A Re-examination of Stanton's Diffusion Estimation with Applications to Financial Model Validation

by

Jianqing Fan
Department of Statistics
University of North Carolina
Chapel Hill, NC 27599
jfan@stat.unc.edu

Chunming Zhang
Department of Statistics
University of Wisconsin
Madison, WI 53706
cmzhang@stat.wisc.edu
A Re-examination of Stanton’s Diffusion Estimations with Applications to Financial Model Validation

Jianqing Fan * Chunming Zhang †

Abstract

Stanton (1997) proposes higher-order approximation formulas for nonparametric drift and diffusion estimators and concludes nonlinearities of the short-term interest rate drift function. Chapman and Pearson (2000) suggest that boundary effect of kernel estimators produce spurious nonlinearity. In this article, we show that the asymptotic variances of Stanton’s estimators escalate nearly exponentially with the order of approximation, under the usual stationary Markovian assumptions. Furthermore, we propose a powerful “generalized likelihood ratio test” combined with Fan and Yao’s estimator to conclude weak evidence against the linear drift of the short rates, while very strong evidence against popular models for the volatility.

Consider the problem of estimating the drift function \( \mu(\cdot) \) and diffusion function \( \sigma(\cdot) \) for continuous-time diffusion process \( \{X_t, t \geq 0\} \) following a stochastic differential equation:

\[
dX_t = \mu(X_t)dt + \sigma(X_t)dW_t,
\]

where \( \{W_t, t \geq 0\} \) is a standard one-dimensional Brownian motion. By “time-homogeneous”, we mean that \( \mu(\cdot) \) and \( \sigma(\cdot) \) are only functions of contemporaneous state value of \( X_t \). This simplified model (1) has been widely employed for describing the stochastic dynamics of the underlying economic variables, including many well-known single-factor models. Examples include the geometric Brownian motion

\[
\text{GBM} : \quad dX_t = \mu X_t dt + \sigma X_t dW_t 
\]

by Osborne (1959) for modeling stock price, and

\[
\begin{align*}
\text{VAS} : & \quad dX_t = (\alpha_0 + \alpha_1 X_t) dt + \sigma dW_t, \\
\text{CIR SR} : & \quad dX_t = (\alpha_0 + \alpha_1 X_t) dt + \sigma \sqrt{X_t} dW_t, \\
\text{CIR VR} : & \quad dX_t = \sigma \sqrt{X_t}^3 dW_t, \\
\text{CKLS} : & \quad dX_t = (\alpha_0 + \alpha_1 X_t) dt + \sigma X_t^\gamma dW_t,
\end{align*}
\]

*Jianqing Fan is Professor, Department of Statistics, University of North Carolina at Chapel Hill and Professor of Statistics, Department of Statistics, Chinese University of Hong Kong. His research was partially supported by the NSF grant DMS-9803200 and the RGC grant CUHK4299/00P from the Hong Kong SAR.

†Correspondence author. Chunming Zhang is Assistant Professor, Department of Statistics, University of Wisconsin at Madison, WI 53706-1685.
by Vasicek (1977), Cox, Ingersoll, and Ross (1985), Cox, Ingersoll, and Ross (1980), and Chan, Karolyi, Longstaff and Sanders (1992) for modeling interest rate dynamics, respectively.

Current researches including parametric approaches to estimate $\mu(\cdot)$ and $\sigma(\cdot)$ have been surveyed by Stanton (1997). To relax model assumptions and to reduce possible modeling biases, nonparametric techniques have been recently utilized in many areas as an exploratory tool for data analysis. Pham (1981) and Prakasa Rao (1985) proposed nonparametric drift estimators. Arfi (1995, 1997) showed that the Nadaraya-Watson kernel estimator of drift is uniformly strongly consistent under ergodic conditions, and obtained same conclusion for the kernel regression estimate of diffusion. Fan and Yao (1998) employed local linear regression estimation to the squared residuals for estimating $\sigma^2(\cdot)$ and showed the proposed approach is efficient. Using infinitesimal generator and Taylor series expansion, Stanton (1997) constructed first, second and third-order approximation formulas for estimating $\mu(\cdot)$ and $\sigma(\cdot)$. These formulas contain unknown conditional expectations which are estimated by the N-W kernel estimator. Stanton’s approach to estimating diffusion function $\sigma(\cdot)$ ignores the drift function $\mu(\cdot)$. This makes his method simple and attractive.

Stanton’s approach encounters both exogenous and endogenous problems. Recently, Chapman and Pearson (2000) studied the finite sample properties of Stanton’s estimator based on the first-order approximation. By applying this procedure to simulated sample paths of a square-root diffusion, they found that Stanton’s estimator produces spurious nonlinearity when the true underlying drift function is indeed linear. As the authors have already nicely concluded, this is Stanton’s first-order approximation couples with the corresponding kernel regression estimator. The latter’s problematic “boundary effects” translate to the former. As a result, artificial pattern of nonlinearity displays noticeably near the boundary regions. Meanwhile, two sensible questions arise naturally: do higher-order approximations outperform lower-order counterparts and are there any reasonable formal procedures in testing whether the nonlinearity in the drift is real or artificial?

In an attempt to answer the first question on the order of approximations, we derive explicitly the formulae for higher order approximations that generalize the idea of Stanton’s. We then compute explicitly the asymptotic variances of nonparametric estimators based on higher order approximations. A striking feature of our study is that higher order approximations will reduce the numerical approximation errors in asymptotic biases, but escalate nearly exponentially the asymptotic variance. This variance inflation phenomenon is not only an artifact of nonparametric fitting. It also applies to parametric models. Therefore, little benefit can be obtained by employing higher order approximations when data are collected with reasonable frequencies. Our conclusion is somewhat at odds with Stanton’s advocacy of using higher order approximations.

Interestingly under a completely distinct scenario, Bai, Russel and Tiao (1999) examined the precision of volatility estimation using financial data sampled at increasingly higher frequencies. While Merton (1980)’s seminal work on ideal continuous-time model of geometric Brownian motion suggested that volatility estimates can be made arbitrarily precise provided that the sampling interval shrinks to zero, they found that with practical discrete-time model large amounts of high frequency
data do not necessarily translate into precise estimates of the volatility. This results from the fact that higher sampling frequency increases temporal dependence as well as normality departure of the data structure, and therefore reduces estimation efficiency. The finding has some spirit in common with our study in parameter estimation. Unlike our analytical results, they have not yet derived asymptotic relationships between estimation accuracy and sampling frequency.

Stanton’s work raises other interesting aspects. Is the drift function in the short-term rate models nonlinear? Or more generally, does a parametric model fit a given economic or financial data? An example of this is whether models (2) through (6) fit adequately short-term rate data. Chapman and Pearson (2000) suggested that the nonlinearity of the drift function might be spurious. His method is based on simulated data with linear drift function and inspects whether the estimated drift looks linear or not. This graphical procedure is informal, but useful. To produce formal statistical tests, an alternative hypothesis (model) is needed. Since we usually don’t have strong preference for alternative competing models, the nonparametric model (1) serves as a natural candidate. The hypothesis testing problem becomes testing a parametric (or semiparametric) null hypothesis against a nonparametric alternative. We will extend the idea of the generalized likelihood ratio (GLR) statistics developed by Fan, Zhang and Zhang (2000) and apply it to the time-homogeneous diffusion model. Our simulation result shows that our tests are indeed powerful and give correct size of the tests. They provide useful tools for testing various models in economics and finance.

The remainder of this paper is organized as follows. In Section I, we discuss distributional properties of Stanton’s estimators of drift and diffusion functions and also derive explicit expressions of asymptotic bias and variance for higher-order approximation formulas. To justify on empirical grounds of our analyses, simulations are conducted in Section II. Section III proposes model validation methods based on the GLR test together with Fan and Yao’s estimation approach, i.e. first-order approximation combined with local linear estimation. Power simulations of GLR test and real data applications are demonstrated. Section IV briefly concludes the paper. Outlines of the proofs are collected in the Appendix.

I Discussion of Stanton’s Method

A Calculations of Conditional Means and Conditional Variances

This section begins with a description of Stanton’s approach. Although his initial construction is based solely upon the first, second and third order approximations, we can build, with some extra efforts, a more general framework that allows us to examine the impacts of higher-order approximations.

Following Stanton’s notation, under appropriate conditions on $\mu(\cdot)$ and $\sigma(\cdot)$, and an arbitrary function $f(\cdot, \cdot)$ (Hille and Phillips (1957), Chapter 11), the conditional expectation $E_t\{f(X_{t+\Delta}, t+\Delta)\}$
can be expressed in forms of Taylor series expansion,

\[ E_t\{ f(X_{t+\Delta}, t+\Delta) \} = f(X_t, t) + \mathcal{L} f(X_t, t) \Delta + \frac{1}{2} \mathcal{L}^2 f(X_t, t) \Delta^2 + \cdots + \frac{1}{n!} \mathcal{L}^n f(X_t, t) \Delta^n + O(\Delta^{n+1}), \]

where \( E_t \) denotes the conditional expectation given \( X_t \) and the infinitesimal generator \( \mathcal{L} \) (see Oksendal 1985) of the process \( \{X_t\} \) is defined as

\[
\mathcal{L} f(x, t) = \lim_{\tau \to t} \frac{E \{ f(X_\tau, \tau) | X_t = x \} - f(x, t)}{\tau - t} = \frac{\partial f(x, t)}{\partial t} + \frac{\partial f(x, t)}{\partial x} \mu(x) + \frac{1}{2} \frac{\partial^2 f(x, t)}{\partial^2 x} \sigma^2(x). \tag{7}
\]

Thus, the first-order approximation formula for \( \mathcal{L} f(X_t, t) \) is

\[
\Delta^{-1} E_t\{ f(X_{t+\Delta}, t+\Delta) - f(X_t, t) \} = \mathcal{L} f(X_t, t) + \frac{1}{2} \mathcal{L}^2 f(X_t, t) \Delta + \cdots + \frac{1}{n!} \mathcal{L}^n f(X_t, t) \Delta^{n-1} + O(\Delta^n). \tag{8}
\]

By taking \( f(x, t) = x \) (or \( f(x, t) = x - X_t \)), we obtain \( \mathcal{L} f(x, t) = \mu(x) \); likewise, \( f(x, t) = (x - X_t)^2 \) implies \( \mathcal{L} f(x, t) = 2(x - X_t) \mu(x) + \sigma^2(x) \) which equals to \( \sigma^2(X_t) \) at \( x = X_t \). Thus these two special functions \( f(\cdot, \cdot) \) exactly recover \( \mu(X_t) \) and \( \sigma^2(X_t) \). Estimating the left-hand side of equation (8) by the N-W kernel method leads to Stanton’s estimators for \( \mu(x) \) and \( \sigma^2(x) \) based on the first-order approximation.

Higher-order approximations can be achieved through linear combinations on the left-hand side of equation (8). More precisely, for any fixed integer \( k \geq 1 \), any sequence of constants \( \{a_{k,j}\}_{j=1}^{k} \), and any discretely observed time steps \( j \Delta, j = 1, \cdots, k \), we shall consider the following linear combination:

\[
\Delta^{-1} \sum_{j=1}^{k} a_{k,j} E_t\{ f(X_{t+j\Delta}, t+j\Delta) - f(X_t, t) \}
\]

\[
= \left\{ \sum_{j=1}^{k} ja_{k,j} \right\} \mathcal{L} f(X_t, t) + \left\{ \sum_{j=1}^{k} j^2 a_{k,j} \right\} \frac{\mathcal{L}^2 f(X_t, t)}{2} \Delta + \cdots + \left\{ \sum_{j=1}^{k} j^k a_{k,j} \right\} \frac{\mathcal{L}^k f(X_t, t)}{k!} \Delta^{k-1} + \left\{ \sum_{j=1}^{k} j^{k+1} a_{k,j} \right\} \frac{\mathcal{L}^{k+1} f(X_t, t)}{(k+1)!} \Delta^k + O(\Delta^{k+1}).
\]

Clearly, a \( k \)-th order approximation

\[
\Delta^{-1} \sum_{j=1}^{k} a_{k,j} E_t\{ f(X_{t+j\Delta}, t+j\Delta) - f(X_t, t) \} = \mathcal{L} f(X_t, t) + O(\Delta^k)
\]

is achieved by choosing coefficients \( \{a_{k,j}\}_{j=1}^{k} \) to satisfy the system of equations

\[
\begin{cases}
\sum_{j=1}^{k} j a_{k,j} = 1, \\
\sum_{j=1}^{k} j^2 a_{k,j} = 0, \\
\vdots \\
\sum_{j=1}^{k} j^k a_{k,j} = 0.
\end{cases} \tag{9}
\]
The general form of solution \( \{a_{k,j}\}_{j=1}^{k} \) is presented in Theorem 1 below, whose proof is given in the Appendix. Apparently, when \( k = 1, k = 2, \) and \( k = 3, \) values of \( \{a_{k,j}\}_{j=1}^{k} \) coincide with those derived in Stanton (1997).

**Theorem 1** For each fixed integer \( k \geq 1, \) the unique solution to equation (9) is given by:

\[
a_{k,j} = (-1)^{j+1} \binom{k}{j} / j, \quad j = 1, \ldots, k. \tag{10}
\]

Further, with this choice of \( a_{k,j}, \) we have

\[
\sum_{j=1}^{k} j^{k+1} a_{k,j} = (-1)^{k+1} k!. \]

Therefore with the above unique solution \( \alpha_{k} = [a_{k,1}, \ldots, a_{k,k}]^{T}, \) we obtain the general \( k \)-th order approximation formula for \( \mathcal{L}f(X_{t}, t) : \)

\[
\Delta^{-1} \sum_{j=1}^{k} a_{k,j} E_{t}\{f(X_{t+j\Delta}, t + j\Delta) - f(X_{t}, t)}\}, \tag{11}
\]

with the error term

\[
(-1)^{k+1} \frac{\mathcal{L}^{k+1} f(X_{t}, t)}{(k+1)} \Delta^{k} + O(\Delta^{k+1}).
\]

By taking \( f_{1}(x, t) = x, \) equation (11) implies that

\[
\Delta^{-1} \sum_{j=1}^{k} a_{k,j} E_{t}(X_{t+j\Delta} - X_{t}) = \mu(X_{t}) + \left[(-1)^{k+1} \frac{\mathcal{L}^{k+1} f_{1}(X_{t}, t)}{(k+1)} \Delta^{k} + O(\Delta^{k+1})\right]. \tag{12}
\]

Similarly, by taking \( f_{2}(x, t) = (x - X_{t})^{2}, \) equation (11) yields that

\[
\Delta^{-1} \sum_{j=1}^{k} a_{k,j} E_{t}(X_{t+j\Delta} - X_{t})^{2} = \sigma^{2}(X_{t}) + \left[(-1)^{k+1} \frac{\mathcal{L}^{k+1} f_{2}(X_{t}, t)}{(k+1)} \Delta^{k} + O(\Delta^{k+1})\right]. \tag{13}
\]

Consequently from equation (13), taking square-root operation yields the same \( k \)-th order approximation formula to diffusion function \( \sigma(X_{t}), \) i.e.

\[
\sigma(X_{t}) = \sqrt{\Delta^{-1} \sum_{j=1}^{k} a_{k,j} E_{t}(X_{t+j\Delta} - X_{t})^{2} + O(\Delta^{k})}. \tag{14}
\]

In addition, for both choices of \( f_{\ell}(x, t), \ \ell = 1, 2, \) the term \( \mathcal{L}^{k+1} f_{\ell}(X_{t}, t) \) does not involve variable \( t \) simply because of equation (7). Therefore, the resulting numerical approximation errors for \( \mu(\cdot), \) \( \sigma^{2}(\cdot) \) and \( \sigma(\cdot) \) are all of order \( O(\Delta^{k}) \) for an arbitrary order \( k. \)

With higher-order approximation formulas to \( \mu(\cdot) \) and \( \sigma(\cdot), \) it still remains to estimate the involved conditional expectations. Given the initial calendar time point \( t_{0} \) and a series of data
\{X_{t+\Delta}, i = 1, \cdots, n\} observed at equally spaced sampling intervals of length \(\Delta > 0\), our first step is to form \((n - k)\) pairs of synthetic data

\[
\left( X_{t+\Delta}^*, \Delta^{-1} \sum_{j=1}^{k} a_{i,j} [X_{t+(i+j)\Delta} - X_{t+\Delta}] \right) =: (X_{i\Delta}^*, Y_{i\Delta}^*), \quad i = 1, \cdots, n - k,
\]

(15)

together with

\[
\left( X_{t+\Delta}^*, \Delta^{-1} \sum_{j=1}^{k} a_{i,j} [X_{t+(i+j)\Delta} - X_{t+\Delta}]^2 \right) =: (X_{i\Delta}^*, Z_{i\Delta}^*), \quad i = 1, \cdots, n - k.
\]

(16)

It follows immediately that both \(\{(X_{i\Delta}^*, Y_{i\Delta}^*)\}_{i=1}^{n-k}\) and \(\{(X_{i\Delta}^*, Z_{i\Delta}^*)\}_{i=1}^{n-k}\) are bivariate strictly stationary processes provided that \(\{X_t, t \geq 0\}\) is. Our second step is to perform the pointwise nonparametric regression estimators \(\hat{\mu}_{1,\Delta}(x_0)\) and \(\hat{\mu}_{2,\Delta}(x_0)\) for conditional expectations

\[
E(Y_{i\Delta}^*|X_{i\Delta}^* = x_0) = \mu(x_0) + O(\Delta^k), \quad \text{and} \quad E(Z_{i\Delta}^*|X_{i\Delta}^* = x_0) = \sigma^2(x_0) + O(\Delta^k),
\]

(17)

[from equations (12) and (13)] respectively.

There are many nonparametric approaches to estimating conditional means in equation (17). The N-W estimator is of the simplest form. It can be improved by local polynomial techniques (Fan and Gijbels, 1996). Therefore our subsequent analytical discussions are concentrated on \(\hat{\mu}_{1,\Delta}(x_0)\) and \(\hat{\mu}_{2,\Delta}(x_0)\) via q-th order local polynomial estimates. The N-W estimator corresponds to the local constant model with order \(q = 0\). We now briefly describe the technique. For \(x\) in a neighborhood of \(x_0\), by Taylor series expansion, any smooth function \(\mu(x)\) can be locally approximated by a q-th order polynomial, i.e.

\[
\mu(x) \approx \mu(x_0) + (x - x_0)\mu'(x_0) + \cdots + \frac{1}{q!}(x - x_0)^q\mu^{(q)}(x_0).
\]

Denote the vector \(\beta(x_0) = [\mu(x_0), \mu'(x_0), \cdots, \mu^{(q)}(x_0)/q!]^T = [\beta_1, \cdots, \beta_{q+1}]^T\), then the q-th order local polynomial estimator \(\hat{\beta}(x_0)\) is the minimizer of the residual sums of squares between \(Y_{i\Delta}^*\) and the local model on \(\mu(X_{i\Delta}^*)\), weighted by the distance of \(X_{i\Delta}^*\) from the target point \(x_0\). Formally, \(\hat{\beta}(x_0)\) minimizes

\[
\sum_{i=1}^{n-k}[Y_{i\Delta}^* - \beta_1 - (X_{i\Delta}^* - x_0)\beta_2 - \cdots - (X_{i\Delta}^* - x_0)^q\beta_{q+1}]^2 K_h(X_{i\Delta}^* - x_0),
\]

where \(K_h(\cdot) = K(\cdot/h)/h\). Here \(K(\cdot)\) and \(h\) are referred to as a kernel function and a bandwidth (or smoothing parameter), respectively. The first component of \(\hat{\beta}(x_0)\) is the q-th order local polynomial estimator \(\hat{\mu}_{1,\Delta}(x_0)\) of \(\mu(x_0)\). For practical applications, Fan and Gijbels (1996) recommended using the local linear fit (\(q = 1\)). A similar procedure applies to the local polynomial estimator \(\hat{\mu}_{2,\Delta}(x_0)\) of \(\sigma^2(x_0)\).

Choices of kernel functions purely depend upon individual preferences. Throughout the rest of this paper, we use Epanechnikov kernel function defined by

\[
K(u) = \frac{3}{4}(1 - u^2)I_{\{|u| \leq 1\}}.
\]

(18)
For the given kernel function, the bandwidth parameter $h$ is very important to the performance of a nonparametric regression estimator. It is often chosen via, either visual inspection or a data-driven rule. Popular data-driven approaches include cross-validation (Allen, 1974, and Stone, 1974), generalized cross-validation (Wahba, 1977), the pre-asymptotic substitution method (Fan and Gijbels, 1995), the plug-in method (Ruppert, Wand, and Sheather, 1995) and empirical bias method (Ruppert, 1997). These methods provide various useful techniques for automatic bandwidth selection, but involve intensive computation and require efforts to program. Further discussions regarding theoretical properties and implementations, including automatic selection of bandwidth, can be found in Fan and Gijbels (1996).

Since any nonparametric procedures are in essence a weighted average of local data, their performance depends always on the local variation, namely the conditional variance. For our current applications based on the synthetic data, the corresponding conditional variances at any value $x_0$ are

$$
\sigma^2_{1,\Delta}(x_0) = \text{Var}(Y_{i_1}^* | X_{i_1}^* = x_0), \quad \text{and} \quad \sigma^2_{2,\Delta}(x_0) = \text{Var}(Z_{i_2}^* | X_{i_2}^* = x_0). \tag{19}
$$

Their orders of magnitude can be summarized as follows, whose proof can be found in the Appendix.

**Theorem 2** Let $A_{1,k}$ and $A_{2,k}$ be $k \times k$ matrices with $(i, j)$-th entry equal to $\min(i, j)$ and $\min(i^2, j^2)$, respectively. Assume that $\{X_i\}$ is a stationary Markov process. Then as $\Delta \to 0$, the conditional variance of $\hat{\mu}_{1,\Delta}(x_0)$ for estimating the drift is given by

$$
\sigma^2_{1,\Delta}(x_0) = \sigma^2(x_0) V_1(k) \Delta^{-1} \{1 + O(\Delta)\}, \tag{20}
$$

and the conditional variance of $\hat{\mu}_{2,\Delta}(x_0)$ for estimating the squared volatility is given by

$$
\sigma^2_{2,\Delta}(x_0) = 2\sigma^4(x_0) V_2(k) \{1 + O(\Delta)\}, \tag{21}
$$

where $V_1(k) = \alpha_k^TA_{1,k}\alpha_k$, $V_2(k) = \alpha_k^TA_{2,k}\alpha_k$, and $\alpha_k$ is a $k \times 1$ vector whose $j$-th element is given by equation (10).

Put Table I about here.

The factors $V_1(k)$ and $V_2(k)$ reflect the premium that higher order approximations have to pay. For this reason, we call them the variance inflation factors by using higher order approximations. To give us some numerical impression, Table I summarizes the numerical values of $V_1(k)$ and $V_2(k)$ for the first 10 orders of approximations. For visual assessment, we also present in Figure 1 the plots of $\log\{V_1(k)\}$ and $\log\{V_2(k)\}$ with respect to order $k$. The overall impacts of higher-order approximations on variance inflation are striking.

Table I and Figure 1 also suggest that the variance inflation factors grow exponentially fast as $k$ increases. This can indeed be verified analytically, as shown in the following theorem.
\textbf{Theorem 3} \hspace{0.2em} (a) $V_1(k)$ is between
\[
\frac{k^2 - 3k - 2}{k(k + 1)^3} \binom{2k}{k} + \frac{2}{k} + 2 \sum_{j=1}^{k} \frac{1}{j} \cdot \frac{2k^2 + 4k + 3}{(k + 1)^2} \approx \frac{4^k}{\pi^{1/2}k^{5/2}}
\]
and
\[
\frac{5k^2 - k - 2}{k(k + 1)^3} \binom{2k}{k} + \frac{2}{k} + 2 \sum_{j=1}^{k} \frac{1}{j} \cdot \frac{3k^2 + 6k + 5}{(k + 1)^2} \approx \frac{5 \times 4^k}{\pi^{1/2}k^{5/2}}.
\]

(b) For $k > 1$,
\[
V_2(k) = \frac{\binom{2k}{k} - (k + 1)}{k - 1} \approx \frac{4^k}{\pi^{1/2}k^{3/2}}.
\]

\section{Asymptotic Distribution}

The asymptotic normal distributions of the pointwise drift estimator $\hat{\mu}_{1,\Delta}(x_0)$ and squared diffusion estimator $\hat{\sigma}_{2,\Delta}(x_0)$ are established in Theorem 4.

\textbf{Theorem 4} Under regularity conditions of Banon (1978) and conditions of Theorem 6.1 in Fan and Gijbels (1996), by using local linear fit at each fixed value $x_0$, as $n \rightarrow \infty$, $h = O(n^{-1/5})$, $\Delta \rightarrow 0$, $n\Delta \rightarrow \infty$ and $h = o(\Delta)$
\[
d_{t,\Delta} \sqrt{n}h \left( \hat{\mu}_{t,\Delta}(x_0) - \mu_{t,\Delta}(x_0) - \frac{\mu_2}{2} h^2 \sigma_{t,\Delta}^2(x_0) \right) \xrightarrow{\mathcal{L}} N \left( 0, \sigma_t^2(x_0) p^{-1}(x_0) \nu_0 \right), \quad t = 1, 2,
\]
where $\mu_2 = \int u^2 K(u) du$, $\nu_0 = \int K^2(u) du$, and
\begin{align*}
d_{t,\Delta} &= \begin{cases} \Delta^{1/2}, & \text{if } t = 1; \\ 1, & \text{if } t = 2, \end{cases} \quad (22) \\
\mu_{t,\Delta}(x_0) &= \begin{cases} \mu(x_0) + O(\Delta^k), & \text{if } \ell = 1; \\ \sigma^2(x_0) + O(\Delta^k), & \text{if } \ell = 2, \end{cases} \quad (23) \\
\sigma_t^2(x_0) &= \lim_{\Delta \to 0} \sigma_{t,\Delta}^2(x_0) d_{t,\Delta}^2 = \begin{cases} \sigma^2(x_0) V_1(k), & \text{if } \ell = 1; \\ 2 \sigma^2(x_0) V_2(k), & \text{if } \ell = 2. \end{cases} \quad (24)
\end{align*}

Similar results hold for higher-order local polynomial regression estimators.

At any time $t = t_0 + i\Delta$ ($i = 1, \cdots, n - k$), equations (12) and (23) indicate that the leading term in the asymptotic bias of $\hat{\mu}_{1,\Delta}(x_0)$ is
\[
(-1)^{k+1} \frac{L^{k+1} f_1(x_0,t)}{(k + 1)} \Delta^k + \frac{\mu_2}{2} h^2 \mu''(x_0),
\]
which is composed of numerical approximation error and nonparametric estimation bias. Comparatively, for N-W kernel regression estimator employed by Stanton, the asymptotic bias of $\hat{\mu}_{1,\Delta}(x_0)$ is
\[
(-1)^{k+1} \frac{L^{k+1} f_1(x_0,t)}{(k + 1)} \Delta^k + \frac{\mu_2}{2} h^2 \left( \mu''(x_0) + 2 \mu'(x_0) \frac{p'(x_0)}{p(x_0)} \right).
\]
Similarly, equations (13) and (23) indicate that the leading term in the asymptotic bias of \( \tilde{\mu}_{2,\Delta}(x_0) \) is
\[
(-1)^{k+1} \frac{L^{k+1} f_2(x_0, t)}{(k + 1)} \Delta^k + \frac{\mu_2}{2} h^2 (\sigma^2)''(x_0). \tag{27}
\]
This should be compared with the asymptotic bias of \( \tilde{\mu}_{2,\Delta}(x_0) \) using the the N-W kernel regression estimator:
\[
(-1)^{k+1} \frac{L^{k+1} f_2(x_0, t)}{(k + 1)} \Delta^k + \frac{\mu_2}{2} h^2 \left( (\sigma^2)''(x_0) + 2(\sigma^2)'(x_0) \frac{p'(x_0)}{p(x_0)} \right). \tag{28}
\]

II Simulations

Realistically, it is unknown whether the stationary Markovian assumption remains valid for financial data recorded at discrete time points. Neither do we know whether the asymptotic results reflect reality. Nevertheless, we could still pursue the drift and diffusion estimations using higher-order approximations and nonparametric techniques. This will enable us to assess empirically how our asymptotic results are reflected in finite samples simulated from the true given diffusion models. Our simulation results confirm the fact that the asymptotic results are reflected in finite sample situations.

A Geometric Brownian Motion

As a first illustration, we consider the familiar example of geometric Brownian motion:
\[
dX_t = (\mu + 2^{-1} \sigma^2) X_t dt + \sigma X_t dW_t, \quad 0 \leq t \leq T. \tag{29}
\]

Apparently by the construction, both the drift and diffusion of the process are linear, but the technical assumption of stationarity is violated. One routine option of simulating sample paths of length \( n \) is utilizing discrete-time analog of model (29) when the sampling time interval \( \Delta = T/n \) is small:
\[
X_{t_{i+\Delta}} - X_{t_i} \approx (\mu + 2^{-1} \sigma^2) X_{t_i} \Delta + \sigma X_{t_i} \sqrt{\Delta} \varepsilon_{t_i}, \quad \text{where} \quad \varepsilon_{t_i} \stackrel{iid}{\sim} N(0, 1). \tag{30}
\]

The initial value \( X_0 \) is set to be 1. This method is simple, but introduces discretization errors. Alternatively, we use directly the analytical solution \( X_t = \exp\{\mu t + \sigma W_t\} \) of the underlying process (29). Within each simulation trial, we compute Stanton's kernel drift estimator \( \tilde{\mu}_{1,\Delta}(x_0) \) and squared diffusion estimator \( \tilde{\mu}_{2,\Delta}(x_0) \) with increasingly higher order \( k \). The simulations are repeated 1000 times, then the sample variances of \( \{\tilde{\mu}_{1,\Delta}(x_0)\} \) and \( \{\tilde{\mu}_{2,\Delta}(x_0)\} \) are calculated across these 1000 simulations respectively.

To facilitate programming and to expedite computation, we adopt the following simple rule of thumb formula
\[
h = \text{constant} \; \tilde{s}_n n^{-1/5}, \tag{31}
\]
where \( \tilde{s}_n \) is the sample standard deviation of design points \( \{X_{i\Delta}^*\}_{i=1}^{n-k} \) defined in equation (15). The attached constant is chosen differently to reflect different set-ups and to ensure sufficient amount of local data are available for conducting local regression. Note that the variance inflation is primarily due to higher-order difference, and relatively less affected by the choices of bandwidth. Thus for purpose of illustration, we set the constant in equation (31) to be 3. We have also experimented with other scales of bandwidth, and obtained similar variance inflation phenomena.

Three parameter values of \( \mu = 0.087, \sigma = 0.178, \) and \( T = 10 \) are specified. Our calculations are focused on point \( x_0 = 1.0, \) simply because more data points fall within its local region. The natural logarithms of the simulated variance ratios of \( \hat{\mu}_{1,\Delta}(1.0) \) and \( \hat{\mu}_{2,\Delta}(1.0) \) against order \( k \) are displayed in Figure 2, where plot (a) is based on samples generated from exact solution, and plot (b) is based on the discretization scheme. Meanwhile for purpose of comparison, we also calculate corresponding results by local linear method in bottom plots (a') and (b'). Except in amplitudes, the tendencies of plots (a), (b), (a') and (b') mimic well our theoretical results presented in Figure 1.

### B CIR Square-Root Diffusion

We include another well-known CIR model for interest rates term structure:

\[
dX_t = \kappa (\theta - X_t) dt + \sigma \sqrt{X_t} \, dW_t, \quad t \geq t_0, \tag{32}
\]

where the spot rate moves around its long-run equilibrium level \( \theta \) at speed \( \kappa \). When the condition \( 2\kappa \theta \geq \sigma^2 \) holds, this process is shown to be positive and stationary. This example differs from the previous one in that the explicit pathwise solution of \( \{X_t\} \) is not readily available. One could always carry out the discretization algorithm to generate sample paths. Alternatively, we use the transitional density properties of the process (Cox, Ingersoll, and Ross, 1985). Given the current interest rate \( X_t = x \) at time \( t \), \( 2eX_s \) is conditionally distributed as noncentral chi-squared with \( 2q+2 \) degrees of freedom and noncentrality parameter \( 2u \), where

\[
q = \frac{2\kappa \theta}{\sigma^2} - 1, \quad u = cxe^{-\kappa (s-t)}, \quad c = \frac{2\kappa}{\sigma^2(1 - e^{-\kappa (s-t)})}.
\]

The initial value of \( X_{t_0} \) can be generated from the steady state Gamma distribution. Here the values of parameters \( \kappa, \theta, \sigma, \Delta \) are cited from Chapman and Pearson (2000), i.e. \( \kappa = 0.21459, \theta = 0.08571, \sigma = 0.07830, \) and \( \Delta = 1/250 \). Simulated sample paths of length 10000 are replicated for 1000 times, with the constant in equation (31) taken to be 6. Again, this number serves for the purposes of illustration. To differentiate effects of higher-order approximation from interaction with “boundary effects” primarily due to data sparsity, we focus on the interior point \( x_0 = 0.1 \). Figure 3 displays the natural logarithms of the simulated variance inflations. For comparison, plots (a) and (a') are based on data generated from the conditional chi-squared distribution, and plots (b) and (b') are based on the discretization scheme. Again, all plots support our theoretical results in Figure 1.
C Local Linear Fit: Boundary Correction

Overall, the previous simulation studies provide convincing evidences that at least for models similar to those two, Stanton’s higher-order approximations amplify variances substantially. As mentioned before, this phenomenon is independent of whichever nonparametric estimation methods have been used. Chapman and Pearson (2000) reported that based on kernel regression estimators, Stanton’s approach produces abnormal behaviours at boundary regions, which result in potentially misleading interpretations of the true drift and diffusion functions. Compared with kernel estimators in estimating regression curves, local linear estimators enjoy the theoretical advantages of design-adaptive, automatic boundary correction, and minimax efficiency. Also, fast computing algorithms are available for implementations. See the monograph by Fan and Gijbels (1996) for details. This naturally leads one to replace kernel estimation by local linear estimation. This modified approach was used by Fan and Yao (1998) based on the first-order difference, which avoids the variance inflation due to higher order approximation. For simplicity, we will refer to this remedial procedure as the “Fan and Yao approach”.

To examine the performance of Fan and Yao approach in diffusion models, we revisit the CIR square-root diffusion model in Subsection B. Precisely the same values of parameters $(\kappa, \theta, \sigma)$ and bandwidth scale constant are inherited to generate samples with daily sampling frequency over a 10 year period. Two feasible methods will provide useful assessments. The first method compares estimated curves based on any single realization of simulated sample path, and in most of cases we observe that the Fan and Yao approach performs better than Stanton’s method. On the other hand, the sample ranges of $\{X_t\}$ vary considerably across simulations. Extremely high levels of $\{X_t\}$ (for e.g., 0.20) can rarely be achievable in both reality and practical simulations. This makes the local regression estimation over broader or entire range of $\{X_t\}$ difficult. To conduct more sensible comparisons, our second method only considers estimated curves for levels of $\{X_t\}$ restricted to the range $[0.03, 0.15]$. Keep a simulated sample path if its range interval covers this fixed range, otherwise this path is not counted. Across 101 qualified simulations, the estimated drift and diffusion curves with medium performance measured respectively by the sum of squared errors at 50 grid points evenly spaced in $[0.03, 0.15]$ are presented in Figure 4. We find that local linear diffusion estimator achieves substantial gains in alleviating the impact of “boundary effects” than the kernel estimator. Comparatively for the drift estimation, we do not anticipate consistently better performances from either the kernel or local linear estimator. This can be easily understood from Theorem 4, which states that the drift estimator is far more variable than the diffusion estimator.

Another nonignorable factor that might influence the behaviours of the estimated drift and diffusion functions is the amount of historical data given the same daily sampling frequency. To evaluate this impact, we simply use the discretization scheme to generate 101 sample realizations (described above) from the CIR square-root diffusion model with same values of parameters as in Section B over 20 year, 30 year, 40 year, till 50 year period respectively, and present the estimated drift and diffusion curves in Figure 5. Clearly, diffusion estimators are more robust to sample size than drift
estimators. In contrast, the unstable performances of drift estimators using both kernel and local linear fit can not be significantly improved if a larger bandwidth is used. This is due to the fact that larger bandwidths force resulting estimators to behave more like least-squares lines with either intercept or slope or both different from their true values. See numerical examples in Subsection A below.

III Model Validation

Model diagnosis plays an important role in examining the relevance of specific assumptions underlying the modeling process and to identifying unusual features of the data that may influence conclusions. Despite a wide variety of well-known parametric models imposed on the short-term interest rates and stock price indices, relatively little is known about how these models capture the actual stochastic dynamics of the underlying processes. Among them, a majority of the useful models have been studied and compared under a unified framework

$$dX_t = (\alpha + \beta X_t)dt + \sigma X_t^\gamma dW_t,$$  \hspace{1cm} (33)

in Chan, Karolyi, Longstaff, and Sanders (1992) based on "generalized methods of moments" of Hansen (1982). However, the question arises frequently whether the model (33) itself correctly captures the stochastic dynamics of a given economic data. To answer this question, we need an alternative family of stochastic models. Nonparametric models offer a very nice solution to this problem. Depending on the cases and the nature of model validation, the nonparametric alternative models can be

$$dX_t = \mu(X_t)dt + \sigma X_t^\gamma dW_t,$$ \hspace{1cm} (34)
$$dX_t = (\alpha + \beta X_t)dt + \sigma(X_t) dW_t,$$ \hspace{1cm} (35)

or more generic model (1) without particular restrictions governing either the structural shift or volatility. While these kinds of hypothesis testing problems arise very often, we have not found other existing testing procedures having successfully tackled these types of model checking problems.

In this section, we first describe approaches used in estimating parameters of models (33) through (35). To defend the validity of these models, we take model (1) as our alternative hypothesis. New hypothesis-testing procedures are proposed based on the "generalized likelihood ratio" by Fan, Zhang and Zhang (2000). The GLR tests' explanatory powers and versatilities are demonstrated by simulations and two real data applications.

A Parametric Estimation

For simplicity, we proceed from parametric model (33). Given data $\{X_{t_i}, i = 1, \cdots, n\}$ driven from model (33), denote $t_{i+1} - t_i = \Delta_i$ and $Y_{ti} = X_{t_{i+1}} - X_{t_i}$, then the parameters $(\alpha, \beta, \sigma, \gamma)$ can be
estimated through a discrete-time specification

\[ Y_{ti} \approx (\alpha + \beta X_{ti})\Delta_i + \sigma X_{ti}^\gamma \sqrt{\Delta_i} \varepsilon_{ti}, \quad i = 1, \ldots, n-1 \]  

(36)

where \( \varepsilon_{ti} = (W_{t_{i+1}} - W_{ti})/\sqrt{\Delta_i} \overset{iid}{\sim} N(0, 1) \). Three steps summarize the estimation procedure:

**Step I**: Pretending model (36) as homoscedastic, obtain the least-squares estimates \((\hat{\alpha}^{(1)}, \hat{\beta}^{(1)})\) of \((\alpha, \beta)\).

**Step II**: Write \( \hat{\varepsilon}_{ti} = \{Y_{ti} - (\hat{\alpha}^{(1)} + \hat{\beta}^{(1)} X_{ti})\Delta_i\}/\sqrt{\Delta_i} \), then from equation (36),

\[ \log(\hat{\varepsilon}_{ti}^2) \approx \log(\sigma^2) + \gamma \log(X_{ti}^\gamma) + \log(\varepsilon_{ti}^2), \quad i = 1, \ldots, n-1. \]  

(37)

Obtain least-squares estimates \((\hat{\sigma}^{(1)}, \hat{\gamma}^{(1)})\) of \((\sigma, \gamma)\) after subtracting \(E\{\log(Z^2)\} \approx -1.270362845\) \((Z \sim N(0, 1))\) from both sides of equation (37).

**Step III** (optional): Substitute \((\hat{\sigma}^{(1)}, \hat{\gamma}^{(1)})\) into model (36) and get weighted least-squares estimates \((\hat{\alpha}^{(2)}, \hat{\beta}^{(2)})\). Meanwhile, update \((\hat{\sigma}^{(2)}, \hat{\gamma}^{(2)})\) in Step II.

This approach can be modified flexibly. For model (34), parameters \((\sigma, \gamma)\) could be estimated directly from Step II by setting \(\hat{\varepsilon}_{ti}\) in equation (37) to \(\{Y_{ti} - \hat{\mu}(X_{ti})\Delta_i\}/\sqrt{\Delta_i}\), where \(\hat{\mu}(X_{ti})\) is estimated nonparametrically by Fan and Yao approach. Denote the resulting estimators of \((\sigma, \gamma)\) by \((\hat{\sigma}^{(3)}, \hat{\gamma}^{(3)})\). Similar modifications apply to model (35).

*Put Table II about here*

To assess the efficiency of the parametric estimators \((\hat{\alpha}^{(\ell)}, \hat{\beta}^{(\ell)}, \hat{\sigma}^{(\ell)}, \hat{\gamma}^{(\ell)}), \ell = 1, 2, \) and \((\hat{\sigma}^{(3)}, \hat{\gamma}^{(3)})\), we generate sample paths of length \(10^3\) and \(10^4\) successively from CIR model \(dX_t = (0.0183925 - 0.21459X_t)dt + 0.0783\sqrt{X_t}dW_t\). Based on 10000 simulation runs, the sample means and sample standard deviations of these estimators are reported in Table II. Obviously, \(\sigma\) and \(\gamma\) can be estimated far more efficiently than \(\alpha\) and \(\beta\). Also, the improvements of the weighted least-squares estimators over the unweighted estimators are negligible.

### B  Generalized Likelihood Ratio Test

The interest rate volatility plays a key role in valuing contingent claims and hedging interest rate risk. Therefore for sake of brevity, we describe how to test model (34) against nonparametric alternative (1). Let \(E_{ti} = \{Y_{ti} - \hat{\mu}(X_{ti})\Delta_i\}/\sqrt{\Delta_i}\) and \(Y_{ti}^{(1)} = \log(E_{ti}^2)\). Then similar to equations (36) and (37), we have approximately

\[ \hat{E}_{ti} \approx \sigma(X_{ti})\varepsilon_{ti}, \quad i = 1, \ldots, n-1, \]

\[ Y_{ti}^{(1)} \approx \log\{\sigma^2(X_{ti})\} + \log(\varepsilon_{ti}^2), \quad i = 1, \ldots, n-1. \]  

(38)

This translates the original testing problem to testing

\[ H_0 : \sigma(X_t) = \sigma X_t^\gamma \quad \text{versus} \quad H_1 : \sigma(X_t) \neq \sigma X_t^\gamma. \]  

(39)
Under the null hypothesis in (39), let \((\hat{\sigma}, \hat{\gamma})\) be the estimated parameters discussed in the previous section. Under the alternative model (1), let \(\hat{\sigma}(x)\) be the estimated diffusion function based on Fan and Yao's approach. The GLR test statistic, proposed in Fan, Zhang and Zhang (2000), is basically given by

\[
\lambda_n(h) = \frac{n - 1}{2} \cdot \frac{\text{RSS}_0 - \text{RSS}_1(h)}{\text{RSS}_1(h)},
\]

where \(\text{RSS}_0\) and \(\text{RSS}_1\) are the residual sums of squares of model (38) under the null and the alternative hypothesis in (39), respectively. Under \(H_0\), there will be little difference in the size of \(\text{RSS}_0\) and \(\text{RSS}_1\); under the alternative, \(\text{RSS}_0\) should become systematically larger than \(\text{RSS}_1\) and the GLR test statistic will tend to take large positive values. Hence, a high value of test statistic \(\lambda_n(h)\) indicates that the null hypothesis should be rejected. This procedure applies similarly to testing other forms of drift or diffusion functions.

In the nonparametric regression model with independent data, Fan, Zhang and Zhang (2000) showed the Wilks type of result: \(r_K \lambda_n(h)\) is asymptotically distributed as \(\chi^2_{d_n(h)}\) under the null hypotheses. Here, \(r_K\) is a normalizing constant, depending only on the kernel function \(K\) and \(d_n(h)\) is the degrees of freedom, relating to \(n\) and \(h\). Although the GLR statistic applied to our current setup (38) involves more complicated stochastic errors and requires more detailed technical justifications, we believe similar Wilks type of results also hold for problem (39). Indeed, we have conducted substantial simulations which provide stark evidences to support this claim.

\section{Power Simulations}

One advantage of nonparametric regression is attributed to their flexibility in model assumptions. This broadens the scope of the applications. As a result, nonparametric tests, while gaining significant flexibility, may result in loss of powers compared with parametric counterparts when the parametric assumptions provide a suitable description of the true pattern. To gauge the power of our proposed GLR test, we perform the following simulation studies.

Firstly, we compute the empirical critical values of GLR statistics for testing one of the four typical forms of null hypotheses

\[
\begin{align*}
H_0^{(1)} & : \mu(X_t) = \alpha_0 + \beta_0 X_t, \quad \sigma(X_t) = c_0 X_t^{0.5}, \\
H_0^{(2)} & : \mu(X_t) = \alpha_0 + \beta_0 X_t, \quad \sigma(X_t) = c_1 X_t^{1.0}, \\
H_0^{(3)} & : \mu(X_t) = 0, \quad \sigma(X_t) = c_2 X_t^{1.5}, \\
H_0^{(4)} & : \mu(X_t) = \alpha_0 + \beta_0 X_t, \quad \sigma(X_t) = \sigma X_t^\gamma.
\end{align*}
\]

against their nonparametric alternatives (1). Here we set \(\alpha_0 = 0.00739\) and \(\beta_0 = -0.11798\), which result from the weighted least-squares estimates of the 3-month interest rate data (described at the beginning of Section D). The constants \(c_0 = 0.05596, c_1 = 0.23103\) and \(c_2 = 0.90114\) are used. The parameters \(\sigma\) and \(\gamma\) in null hypothesis (44) are unknown. 1000 simulated samples of length 2400
are generated with weekly frequency from each of the above hypothetical models with the initial value \( x_0 = 0.013 \), the first observation of the interest rate data. For simulated samples generated from model (44), the diffusion function \( \sigma(X_t) \) is taken to be the parametric fitting with weighted least-squares estimates \( \hat{\sigma} = 0.071258 \) and \( \hat{\gamma} = 0.72957 \) from the interest rate data. Some necessary adjustments are incorporated in programming to prevent divergent sequences (e.g. \( X_t \) falling below 0 or above 1) during simulations.

To perform the GLR test combined with Fan and Yao’s approach, we adopt the empirical formula for the initial bandwidth

\[
h_0 = 4 \text{std}(\{X_{t_1}, X_{t_2}, \ldots, X_{t_n}\}) n^{-2/9},
\]

where \( \{X_{t_i}\}_{i=1}^n \) are simulated sample path. We also consider the GLR test statistics at three different scales of bandwidth \( h_j = 1.5^{j-1}h_0 \), \( j = 1, 2, 3 \) to simultaneously assess the impact of bandwidth choices. To expedite the computations, we calculate the local linear estimates at 200 grid points evenly spaced within the ranges of simulated samples, and then take linear interpolation to obtain the estimates at all 2400 data points. The critical values do not very sensitively depend on the true values of parameters in the null hypotheses, though they should depend on the choice of bandwidths. The results are summarized in Table III.

*Put Table III about here*

*Put Table IV about here*

Secondly, to understand the power of the test statistics \( \lambda_n(h_j), j = 1, 2, 3 \), we consider testing the CIR model (41) against the nonparametric alternative (1). We evaluate the power of the tests at 5% level for the specific models \( H^2_0 \), \( \ell = 1, 2, 3, 4 \), based on 400 simulations. Figure 6 depicts the degree of departures of the volatility functions \( c_1 X^{1.0}_t \), \( c_2 X^{1.5}_t \) and \( \sigma X^{2}_t \) from the hypothetical null model \( c_0 X^{0.5}_t \). Thus, the GLR tests, as shown in Table IV, are powerful in detecting slight deviations, in addition to keeping the right size.

### D Testing Commonly-used Short Rate Models

The Treasury bill data set for our study consists of 2400 weekly observations, from January 8, 1954 up to December 31, 1999. Treasury bill secondary market rates are the averages of the bid rates quoted on a bank discount basis by a simple of primary dealers who report to the federal reserve bank of New York. The rates reported are based on quotes at the official close of the U.S. government securities market for each business day. Figure 7 shows the estimated drift and volatility curves based on Fan and Yao’s approach. The estimated drift function exhibitss strong nonlinearities at right boundary region, also the volatility behaves like CIR VR form.

We first address Chapman and Pearson (2000)’s question “Is the short-rate drift actually nonlinear?”, which becomes tantamount to testing the model (35) versus model (1). This problem is very hard, because with the larger magnitude of noises contaminated, distinguishing signal patterns becomes even challenging than signal extraction. Despite Chapman and Pearson’s full coverage and great
efforts in explaining the seemingly nonlinear drift function, there still lacks convincing procedures to formally justify whether the deviation from linearity is due to chance variations. With the aid of the powerful GLR test, we are able to compute the associated $P$-value based on regression bootstrap method for approximating empirical null distribution of GLR test statistics. Complete procedure is composed of the following steps.

**Step 1:** For original T-bill data \( \{X_{t_i}\}_{i=1}^n \), denote \( Y_{t_i} = X_{t_{i+1}} - X_{t_i} \). From \( \{(X_{t_i}, Y_{t_i})_{i=1}^{n-1}\} \), obtain least-squares estimates \((\hat{\alpha}, \hat{\beta})\), and \( \text{RSS}_0 = \sum_{i=1}^{n-1} \{Y_{t_i}/\Delta - \hat{\alpha} - \hat{\beta}X_{t_i}\}^2 \). Use Fan and Yao’s approach with bandwidth \( h \) to obtain \( \tilde{\mu}(X_{t_i}), \tilde{\sigma}(X_{t_i}) \), and \( \text{RSS}_1(h) = \sum_{i=1}^{n-1} \{Y_{t_i}/\Delta - \tilde{\mu}(X_{t_i})\}^2 \). Compute the observed value of test statistic \( \lambda_{n, \text{obs}}(h) = \frac{n-1}{2} \frac{\text{RSS}_0 - \text{RSS}_1(h)}{\text{RSS}_1(h)} \). Obtain the standardized residuals \( \tilde{e}_{t_i} = \frac{Y_{t_i} - \tilde{\mu}(X_{t_i})\Delta}{\tilde{\sigma}(X_{t_i})\sqrt{\Delta}} \).

**Step 2:** Obtain bootstrap sample \( \{\tilde{e}_{t_i}^{(b)}\} \) from \( \{\tilde{e}_{t_i}\} \) and construct the bootstrap sample \( Y_{t_i}^{(b)} = (\hat{\alpha} + \hat{\beta}X_{t_i})\Delta + \tilde{\sigma}(X_{t_i})\sqrt{\Delta}\tilde{e}_{t_i}^{(b)} \). Use the bootstrapped data \( \{(X_{t_i}, Y_{t_i}^{(b)})_{i=1}^{n-1}\} \) to get the bootstrap test statistic \( \lambda_{n}^{(b)} \).

**Step 3:** Repeat Step 2 many times (1000, say) and compute the proportion of times that \( \lambda_{n}^{(b)} \) exceeds \( \lambda_{n, \text{obs}}(h) \). This yields the $P$-value of the GLR test.

Using the bootstrap procedure just described, we obtain the $P$-values, shown in the second row of Table V, for the three different bandwidths \( h_j \) as in previous Subsection C. Thus, there is no strong evidence against the null hypothesis of linear drift. Our proposed test provides formal proofs to reinforce Chapman and Pearson’s study.

*Put Table V about here*

*Put Table VI about here*

We also apply similar procedures to testing the validity of some previously established hypotheses regarding the variance nature, in particular, competing forms (2) through (6) for volatility functions. The $P$-values are displayed in Table VI. Surprisingly, strong evidence indicates that these assumptions on volatility function can not be validated by our GLR tests. This is consistent with the results reported by Gallant, Long and Tauchen (1997).

To calibrate the GLR tests’ abilities in correctly rejecting null hypotheses, we simulate 100 data sets each of length 2400 from CIR square-root model (41). Based on the level 5% critical values of the above bootstrapped null distributions, a decision on rejecting the linear drift or not can be made for each sample. The proportion of rejection across 100 experiments are supplemented in third row of Table V. Similar results concerning volatility functions are listed in Table VI. Therefore, both Table V and Table VI strengthen the assertion that our bootstrap procedures are powerful in correctly accepting or rejecting the null hypotheses.

*Put Table VII about here*

*Put Table VIII about here*
E Testing Models for S&P 500 Index

In addition to the interest rate application, let’s investigate the significance of structural shifts of S&P 500 data from previously studied models. This data set contains 6890 daily observations, from January 4, 1971 up to April 8, 1998 on the Standard and Poor’s composite price index. Following the conventional practice in finance research, we first take logarithm transformation of the price index. The estimated drift and volatility based on Fan and Yao’s approach are displayed in Figure 8. The associated bootstrap P-values are presented in Table VII and Table VIII. Clearly, there is no strong evidence against the hypothesis on the linear drift. For the volatility function, our testing results tend to pick up the Vasicek model for the logarithm of the index. This in turn suggests that the S&P 500 index follows the GBM. Therefore, we formally verify the most critical assumption that is used to derive the celebrated option pricing formula of Black and Scholes (1973). However, this conclusion is drawn within the context of time-homogeneous model (1). Note also that our test also validates the CKLS model for the volatility of the S&P 500 index. This is natural since the submodel “Vasicek” has already been accepted.

IV Conclusion

In his paper, Stanton (1997) claimed (in page 1982) by stating “The higher the order of the approximation, the faster it will converge to the true drift and diffusion of the process given in equation (1), as we observe the variable $X_t$ at finer and finer time intervals. Eventually, if we can sample arbitrarily often, higher order approximations must outperform lower order approximations”, and reiterated (in page 1983) that “…even with daily or weekly data, we can achieve gains by using higher order approximations compared with the traditional first order discretizations”. Actually, these claims are correct but somewhat misleading. They ignore the variance inflation in statistical estimation due to higher order approximation. This variance inflation phenomenon is not an artifact of nonparametric fitting. It also applies to parametric models. In this paper, we emphasize that the opposite side of Stanton’s claim has been neglected. With the tool of asymptotic analysis, we show that higher-order approximations benefit from reducing the numerical approximation error within asymptotic bias, a statement correctly made by Stanton (1997), but nevertheless are penalized by escalating asymptotic variance nearly exponentially with the order of approximations. This shadows higher-order approximation scheme. This phenomenon can be accounted for by the stochastic nature of the Taylor series expansion in equation (8) accumulated with higher-order linearization scheme (11). Cautions should be taken when using higher-order formulas. This bias and variance trade-off phenomenon yields general and insightful understandings of the estimators, and also provides useful guidance for determining optimal strategy for order of approximation as well as proposing more efficient estimators awaiting future research endeavors.

Encouragingly, by using Fan and Yao’s approach, spurious “boundary effects” are ameliorated especially for estimating diffusion functions. This nonparametric estimation approach could also
be incorporated with the “GLR statistic” to test a wide variety of parametric time-homogeneous diffusion models, and to formally check nonlinearity of the short-rate drift. Our simulation shows that our procedures are indeed powerful and have nearly correct size of the test. Our test formally verifies the GBM assumption for the S&P 500 data, while rejects all of the commonly-used models for the short-term interest rates. The procedures are useful for verifying various models in Finance and Economics.

Appendix

Proof of Theorem 1. Using the matrix notation, the system of equation in (9) can be written as $Ax = b$, where

$$A = \begin{bmatrix} 1 & 2 & \cdots & j & \cdots & k \\ 1 & 2^2 & \cdots & j^2 & \cdots & k^2 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & 2^k & \cdots & j^k & \cdots & k^k \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Thus, the solution $x = [x_1, \cdots, x_k]^T$ is uniquely determined by

$$x = |A|^{-1} A^* b, \quad (46)$$

where $A^*$ and $|A|$ denote respectively the adjoint matrix and the determinant of the matrix $A$. Namely $x$ is the first column of $A^{-1}$. Using the property of Vandermonde matrix, we obtain the determinant of matrix $A$

$$|A| = 2 \times 3 \times \cdots \times k \times \begin{vmatrix} 1 & 1 & \cdots & 1 & \cdots & 1 \\ 1 & 2 & \cdots & j & \cdots & k \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & 2^{k-1} & \cdots & j^{k-1} & \cdots & k^{k-1} \end{vmatrix} = k! \prod_{1 \leq l_1 < l_2 \leq k} (l_2 - l_1).$$

Using this property again, the $j$-th element in the first column of matrix $A^*$ is

$$A^*(j, 1) = (-1)^{j+1} \frac{(k!)^2}{j^2} \prod_{1 \leq l_1 < l_2 \leq k, \ l_1 \neq j, \ l_2 \neq j} (l_2 - l_1).$$

Hence the solution $x_j (1 \leq j \leq k)$ in equation (46) can be simplified as:

$$x_j = (-1)^{j+1} \frac{(k!)^2}{j^2} \prod_{1 \leq l_1 < l_2 \leq k} (l_2 - l_1) = \frac{(-1)^{j+1}k!}{j^2 (j-1)! (k-j)!} = \frac{(-1)^{j+1}k!}{j^2 (j-1)! (k-j)!} = \frac{(-1)^{j+1}}{j^2} \binom{k}{j}/j.$$ 

This proves the first statement. We now prove the second statement. The proof is based on the recursion relation, which we now derive. For any $1 \leq j \leq k$, $\binom{k}{j}j = \binom{k-1}{j-1}k$. Hence, by the first
statement, we have
\[
\sum_{j=1}^{k} j^{k+1} a_{k,j} = \sum_{j=1}^{k} (-1)^{j+1} \binom{k}{j} j^{k} = -k \left[ (-1) + \sum_{j=1}^{k-1} (-1)^{j+1} \binom{k-1}{j} (j+1)^{k-1} \right].
\]
Using the Binomial expansion for the factor \((j + 1)^{k-1}\) and exchanging the summation, we obtain
\[
\sum_{j=1}^{k} j^{k+1} a_{k,j} = -k \left[ (-1) + \sum_{l=0}^{k-1} \binom{k-1}{l} \sum_{j=1}^{l} j^{l+1} a_{k-l,j} \right]
\]
This together with (9) yield
\[
\sum_{j=1}^{k} j^{k+1} a_{k,j} = -k \left[ (-1) + 1 + \sum_{j=1}^{k-1} j^{k} a_{k-1,j} \right] = -k \sum_{j=1}^{k-1} j^{k} a_{k-1,j}.
\]
The conclusion follows from the above inductive formula. \(\blacksquare\)

Before we derive the asymptotic variances in Theorem 2, we need the following lemma.

**Lemma 1** For each fixed \(x_0\), as \(\Delta \to 0\),

\[
(a) \quad \mathbb{E}\{(X_{t+\Delta} - X_t)|X_t = x_0\} = \mu(x_0) \Delta + O(\Delta^2), \tag{47}
\]
\[
(b) \quad \mathbb{E}\{(X_{t+\Delta} - X_t)^2|X_t = x_0\} = \sigma^2(x_0) \Delta + O(\Delta^2), \tag{48}
\]
\[
(c) \quad \mathbb{E}\{(X_{t+\Delta} - X_t)^3|X_t = x_0\} = 3\sigma^2(x_0) \{\mu(x_0) + 2^{-1} (\sigma^2)'(x_0)\} \Delta^2 + O(\Delta^3), \tag{49}
\]
\[
(d) \quad \mathbb{E}\{(X_{t+\Delta} - X_t)^4|X_t = x_0\} = 3\sigma^4(x_0) \Delta^2 + O(\Delta^3), \tag{50}
\]
\[
(e) \quad \mathbb{E}\{(X_{t+\Delta} - X_t)^5|X_t = x_0\} = \mu^2(x_0) + \mu'(x_0) \sigma^2(x_0) \Delta + O(\Delta^2), \tag{51}
\]
\[
(f) \quad \mathbb{E}\{(X_{t+\Delta} - X_t)^6|X_t = x_0\} = \sigma^2(x_0) \Delta + O(\Delta^2), \tag{52}
\]
\[
(g) \quad \mathbb{E}\{(X_{t+\Delta} - X_t)^3|X_t = x_0\} = O(\Delta^2). \tag{53}
\]

**Proof of Lemma 1.** To show results (a) through (g), we choose corresponding functions \(f_1(x, t) = (x - X_t)\), \(f_2(x, t) = (x - X_t)^2\), \(f_3(x, t) = (x - X_t)^3\), \(f_4(x, t) = (x - X_t)^4\), \(f_5(x, t) = (x - X_t)\mu(x)\), \(f_6(x, t) = (x - X_t)^2\sigma^2(x)\), and \(f_7(x, t) = (x - X_t)^3\mu(x)\) respectively. Straightforward calculations using the differential operator \(\mathcal{L}\) defined in equation (7) give the following relations

\[
\mathcal{L} f_1(x, t) = \mu(x),
\]
\[
\mathcal{L}^2 f_1(x, t) = \mu'(x) \mu(x) + 2^{-1} \mu''(x) \sigma^2(x),
\]
\[
\mathcal{L} f_2(x, t) = 2(x - X_t) \mu(x) + \sigma^2(x),
\]
\[
\mathcal{L}^2 f_2(x, t) = \{2\mu(x) + 2(x - X_t) \mu'(x) + (\sigma^2)'(x)\} \mu(x)
+ 2^{-1} \{4\mu'(x) + 2(x - X_t) \mu''(x) + (\sigma^2)''(x)\} \sigma^2(x),
\]
\[
\mathcal{L} f_3(x, t) = 3(x - X_t)^2 \mu(x) + 3(x - X_t) \sigma^2(x),
\]
\[
\mathcal{L}^2 f_3(x, t) = \left\{ \frac{6(x - X_t) \mu(x) + 3(x - X_t)^2 \mu'(x) + 3 \sigma^2(x)}{\mu(x)} + 3(x - X_t) \sigma^2(x) \right\} \mu(x) + 2^{-1} \sigma^2(x)
\]
\[
\mathcal{L} f_4(x, t) = 4(x - X_t)^3 \mu(x) + 6(x - X_t)^2 \sigma^2(x),
\]
\[
\mathcal{L}^2 f_4(x, t) = \left\{ \frac{12(x - X_t)^2 \mu(x) + 4(x - X_t)^2 \mu'(x) + 12(x - X_t) \sigma^2(x) + 6(x - X_t)^2 \sigma^2(x)}{\mu(x)} \right\} \mu(x)
\]
\[
\quad + 2^{-1} \left\{ \frac{24(x - X_t) \mu(x) + 24(x - X_t)^2 \mu'(x) + 4(x - X_t)^3 \mu''(x)}{\sigma^2(x)} \right\} \sigma^2(x) + 12 \sigma^2(x) + 24(x - X_t)(\sigma^2(x) + (x - X_t)^2 \sigma^2(x)),
\]
\[
\mathcal{L} f_5(x, t) = \left\{ \mu(x) + (x - X_t) \mu'(x) \right\} \mu(x) + 2^{-1} \left\{ \mu''(x) + \mu'(x) + (x - X_t) \mu''(x) \right\} \sigma^2(x),
\]
\[
\mathcal{L} f_6(x, t) = \left\{ \frac{2(x - X_t) \sigma^2(x) + (x - X_t)^2 (\sigma^2)'(x)}{\mu(x)} \right\} \mu(x)
\]
\[
\quad + 2^{-1} \left\{ \frac{2 \sigma^2(x) + 4(x - X_t) (\sigma^2)'(x) + (x - X_t)^2 (\sigma^2)''(x)}{\sigma^2(x)} \right\} \sigma^2(x),
\]
\[
\mathcal{L} f_7(x, t) = \left\{ \frac{3(x - X_t)^2 \mu(x) + (x - X_t)^3 \mu'(x)}{\mu(x)} \right\} \mu(x)
\]
\[
\quad + 2^{-1} \left\{ \frac{6(x - X_t) \mu(x) + 6(x - X_t)^2 \mu'(x) + (x - X_t)^3 \mu''(x)}{\sigma^2(x)} \right\} \sigma^2(x).
\]

The proofs are completed by using Taylor series expansion in equation (8). ■

**Proof of Theorem 2.** We start by considering the conditional variance for the drift estimator. From the definitions in equations (15) and (19) as well as the stationarity of \( \{X_t, t \geq 0\} \), we have at any time location \( t = t_0 + i \Delta, i = 1, \ldots, n - k \),

\[
\sigma^2_t(x_0) = \Delta^{-2} \left\{ \sum_{1 \leq j \leq k} a_{k,j} \text{Var}\{(X_{t+j \Delta} - X_t)|X_t = x_0\} + \right.
\]
\[
\left. 2 \sum_{1 \leq i < j \leq k} a_{i,j} a_{k,j} \text{Cov}\{(X_{t+i \Delta} - x_0, X_{t+j \Delta} - x_0)|X_t = x_0\} \right\}. \quad (54)
\]

For \( j \geq 1 \), equations (20) and (21) imply that

\[
\text{Var}\{(X_{t+j \Delta} - X_t)|X_t = x_0\} = \text{E}\{(X_{t+j \Delta} - X_t)^2|X_t = x_0\} - \text{E}\{(X_{t+j \Delta} - X_t)|X_t = x_0\}^2
\]
\[
= \sigma^2(x_0) j \Delta + O(\Delta^2). \quad (55)
\]

For \( 1 \leq i < j \leq k \), under the regularity conditions that \( \{X_t, t \geq 0\} \) is a Markov process (Wong, 1971), we have from equations (47), (48) and (51),

\[
\text{E}\{(X_{t+i \Delta} - x_0)(X_{t+j \Delta} - x_0)|X_t = x_0\}
\]
\[
= \text{E}\{(X_{t+i \Delta} - x_0)E[X_{t+j \Delta} - x_0|X_{t+i \Delta}]|X_t = x_0\} \quad \text{(Markovian property)}
\]
\[
= \text{E}\{(X_{t+i \Delta} - x_0)(X_{t+j \Delta} - x_0) + \mu(X_{t+i \Delta})(j - i) \Delta + O(\Delta^2)|X_t = x_0\}
\]
\[
= \text{E}\{(X_{t+i \Delta} - x_0)^2 + (X_{t+i \Delta} - x_0)\mu(X_{t+i \Delta})(j - i) \Delta + (X_{t+i \Delta} - x_0)O(\Delta^2)|X_t = x_0\}
\]
\[
= \sigma^2(x_0) i \Delta + O(\Delta^2). \quad (56)
\]
We also obtain from equation (20),

\[
\begin{align*}
\mathbb{E}\{(X_{t+\Delta} - x_0)|X_t = x_0\} &= \mathbb{E}\{(X_{t+j\Delta} - x_0)|X_t = x_0\} \\
&= \{\mu(x_0)i\Delta + O(\Delta^2)\} \{\mu(x_0)j\Delta + O(\Delta^2)\} \\
&= O(\Delta^2).
\end{align*}
\]  

(57)

The expression (20) follows directly from the combination of equations (54), (55), (56) and (57).

We now consider the conditional variance for the squared diffusion estimator. In the same vein, we have

\[
\sigma^2_{2,\Delta}(x_0) = \Delta^2 \left\{ \sum_{1 \leq j \leq k} a_{k,j}^2 \text{Var}\{ (X_{t+j\Delta} - X_t)^2 | X_t = x_0 \} + 2 \sum_{1 \leq i < j \leq k} a_{k,i}a_{k,j} \text{Cov}\{ (X_{t+i\Delta} - x_0)^2, (X_{t+j\Delta} - x_0)^2 | X_t = x_0 \} \right\}.
\]

(58)

For \( j \geq 1 \), equations (21) and (23) imply that

\[
\text{Var}\{ (X_{t+j\Delta} - X_t)^2 | X_t = x_0 \} = \mathbb{E}\{ (X_{t+j\Delta} - X_t)^4 | X_t = x_0 \} - \mathbb{E}\{ (X_{t+j\Delta} - X_t)^2 | X_t = x_0 \}^2 = 2\sigma^4(x_0)(j\Delta)^2 + O(\Delta^3).
\]

(59)

For \( 1 \leq i < j \leq k \), under the regularity conditions that \( \{X_t, t \geq 0\} \) is a Markov process (Wong, 1971), we have from equations (50), (52) and (53),

\[
\begin{align*}
\mathbb{E}\{(X_{t+i\Delta} - x_0)^2(X_{t+j\Delta} - x_0)^2 | X_t = x_0 \} &= \mathbb{E}\{(X_{t+i\Delta} - x_0)^2|X_{t-i\Delta} = x_0\} \mathbb{E}\{(X_{t+j\Delta} - x_0)^2|X_{t-j\Delta} = x_0\} \quad \text{(Markovian property)} \\
&= \mathbb{E}\{(X_{t+i\Delta} - x_0)^2[(X_{t+i\Delta} - x_0)^2 + 2(X_{t+i\Delta} - x_0)\mu(X_{t+i\Delta}) + \sigma^2(X_{t+i\Delta})(j - i)\Delta + O(\Delta^3)] | X_t = x_0 \} \\
&+ \mathbb{E}\{(X_{t+j\Delta} - x_0)^2[(X_{t+j\Delta} - x_0)^2 + 2(X_{t+j\Delta} - x_0)\mu(X_{t+j\Delta}) + \sigma^2(X_{t+j\Delta})(j - i)\Delta + O(\Delta^3)] | X_t = x_0 \} \\
&= 3\sigma^4(x_0)(i\Delta)^2 + O(\Delta^3) + \sigma^4(x_0)i\Delta(j - i)\Delta + O(\Delta^3) \\
&= 2\sigma^4(x_0)(i\Delta)^2 + \sigma^4(x_0)ij\Delta^2 + O(\Delta^3).
\end{align*}
\]

(60)

We also obtain from equation (21),

\[
\begin{align*}
\mathbb{E}\{(X_{t+i\Delta} - x_0)^2 | X_t = x_0 \} &= \mathbb{E}\{(X_{t+j\Delta} - x_0)^2 | X_t = x_0 \} \\
&= \{\sigma^2(x_0)i\Delta + O(\Delta^2)\} \{\sigma^2(x_0)j\Delta + O(\Delta^2)\} \\
&= \sigma^4(x_0)ij\Delta^2 + O(\Delta^3).
\end{align*}
\]

(61)

The equality (21) follows directly from equations (58), (59), (60) and (61).

**Proof of Theorem 3.** The proofs below are based on some combinatorial relations. Let \( \gamma = \lim_{n \to \infty} \left\{ \sum_{k=1}^n k^{-1} - \log(n) \right\} \approx 0.577216 \) be the Euler’s constant and function \( \psi(z) = \Gamma'(z)/\Gamma(z) \), where
\[ \Gamma(z) = \int_0^\infty u^{z-1}e^{-u}du, \quad z > 0. \] First we consider part (a). With the aid of Mathematica, we obtain the identities,

\begin{align*}
\sum_{j=1}^{k} \binom{k}{j}^2 \frac{(j+2)}{(j+1)^2} &= \frac{(2k+1)!}{((k+1)!)^2} + \frac{4^{k+1}\Gamma(3/2 + k)}{(k+1)^3 \sqrt{\pi} k!} - \frac{2k^2 + 4k + 3}{(k+1)^2}, \quad (62) \\
\sum_{j=1}^{k} \binom{k}{j}^2 \frac{(j+3)}{(j+1)^2} &= \frac{(2k+1)!}{((k+1)!)^2} + \frac{2^{2k+3}\Gamma(3/2 + k)}{(k+1)^3 \sqrt{\pi} k!} - \frac{3k^2 + 6k + 5}{(k+1)^2}, \quad (63)
\end{align*}

and

\begin{equation}
\sum_{j=2}^{k} \frac{\left\{ \sum_{i=1}^{j-1}(-1)^{i+1}\binom{k}{i}\right\}(-1)^{j+1}\binom{k}{j}}{j} = \frac{1 + \gamma}{k} - \frac{1}{k} \binom{2k}{k} + \psi(k+1). \quad (64)
\end{equation}

Consequently by using \(a_{k,j} = (-1)^{j+1}\binom{k}{j}/j\) and simplifying the right hand sides of equations (62) and (63), we have

\begin{align*}
\sum_{j=1}^{k} j a_{k,j}^2 &> \sum_{j=1}^{k} \binom{k}{j}^2 \frac{(j+2)}{(j+1)^2} = \frac{(2k+1)(k+3)}{(k+1)^3} \binom{2k}{k} - \frac{2k^2 + 4k + 3}{(k+1)^2}, \quad (65) \\
\sum_{j=1}^{k} j a_{k,j}^2 &< \sum_{j=1}^{k} \binom{k}{j}^2 \frac{(j+3)}{(j+1)^2} = \frac{(2k+1)(k+5)}{(k+1)^3} \binom{2k}{k} - \frac{3k^2 + 6k + 5}{(k+1)^2}. \quad (66)
\end{align*}

Using equation (64) and \(\psi(n) = \sum_{j=1}^{n-1} j^{-1} - \gamma\) for integer \(n \geq 2\), we also obtain

\begin{equation}
\sum_{1 \leq i < j \leq k} i a_{k,i} a_{k,j} = \frac{1}{k} + \sum_{j=1}^{k} \frac{1}{j} - \frac{1}{k} \binom{2k}{k}. \quad (67)
\end{equation}

Hence equations (54), (20), (65), (66) and (67) imply that, \(V_1(k)\) has a lower bound

\begin{equation}
\frac{k^2 - 3k - 2}{k(k+1)^3} \binom{2k}{k} + \frac{2}{k} + 2 \sum_{j=1}^{k} \frac{1}{j} - \frac{2k^2 + 4k + 3}{(k+1)^2}, \quad (68)
\end{equation}

and an upper bound

\begin{equation}
\frac{5k^2 - k - 2}{k(k+1)^3} \binom{2k}{k} + \frac{2}{k} + 2 \sum_{j=1}^{k} \frac{1}{j} - \frac{3k^2 + 6k + 5}{(k+1)^2}. \quad (69)
\end{equation}

The conclusion follows from applying Stirling formula \(n! = \sqrt{2\pi n} \left(\frac{n}{e}\right)^n \exp\left(\frac{\theta}{12n}\right)\) for some \(0 < \theta < 1\) to the first dominating terms of equations (68) and (69).

Next we consider part (b). For \(k \geq 1\), it follows directly that

\begin{equation}
\sum_{j=1}^{k} j^2 a_{k,j}^2 = \binom{2k}{k} - 1. \quad (70)
\end{equation}
With the aid of Mathematica again, for \( k > 1 \) and \( 2 \leq j \leq k \), we obtain the identity

\[
\sum_{i=1}^{j-1} (-1)^{i+1} \binom{k}{i} = \frac{(-1)^j j \Gamma(k)}{\Gamma(j) \Gamma(k - j + 1)} - \frac{(-1)^j \Gamma(k - 1)}{\Gamma(j) \Gamma(k - j)},
\]

which implies that

\[
\sum_{1 \leq i < j \leq k} i^2 a_{k,i} a_{k,j} = \frac{1}{k - 1} \sum_{j=2}^{k} \binom{k-1}{j} \binom{k}{j} - \sum_{j=2}^{k} \binom{k-1}{k-j} \binom{k}{j} = \frac{(2k-1)(k-2)+1}{k-1}.
\]

The conclusion (b) follows from equations (58), (21), (70), (72) and Stirling formula. ■

**Proof of Theorem 4.** The outline of the proof is to apply Banon (1978)'s regularity conditions for stationarity of \( \{X_t\}_{t \geq 0} \) to \( \{(X_{i\Delta}^*, Y_{i\Delta}^*)\}_{i=1}^{n-k} \) and \( \{(X_{i\Delta}^*, Z_{i\Delta}^*)\}_{i=1}^{n-k} \), then apply the usual “big-block, small-block” technical arguments in the proof of Theorem 6.1 in Fan and Gijbels (1996). The length proof is omitted here. ■
REFERENCES


Bai, X.Z., Russell, J.R., and Tiao, G.C., 1999, Beyond Merton's Utopia; effects of non-normality and dependence on the precision of variance estimates using high-frequency financial data, manuscript, Graduate School of Business, University of Chicago.


Figure 1: Theoretical values of $\log\{V_j(k)\}$ versus order $k$. $j = 1$ refers to drift estimator $\hat{\mu}_{1,\Delta}(\cdot)$; $j = 2$ refers to squared diffusion estimator $\hat{\mu}_{2,\Delta}(\cdot)$. 
Figure 2: Simulated values of $\log\{V_j(k)\}$ versus order $k$ for geometric Brownian motion $dX_t = (0.087 + 2^{-1}0.178^2)X_t dt + 0.178X_t dW_t$. $j = 1$ refers to drift estimator $\hat{\mu}_{1,\Delta}(1.0)$; $j = 2$ refers to squared diffusion estimator $\hat{\mu}_{2,\Delta}(1.0)$. Plots (a) and (a') are based on the same simulated sample paths generated by exact solution $X_t = \exp\{0.087t + 0.178W_t\}$, and plots (b) and (b') are based on the same simulated sample paths generated by discretization scheme (30).
Figure 3: Simulated values of $\log\{V_j(k)\}$ versus order $k$ for CIR model $dX_t = 0.21459(0.08571 - X_t)dt + 0.07830\sqrt{X_t}dW_t$. $j = 1$ refers to drift estimator $\hat{\mu}_{1,\Delta}(0.1)$; $j = 2$ refers to squared diffusion estimator $\hat{\mu}_{2,\Delta}(0.1)$. Plots (a) and (a') are based on the same simulated sample paths generated by noncentral chi-squared distribution, and plots (b) and (b') are based on the same simulated sample paths generated by the discretization scheme.
Figure 4: Estimated drift and diffusion functions for CIR model 
\[ dX_t = 0.21459(0.08571 - X_t)dt + 0.07830\sqrt{X_t}dW_t. \] 
Solid line is the true function. The dashed line is Fan and Yao’s local linear estimate, and the dash dotted line is Stanton’s kernel estimate. The first row of graphs are based on simulated sample paths generated by noncentral chi-squared distribution, and the second row of graphs are based on simulated sample paths generated by discretization scheme. Sample paths are generated over 10 year period.
Figure 5: Estimated drift and diffusion functions for CIR model \( dX_t = 0.21459(0.08571 - X_t)dt + 0.07830\sqrt{X_t}dW_t \). The left column of graphs reports estimated drift functions over 20 year, 30 year, 40 year, and 50 year period; and the right column of graphs reports the corresponding estimated diffusion functions. Sample paths are generated by discretization scheme. Solid line is the true function. The dashed line is Fan and Yao’s local linear estimate, and the dash dotted line is Stanton’s kernel estimate.
Figure 6: **Comparison of four different volatility curves.** The constants are $c_0 = 0.05596$, $c_1 = 0.23103$, and $c_2 = 0.90114$.

Figure 7: **Estimated drift and volatility of short rate.** Estimated drift and volatility functions based on Fan and Yao's approach, calculated using weekly data, January 8, 1954 to December 31, 1999. For $j = 1, 2, 3$, the bandwidths $h_j = 1.5^{j-1}h_0$, where $h_0 = 0.01984$ from formula (45).
Figure 8: Estimated drift and volatility of the S&P 500 Index. Estimated drift and volatility functions based on Fan and Yao's approach, calculated using daily data, January 4, 1971 to April 8, 1998. For $j = 1, 2, 3$, the bandwidths $h_j = 1.5^{j-1}h_0$, where $h_0 = 0.4019$ from formula (45).
Table I: Variance Inflation Factors by Using Higher Order Differences

<table>
<thead>
<tr>
<th>Order $k$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_1(k)$</td>
<td>1.00</td>
<td>2.50</td>
<td>4.83</td>
<td>9.25</td>
<td>18.95</td>
<td>42.68</td>
<td>105.49</td>
<td>281.65</td>
<td>798.01</td>
<td>2364.63</td>
</tr>
<tr>
<td>$V_2(k)$</td>
<td>1.00</td>
<td>3.00</td>
<td>8.00</td>
<td>21.66</td>
<td>61.50</td>
<td>183.40</td>
<td>570.66</td>
<td>1837.28</td>
<td>6076.25</td>
<td>20527.22</td>
</tr>
</tbody>
</table>

Table II: Parametric Estimation of CIR model: $dX_t = (0.0183925 - 0.21459X_t)dt + 0.0783\sqrt{X_t}dW_t$

<table>
<thead>
<tr>
<th></th>
<th>$\hat{\alpha}^{(1)}$</th>
<th>$\hat{\alpha}^{(2)}$</th>
<th>$\hat{\beta}^{(1)}$</th>
<th>$\hat{\beta}^{(2)}$</th>
<th>$\bar{\sigma}^{(1)}$</th>
<th>$\bar{\sigma}^{(2)}$</th>
<th>$\bar{\sigma}^{(3)}$</th>
<th>$\bar{\gamma}^{(1)}$</th>
<th>$\bar{\gamma}^{(2)}$</th>
<th>$\bar{\gamma}^{(3)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 10^3$ Mean</td>
<td>0.1288</td>
<td>0.1277</td>
<td>-1.5137</td>
<td>-1.5035</td>
<td>0.0908</td>
<td>0.0907</td>
<td>0.0903</td>
<td>0.4954</td>
<td>0.4951</td>
<td>0.4935</td>
</tr>
<tr>
<td>Std.</td>
<td>0.1114</td>
<td>0.1110</td>
<td>1.1635</td>
<td>1.1607</td>
<td>0.0793</td>
<td>0.0728</td>
<td>0.0830</td>
<td>0.2266</td>
<td>0.2264</td>
<td>0.2267</td>
</tr>
<tr>
<td>$n = 10^4$ Mean</td>
<td>0.0284</td>
<td>0.0275</td>
<td>-0.3367</td>
<td>-0.3266</td>
<td>0.0785</td>
<td>0.0785</td>
<td>0.0784</td>
<td>0.4998</td>
<td>0.4999</td>
<td>0.4996</td>
</tr>
<tr>
<td>Std.</td>
<td>0.0131</td>
<td>0.0126</td>
<td>0.1544</td>
<td>0.1496</td>
<td>0.0063</td>
<td>0.0063</td>
<td>0.0064</td>
<td>0.0317</td>
<td>0.0317</td>
<td>0.0318</td>
</tr>
</tbody>
</table>
Table III: Upper $\alpha$-Quantiles of Test Statistics $\lambda_n(h_j)$, $j = 1, 2, 3$ under Models $H^{(\ell)}_0$, $\ell = 1, 2, 3, 4$

<table>
<thead>
<tr>
<th></th>
<th>$\alpha = 0.01$</th>
<th>$\alpha = 0.05$</th>
<th>$\alpha = 0.10$</th>
<th>$\alpha = 0.15$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H^{(1)}_0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda_n(h_1)$</td>
<td>153.7</td>
<td>105.2</td>
<td>77.2</td>
<td>62.3</td>
</tr>
<tr>
<td>$\lambda_n(h_2)$</td>
<td>136.9</td>
<td>92.6</td>
<td>71.4</td>
<td>55.7</td>
</tr>
<tr>
<td>$\lambda_n(h_3)$</td>
<td>138.7</td>
<td>94.1</td>
<td>69.0</td>
<td>56.8</td>
</tr>
<tr>
<td>$H^{(2)}_0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda_n(h_1)$</td>
<td>135.4</td>
<td>89.7</td>
<td>68.8</td>
<td>55.3</td>
</tr>
<tr>
<td>$\lambda_n(h_2)$</td>
<td>132.5</td>
<td>93.6</td>
<td>66.2</td>
<td>49.6</td>
</tr>
<tr>
<td>$\lambda_n(h_3)$</td>
<td>116.0</td>
<td>77.6</td>
<td>59.2</td>
<td>46.4</td>
</tr>
<tr>
<td>$H^{(3)}_0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda_n(h_1)$</td>
<td>134.1</td>
<td>95.1</td>
<td>76.1</td>
<td>62.5</td>
</tr>
<tr>
<td>$\lambda_n(h_2)$</td>
<td>122.4</td>
<td>91.1</td>
<td>69.2</td>
<td>54.9</td>
</tr>
<tr>
<td>$\lambda_n(h_3)$</td>
<td>125.5</td>
<td>93.9</td>
<td>68.5</td>
<td>52.4</td>
</tr>
<tr>
<td>$H^{(4)}_0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda_n(h_1)$</td>
<td>138.2</td>
<td>92.3</td>
<td>71.7</td>
<td>54.6</td>
</tr>
<tr>
<td>$\lambda_n(h_2)$</td>
<td>128.6</td>
<td>93.3</td>
<td>69.0</td>
<td>56.9</td>
</tr>
<tr>
<td>$\lambda_n(h_3)$</td>
<td>124.0</td>
<td>82.6</td>
<td>65.9</td>
<td>52.5</td>
</tr>
</tbody>
</table>

Table IV: Simulated Rejection Rates against Models $H^{(\ell)}_0$, $\ell = 1, 2, 3, 4$

<table>
<thead>
<tr>
<th></th>
<th>$\lambda_n(h_1)$</th>
<th>$\lambda_n(h_2)$</th>
<th>$\lambda_n(h_3)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H^{(1)}_0$</td>
<td>0.0500</td>
<td>0.0475</td>
<td>0.0400</td>
</tr>
<tr>
<td>$H^{(2)}_0$</td>
<td>0.3075</td>
<td>0.3275</td>
<td>0.2925</td>
</tr>
<tr>
<td>$H^{(3)}_0$</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>$H^{(4)}_0$</td>
<td>0.9125</td>
<td>0.9350</td>
<td>0.9575</td>
</tr>
</tbody>
</table>

Table V: Testing Linear Drift Function for Short Rate

<table>
<thead>
<tr>
<th></th>
<th>$\lambda_n(h_1)$</th>
<th>$\lambda_n(h_2)$</th>
<th>$\lambda_n(h_3)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bootstrap P-value</td>
<td>0.151</td>
<td>0.112</td>
<td>0.103</td>
</tr>
<tr>
<td>Rejection Rate</td>
<td>0.08</td>
<td>0.10</td>
<td>0.09</td>
</tr>
</tbody>
</table>
Table VI: Testing Forms of Volatility Function for Short Rate

<table>
<thead>
<tr>
<th></th>
<th>GBM</th>
<th>VAS</th>
<th>CIR SR</th>
<th>CIR VR</th>
<th>CKLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_n(h_1)$</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Bootstrap P-value</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda_n(h_2)$</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>$\lambda_n(h_3)$</td>
<td>0.000</td>
<td>0.000</td>
<td>0.003</td>
<td>0.000</td>
<td>0.021</td>
</tr>
<tr>
<td>Rejection Rate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda_n(h_1)$</td>
<td>1</td>
<td>1</td>
<td>0.09</td>
<td>1</td>
<td>0.10</td>
</tr>
<tr>
<td>$\lambda_n(h_2)$</td>
<td>1</td>
<td>1</td>
<td>0.10</td>
<td>1</td>
<td>0.08</td>
</tr>
<tr>
<td>$\lambda_n(h_3)$</td>
<td>1</td>
<td>1</td>
<td>0.07</td>
<td>1</td>
<td>0.06</td>
</tr>
</tbody>
</table>

Table VII: Testing Linear Drift Function for logarithm of S&P 500 Index

<table>
<thead>
<tr>
<th></th>
<th>$\lambda_n(h_1)$</th>
<th>$\lambda_n(h_2)$</th>
<th>$\lambda_n(h_3)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bootstrap P-value</td>
<td>0.815</td>
<td>0.541</td>
<td>0.556</td>
</tr>
</tbody>
</table>

Table VIII: Testing Forms of Volatility Function for logarithm of S&P 500 Index

<table>
<thead>
<tr>
<th></th>
<th>GBM</th>
<th>VAS</th>
<th>CIR SR</th>
<th>CIR VR</th>
<th>CKLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_n(h_1)$</td>
<td>0</td>
<td>0.002</td>
<td>0.000</td>
<td>0</td>
<td>0.027</td>
</tr>
<tr>
<td>Bootstrap P-value</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda_n(h_2)$</td>
<td>0</td>
<td>0.311</td>
<td>0.003</td>
<td>0</td>
<td>0.440</td>
</tr>
<tr>
<td>$\lambda_n(h_3)$</td>
<td>0</td>
<td>0.486</td>
<td>0.189</td>
<td>0</td>
<td>0.567</td>
</tr>
</tbody>
</table>