}{2T}\int_0^T (\beta_{10} + \beta_{11}X(t) - \beta_{10,0} - \beta_{11,0}X(t))^2 I(X(t) \le r_0)dt \\
+ & (|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{1,0}| + |\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{2,0}| + |\boldsymbol{\beta}_2 - \boldsymbol{\beta}_{2,0}|)o_p(1) \\
\le & -\frac{1}{2T}|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{1,0}|^2 \int_0^T (\frac{\beta_{11} - \beta_{11,0}}{|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{1,0}|}X(t) + \frac{\beta_{10} - \beta_{10,0}}{|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{1,0}|})^2 I(X(t) \le r_0)dt \\
+ & (|\boldsymbol{\beta}_{1,0} - \boldsymbol{\beta}_{2,0}| + 3|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{1,0}|)o_p(1) \\
\le & -\frac{1}{2}|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{1,0}|^2(\zeta + o_p(1)) + (|\boldsymbol{\beta}_{1,0} - \boldsymbol{\beta}_{2,0}| + 3|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{2,0})|)o_p(1) \qquad \text{(S2)}
\end{aligned}$$

Thus, $\exists M > 0$ such that for $|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{1,0}| > M$, (S2) is negative with probability going to 1 as $T \to \infty$. Similar inequalities can be established for the case $a \le r \le r_0$ and/or $|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_{2,0}| \ge |\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{1,0}|$. Consequently, it holds with probability approaching 1 as $T \to \infty$ that $\frac{l(\boldsymbol{\theta}) - l(\boldsymbol{\theta}_0)}{T} < 0$ for $\boldsymbol{\theta} \notin C_1$, for a suitably chosen finite $M > 0$.

8.4. Proof of Theorem 2

Without loss of generality, assume $\boldsymbol{\theta} \in C_1$. By the boundedness of C_1 and Lemma 1,

$$\frac{1}{T}(l(\boldsymbol{\theta}) - l(\boldsymbol{\theta}_0)) \to E[\frac{1}{T}(l(\boldsymbol{\theta}) - l(\boldsymbol{\theta}_0))] \text{ in probability, uniformly for } \boldsymbol{\theta} \in C_1.$$

Then it remains to show that $E(l(\boldsymbol{\theta}))$ is maximized at $\boldsymbol{\theta}_0$ and it is a well-separated maximum. Let X be a random variable with the same marginal stationary distribution of $\{X(t)\}$. For $r > r_0$,

$$\begin{aligned}
H(\boldsymbol{\theta}) &= E(\frac{l(\boldsymbol{\theta}) - l(\boldsymbol{\theta}_0)}{T}) \\
&= -\frac{1}{2}E\{(\beta_{10} + \beta_{11}X - \beta_{10,0} - \beta_{11,0}X)^2 I(X \le r_0)\} \\
&= \quad -\frac{1}{2}E\{(\beta_{10} + \beta_{11}X - \beta_{20,0} - \beta_{21,0}X)^2 I(r_0 < X \le r)\} \\
&= \quad -\frac{1}{2}E\{(\beta_{20} + \beta_{21}X - \beta_{20,0} - \beta_{21,0}X)^2 I(X > r)\} \\
&< 0 \text{ if } \boldsymbol{\theta} \ne \boldsymbol{\theta}_0.
\end{aligned}$$

It can be shown similarly that $H(\boldsymbol{\theta}) < 0$ if $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ for the case $r \leq r_0$. As the function $H(\cdot)$ is continuous in $\boldsymbol{\theta}$, for all sufficiently small $\epsilon > 0$

$$\min_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\geq\epsilon, \boldsymbol{\theta}\in C_1} H(\boldsymbol{\theta}) < 0.$$

Thus, $\boldsymbol{\theta}$ is a well-separated maximum and we have $\hat{\boldsymbol{\theta}}_T \to \boldsymbol{\theta}_0$, in probability as $T \to \infty$, c.f. Van der Vaart (2000, p.45).

8.5. Proof of Lemma 3

First we show that $\{I(0 < X(t) < r)\}$ is $\rho-$mixing. Below, $\pi(x)$ is the stationary density function of $X(t)$, and $p^{t-s}(x, y)$ is the conditional density function of $X(t)$ at y given $X(s) = x$, with $s \leq t$. Assumption (A2) implies that, for some $\gamma < 1$ and an integrable non-negative function h, $\int_{-\infty}^{\infty} |p^{t-s}(x, y) - \pi(y)|dy < \gamma^{t-s}h(x)$ (Cline and Pu, 1999).

$$\begin{aligned}
& |\mathrm{cov}(I(0 < X(t) < r), I(0 < X(s) < r)| \\
= & |E(I(0 < X(t) < r)I(0 < X(s) < r)) - E^2[I(0 < X(t) < r)]| \\
= & |\int_0^r \int_0^r \pi(x)p^{t-s}(x, y)dxdy - E^2[I(0 < X(t) < r)]| \\
= & |\int_0^r \int_0^r \pi(x)(p^{t-s}(x, y) - \pi(y) + \pi(y))dxdy - E^2[I(0 < X(t) < r)]| \\
= & |\int_0^r \int_0^r \pi(x)(p^{t-s}(x, y) - \pi(y))dxdy| \\
\leq & \gamma^{t-s} \int_0^r h(x)\pi(x)dx,
\end{aligned}$$

which verifies the $\rho-$mixing property for $\{I(0 < X(t) < r)\}$.

Then we show that $f(X(t))I(0 < X(t) < r)$ is ρ-mixing following similar reasoning.
$|\mathrm{Cov}(f(X(t))I(0 < X(t) < r), f(X(s))I(0 < X(s) < r))|$
$= |E(f(X(t))I(0 < X(t) < r)f(X(s))I(0 < X(s) < r)) - E^2(f(X(t))I(0 < X(t) < r))|$
$= |\int_0^r \int_0^r f(x)f(y)\pi(x)p^{t-s}(x, y)dxdy - E^2(f(X(t))I(0 < X(t) < r))|$
$= |\int_0^r \int_0^r f(x)f(y)\pi(x)(p^{t-s}(x, y) - \pi(y) + \pi(y))dxdy - E^2(f(X(t))I(0 < X(t) < r))|$
$= |\int_0^r \int_0^r f(x)f(y)\pi(x)(p^{t-s}(x, y) - \pi(y))dxdy|$
$\leq c\gamma^{t-s} \int_0^r f(x)h(x)\pi(x)dx$, for some constant c because $f(\cdot)$ is bounded over compact sets.

8.6. Proof of Theorem 3

By the consistency of the quasi-likelihood estimator, we may and shall assume that $\boldsymbol{\theta} \in C_2 = \{\boldsymbol{\theta} : |\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}| < c, |\boldsymbol{\beta}_2 - \boldsymbol{\beta}_{20}| < c, |r - r_0| < \Delta\}$

with $1 \geq c, \Delta > 0$ to be determined below. For simplicity, assume $r_0 = 0$. It suffices to show that for all $\epsilon > 0$, $\exists K > 0$, such that with probability greater than $1 - \epsilon$, $1 > |r| > K/T$ implies $l(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, r) - l(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, 0) < 0$.

Define $M_{\boldsymbol{\beta}}(X(t)) = -(\beta_0 + \beta_1 X(t) - \beta_{20,0} - \beta_{21,0} X(t))^2$. Note that for any $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, $|M_{\boldsymbol{\beta}_i}(X(t)) - M_{\boldsymbol{\beta}_j}(X(t))| \leq |\boldsymbol{\beta}_i - \boldsymbol{\beta}_j| \Lambda(X(t))$ where $\Lambda(X(t)) = (2 + |\boldsymbol{\beta}_{1,0} - \boldsymbol{\beta}_{2,0}|)(1 + X^2(t))$. Below, we only consider the case $r > 0$ as the case $r \leq 0$ can be proved similarly.
Define $Q(r) = E[I(0 < X(t) < r)]$. Consider

$$\begin{aligned}
& \frac{1}{TQ(r)}(l(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, r) - l(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, 0)) \\
= & -\frac{1}{2TQ(r)} \int (\beta_{10} + \beta_{11} X(t) - \beta_{20,0} - \beta_{21,0} X(t))^2 I(0 < X(t) \leq r) dt \\
& + \frac{1}{TQ(r)} \int (\beta_{20} + \beta_{21} X(t) - \beta_{10} - \beta_{11} X(t)) I(0 < X(t) \leq r) \sigma dW(t) \\
= & \frac{1}{2TQ(r)} \int [M_{\boldsymbol{\beta}_1}(X(t)) I(0 < X(t) \leq r) - M_{\boldsymbol{\beta}_2}(X(t)) I(0 < X(t) \leq r)] dt \\
& + \frac{1}{TQ(r)} \int (\beta_{20} + \beta_{21} X(t) - \beta_{10} - \beta_{11} X(t)) I(0 < X(t) \leq r) \sigma dW(t) \\
= & \frac{1}{2TQ(r)} \int [M_{\boldsymbol{\beta}_1}(X(t)) - M_{\boldsymbol{\beta}_{1,0}}(X(t))] I(0 < X(t) \leq r) dt \\
& + \frac{1}{2TQ(r)} \int [M_{\boldsymbol{\beta}_{2,0}}(X(t)) - M_{\boldsymbol{\beta}_2}(X(t))] I(0 < X(t) \leq r) dt \\
& + \frac{1}{2TQ(r)} \int [M_{\boldsymbol{\beta}_{1,0}}(X(t)) - M_{\boldsymbol{\beta}_{2,0}}(X(t))] I(0 < X(t) \leq r) dt \\
& + \frac{1}{TQ(r)} \int (\beta_{20} + \beta_{21} X(t) - \beta_{10} - \beta_{11} X(t)) I(0 < X(t) \leq r) \sigma dW(t) \\
\leq & (|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{1,0}| + |\boldsymbol{\beta}_2 - \boldsymbol{\beta}_{2,0}|) \frac{1}{2TQ(r)} \int_0^T \Lambda(X(t)) I(0 < X(t) < r) dt \\
& + \frac{1}{2TQ(r)} \int_0^T M_{\boldsymbol{\beta}_{1,0}}(X(t)) I((0 < X(t) < r) dt \\
& + |\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2| \frac{1}{TQ(r)} \sum_{j=0}^{1} | \int_0^T X^j(t) I(0 < X(t) \leq r) \sigma(X(t)) dW(t)|
\end{aligned}$$

Let $f(\cdot)$ be a real-valued function that is bounded over compact sets. Let $\Delta > 0$ be fixed. We claim that for any $\epsilon > 0, C > 0$, $\exists K > 0$ such that for T sufficiently large

$$P(\sup_{K/T\le r<\Delta}|\int_0^T \frac{I(0<X(t)<r)}{TQ(r)}dt-1|<C)>1-\epsilon \tag{S3}$$

$$P(\sup_{K/T\le r<\Delta}|\int_0^T \frac{f(X(t))I(0<X(t)<r)-E(f(X(t))I(0<X(t)<r))}{TQ(r)}dt|<C)>1-\epsilon \tag{S4}$$

$$P(\sup_{K/T\le r<\Delta}|\frac{\int_0^T X^j(t)I(0<X(t)<r)\sigma(X(t))dW(t)dt}{TQ(r)}|<C)>1-\epsilon, \qquad j=0,1. \tag{S5}$$

Assuming the validity of this claim, we proceed as follows. Let $\kappa = -(\beta_{10,0}-\beta_{20,0})^2/2<0$, by assumption (A1). Note that $M_{\boldsymbol{\beta}_{1,0}}(0)=2\kappa$ and hence $M_{\boldsymbol{\beta}_{1,0}}(X(t))<\kappa$ for $X(t)\in(0,\Delta]$ if Δ is sufficiently small, which is assumed to be the case henceforth. Let M_1 be a finite upper bound of $E(\Lambda(X(t)))$ for $X(t)\in[0,\Delta]$. Then, with probability greater than $1-4\epsilon$, for $\Delta>r>K/T$ and T sufficiently large:
$\frac{1}{TQ(r)}(l(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,r)-l(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,0))$
$<(|\boldsymbol{\beta}_{1,0}-\boldsymbol{\beta}_1|+|\boldsymbol{\beta}_{2,0}-\boldsymbol{\beta}_2|)(C+M_1)+C+\kappa+C|\boldsymbol{\beta}_1-\boldsymbol{\beta}_2|$
$\le 2c(C+M_1)+C+\kappa+(2c+|\boldsymbol{\beta}_{1,0}-\boldsymbol{\beta}_{2,0}|)C$
which is <0 if we choose $c, C>0$ to be sufficiently small. Hence, with probability approaching 1 as $T\to\infty$,

$$\sup_{K/T\le r<\Delta}(l(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,r)-l(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,0))<0.$$

Now we verify (S3) and (S4). As the stationary density function of $\{X(t)\}$ is continuous and positive at $r_0=0$, for Δ small, $\exists\ 0<m<M<\infty$, such that $mr<Q(r)<Mr$, for $r\in(-\Delta,\Delta)$. Note that $\text{var}(I(0<X(t)<r))=Q(r)(1-Q(r))\le Q(r)(1-mr)$.
Because $f(x)$ is bounded over compact sets, $\exists$ a constant $H>0$, such that
$E(f(X(t))I(r_1<X(t)\le r_2))\le H(Q(r_2)-Q(r_1))$
$\text{var}((X(t))I(r_1<X(t)\le r_2))\le H(Q(r_2)-Q(r_1))$
Define:

$$\begin{aligned}
Q_T(r) &= \int_0^T \frac{I(0<X(t)\le r)}{T}dt;\\
R_T(r) &= \int_0^T \frac{f(X(t))I(0<X(t)\le r)}{T}dt;\\
R_T(r_1,r_2) &= \int_0^T \frac{f(X(t))I(r_1<X(t)\le r_2)}{T}dt;
\end{aligned}$$

Then $\exists$ a constant $H > 0$ such that

$$\mathrm{var}(TQ_T(r)) = \int_0^T \int_0^T E((I(0 < X(t) \leq r) - Q(r))(I(0 < X(s) \leq r) - Q(r)))dt \leq THQ(r),$$

$$\mathrm{var}(TR_T(r)) \leq THQ(r),$$

$$\mathrm{var}(TR_T(r_1, r_2)) \leq TH\{Q(r_2) - Q(r_1)\}$$

as the the process $f(X(t))I(r_1 < X(t) < r_2)$ is ρ-mixing and $f(\cdot)$ is bounded over compact sets. Then (S3) and (S4) can be verified using the above results and an argument in Chan (1993, Proposition 1).

It remains to verify (S5), which can be similarly proved by noting that, for $j = 0, 1$, there exists a constant H such that $\mathrm{var}(\int_0^T X^j(t)I(r_1 < X(t) \leq r_2)\sigma(X(t))dW(t)) \leq TH(Q(r_2) - Q(r_1))$ for some finite constant $H > 0$, thanks to the Burkholder-Davis-Gundy inequality.

8.7. Proof of Lemma 4

Recall that $l(\boldsymbol{\theta})$ is the quasi-likelihood function of $\boldsymbol{\theta} = (\boldsymbol{\delta}^\top, r)^\top$ where $\boldsymbol{\delta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)^\top$. $l(., r)$ is maximized globally at $\hat{\boldsymbol{\delta}}_r = (\hat{\boldsymbol{\beta}}_{1,r}, \hat{\boldsymbol{\beta}}_{2,r})$. We aim to show that $l(\boldsymbol{\delta}, r)$ attains maximum at $|\hat{\boldsymbol{\delta}}_r - \hat{\boldsymbol{\delta}}_{r_0}| = o_p(1/\sqrt{T})$ for $|r - r_0| \leq K/T$.

By the consistency result, the parameter space can be restricted to a neighborhood of $\boldsymbol{\theta}_0$, say $E = \{|\boldsymbol{\beta}_i - \boldsymbol{\beta}_{i,0}| < 1, |r - r_0| < 1, i = 1, 2\}$.
First, consider the case $r > r_0$. Then,

$$\begin{aligned}
& l(\boldsymbol{\delta}, r) - l(\hat{\boldsymbol{\delta}}_{r_0}, r) \\
= & -\frac{1}{2}\int_0^T [((\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{1,0})^\top Y(t))^2 - ((\hat{\boldsymbol{\beta}}_{1,0} - \boldsymbol{\beta}_{1,0})^\top Y(t))^2] I(X(t) \leq r_0)dt \\
- & \frac{1}{2}\int_0^T [((\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{2,0})^\top Y(t))^2 - ((\hat{\boldsymbol{\beta}}_{1,0} - \boldsymbol{\beta}_{2,0})^\top Y(t))^2] I(r_0 < X(t) \leq r)dt \\
- & \frac{1}{2}\int_0^T [((\boldsymbol{\beta}_2 - \boldsymbol{\beta}_{2,0})^\top Y(t))^2 - ((\hat{\boldsymbol{\beta}}_{2,0} - \boldsymbol{\beta}_{2,0})^\top Y(t))^2] I(X(t) > r)dt \\
+ & \int_0^T (\boldsymbol{\beta}_1^\top Y(t) - \hat{\boldsymbol{\beta}}_{1,0}^\top Y(t)) I(X(t) \leq r_0)\sigma(X(t))dW(t) \\
+ & \int_0^T (\boldsymbol{\beta}_1^\top Y(t) - \hat{\boldsymbol{\beta}}_{2,0}^\top Y(t)) I(r_0 < X(t) \leq r)\sigma(X(t))dW(t) \\
+ & \int_0^T (\boldsymbol{\beta}_2^\top Y(t) - \hat{\boldsymbol{\beta}}_{2,0}^\top Y(t)) I(X(t) > r)\sigma(X(t))dW(t)
\end{aligned}$$

Because $l(\delta, r)$ is a quadratic function of $\boldsymbol{\delta}$,

$$\begin{aligned} & l(\boldsymbol{\delta}, r) - l(\hat{\boldsymbol{\delta}}_{r_0}, r) \\ = \ & (\boldsymbol{\delta}_r - \hat{\boldsymbol{\delta}}_{r_0})^\top \dot{l}(\hat{\boldsymbol{\delta}}_{r_0}, r) + (1/2)(\boldsymbol{\delta}_r - \hat{\boldsymbol{\delta}}_{r_0})^\top \ddot{l}(\hat{\boldsymbol{\delta}}_{r_0}, r)(\boldsymbol{\delta}_r - \hat{\boldsymbol{\delta}}_{r_0}) \end{aligned} \tag{S6}$$

where $\dot{l}(\boldsymbol{\delta}, r)$ and $\ddot{l}(\boldsymbol{\delta}, r)$ are the first and second partial derivative of l w.r.t. δ. $\dot{l}(\boldsymbol{\delta}, r_0)$ would be 0 at $\boldsymbol{\delta} = \hat{\boldsymbol{\delta}}_{r_0}$, so:

$$\dot{l}(\hat{\boldsymbol{\delta}}_{r_0}, r) = \dot{l}(\hat{\boldsymbol{\delta}}_{r_0}, r) - \dot{l}(\hat{\boldsymbol{\delta}}_{r_0}, r_0) = \begin{pmatrix} -\int_0^T (\hat{\boldsymbol{\beta}}_{1,r_0} - \boldsymbol{\beta}_{2,0})^\top Y(t)Y(t)I(r_0 < X(t) \le r)dt \\ -\int_0^T (\hat{\boldsymbol{\beta}}_{2,r_0} - \boldsymbol{\beta}_{2,0})^\top Y(t)Y(t)I(r_0 < X(t) \le r)dt \end{pmatrix}$$

Note that, for $r_0 < r \le r_0 + K/T$, $\dot{l}(\hat{\boldsymbol{\delta}}_{r_0}, r)$ is bounded in magnitude by $2\int_0^T (1 + X^2(t))I(|X(t) - r_0| \le K/T)dt$ whose expectation is $O(1)$. This bound can be similarly shown to hold for the case when $r_0 \ge r \ge r_0 - K/T$. Thus, $|\dot{l}(\hat{\boldsymbol{\delta}}_{r_0}, r)| = O_p(1)$, uniformly for $|r - r_0| \le K/T$. On the other hand, $\ddot{l}(\hat{\boldsymbol{\delta}}_{r_0}, r) = [\ddot{l}(\hat{\boldsymbol{\delta}}_{r_0}, r) - \ddot{l}(\hat{\boldsymbol{\delta}}_{r_0}, r_0)] + \ddot{l}(\hat{\boldsymbol{\delta}}_{r_0}, r_0)$, where the first term is an $O_p(1)$ term using similar reasoning as above. Denote $I_1(t, r_0) = I(X(t) \le r_0), I_2(t, r_0) = 1 - I_1(t, r_0)$,

$$A_i(T) = \begin{pmatrix} -\int_0^T I_i(t, r_0)dt & -\int_0^T X(t)I_i(t, r_0)dt \\ -\int_0^T X(t)I_i(t, r_0)dt & -\int_0^T X(t)^2 I_i(t, r_0)dt \end{pmatrix}, i = 1, 2.$$

The second derivative $\ddot{l}(\hat{\boldsymbol{\delta}}_{r_0}, r_0)$ is a block diagonal matrix consisting of $A_1(T)$ and $A_2(T)$ as the first and second block diagonal matrices, respectively. By ergodicity, $\frac{1}{T}A_i(T) \to -E\{(I_i(t, r_0), X(t)I_i(t, r_0))^\top (I_i(t, r_0), X(t)I_i(t, r_0))\}$, which are negative definite and hence their eigenvalues are less than $-\lambda$ for some $\lambda > 0$. So $\ddot{l}(\hat{\boldsymbol{\delta}}, r_0) \le -T(2\lambda - o_p(1)) \times I_4$ where I_4 is a 4×4 identity matrix. Plugging in these inequalities back to $l(\boldsymbol{\delta}, r) - l(\hat{\boldsymbol{\delta}}_{r_0}, r)$, then $\forall K > 0$, $|r - r_0| < K/T$, and for $\boldsymbol{\delta}$ on the boundary of the open neighborhood of radius $a_T = O(1/T^\gamma), 1 > \gamma > 1/2$ centered at $\hat{\boldsymbol{\delta}}_{r_0}$,

$$l(\boldsymbol{\delta}, r) - l(\hat{\boldsymbol{\delta}}_{r_0}, r) < a_T \times \{O(1) - (\lambda + o_p(1))Ta_T/2\} < 0,$$

with probability approaching 1 as $T \to \infty$. Thus, $l(\boldsymbol{\delta}, r)$ would only attain its global maximum within the $o_p(1/\sqrt{T})$ neighborhood of $\hat{\boldsymbol{\delta}}_{r_0}$, which completes the proof.

8.8. Proof of Lemma 5

We shall only give the proof for the case $\kappa > 0$ as the case $\kappa \leq 0$ is similar. Let M_β be as defined in the beginning of the proof of Theorem 3.

$$\begin{aligned}
& \tilde{l}(\kappa) \\
= \; & l(\hat{\delta}_{r_0+\kappa/T}, r_0 + \kappa/T) - l(\hat{\boldsymbol{\delta}}_{r_0}, r_0 + \kappa/T) + l(\hat{\boldsymbol{\delta}}_{r_0}, r_0 + \kappa/T) - l(\hat{\boldsymbol{\delta}}, r_0) \\
= \; & o_p(1) + \frac{1}{2}\int_0^T (M_{\hat{\boldsymbol{\beta}}_{1,r_0}}(X(t)) - M_{\hat{\boldsymbol{\beta}}_{2,r_0}}(X(t)))I(0 < X(t) \leq \kappa/T)dt \\
+ \; & O_p(1/\sqrt{T}),
\end{aligned} \tag{S7}$$

where the terms $o_p(1)$ and $O_p(1/\sqrt{T})$ hold uniformly for $|\kappa| < K$, by making use of (S6) and Lemma 4and employing arguments similar to those employed in the proof of Lemma 4.On the other hand,

$$\begin{aligned}
& \int_0^T (M_{\hat{\boldsymbol{\beta}}_{1,r_0}}(X(t)) - M_{\hat{\boldsymbol{\beta}}_{2,r_0}}(X(t)))I(0 < X(t) \leq \kappa/T)dt \\
= \; & \int_0^T (M_{\boldsymbol{\beta}_{1,0}}(X(t)) - M_{\boldsymbol{\beta}_{2,0}}(X(t)))I(r_0 < X(t) \leq r_0 + \kappa/T)dt \\
+ \; & \int_0^T (M_{\hat{\boldsymbol{\beta}}_{1,r_0}}(X(t)) - M_{\boldsymbol{\beta}_{1,r_0}}(X(t)))I(r_0 < X(t) \leq r_0 + \kappa/T)dt \\
+ \; & \int_0^T (M_{\hat{\boldsymbol{\beta}}_{2,r_0}}(X(t)) - M_{\boldsymbol{\beta}_{2,r_0}}(X(t)))I(r_0 < X(t) \leq r_0 + \kappa/T)dt
\end{aligned}$$

where the last two terms can easily be shown to be $o_p(1)$ terms, owing to Lemma 4 and using similar argument as we did in the proof the Lemma 4. This completes the proof.

8.9. Proof of Theorem 5

Given r_0, the quasi-likelihood function $l(\boldsymbol{\delta}, r_0)$ is continuous and twice differentiable in $\boldsymbol{\delta}$, so an application of Van der Vaart (2000, Theorem 5.21) implies that $\sqrt{T}(\hat{\boldsymbol{\delta}} - \delta_0)$ is asymptotically normal with zero mean and covariance matrix $V(\boldsymbol{\delta}_0) = E(\ddot{l}_{\boldsymbol{\theta}_0}))^{-1}E(\dot{l}_{\boldsymbol{\theta}_0}\dot{l}_{\boldsymbol{\theta}_0}^\top)((E(\ddot{l}_{\boldsymbol{\theta}_0}))^{-1})^\top$. Because $\hat{r}$ is T-consistent and $\hat{\boldsymbol{\beta}}_i = \hat{\boldsymbol{\beta}}_{i,r_0} + o_p(1/\sqrt{T})$, hence, $\hat{\boldsymbol{\beta}}_i$ and $\hat{\boldsymbol{\beta}}_{i,r_0}$ follows the same asymptotic distribution. As a special case, when we have a constant σ, the asymptotic distribution coincides with the covariance matrix of the maximum likelihood estimators.