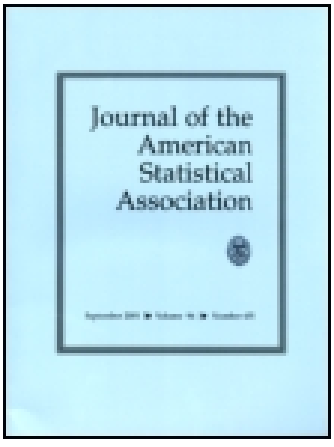

Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uasa20

A Reexamination of Diffusion Estimators With Applications to Financial Model Validation

Jianqing Fan[a] & Chunming Zhang[a]

[a] Jianqing Fan is Professor, Department of Statistics, University of North Carolina, Chapel Hill, NC and Professor of Statistics, Department of Statistics, Chinese University of Hong Kong . Fan's research was partially supported by National Science Foundation grants DMS-0196041 and DMS-0204329, a Research Grants Council direct grant, and the Research Grants Council grant CUHK4299/00P from the Hong Kong SAR. Chunming Zhang is Assistant Professor, Department of Statistics, University of Wisconsin, Madison, WI 53706 . The authors thank the editor, the associate editor, and the referee for their constructive comments and suggestions that have led to signi. cant improvement of this article.

Published online: 31 Dec 2011.

To cite this article: Jianqing Fan & Chunming Zhang (2003) A Reexamination of Diffusion Estimators With Applications to Financial Model Validation, Journal of the American Statistical Association, 98:461, 118-134, DOI: 10.1198/016214503388619157

To link to this article: http://dx.doi.org/10.1198/016214503388619157

A Reexamination of Diffusion Estimators With Applications to Financial Model Validation

Jianqing FAN and Chunming ZHANG

Time-homogeneous diffusion models have been widely used for describing the stochastic dynamics of the underlying economic variables. Recently, Stanton proposed drift and diffusion estimators based on a higher-order approximation scheme and kernel regression method. He claimed that "higher order approximations must outperform lower order approximations" and concluded nonlinearity in the instantaneous return function of short-term interest rates. To examine the impact of higher-order approximations, we develop general and explicit formulas for the asymptotic behavior of both drift and diffusion estimators. We show that these estimators will reduce the numerical approximation errors in asymptotic biases, but their asymptotic variances escalate nearly exponentially with the order of approximation. Simulation studies also confirm our asymptotic results. This variance inflation problem arises not only from nonparametric fitting, but also from parametric fitting. Stanton's work also postulates the interesting question of whether the short-term rate drift is nonlinear. Based on empirical simulation studies, Chapman and Pearson suggested that the nonlinearity might be spurious, due partially to the boundary effect of kernel regression. This prompts us to use the local linear fit based on the first-order approximation, proposed by Fan and Yao, to ameliorate the boundary effect and to construct formal tests of parametric financial models against the nonparametric alternatives. Our simulation results show that the local linear method indeed outperforms the kernel approach. Furthermore, our nonparametric "generalized likelihood ratio tests" are indeed versatile and powerful in detecting nonparametric alternatives. Using this formal testing procedure, we show that the evidence against the linear drift of the short-term interest rates is weak, whereas evidence against a family of popular models for the volatility function is very strong. Application to Standard & Poor 500 data is also illustrated.

KEY WORDS: Goodness of fit; Local polynomial regression; Markovian; Stochastic differential equation; Variance estimation.

1. INTRODUCTION

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

$$dX_t = \mu(X_t)\,dt + \sigma(X_t)\,dW_t, \tag{1}$$

where $\{W_t, a \le t \le T\}$ is a standard one-dimensional Brownian motion. Sufficient conditions due to Itô imposed on $\mu(\cdot)$ and $\sigma(\cdot)$ for the existence, uniqueness and a measurable Markov process of the diffusion solution, $\{X_t\}$, have been given by, for example, Wong (1971) and Kloeden and Platen (1992). Further regularity conditions for the stationarity of $\{X_t\}$ have been established by Banon (1978). This time-homogeneous diffusion model has been widely used for describing the stochastic dynamics of the underlying economic variables of many well-known single-factor financial models. Examples include the geometric Brownian motion (GBM),

$$dX_t = \mu X_t\,dt + \sigma X_t\,dW_t, \tag{2}$$

by Osborne (1959) for modeling stock price and models

$$\text{VAS}: dX_t = (\alpha_0 + \alpha_1 X_t)\,dt + \sigma\,dW_t, \tag{3}$$

$$\text{CIR SR}: dX_t = (\alpha_0 + \alpha_1 X_t)\,dt + \sigma X_t^{1/2}\,dW_t, \tag{4}$$

$$\text{CIR VR}: dX_t = \sigma X_t^{3/2}\,dW_t, \tag{5}$$

and

$$\text{CKLS}: dX_t = (\alpha_0 + \alpha_1 X_t)\,dt + \sigma X_t^{\gamma}\,dW_t, \tag{6}$$

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.

Current research including parametric approaches to estimating $\mu(\cdot)$ and $\sigma(\cdot)$ has been surveyed by Stanton (1997). To relax model assumptions and reduce possible modeling biases, nonparametric regression techniques have recently been studied in this area. Pham (1981) and Prakasa Rao (1985) proposed nonparametric drift estimators. Arfi (1995, 1998) showed that the Nadaraya–Watson (N-W) kernel estimator of drift is uniformly strongly consistent under ergodic conditions and reached the same conclusion for the kernel regression estimate of the diffusion function. Fan and Yao (1998) used local linear regression to the squared residuals for estimating $\sigma^2(\cdot)$ and showed that the proposed approach is efficient. Aït-Sahalia (1996) proposed a semiparametric procedure for estimating the diffusion function, under the parametric specification of the drift function. Jiang and Knight (1997) developed a nonparametric kernel estimator for the diffusion function, and then derived a consistent nonparametric drift estimator. Using an infinitesimal generator and Taylor series expansion, Stanton (1997) constructed the first-, second-, and third-order approximation formulas for $\mu(\cdot)$ and $\sigma(\cdot)$ and further claimed the superiority of higher-order approximations. These formulas contain unknown conditional expectations estimated by N-W kernel regression. Stanton's approach can estimate the diffusion function $\sigma(\cdot)$ separately without knowing or estimating $\mu(\cdot)$ a priori. This feature makes his method simple and attractive.

Jianqing Fan is Professor, Department of Statistics, University of North Carolina, Chapel Hill, NC and Professor of Statistics, Department of Statistics, Chinese University of Hong Kong (E-mail: *jfan@stat.unc.edu*). Fan's research was partially supported by National Science Foundation grants DMS-0196041 and DMS-0204329, a Research Grants Council direct grant, and the Research Grants Council grant CUHK4299/00P from the Hong Kong SAR. Chunming Zhang is Assistant Professor, Department of Statistics, University of Wisconsin, Madison, WI 53706 (E-mail: *cmzhang@stat.wisc.edu*). The authors thank the editor, the associate editor, and the referee for their constructive comments and suggestions that have led to significant improvement of this article.

Journal of the American Statistical Association
March 2003, Vol. 98, No. 461, Theory and Methods
DOI 10.1198/016214503388619157

Stanton's approach has some problems. Chapman and Pearson (2000) studied the finite-sample properties of Stanton's estimator. By applying his procedure to simulated sample paths of a squared-root diffusion, they found that Stanton's estimator produces spurious nonlinearity when the underlying drift function is truly linear. Chapman and Pearson nicely concluded that the "mean reversion" and small sample at the boundary create artificial patterns of nonlinearity displayed noticeably near the boundary regions. Meanwhile, two sensible questions naturally arise: (1) Do higher-order approximations outperform their lower-order counterparts? and (2) Are there any reasonable and formal procedures that help determine whether the observed nonlinearity in the drift is real or due to chance variation?

In an attempt to answer the first question on the order of approximations, we derive explicitly the formulas of a higher-order approximation scheme that generalizes Stanton's idea. We then compute explicitly the asymptotic variances of nonparametric estimators based on higher-order approximations. A striking result from our asymptotic study is that higher-order approximations will reduce the numerical approximation errors in asymptotic biases but escalate (nearly exponentially) the asymptotic variances. This variance inflation phenomenon is not only an artifact of nonparametric fitting—it also applies to parametric modeling. The issue of a trade-off between bias reduction and variance increment is made explicit in Theorem 4 (Sec. 2).

Stanton's work raises some other interesting issues. Is the drift function in the short-term rate model nonlinear? Or, more generally, does a parametric model fit a given set of economic or financial data? An example of this is whether models (2)–(6) adequately fit short-term rate data. Chapman and Pearson (2000) suggested that the nonlinearity of the drift function might be spurious. Their method is based on simulated data from diffusion models with a linear drift function and evaluates whether the estimated drift looks linear. This graphical procedure is useful, but informal. To set up formal statistical tests, an alternative hypothesis (model) is needed. Because we usually do not have strong preference for alternative competing models, the nonparametric model (1) serves as a natural candidate. The hypothesis testing problem becomes one of testing a parametric (or semiparametric) null hypothesis against a nonparametric alternative. The latter half of this article is thus devoted to model validation. There we extend the idea of the *generalized likelihood ratio* (GLR) statistic, developed by Fan, Zhang, and Zhang (2001), and apply it to the time-homogeneous diffusion models. Our simulation results show that GLR tests are indeed powerful and give the correct test size. They provide useful tools for assessing the validity of various models in economics and finance.

The remainder of the article is organized as follows. In Section 2 we discuss the distributional properties of Stanton's drift and diffusion estimators, and also derive explicit expressions of asymptotic biases and variances for higher-order approximations. To justify our analyses on empirical grounds, we report on simulations in Section 3. In Section 4 we propose model validation methods using the GLR test, based on the first-order approximation combined with the local linear estimation. Simulations of the GLR test and real data applications are also demonstrated. In Section 5 we briefly summarize our conclusions. We collect outlines of the proofs in the Appendix.

2. HIGHER-ORDER APPROXIMATIONS

This section begins with a description of Stanton's approach. Although his initial construction is based solely on the first-, second-, and third-order approximations, we can build, with some extra effort, a more general framework that gives us the flexibility to examine the impact of higher-order approximations.

2.1 Conditional Means and Conditional Variances of Higher-Order Differences

Following Stanton's notations, under appropriate conditions on $\mu(\cdot), \sigma(\cdot)$, and an arbitrary bivariate function $f(\cdot,\cdot)$, the conditional expectation $E_t\{f(X_{t+\Delta}, t+\Delta)\}$ can be expressed in the form of a Taylor series expansion,

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

as time increment $\Delta \downarrow 0$. Here the symbol E_t denotes the conditional expectation given X_t, and the infinitesimal generator, $\mathcal{L}$, of the process $\{X_t\}$, is defined by

$$\begin{aligned}\mathcal{L}f(x,t) &= \lim_{\tau \downarrow t} \frac{E\{f(X_\tau,\tau)|X_t = x\} - f(x,t)}{\tau - t} \\ &= \frac{\partial f(x,t)}{\partial t} + \frac{\partial f(x,t)}{\partial x}\mu(x) \\ &\quad + \frac{1}{2}\frac{\partial^2 f(x,t)}{\partial^2 x}\sigma^2(x) \end{aligned} \tag{7}$$

(see Øksendal 1985 for more details). Thus the first-order approximation formula for the target function, $\mathcal{L}f(X_t,t)$, is given by

$$\Delta^{-1}E_t\{f(X_{t+\Delta}, t+\Delta) - f(X_t,t)\} = \mathcal{L}f(X_t,t) + O(\Delta). \tag{8}$$

In particular, setting $f(x,t) = x$ (or $f(x,t) = x - X_t$) gives $\mathcal{L}f(x,t) = \mu(x)$; likewise, taking $f(x,t) = (x - X_t)^2$ implies $\mathcal{L}f(x,t) = 2(x - X_t)\mu(x) + \sigma^2(x)$, which at $x = X_t$ equals $\sigma^2(X_t)$. Hence these two special functions $f(\cdot,\cdot)$ can exactly recover $\mu(X_t)$ and $\sigma^2(X_t)$. In such cases, estimating the left side of (8) by the N-W kernel method leads to Stanton's estimates for $\mu(x)$ and $\sigma^2(x)$, based on the first-order approximation.

Higher-order approximations (or differences) can be achieved through a linear combination of terms on the left side of (8). More precisely, for any fixed integer $k \geq 1$, any sequence of constants $\{a_{k,j},\ j = 1,\ldots,k\}$, and any discretely observed time steps $j\Delta$, for $j = 1,\ldots,k$, we consider the

following linear combination:

$$\begin{aligned}\Delta^{-1}\sum_{j=1}^{k} a_{k,j}E_t\{f(X_{t+j\Delta}, t+j\Delta) - f(X_t, t)\} \\ = \left\{\sum_{j=1}^{k} j a_{k,j}\right\}\mathcal{L}f(X_t, t) + \left\{\sum_{j=1}^{k} j^2 a_{k,j}\right\}\frac{\mathcal{L}^2 f(X_t, t)}{2}\Delta \\ + \cdots + \left\{\sum_{j=1}^{k} j^k a_{k,j}\right\}\frac{\mathcal{L}^k f(X_t, t)}{k!}\Delta^{k-1} \\ + \left\{\sum_{j=1}^{k} j^{k+1} a_{k,j}\right\}\frac{\mathcal{L}^{k+1} f(X_t, t)}{(k+1)!}\Delta^k + O(\Delta^{k+1}).\end{aligned}$$

It is readily seen that a kth order approximation scheme,

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

is obtained 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 the solutions, $\{a_{k,j},\ j = 1, \ldots, k\}$, is presented in Theorem 1, the proof of which is given in the Appendix. Apparently, with orders $k = 1, 2, 3$, the values of $\{a_{k,j},\ j = 1, \ldots, k\}$ coincide with those derived by Stanton (1997)—namely, $\{1\}$ for $k = 1$, $\{2, -1/2\}$ for $k = 2$, and $\{3, -3/2, 1/3\}$ for $k = 3$.

Theorem 1. For each fixed integer $k \geq 1$, the unique solutions to the system of (9) are given by

$$a_{k,j} = (-1)^{j+1}\binom{k}{j}\Big/ j, \qquad j = 1, \ldots, k. \tag{10}$$

Furthermore, with these choices of $\{a_{k,j}\}_{j=1}^{k}$, we have

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

Therefore, using the foregoing unique solutions $(a_{k,1}, \ldots, a_{k,k})$, we obtain for $\mathcal{L}f(X_t, t)$ a general form of the kth order approximation formula,

$$\Delta^{-1}\sum_{j=1}^{k} a_{k,j}E_t\{f(X_{t+j\Delta}, t+j\Delta) - f(X_t, t)\}, \tag{11}$$

with the approximation error term expressed as

$$(-1)^{k+1}\frac{\mathcal{L}^{k+1} f(X_t, t)}{(k+1)}\Delta^k + O(\Delta^{k+1}). \tag{12}$$

Equations (11) and (12) imply that

$$\begin{aligned}\Delta^{-1}\sum_{j=1}^{k} a_{k,j}E_t(X_{t+j\Delta} - X_t) \\ = \mu(X_t) + \left[(-1)^{k+1}\frac{\mathcal{L}^{k+1} f_1(X_t, t)}{(k+1)}\Delta^k + O(\Delta^{k+1})\right]\end{aligned} \tag{13}$$

with the choice $f_1(x, t) = x$, and that

$$\begin{aligned}\Delta^{-1}\sum_{j=1}^{k} a_{k,j}E_t(X_{t+j\Delta} - X_t)^2 \\ = \sigma^2(X_t) + \left[(-1)^{k+1}\frac{\mathcal{L}^{k+1} f_2(X_t, t)}{(k+1)}\Delta^k + O(\Delta^{k+1})\right]\end{aligned} \tag{14}$$

with the choice $f_2(x, t) = (x - X_t)^2$. From (14), one can simply take the square root operation to obtain the kth-order approximation formula for the function $\sigma(X_t)$, such that

$$\sigma(X_t) = \left\{\Delta^{-1}\sum_{j=1}^{k} a_{k,j}E_t(X_{t+j\Delta} - X_t)^2\right\}^{1/2} + O(\Delta^k). \tag{15}$$

In addition, for each of the choices $f_\ell(x, t)$, $\ell = 1, 2$, the term $\mathcal{L}^{k+1} f_\ell(X_t, t)$ does not vanish and is independent of the time variable t. Therefore, the resulting numerical approximation errors for $\mu(\cdot), \sigma^2(\cdot)$, and $\sigma(\cdot)$ maintain, for any integer $k \geq 1$, the same convergence rates, $O(\Delta^k)$, to 0. Simulation comparisons of the first three order approximations with the true drift and diffusion functions, for the processes (3) and (4), were demonstrated in tables I–IV of Stanton (1997), whereas numerical comparisons conducted for the interest rate data were shown in his figures 4–7, along with the pointwise 95% confidence bands based only on the first-order approximation.

With the kth-order approximation formulas available for $\mu(\cdot)$ and $\sigma^2(\cdot)$, the involved conditional expectations remain to be estimated. Given the initial calendar time point t_0 and time series data $\{X_{t_0+i\Delta}, i = 1, \ldots, n\}$ observed at equally spaced time points, our first step is to form $(n-k)$ pairs of synthetic data,

$$\left(X_{t_0+i\Delta}, \Delta^{-1}\sum_{j=1}^{k} a_{k,j}\{X_{t_0+(i+j)\Delta} - X_{t_0+i\Delta}\}\right) \equiv (X^*_{i\Delta}, Y^*_{i\Delta}), \quad i = 1, \ldots, n-k, \tag{16}$$

for estimating $\mu(\cdot)$, together with

$$\left(X_{t_0+i\Delta}, \Delta^{-1}\sum_{j=1}^{k} a_{k,j}\{X_{t_0+(i+j)\Delta} - X_{t_0+i\Delta}\}^2\right) \equiv (X^*_{i\Delta}, Z^*_{i\Delta}), \quad i = 1, \ldots, n-k, \tag{17}$$

for estimating $\sigma^2(\cdot)$. Our second step is to use appropriate pointwise nonparametric regression estimators, $\hat{\mu}_{1,\Delta}(x_0)$ and $\hat{\mu}_{2,\Delta}(x_0)$, for estimating the conditional expectations

$$E(Y^*_{i\Delta}|X^*_{i\Delta} = x_0) = \mu(x_0) + O(\Delta^k) \quad \text{and} \quad E(Z^*_{i\Delta}|X^*_{i\Delta} = x_0) = \sigma^2(x_0) + O(\Delta^k), \tag{18}$$

from (13) and (14).

Table 1. Variance Inflation Factors Using Higher-Order Differences

	Order k									
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
$V_1(k)$	1.00	2.50	4.83	9.25	18.95	42.68	105.49	281.65	798.01	2,364.63
$V_2(k)$	1.00	3.00	8.00	21.66	61.50	183.40	570.66	1,837.28	6,076.25	20,527.22

There are many nonparametric methods for estimating the conditional expectations in (18); the N-W estimator is the simplest. 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 $\hat{\mu}_{2,\Delta}(x_0)$ for an interior point x_0, via the qth-degree local polynomial estimation ($q \geq 0$); the N-W estimator corresponds to the local constant method with degree $q = 0$. We now briefly describe the technique for estimating $E(Y^*_{i\Delta}|X^*_{i\Delta} = x_0)$. By a Taylor series expansion, a smooth function $m(x) = E(Y^*_{i\Delta}|X^*_{i\Delta} = x)$, with x located in a neighborhood of x_0, can be locally approximated by a qth-degree polynomial, that is,

$$m(x) \approx m(x_0) + (x - x_0)m'(x_0) + \cdots + (x - x_0)^q m^{(q)}(x_0)/q!.$$

Denote the coefficient vector by $\boldsymbol{\beta}(x_0) = (m(x_0), m'(x_0), \ldots, m^{(q)}(x_0)/q!)^T = (\beta_0, \beta_1, \ldots, \beta_q)^T$. Then the local polynomial estimator $\hat{\boldsymbol{\beta}}(x_0)$, of the qth degree, is determined by the minimizer of the residual sum of squares between $Y^*_{i\Delta}$ and the local model on $m(X^*_{i\Delta})$, weighted by the distance of $X^*_{i\Delta}$ from the fitting point x_0. Formally, $\hat{\boldsymbol{\beta}}(x_0)$ minimizes the objective function

$$\sum_{i=1}^{n-k}\{Y^*_{i\Delta} - \beta_0 - (X^*_{i\Delta} - x_0)\beta_1 - \cdots - (X^*_{i\Delta} - x_0)^q\beta_q\}^2 K_h(X^*_{i\Delta} - x_0) \quad (19)$$

over values of $\boldsymbol{\beta}(x_0)$, where $K_h(\cdot) = K(\cdot/h)/h$. Here $K(\cdot)$ and h are referred to as the kernel function and the bandwidth (or smoothing parameter). The first component of the vector $\hat{\boldsymbol{\beta}}(x_0)$ gives $\hat{\mu}_{1,\Delta}(x_0)$, the qth degree local polynomial estimate of $E(Y^*_{i\Delta}|X^*_{i\Delta} = x_0)$. A similar procedure can be applied to obtain the qth degree local polynomial estimate $\hat{\mu}_{2,\Delta}(x_0)$ of $E(Z^*_{i\Delta}|X^*_{i\Delta} = x_0)$. For practical application, Fan and Gijbels (1996) recommended the use of local linear fit ($q = 1$).

Because any nonparametric regression procedure is in essence a weighted average of local data, its performance always depends on the local variation, namely the conditional variance. For our current applications, based on the synthetic data, the corresponding conditional variances are

$$\sigma^2_{1,\Delta}(x_0) = \mathrm{var}(Y^*_{i\Delta}|X^*_{i\Delta} = x_0) \quad \text{and} \quad \sigma^2_{2,\Delta}(x_0) = \mathrm{var}(Z^*_{i\Delta}|X^*_{i\Delta} = x_0). \quad (20)$$

Theorem 2, proved in the Appendix, summarizes the magnitudes of $\sigma^2_{1,\Delta}(x_0)$ and $\sigma^2_{2,\Delta}(x_0)$. Note that some regularity conditions (see, e.g., Wong 1971, chapter 4, prop. 4.1) put on $\mu(\cdot), \sigma(\cdot)$, and X_{t_0} for the unique existence and Markov process of $\{X_t\}$ in (1) are always assumed implicitly in Theorems 2 and 4.

Theorem 2. Assume that $\{X_t\}$ is a Markov process. Let $\boldsymbol{A}_{1,k}$ and $\boldsymbol{A}_{2,k}$ be $k \times k$ matrices with (i, j)th entry equal to $\min(i, j)$ and $\min(i^2, j^2)$, and let $\boldsymbol{\alpha}_k$ be a $k \times 1$ vector, the jth element of which is given in (10). Denote $V_1(k) = \boldsymbol{\alpha}_k^T \boldsymbol{A}_{1,k}\boldsymbol{\alpha}_k$ and $V_2(k) = \boldsymbol{\alpha}_k^T \boldsymbol{A}_{2,k}\boldsymbol{\alpha}_k$. Then as $\Delta \to 0$, the conditional variance of the kth-order difference formula for $\mu(x_0)$ is given by

$$\sigma^2_{1,\Delta}(x_0) = \sigma^2(x_0)V_1(k)\Delta^{-1}\{1 + O(\Delta)\}, \quad (21)$$

whereas the conditional variance of the kth order difference formula for $\sigma^2(x_0)$ is given by

$$\sigma^2_{2,\Delta}(x_0) = 2\sigma^4(x_0)V_2(k)\{1 + O(\Delta)\}. \quad (22)$$

The factors $V_1(k)$ and $V_2(k)$ reflect the premium that higher-order approximations must pay. For this reason, we call them the *variance inflation factors* for using higher-order approximations. To provide some numerical impression, Table 1 summarizes the numerical values of $V_1(k)$ and $V_2(k)$ for approximations of orders up to the 10th. For visual assessment, Figure 1 contains plots of $\log\{V_1(k)\}$ and $\log\{V_2(k)\}$ versus order k. The overall impacts of higher-order approximations on variance inflation are striking.

It is also notable from Table 1 and Figure 1 that the variance inflation factors grow nearly exponentially fast as the order k increases. This relation can indeed be verified analytically, as shown in the following theorem.

Theorem 3. (a) For $k \geq 1$, the factor $V_1(k)$ in (21) is bounded below by

$$\frac{k^2 - 3k - 2}{k(k+1)^3}\binom{2k}{k} + \frac{2}{k} + 2\sum_{j=1}^{k}\frac{1}{j} - \frac{2k^2 + 4k + 3}{(k+1)^2} \approx \frac{4^k}{\pi^{1/2}k^{5/2}},$$

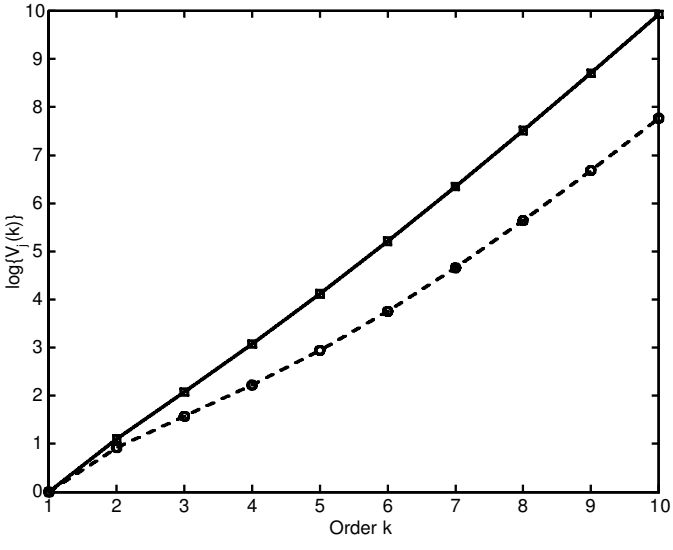

Figure 1. Theoretical Values of $\log\{V_j(k)\}$ Versus Order k. The factors $V_j(k)$ are given in Theorem 2, where $j = 1$ (-∘-) refers to drift $\mu(\cdot)$ and $j = 2$ (–⊟–) refers to squared diffusion $\sigma^2(\cdot)$.

and bounded above by

$$\frac{5k^2-k-2}{k(k+1)^3}\binom{2k}{k}+\frac{2}{k}+2\sum_{j=1}^{k}\frac{1}{j}-\frac{3k^2+6k+5}{(k+1)^2}\approx\frac{5\times 4^k}{\pi^{1/2}k^{5/2}}.$$

(b) For $k>1$, the factor $V_2(k)$ in (22) is given by

$$V_2(k)=\frac{\binom{2k}{k}-(k+1)}{k-1}\approx\frac{4^k}{\pi^{1/2}k^{3/2}}.$$

2.2 Asymptotic Behavior of Nonparametric Estimators

The asymptotic bias and variance of the pointwise drift estimator $\hat{\mu}_{1,\Delta}(x_0)$ and the squared diffusion function estimator $\hat{\mu}_{2,\Delta}(x_0)$, based on the kth order approximation scheme and the qth degree local polynomial fitting, are presented in Theorem 4. The results demonstrate that higher-order differences result in reductions of the asymptotic bias, while translating the variance inflation into the asymptotic variance of the associated nonparametric drift and diffusion estimators.

We first introduce some notations and definitions. Set $f_1(x,t)=x, f_2(x,t)=(x-X_t)^2, \mu_j=\int u^jK(u)\,du, \nu_j=\int u^jK^2(u)\,du, \boldsymbol{e}_1=(1,0,\ldots,0)^T, S=(\mu_{i+j-2})_{i,j=1,\ldots,q+1}$, $S^*=(\nu_{i+j-2})_{i,j=1,\ldots,q+1}, \boldsymbol{c}_q=(\mu_{q+1},\ldots,\mu_{2q+1})^T$, and $\tilde{\boldsymbol{c}}_q=(\mu_{q+2},\ldots,\mu_{2q+2})^T$. For integers $\ell>0$, let $p_\ell(y|x)$ denote the conditional probability density of $X_{t_0+(\ell+1)\Delta}$ given $X_{t_0+\Delta}$.

Theorem 4. Let $\{X_{t_0+i\Delta}, i=1,\ldots,n-k\}$ be a sequence of observations on a stationary Markov process with a bounded continuous density $p(\cdot)$. Assume that $p_\ell(y|x)$ is continuous in the variables (y,x) and is bounded by a constant independent of ℓ. The sequence $\{X_{t_0+i\Delta}, i=1,\ldots,n-k\}$ satisfies the stationarity conditions of Banon (1978) and the G_2 condition of Rosenblatt (1970) on the transition operator. Assume that the kernel K is a bounded symmetric probability density function with bounded support. Suppose that x_0 is any given point in the interior of the support of p where $p(x_0)>0, \sigma^2(x_0)>0$, and that $\mu^{(q+1)}(\cdot)$ and $(\sigma^2)^{(q+1)}(\cdot)$ are continuous in a neighborhood of x_0. Put $\sigma_1^2(x_0;k)=\sigma^2(x_0)V_1(k)$ and $\sigma_2^2(x_0;k)=2\sigma^4(x_0)V_2(k)$. Let $n\to\infty$, such that $h\to 0$ and $nh\to\infty$, and $\Delta\to 0$, then at any time $t=t_0+i\Delta, i=1,\ldots,n-k$,

(a) The asymptotic bias of $\hat{\mu}_{1,\Delta}(x_0)$ for odd degrees q is given by

$$(-1)^{k+1}\frac{\mathcal{L}^{k+1}f_1(x_0,t)}{(k+1)}\Delta^k+O(\Delta^{k+1})+\boldsymbol{e}_1^TS^{-1}\boldsymbol{c}_q\frac{\mu^{(q+1)}(x_0)}{(q+1)!}h^{q+1}+o_P(h^{q+1}),\quad(23)$$

whereas for even degrees q, the last two terms in (23) become

$$\frac{\boldsymbol{e}_1^TS^{-1}\tilde{\boldsymbol{c}}_q}{(q+2)!}\{\mu^{(q+2)}(x_0)+(q+2)\mu^{(q+1)}(x_0)\times p'(x_0)/p(x_0)\}h^{q+2}+o_P(h^{q+2}),\quad(24)$$

provided that $p'(\cdot)$ and $\mu^{(q+2)}(\cdot)$ are continuous in a neighborhood of x_0 and $nh^3\to\infty$. Assume further that $h=O(\Delta^{1/2})$; then the asymptotic variance is

$$(nh\Delta)^{-1}\boldsymbol{e}_1^TS^{-1}S^*S^{-1}\boldsymbol{e}_1\sigma_1^2(x_0;k)/p(x_0)\{1+o(1)\}.\quad(25)$$

(b) The asymptotic bias of $\hat{\mu}_{2,\Delta}(x_0)$ for odd degrees q is given by

$$(-1)^{k+1}\frac{\mathcal{L}^{k+1}f_2(x_0,t)}{(k+1)}\Delta^k+O(\Delta^{k+1})+\boldsymbol{e}_1^TS^{-1}\boldsymbol{c}_q\frac{(\sigma^2)^{(q+1)}(x_0)}{(q+1)!}h^{q+1}+o_P(h^{q+1}),\quad(26)$$

whereas for even degrees q, the last two terms in (26) become

$$\frac{\boldsymbol{e}_1^TS^{-1}\tilde{\boldsymbol{c}}_q}{(q+2)!}\{(\sigma^2)^{(q+2)}(x_0)+(q+2)(\sigma^2)^{(q+1)}(x_0)\times p'(x_0)/p(x_0)\}h^{q+2}+o_P(h^{q+2}),\quad(27)$$

provided that $p'(\cdot)$ and $(\sigma^2)^{(q+2)}(\cdot)$ are continuous in a neighborhood of x_0 and $nh^3\to\infty$. Assume further that $h=O(\Delta^{1/4})$; then the asymptotic variance is

$$(nh)^{-1}\boldsymbol{e}_1^TS^{-1}S^*S^{-1}\boldsymbol{e}_1\sigma_2^2(x_0;k)/p(x_0)\{1+o(1)\}.\quad(28)$$

It is clearly observed from (23) that the bias of $\hat{\mu}_{1,\Delta}(x_0)$ is composed of a numerical approximation error, expressed by $E(Y_{i\Delta}^*|X_{i\Delta}^*=x_0)-\mu(x_0)$, in addition to the usual nonparametric estimation bias, $\hat{\mu}_{1,\Delta}(x_0)-E(Y_{i\Delta}^*|X_{i\Delta}^*=x_0)$. Results of (23) and (24) indicate that for the kernel estimator used by Stanton (1997), the leading term of its asymptotic bias is

$$(-1)^{k+1}\frac{\mathcal{L}^{k+1}f_1(x_0,t)}{(k+1)}\Delta^k+\frac{\mu_2}{2}h^2\{\mu''(x_0)+2\mu'(x_0)p'(x_0)/p(x_0)\},\quad(29)$$

whereas for the local linear method, the second term becomes $2^{-1}\mu_2h^2\mu''(x_0)$. A similar comparison can be made for $\hat{\mu}_{2,\Delta}(x_0)$.

Remark 1. As shown by Banon and Nguyen (1981, lemma 2.1), a stationary Markov process satisfying a certain mixing condition, namely the G_2 condition of Rosenblatt (1970), is asymptotically uncorrelated (Rosenblatt 1971). Therefore, the "big-block and small-block" arguments similar to those used by Fan and Gijbels (1996, theorem 6.1) can be incorporated to show the asymptotic normality of $\hat{\mu}_{1,\Delta}(x_0)$ and $\hat{\mu}_{2,\Delta}(x_0)$. The lengthy details are omitted here.

Remark 2. The conclusions of Theorems 2 and 3 do not depend on the stationarity condition. The stationarity condition in Theorem 4 is imposed to facilitate technical manipulations; it is not a necessary condition. The stationarity condition possibly can be relaxed.

3. SIMULATIONS

Realistically, we do not know whether the stationary Markovian assumption remains valid for financial data recorded at discrete time points. We also do not know whether the asymptotic results reflect reality. Nevertheless, we can still carry out the drift and diffusion estimations using higher-order approximations and nonparametric regression techniques. This will enable us to assess empirically how our asymptotic results are reflected in finite samples. Our simulation studies show the fact that the variance inflation due to higher-order approximations is reflected in finite samples.

3.1 Cox–Ingersoll–Ross Squared-Root Diffusion

As a first illustration, we consider the well-known Cox–Ingersoll–Ross (CIR) model for interest rate term structure,

$$dX_t = \kappa(\theta - X_t)\,dt + \sigma X_t^{1/2}\,dW_t, \qquad t \geq t_0, \tag{30}$$

where the spot rate, X_t, moves around its long-run equilibrium level θ at speed κ. When the condition $2\kappa\theta \geq \sigma^2$ holds, this process is shown to be positive and stationary. Provided that the time step size Δ is small, we can use the discrete-time order 1.0 strong approximation scheme given in (3.14) of Kloeden, Platen, Schurz, and Sørensen (1996). In this example, the scheme takes the form

$$\begin{aligned} X_{t_{i+1}} \approx\ & X_{t_i} + \{\kappa(\theta - X_{t_i}) - 4^{-1}\sigma^2\}\Delta \\ & + 2^{-1}\sigma\Big[\{X_{t_i} + (\kappa\theta - \kappa X_{t_i} - 4^{-1}\sigma^2)\Delta \\ & \quad + \sigma (X_{t_i})_+^{1/2}\,\varepsilon_i\sqrt{\Delta}\}_+^{1/2} + (X_{t_i})_+^{1/2}\Big]\varepsilon_i\sqrt{\Delta}, \end{aligned} \tag{31}$$

for $1 \leq i \leq n-1$, where $\varepsilon_i \overset{iid}{\sim} N(0,1)$ and $x_+ = \max(x, 0)$. Alternatively, one might use the transition density properties of the process (see Cox et al. 1985). That is, given $X_t = x$ at the current time t, the variable $2cX_s$ at the future time s has a noncentral chi-squared distribution with degrees of freedom $2q+2$ and noncentrality parameter $2u$, where $q = 2\kappa\theta/\sigma^2 - 1$, $u = cxe^{-\kappa(s-t)}$, and $c = \frac{2\kappa}{\sigma^2\{1-e^{-\kappa(s-t)}\}}$. The initial value of X_{t_0} can be generated from the steady-state gamma distribution of $\{X_t\}$, with the probability density $p(y) = \omega^\nu/\Gamma(\nu)y^{\nu-1}e^{-\omega y}$, where $\nu = 2\kappa\theta/\sigma^2$ and $\omega = 2\kappa/\sigma^2$. For each simulation experiment, we generate a sample path of length 10,000 and compute, based on the synthetic data [see (16) and (17)], Stanton's kernel drift estimate $\hat{\mu}_{1,\Delta}(x_0)$, and the squared diffusion estimate $\hat{\mu}_{2,\Delta}(x_0)$. We replicate the experiments 1,000 times, and calculate the sample variances of $\{\hat{\mu}_{1,\Delta}(x_0)\}$ and $\{\hat{\mu}_{2,\Delta}(x_0)\}$ across these 1,000 simulations respectively.

Choices of kernel function depend purely on individual preferences. Throughout our numerical work in this article, we use the Epanechnikov kernel, defined by $K(u) = 3/4(1-u^2)I(|u| \leq 1)$, where $I(\cdot)$ stands for the indicator function. For a given kernel function, the choice of an effective bandwidth parameter is very important to the performance of a nonparametric regression estimator. It is often selected through either visual inspection of the resulting smooths or a data-driven technique. Popular data-dependent approaches include cross-validation (Allen 1974; Stone 1974), generalized cross-validation (Wahba 1977), the preasymptotic substitution method (Fan and Gijbels 1995), the plug-in method (Ruppert, Sheather, and Wand 1995), and the empirical bias method (Ruppert 1997). These techniques provide various useful means for automatic bandwidth selection, but involve intensive computation and extra effort to program. A more detailed look at these methods, regarding theoretical properties and implementations, was given by Fan and Gijbels (1996). Alternatively, a simple rule of thumb bandwidth formula, such as

$$h = \text{constant} \times \text{std}(\{X_\Delta^*, \ldots, X_{(n-k)\Delta}^*\})\, n^{-1/5}, \tag{32}$$

also can be used. To show the occurrence of variance inflation with order k, by finite-sample simulation, an appropriate choice of bandwidth is constant-valued and independent of k, even though the optimal bandwidth may depend on k. For the purpose of illustration, we set $h = .004$ in this example. Other choices of bandwidth have also been tried, and the results have been similar.

In our implementation, the values of the model parameters are cited from Chapman and Pearson (2000), that is, $\kappa = .21459$, $\theta = .08571$, $\sigma = .07830$, and $\Delta = 1/250$. To differentiate the effects of the higher-order approximation scheme from the boundary effects of the kernel estimator, we focus on an interior state point, $x_0 = .1$. The natural logarithms of the simulated variance ratios of $\hat{\mu}_{1,\Delta}(.1)$ and $\hat{\mu}_{2,\Delta}(.1)$, based on higher-order difference, to those of their first-order counterparts, are displayed in Figure 2, where plot (a) is based on sample paths generated from the conditional chi-squared distribution and plot (b) results from the discretization scheme (31). Meanwhile, for the purpose of comparison, we also present, in plots (a′) and (b′), the corresponding results by local linear estimation. All plots mimic (except in amplitude) our theoretical results shown in Figure 1.

3.2 Geometric Brownian Motion

We include another familiar example of geometric Brownian motion determined by

$$dX_t = (\mu + 2^{-1}\sigma^2)X_t\,dt + \sigma X_t\,dW_t, \qquad 0 \leq t \leq T. \tag{33}$$

Apparently from its construction, both the drift and diffusion are linear, and thus $\{X_t\}$ is Markovian (see Wong 1971, prop. 4.1), but the technical assumption of stationarity is violated. This model is incorporated to illustrate that the conclusion of Theorem 4 extends to more general diffusion processes.

For (33), we simulate in time interval $[0, T]$ with $T = 10$, the corresponding approximate process with parameters $\mu = .087$ and $\sigma = .178$ starting at $X_0 = 1$. We choose the order 1.0 scheme

$$\begin{aligned} X_{t_{i+1}} \approx\ & X_{t_i} + (\mu + 2^{-1}\sigma^2)X_{t_i}\Delta + \sigma X_{t_i}\varepsilon_i\sqrt{\Delta} \\ & + 2^{-1}\sigma^2 X_{t_i}(\varepsilon_i^2 - 1)\Delta \end{aligned} \tag{34}$$

given in (3.5) of Kloeden et al. (1996). Alternatively, we could directly use the explicit solution $X_t = X_0\exp\{\mu t + \sigma W_t\}$ for (33). For both schemes, 1,000 sample paths of length 1,000 are generated. The bandwidth parameter, $h = .04$, is used for

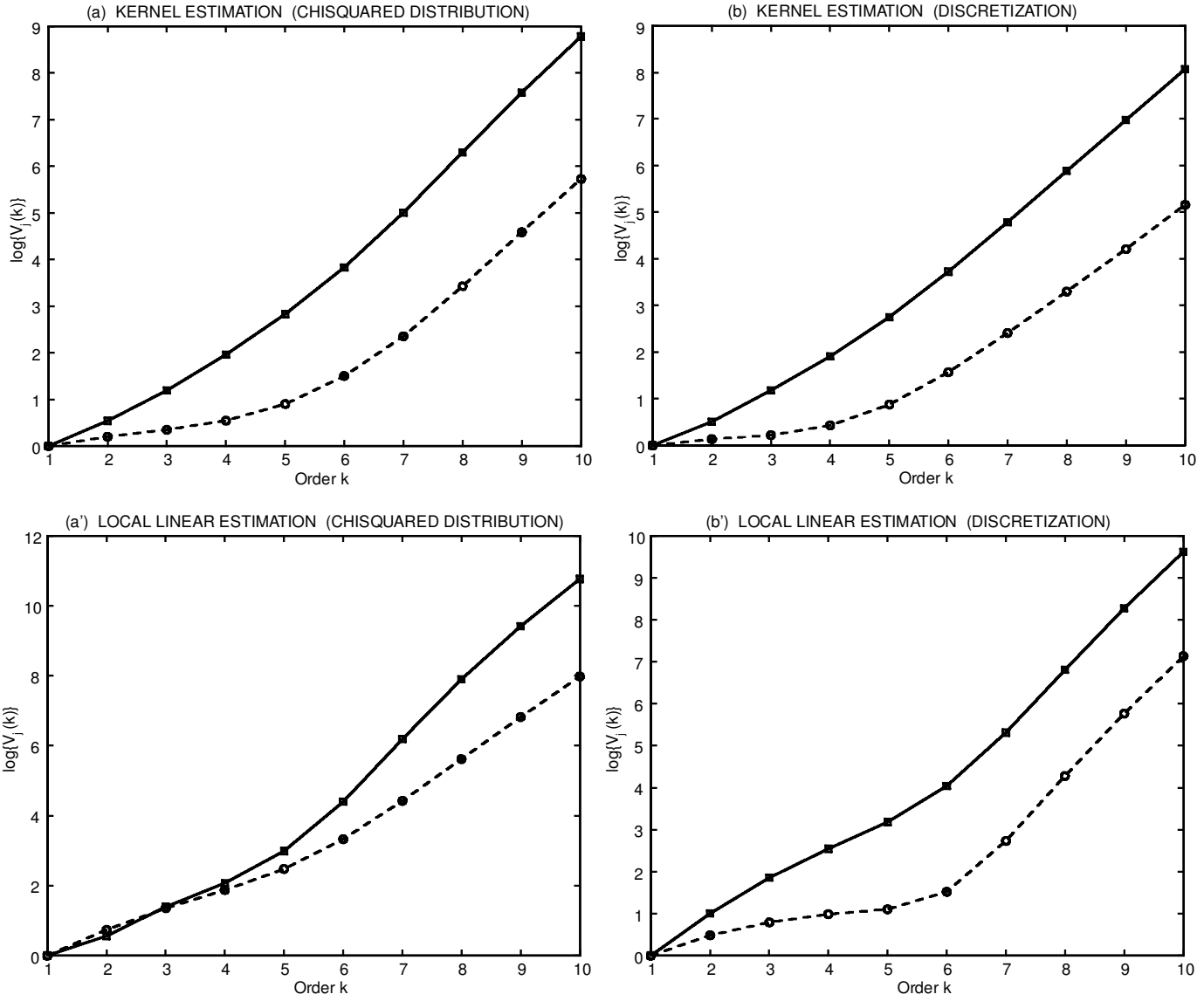

Figure 2. Simulated Values of $\log\{V_j(k)\}$ Versus Order k for CIR Model $dX_t = .21459(.08571 - X_t)\,dt + .07830X_t^{1/2}\,dW_t$. The index $j = 1$ (-○-) refers to the drift estimator $\hat{\mu}_{1,\Delta}(.1)$; $j = 2$ (–⊟) refers to the squared diffusion estimator $\hat{\mu}_{2,\Delta}(.1)$. Plots (a) and (a′) are based on the same sets of sample paths generated by the noncentral chi-squared distribution, whereas plots (b) and (b′) are based on the same sets of sample paths generated by the discretization scheme (31).

local smoothing. Again, this number serves for the sake of illustration. For the same reason stated in the previous example, we restrict attention to the state value $x_0 = 1.0$, simply because more data points fall within its local region. Figure 3 displays similar types of plots as those shown in Figure 2. For comparison, plots (a) and (a′) are based on data generated from the exact solution, and plots (b) and (b′) depend on the discretization scheme (34). Again, all plots in Figure 3 support our theoretical results in Figure 1, although we used a smaller sample size and lower sampling frequency than those in the preceding example of the CIR model.

3.3 Local Linear Fit: Boundary Correction

Overall, the foregoing simulation studies present convincing evidence that, at least for models similar to those two types, the higher-order approximations substantially amplify variances. As discussed in Section 2, this phenomenon always occurs, regardless of the method used for nonparametric regression. It is well known that the kernel regression estimator can create boundary biases. In contrast, the local linear estimator enjoys the theoretical advantages of design adaptation, automatic boundary correction, and minimax efficiency (see Fan and Gijbels 1996 for further details). This naturally leads us to substitute kernel estimation by local linear estimation. A similar application of local linear fit to the first-order approximation of continuous-time diffusion models was used by Fan and Yao (1998), who also suggested correcting the drift term before the variance estimation.

To examine the performance of local linear estimation of diffusion models, we revisit the CIR square-root diffusion model discussed in Section 3.1. We adopt the same values of model parameters κ, θ, and σ to generate, with weekly frequency, sample paths of length 5,000, using the (noncentral chi-squared) transition density. To conduct kernel and local linear fits, based on the first-order synthetic data, a scale constant, 6, is used in the empirical bandwidth formula (32).

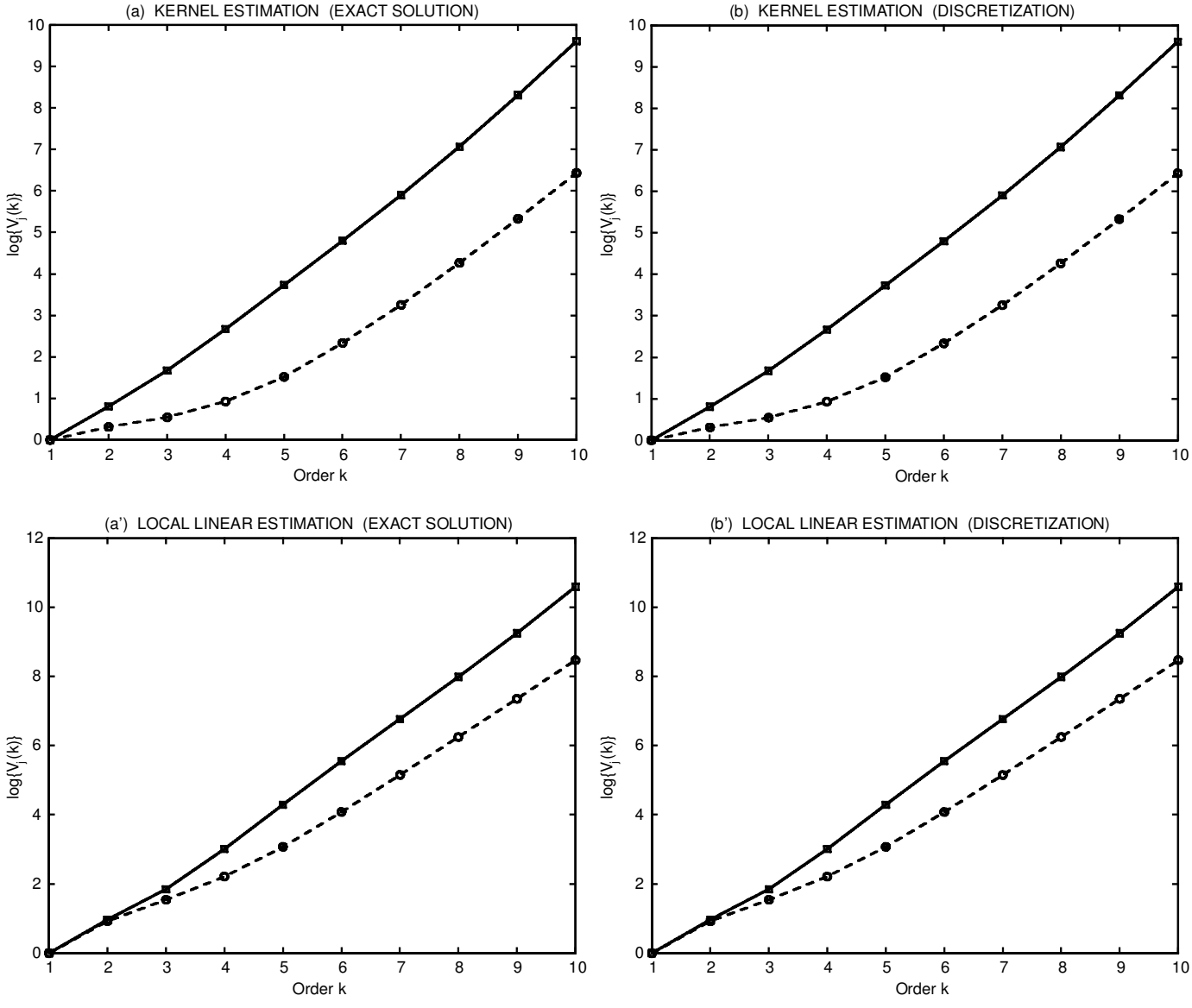

Figure 3. Simulated Values of $\log\{V_j(k)\}$ Versus Order k for Geometric Brownian Motion $dX_t = (.087 + .178^2/2)X_t\,dt + .178X_t\,dW_t$. The index $j = 1$ (-∘-) refers to the drift estimator $\hat{\mu}_{1,\Delta}(1.0)$; $j = 2$ (–⊟) refers to the squared diffusion estimator $\hat{\mu}_{2,\Delta}(1.0)$. Plots (a) and (a′) are based on the same sets of sample paths generated from the exact solution $X_t = X_0 \exp\{.087t + .178W_t\}$, whereas plots (b) and (b′) are based on the same sets of sample paths generated by the discretization scheme (34).

For individual simulated trajectories, we compared the estimated drift and diffusion, for which we observed that in most cases the local linear approach is superior to the kernel method. In fact, according to Fan (1992), the local linear fit has a better bias-correction property than the kernel method. Thus, as the bandwidth gets larger, the outperformance of the local linear fit over the kernel method can become even more dramatic. In contrast, the sample ranges of $\{X_t\}$ vary considerably across different simulations. Extremely high levels of those states x (e.g., .20) rarely occur in reality or are visited in practical simulations. To conduct more sensible comparisons, we simulate 101 sample paths with range interval $\mathcal{I} = [.03, .15]$. The drift and diffusion are estimated for each realization, and the 25th and 75th percentiles (dashed curves) and the median (dash dotted curves) of the estimates, over the 101 realizations, are presented in Figure 4. Similar graphs using discretization schemes such as (31) are omitted here. For the volatility estimates, we find that the local linear method achieves more gains in alleviating the impact of "boundary effects" than the kernel counterpart. The same conclusion applies to estimation of the drift function. The wider bands of the interquartile ranges of the drift estimates compared to those of the diffusion estimates can be easily understood from Theorem 4, which states that the estimates of drift are more variable than the estimates of diffusion. Furthermore, this necessitates the importance of developing formal procedures for model validation.

4. MODEL VALIDATION

Model diagnosis plays an important role in examining the relevance of specific assumptions underlying the modeling process and in 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

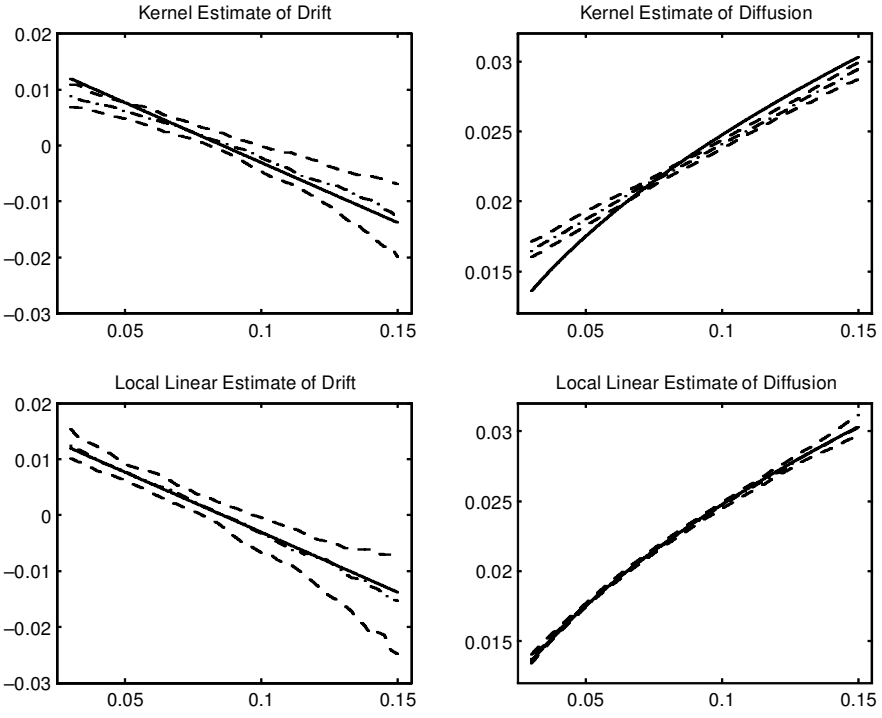

Figure 4. Estimated Drift and Diffusion Functions for CIR Model $dX_t = .21459(.08571 - X_t)\,dt + .07830X_t^{1/2}\,dW_t$. The solid curves are the true functions, the dashed-dotted curves denote the medians of the estimates, and the dashed curves correspond to the 25th and 75th sample percentiles of the estimates, over simulated data (101 replications). The sample paths are generated by the transitional noncentral chi-squared distribution.

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 in terms of their relative performances under a unified parametric framework,

$$dX_t = (\alpha + \beta X_t)\,dt + \sigma X_t^{\gamma}\,dW_t, \tag{35}$$

in Chan et al. (1992). The generalized method of moments of Hansen (1982) is frequently used to estimate the parameters. However, the question frequently arises whether model (35) itself correctly captures the stochastic dynamics of a given set of economic data. To address this issue, we need an alternative family of stochastic models. Nonparametric models offer a very nice solution to this problem. Depending on the cases and the natures of model validation, the alternative nonparametric models can be of the form

$$dX_t = \mu(X_t)\,dt + \sigma X_t^{\gamma}\,dW_t, \tag{36}$$

$$dX_t = (\alpha + \beta X_t)\,dt + \sigma(X_t)\,dW_t, \tag{37}$$

or the more generic model (1), which places no particular restriction on either the structural shift or volatility. These kinds of hypothesis testing problems often arise in financial modeling.

In this section, we first describe approaches used for estimating parameters of models (35)–(37). To testify against these models (null hypotheses), we treat model (1) as our alternative hypothesis. We propose new hypothesis-testing procedures based on the "generalized likelihood ratio" by Fan et al. (2001), and demonstrate the explanatory power and versatility of the GLR tests by simulations and two sets of real data.

4.1 Parametric Estimation

For ease of exposition, we proceed from the parametric model (35). Given discretely sampled observations $\{X_{t_i}, i = 1, \ldots, n\}$ from this model, denote $\Delta_i = t_{i+1} - t_i$ and $Y_{t_i} = X_{t_{i+1}} - X_{t_i}$, for $1 \le i \le n-1$. Then the parameters α, β, σ, and γ can be estimated through a discrete-time specification

$$Y_{t_i} \approx (\alpha + \beta X_{t_i})\Delta_i + \sigma X_{t_i}^{\gamma}\varepsilon_i\sqrt{\Delta_i}, \quad i = 1, \ldots, n-1, \tag{38}$$

where $\varepsilon_i \overset{iid}{\sim} N(0,1)$. Three steps summarize the estimation procedure:

Step I: Pretend that model (38) is homoscedastic, and obtain the least squares estimates of (α, β), denoted by $(\hat{\alpha}^{(1)}, \hat{\beta}^{(1)})$.

Step II: Let $\hat{e}_{t_i} = \{Y_{t_i} - (\hat{\alpha}^{(1)} + \hat{\beta}^{(1)}X_{t_i})\Delta_i\}/\Delta_i^{1/2}$, which transforms model (38) into

$$\log(\hat{e}_{t_i}^2) \approx \log(\sigma^2) + \gamma\log(X_{t_i}^2) + \log(\varepsilon_i^2), \quad i = 1, \ldots, n-1. \tag{39}$$

Obtain least squares estimates $(\hat{\sigma}^{(1)}, \hat{\gamma}^{(1)})$ of (σ, γ) after subtracting $E\{\log(Z^2)\} \approx -1.270362845$ from both sides of model (39), where $Z \sim N(0,1)$.

Step III (optional): Substitute $(\hat{\sigma}^{(1)}, \hat{\gamma}^{(1)})$ into model (38) and get weighted least squares estimates of (α, β), denoted by $(\hat{\alpha}^{(2)}, \hat{\beta}^{(2)})$. Meanwhile, get updated estimates $(\hat{\sigma}^{(2)}, \hat{\gamma}^{(2)})$ at step II.

This approach can be flexibly modified. For instance, the diffusion parameters σ and γ in model (36) could be estimated

Table 2. Parameter Estimates and Standard Errors (%, in brackets) for the CIR Model: $dX_t = (\alpha + \beta X_t)\,dt + \sigma X_t^{\gamma}\,dW_t$, Where $\alpha = .0183925, \beta = -.21459, \sigma = .0783$, and $\gamma = .5$

n	$\hat{\alpha}^{(1)}$	$\hat{\alpha}^{(2)}$	$\hat{\beta}^{(1)}$	$\hat{\beta}^{(2)}$	$\hat{\sigma}^{(1)}$	$\hat{\sigma}^{(2)}$	$\hat{\sigma}^{(3)}$	$\hat{\gamma}^{(1)}$	$\hat{\gamma}^{(2)}$	$\hat{\gamma}^{(3)}$
5,000	.0224	.0217	−.2620	−.2534	.0782	.0783	.0781	.4979	.4983	.4976
	(.72)	(.65)	(8.47)	(7.76)	(.82)	(.81)	(.81)	(4.05)	(4.02)	(4.00)
10,000	.0205	.0200	−.2385	−.2328	.0778	.0779	.0777	.4971	.4974	.4968
	(.45)	(.40)	(5.44)	(4.86)	(.56)	(.57)	(.55)	(2.83)	(2.84)	(2.78)

directly from step II, except for setting $\hat{e}_{t_i}$ in (39) to $\{Y_{t_i} - \hat{\mu}(X_{t_i})\Delta_i\}/\Delta_i^{1/2}$, where $\hat{\mu}(X_{t_i})$ is estimated nonparametrically by the local linear method. Call $(\hat{\sigma}^{(3)}, \hat{\gamma}^{(3)})$ the resulting estimators. Estimation of the drift parameters of model (37) can be accomplished by similar adjustment.

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, with weekly frequency and by the transition density, pathwise samples of lengths 5,000 and 10,000 from the CIR model, $dX_t = (.0183925 - .21459X_t)\,dt + .0783X_t^{1/2}\,dW_t$. The sample means and standard errors of these estimates over 1,000 samples are reported in Table 2. Obviously, σ and γ can be estimated far more efficiently than α and β. This is directly attributed to the lower magnitude of signal compared with that of stochastic noise in (35) or (38). Also, the improvements of the weighted least squares estimators over the unweighted estimators are negligible. This is why we leave step III as optional.

4.2 Generalized Likelihood Ratio Test

Interest rate volatility plays a key role in valuing contingent claims and hedging interest rate risks. For the sake of brevity, we describe how to test model (36) against the nonparametric alternative (1), namely, the following testing problem:

$$H_0 : \sigma(X_t) = \sigma X_t^{\gamma} \quad \text{vs.} \quad H_1 : \sigma(X_t) \neq \sigma X_t^{\gamma}.$$

Let $\widehat{E}_{t_i} = \{Y_{t_i} - \hat{\mu}(X_{t_i})\Delta_i\}/\Delta_i^{1/2}$ and $Y_{t_i}^{(1)} = \log(\widehat{E}_{t_i}^2)$. Then similar to (38) and (39), we have approximately

$$\widehat{E}_{t_i} \approx \sigma(X_{t_i})\,\varepsilon_i, \qquad i = 1, \ldots, n-1$$

and

$$Y_{t_i}^{(1)} \approx \log\{\sigma^2(X_{t_i})\} + \log(\varepsilon_i^2), \qquad i = 1, \ldots, n-1. \quad (40)$$

This transforms the test originally for (36) into that for

$$H_0 : \log\{\sigma^2(X_t)\} = \log(\sigma^2) + \gamma\log(X_t^2) \quad \text{versus}$$
$$H_1 : \log\{\sigma^2(X_t)\} \neq \log(\sigma^2) + \gamma\log(X_t^2), \quad (41)$$

that is, testing the linear relationship of the bivariate data $\{(X_{t_i}, Y_{t_i}^{(1)})_{i=1}^{n-1}\}$. Under the null hypothesis in (41), let $\hat{\sigma}$ and $\hat{\gamma}$ be the parameter estimates outlined in Section 4.1. Under the alternative model (1), let $\hat{\sigma}(\cdot)$ be the estimated diffusion function based on the local linear approach. The GLR test statistic, proposed by Fan et al. (2001), is given by

$$\lambda_n(h) = \frac{n-1}{2}\log\frac{\mathrm{RSS}_0}{\mathrm{RSS}_1(h)}, \quad (42)$$

where RSS_0 and RSS_1 [depending on h through $\hat{\sigma}(\cdot)$] represent the residual sums of squares of model (40) under the null and alternative hypotheses in (41). Under H_0, there will be little difference in size between RSS_0 and RSS_1, whereas under the alternative, RSS_0 should become systematically larger than RSS_1, and the GLR statistic thus will tend to take large positive values. Hence a high value of the test statistic $\lambda_n(h)$ indicates that the null hypothesis should be rejected. This procedure can similarly be applied to testing other forms of drift or diffusion functions.

In the nonparametric regression model with independent data, Fan et al. (2001) showed the Wilks type of result, that $r_K\lambda_n(h)$, under certain types of null hypotheses, is asymptotically distributed as $\chi^2_{d_n(h)}$. Here the normalizing constant is $r_K = \frac{(K-2^{-1}K*K)(0)}{\int(K-2^{-1}K*K)^2(t)\,dt}$, the degrees of freedom is $d_n(h) = r_K c_K|\Omega|h^{-1}$, with $c_K = (K - 2^{-1}K*K)(0)$, and $|\Omega|$ measures the length of the support of the regressor variable. In the same paper, it was shown that λ_n is asymptotically equivalent to a quadratic form, $\sum_{i=1}^n\sum_{j=1}^n W_{ijn}(R_i, R_j)$, in which the variables $\{R_i\}$ are independent. Although the GLR statistic applied to our current setup (40) involves more complicated stochastic errors and requires more detailed technical justifications, we believe that a similar Wilks type of result continues to hold under the null hypothesis in (41). This is due to the fact that the quadratic form is a special case of Hoeffding's U statistic. Probabilistic limit theorems (limit law, convergence rate) on U statistics and von Mises statistics for weakly dependent processes are available (see Denker and Keller 1983). Therefore, with dependent $\{R_i\}$, it is technically feasible to work out the limiting distribution of λ_n. Indeed, we have conducted substantial simulations that provide stark evidence to support this claim. However, rigorous justifications are beyond the scope of this article.

4.3 Power Calculation

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

First, we compute the empirical critical values of the GLR statistics under each form of the following typical null hypotheses:

$$H_0^{(1)} : \mu(X_t) = \alpha_0 + \beta_0 X_t, \quad \sigma(X_t) = c_0, \quad (43)$$
$$H_0^{(2)} : \mu(X_t) = \alpha_0 + \beta_0 X_t, \quad \sigma(X_t) = c_1 X_t^{.5}, \quad (44)$$

$$H_0^{(3)} : \mu(X_t) = 0, \quad \sigma(X_t) = c_2 X_t^{1.5}, \tag{45}$$

and

$$H_0^{(4)} : \mu(X_t) = \alpha_0 + \beta_0 X_t, \quad \sigma(X_t) = \sigma X_t^{\gamma}, \tag{46}$$

against the nonparametric alternative (1). Here we set $\alpha_0 = .00739$ and $\beta_0 = -.11798$, which result from the weighted least squares estimates of the 3-month interest rate data (described at the beginning of Sec. 4.4). The constants $c_0 = .01272$, $c_1 = .05596$, and $c_2 = .90114$ are put in (43), (44), and (45), to match the average height of the local linear estimates of volatility, while the parameters σ and γ in (46) are unknown. We have generated with weekly frequency 1,000 pathwise samples of length 2,400, from each of the four hypothetical models, starting at an initial value of .013, the first observation of the interest rate data. In such instances, we use the scheme (3.14) of Kloeden et al. (1996) for models (44) and (46), and use their scheme (3.5) for models (43) and (45). To simulate realizations from model (46), we take the parametrically fitted diffusion function, for which the weighted least squares estimates, $\hat{\sigma} = .071258$ and $\hat{\gamma} = .72957$, are obtained from the interest rate data.

To perform the GLR test combined with the local linear approach, we adopt the empirical formula for bandwidth. For simplicity, three different scales of bandwidth, $h_j = 1.5^{j-1} h_0$, $j = 1, 2, 3$, are also considered, to evaluate simultaneously the impact of bandwidth choice on the test. These bandwidths are roughly viewed as "smaller," "just right," and "bigger." In particular, we use

$$h_0 = 4\,\mathrm{std}(\{X_{t_1}, X_{t_2}, \ldots, X_{t_n}\})\, n^{-2/9}, \tag{47}$$

where $\{X_{t_i}, i = 1, \ldots, n\}$ denotes the simulated sample path, and the rate, $n^{-2/9}$, was shown by Fan et al. (2001) to be the asymptotically optimal rate of bandwidth such that the GLR test can detect alternatives converging to the null at the optimal rate for nonparametric testing. To expedite the computation, we evaluate the local linear fits at 200 grid points, distributed evenly on the ranges of the simulated samples, and then take linear interpolation to obtain the estimates at all of the 2,400 data points. The results of the quantiles are summarized in Table 3. As can be seen, the empirical critical values of $\lambda_n(h_j)$ do not depend sensitively on the true parameter values of the null models, although they should depend on the choice of bandwidth and significance level α.

Table 3. 100(1 − α)th Percentiles of Test Statistics $\lambda_n(h_j)$, $j = 1,2,3$, Under Models $H_0^{(\ell)}$, $\ell = 1,2,3,4$

		Percentile			
Null	Test statistic	$\alpha = .01$	$\alpha = .025$	$\alpha = .05$	$\alpha = .10$
$H_0^{(1)}$	$\lambda_n(h_1)$	127.6	109.4	85.7	66.2
	$\lambda_n(h_2)$	119.3	105.8	85.0	65.4
	$\lambda_n(h_3)$	121.7	94.0	78.1	65.0
$H_0^{(2)}$	$\lambda_n(h_1)$	132.4	114.6	92.3	74.9
	$\lambda_n(h_2)$	123.4	103.0	90.6	74.0
	$\lambda_n(h_3)$	120.6	106.0	86.2	65.2
$H_0^{(3)}$	$\lambda_n(h_1)$	132.7	109.8	91.0	70.7
	$\lambda_n(h_2)$	139.5	108.0	87.4	67.3
	$\lambda_n(h_3)$	139.3	109.5	84.5	67.6
$H_0^{(4)}$	$\lambda_n(h_1)$	119.5	102.3	83.3	65.6
	$\lambda_n(h_2)$	121.1	99.8	82.6	63.7
	$\lambda_n(h_3)$	120.8	100.7	82.0	63.0

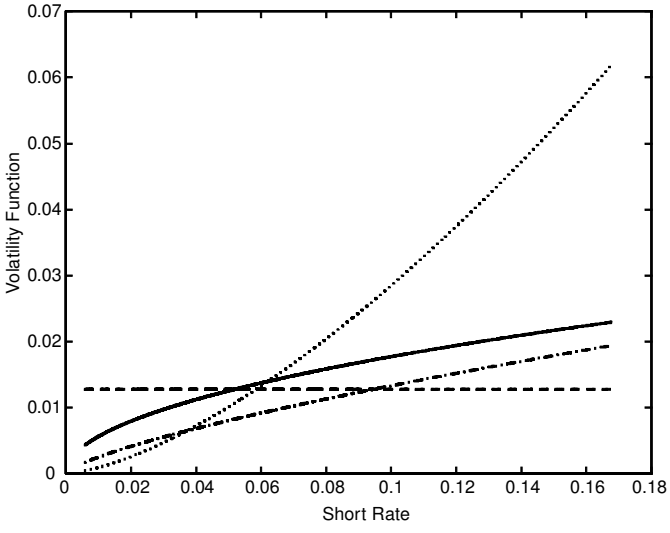

Figure 5. Comparison of Volatility Curves Under Null Hypotheses (44)–(46). The dashed line is c_0; the solid line is $c_1 X_t^{.5}$; the dotted line is $c_2 X_t^{1.5}$; the dash-dotted line is σX_t^{γ}. The constants are $c_0 = .01272$, $c_1 = .05596$, $c_2 = .90114$, $\sigma = .071258$, and $\gamma = .72957$.

Second, to examine the power of the GLR test statistics $\lambda_n(h_j)$, $j = 1, 2, 3$, we consider testing for CIR model (44) against the nonparametric alternative (1). We evaluate the power of the tests at a nominal level 5%, based on 400 datasets simulated from the specific models $H_0^{(\ell)}$, $\ell = 1, 2, 3, 4$. Figure 5 depicts how far apart the volatility functions .01272, $.90114 X_t^{1.5}$, and $.071258 X_t^{.72957}$ deviate from the hypothetical volatility function $.05596 X_t^{.5}$. Thus the GLR tests, as shown in Table 4, are powerful in detecting slight departures from the null, in addition to keeping the right size.

4.4 Testing Commonly Used Short Rate Models

The Treasury bill (T-bill) dataset for our study consists of 2,400 weekly observations covering the period January 8, 1954–December 31, 1999. U.S. Treasury bill secondary market rates are the averages of the bid rates quoted on a bank discount basis by a sample 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 6 shows the estimated drift and volatility curves based on a local linear approach. The estimated drift function exhibits strong nonlinearities at the right boundary region; also, the estimated volatility curve looks like a CIR VR form.

Table 4. Simulated Rejection Rates Against Models $H_0^{(\ell)}$, $\ell = 1,2,3,4$

	Rejection rate			
Test statistic	$H_0^{(1)}$	$H_0^{(2)}$	$H_0^{(3)}$	$H_0^{(4)}$
$\lambda_n(h_1)$	.6175	.0525	1.0000	.9525
$\lambda_n(h_2)$	.6125	.0450	1.0000	.9575
$\lambda_n(h_3)$	.6300	.0375	1.0000	.9475

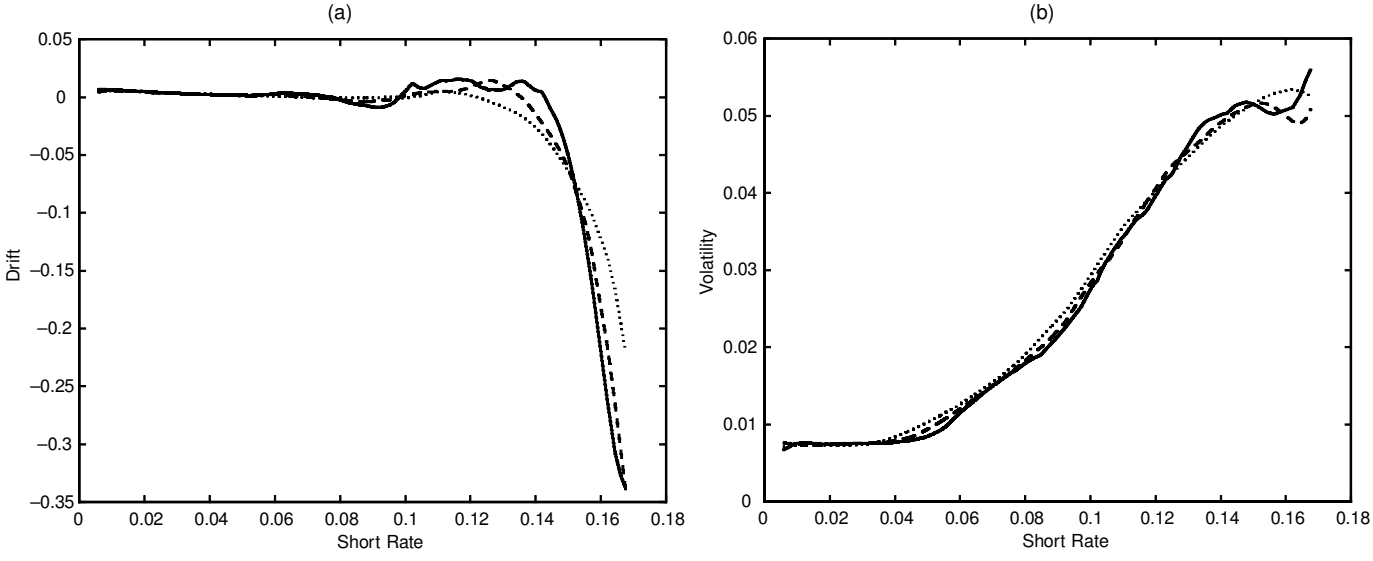

Figure 6. Estimated Drift (a) and Volatility (b) of Short Rate. Estimated drift and volatility functions based on a local linear approach, calculated using weekly data, January 8, 1954–December 31, 1999. The bandwidths are $h_j = 1.5^{j-1}h_0$, $j = 1,2,3$, where $h_0 = .01984$ is calculated from formula (47) (—, h_1; ---, h_2; ····, h_3).

We first address the issue, raised by Chapman and Pearson (2000) of whether the short-rate drift is actually nonlinear, which becomes tantamount to testing model (37) versus model (1). Due to the presence of a larger magnitude of noise, distinguishing the pattern of the signal component from the random-error component becomes very challenging. Despite Chapman and Pearson's full coverage and great efforts in explaining the seemingly nonlinear drift function, there are still no convincing procedures for formally justifying whether the observed deviation from linearity indicates significant departure from model (37). With the aid of the powerful GLR test, we can compute the associated p value, based on a regression bootstrap method for approximating the empirical null distributions of the GLR test statistics. A complete procedure comprises the following steps:

Step 1: For the original T-bill data $\{X_{t_i}, i = 1, \ldots, n\}$, 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 $\mathrm{RSS}_0 = \sum_{i=1}^{n-1}\{Y_{t_i}/\Delta - \hat{\alpha} - \hat{\beta}X_{t_i}\}^2$. Use a local linear approach with bandwidth h to obtain $\hat{\mu}(X_{t_i}), \hat{\sigma}(X_{t_i})$, and $\mathrm{RSS}_1(h) = \sum_{i=1}^{n-1}\{Y_{t_i}/\Delta - \hat{\mu}(X_{t_i})\}^2$. Compute the observed value of the test statistic, $\lambda_{n;\,\mathrm{obs}}(h) = \frac{n-1}{2}\log\frac{\mathrm{RSS}_0}{\mathrm{RSS}_1(h)}$. Get the standardized residuals $\hat{e}_{t_i} = \frac{Y_{t_i} - \hat{\mu}(X_{t_i})\Delta}{\hat{\sigma}(X_{t_i})\Delta^{1/2}}$.

Step 2: Obtain the bootstrap residuals $\{\hat{e}_{t_i}^{(b)}, i = 1, \ldots, n-1\}$ via sampling randomly and with replacement from $\{\hat{e}_{t_j}, j = 1, \ldots, n-1\}$, and define the bootstrap responses, $Y_{t_i}^{(b)} = (\hat{\alpha} + \hat{\beta}X_{t_i})\Delta + \hat{\sigma}(X_{t_i})\Delta^{1/2}\hat{e}_{t_i}^{(b)}$. Use the bootstrap sample $\{(X_{t_i}, Y_{t_i}^{(b)})_{i=1}^{n-1}\}$ to get the bootstrap test statistic $\lambda_n^{(b)}(h)$.

Table 5. Testing Linear Drift Function for T-Bill Short Rate

Test statistic	*Bootstrap p value*	*Rejection rate*
$\lambda_n(h_1)$	.141	.06
$\lambda_n(h_2)$	.104	.11
$\lambda_n(h_3)$	.092	.09

Step 3: Repeat step 2 many times (indexed by superscripts $b = 1, \ldots, 1{,}000$, say), and compute the proportion of times that $\{\lambda_n^{(b)}(h)\}$ exceeds $\lambda_{n;\,\mathrm{obs}}(h)$. This yields the p value of the observed GLR test statistic.

Using this bootstrap procedure, we obtain the p value of the GLR test for model (37) against model (1), shown in the second column of Table 5, with three different bandwidths $\{h_j\}$ as in Section 4.3. Thus there is no strong evidence against the null hypothesis of linear drift. Our proposed test provides formal proofs to reinforce the findings of Chapman and Pearson (2000).

We also apply similar procedures for assessing the adequacy of some previously established hypotheses regarding the variance nature, in particular, competing forms (2)–(6) for volatility functions. The associated p values are displayed in Table 6. Surprisingly, strong evidence indicates that these assumptions on the volatility function cannot be validated by our GLR tests. This is consistent with the results reported by Gallant and Long (1997).

To calibrate the GLR test's ability to correctly reject null hypotheses, we simulate 100 datasets, each containing 2,400 observations from the CIR squared root model (44). Based on the level 5% critical values of the foregoing bootstrapped null distributions, a decision on whether or not to reject the

Table 6. Testing Forms of Volatility Function for T-Bill Short Rate

Test statistic	*GBM*	*VAS*	*CIR SR*	*CIR VR*	*CKLS*
Bootstrap p value					
$\lambda_n(h_1)$	.000	.000	.000	.000	.000
$\lambda_n(h_2)$	.000	.000	.000	.000	.000
$\lambda_n(h_3)$	.000	.000	.002	.000	.015
Rejection rate					
$\lambda_n(h_1)$	1	1	.08	1	.08
$\lambda_n(h_2)$	1	1	.04	1	.06
$\lambda_n(h_3)$	1	1	.04	1	.03

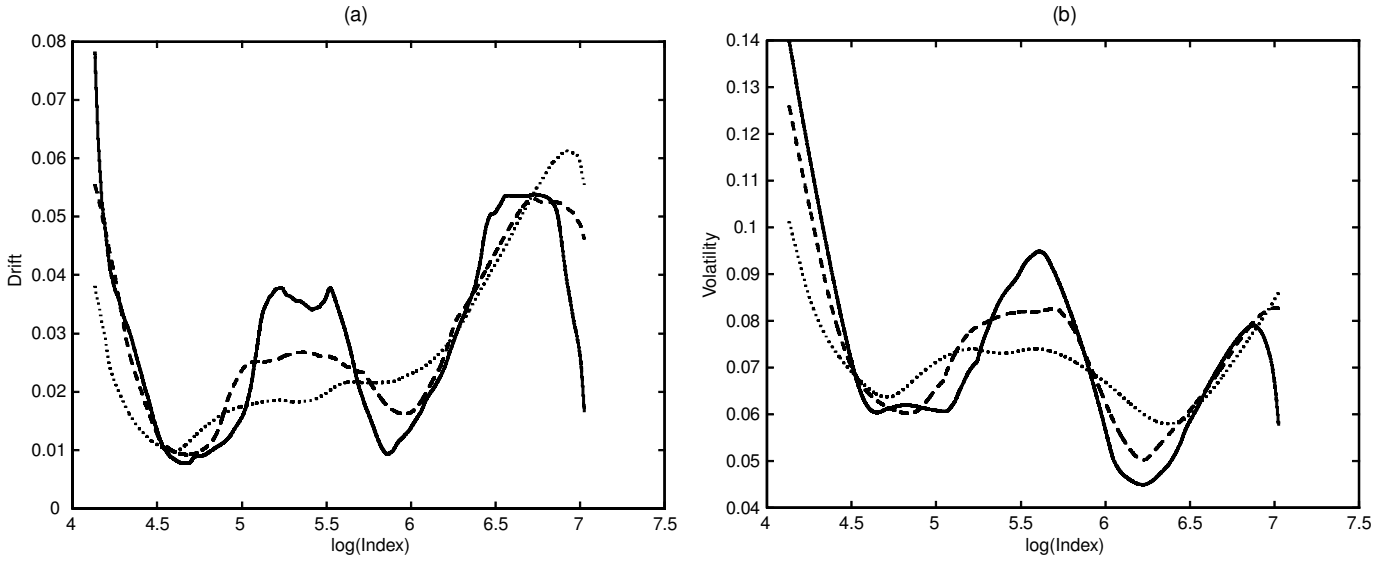

Figure 7. Estimated Drift (a) and Volatility (b) of the S&P 500 Index. Estimated drift and volatility functions based on a local linear approach, calculated using daily data, January 4, 1971–April 8, 1998. The bandwidths are $h_j = 1.5^{j-1}h_0$, $j = 1,2,3$, where $h_0 = .4019$ is calculated from formula (47). (—, h_1; ---, h_2; ····, h_3).

null hypothesis of linear drift can be made with respect to each sample. The proportion of rejections across 100 samples is presented in the third column of Table 5. Similar results concerning volatility functions are listed in Table 6. Therefore, both Table 5 and Table 6 strengthen the assertion that our bootstrap procedures are powerful in correctly accepting or rejecting the null hypotheses.

4.5 Testing Models for Standard & Poor 500 Index

In addition to the interest rate application, we investigate the significance of structural shifts of Standard & Poor (S&P) 500 data from previously studied models. This dataset contains 6,890 daily observations on the S&P composite price index for January 4, 1971–April 8, 1998. Following the conventional practice in finance research, we first take the logarithmic transformation of the price index. The estimated drift and volatility based on a local linear approach are displayed in Figure 7, and the associated bootstrap p values are presented in Tables 7 and 8. Clearly, there is no strong evidence against the hypothesis on the linear drift. For the volatility function, our test suggests that the GBM and CIR VR models do not fit the logarithm of the index. Furthermore, our test also indicates that the VAS, CIR SR, and CKLS models cannot be validated based on the test statistics $\lambda_n(h_j)$, for $j = 1, 2, 3$, together.

5. CONCLUSION

Stanton (1997) proposed drift and diffusion estimators based on a higher-order approximation scheme and a nonparametric kernel estimation. He claimed (p. 1982) that "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 (p. 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. 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 they are penalized by an asymptotic variance escalating nearly exponentially with the order of the approximations. This shadows the higher-order approximation scheme. This phenomenon can be accounted for by the stochastic nature of the Taylor series expansion

Table 7. Testing Linear Drift Function for Logarithms of the S&P 500 Index

Test statistic	Bootstrap p value
$\lambda_n(h_1)$	.814
$\lambda_n(h_2)$	.554
$\lambda_n(h_3)$	.582

Table 8. Testing Forms of Volatility Function for Logarithms of the S&P 500 Index

	Bootstrap p value				
Test statistic	GBM	VAS	CIR SR	CIR VR	CKLS
$\lambda_n(h_1)$	0	.000	.000	0	.031
$\lambda_n(h_2)$	0	.295	.004	0	.418
$\lambda_n(h_3)$	0	.491	.204	0	.576

in (8) accumulated with the linear combination of higher-order differences (11). Caution should be taken when using higher-order formulas. This bias and variance trade-off phenomenon yields general and insightful understandings of the estimators. It also provides useful guidance for determining an optimal strategy for order of approximation, as well as proposing possibly more efficient estimators.

Encouragingly, by using the local linear approach, spurious "boundary effects" from Stanton's kernel estimation are ameliorated, especially for estimating diffusion functions. This local linear estimation approach could also be incorporated with the GLR statistic to test a wide variety of parametric time-homogeneous diffusion models and also to formally check nonlinearity of the short-rate drift. Our simulation shows that our procedures are indeed powerful and have nearly the correct size of the test. The procedures are useful for verifying various models in finance and economics.

APPENDIX: PROOF OF THEOREMS

A.1 Proof of Theorem 1

Using the matrix notation, the system of equations in (9) can be written as $A\mathbf{x} = \mathbf{b}$, where

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

Thus the solution $\mathbf{x} = (x_1, \ldots, x_k)^T$ is uniquely determined by

$$\mathbf{x} = |A|^{-1} A^* \mathbf{b}, \tag{A.1}$$

where A^* and $|A|$ denote the adjoint matrix and the determinant of the matrix A; that is, $\mathbf{x}$ is the first column of A^{-1}. Applying the property of the Vandermonde matrix, we see that the determinant of the matrix A is

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

and that the jth entry in the first column of matrix A^* is

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

Hence in (A.1), the solutions x_j, $j = 1, \ldots, k$, can be simplified as

$$x_j = (-1)^{j+1} \frac{(k!)^2}{j^2} \frac{\prod\prod_{\substack{1 \le l_1 < l_2 \le k \\ l_1 \ne j,\ l_2 \ne j}} (l_2 - l_1)}{k! \prod\prod_{1 \le l_1 < l_2 \le k} (l_2 - l_1)}$$
$$= \frac{(-1)^{j+1} k!}{j^2 (j-1)! (k-j)!} = (-1)^{j+1} \binom{k}{j} \Big/ j.$$

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 $1 \le j \le k$, $\binom{k}{j} j = \binom{k-1}{j-1} k$, which, when applied to the first statement, results in

$$\sum_{j=1}^{k} j^{k+1} a_{k,j} = \sum_{j=1}^{k} (-1)^{j+1} \binom{k}{j} j^k$$
$$= -k \left[(-1) + \sum_{j=1}^{k-1} (-1)^{j+1} \binom{k-1}{j} (j+1)^{k-1} \right].$$

Using the binomial expansion for the factor $(j+1)^{k-1}$ and exchanging the order of summations, we obtain

$$\sum_{j=1}^{k} j^{k+1} a_{k,j} = -k \left[(-1) + \sum_{l=0}^{k-1} \binom{k-1}{l} \sum_{j=1}^{k-1} j^{l+1} a_{k-1,j} \right].$$

This together with (9) yields

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

The conclusion follows from the foregoing inductive formula.

A.2 Proof of Theorem 2

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

Lemma A.1. Assume the same regularity conditions on $\{X_t\}$ as in Theorem 2. For each fixed x_0, as $\Delta \to 0$, it holds that

$$E\{(X_{t+\Delta} - X_t) | X_t = x_0\} = \mu(x_0)\Delta + O(\Delta^2), \tag{A.2}$$
$$E\{(X_{t+\Delta} - X_t)^2 | X_t = x_0\} = \sigma^2(x_0)\Delta + O(\Delta^2), \tag{A.3}$$
$$E\{(X_{t+\Delta} - X_t)^3 | X_t = x_0\}$$
$$= 3\sigma^2(x_0)\{\mu(x_0) + 2^{-1}(\sigma^2)'(x_0)\}\Delta^2 + O(\Delta^3), \tag{A.4}$$
$$E\{(X_{t+\Delta} - X_t)^4 | X_t = x_0\} = 3\sigma^4(x_0)\Delta^2 + O(\Delta^3), \tag{A.5}$$
$$E\{(X_{t+\Delta} - X_t)\mu(X_{t+\Delta}) | X_t = x_0\}$$
$$= \{\mu^2(x_0) + \mu'(x_0)\sigma^2(x_0)\}\Delta + O(\Delta^2), \tag{A.6}$$
$$E\{(X_{t+\Delta} - X_t)^2 \sigma^2(X_{t+\Delta}) | X_t = x_0\} = \sigma^4(x_0)\Delta + O(\Delta^2), \tag{A.7}$$

and

$$E\{(X_{t+\Delta} - X_t)^3 \mu(X_{t+\Delta}) | X_t = x_0\} = O(\Delta^2). \tag{A.8}$$

Proof. To show results (A.2)–(A.8), we choose the 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 $f_7(x,t) = (x - X_t)^3\mu(x)$. Straightforward calculations, applying the differential operator $\mathcal{L}$ defined by (7), give the

following relations:

$$\begin{aligned}
\mathcal{L}f_1(x,t) &= \mu(x),\\
\mathcal{L}^2 f_1(x,t) &= \mu'(x)\mu(x)+2^{-1}\mu''(x)\sigma^2(x),\\
\mathcal{L}f_2(x,t) &= 2(x-X_t)\mu(x)+\sigma^2(x),\\
\mathcal{L}^2 f_2(x,t) &= \{2\mu(x)+2(x-X_t)\mu'(x)+(\sigma^2)'(x)\}\mu(x)\\
&\quad +2^{-1}\{4\mu'(x)+2(x-X_t)\mu''(x)+(\sigma^2)''(x)\}\sigma^2(x);\\
\mathcal{L}f_3(x,t) &= 3(x-X_t)^2\mu(x)+3(x-X_t)\sigma^2(x),\\
\mathcal{L}^2 f_3(x,t) &= \{6(x-X_t)\mu(x)+3(x-X_t)^2\mu'(x)+3\sigma^2(x)\\
&\quad +3(x-X_t)(\sigma^2)'(x)\}\mu(x)+2^{-1}\sigma^2(x)\\
&\quad \times\{6\mu(x)+12(x-X_t)\mu'(x)+3(x-X_t)^2\mu''(x)\\
&\quad +6(\sigma^2)'(x)+3(x-X_t)(\sigma^2)''(x)\},\\
\mathcal{L}f_4(x,t) &= 4(x-X_t)^3\mu(x)+6(x-X_t)^2\sigma^2(x),\\
\mathcal{L}^2 f_4(x,t) &= \{12(x-X_t)^2\mu(x)+4(x-X_t)^3\mu'(x)\\
&\quad +12(x-X_t)\sigma^2(x)+6(x-X_t)^2(\sigma^2)'(x)\}\mu(x)\\
&\quad +2^{-1}\{24(x-X_t)\mu(x)+24(x-X_t)^2\mu'(x)\\
&\quad +4(x-X_t)^3\mu''(x)+12\sigma^2(x)\\
&\quad +24(x-X_t)(\sigma^2)'(x)\\
&\quad +6(x-X_t)^2(\sigma^2)''(x)\}\sigma^2(x);
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{L}f_5(x,t) &= \{\mu(x)+(x-X_t)\mu'(x)\}\mu(x)\\
&\quad +2^{-1}\{\mu'(x)+\mu'(x)+(x-X_t)\mu''(x)\}\sigma^2(x),\\
\mathcal{L}f_6(x,t) &= \{2(x-X_t)\sigma^2(x)+(x-X_t)^2(\sigma^2)'(x)\}\mu(x)\\
&\quad +2^{-1}\{2\sigma^2(x)+4(x-X_t)(\sigma^2)'(x)\\
&\quad +(x-X_t)^2(\sigma^2)''(x)\}\sigma^2(x),\\
\mathcal{L}f_7(x,t) &= \{3(x-X_t)^2\mu(x)+(x-X_t)^3\mu'(x)\}\mu(x)\\
&\quad +2^{-1}\{6(x-X_t)\mu(x)+6(x-X_t)^2\mu'(x)\\
&\quad +(x-X_t)^3\mu''(x)\}\sigma^2(x).
\end{aligned}$$

The proof of Lemma A.1 is completed by using a Taylor series expansion in (8).

To show Theorem 2, we start by considering the conditional variance of the drift estimator. Write $t = t_0 + \ell\Delta$ for any fixed index $\ell = 1, \ldots, n-k$, throughout the following derivations. From the definitions in (16) and (20), we have

$$\begin{aligned}
&\sigma^2_{1,\Delta}(x_0)\\
&\quad = \Delta^{-2}\Bigg[\sum_{1\le j\le k} a^2_{k,j}\operatorname{var}\{(X_{t+j\Delta}-X_t)|X_t=x_0\}+2\sum_{1\le i<j\le k}\sum a_{k,i}a_{k,j}\\
&\qquad \times\operatorname{cov}(X_{t+i\Delta}-x_0, X_{t+j\Delta}-x_0|X_t=x_0)\Bigg]. \qquad (A.9)
\end{aligned}$$

For $j \ge 1$, (A.2) and (A.3) imply that

$$\begin{aligned}
&\operatorname{var}\{(X_{t+j\Delta}-X_t)|X_t=x_0\}\\
&\quad = E\{(X_{t+j\Delta}-X_t)^2|X_t=x_0\}-[E\{(X_{t+j\Delta}-X_t)|X_t=x_0\}]^2\\
&\quad = \sigma^2(x_0)j\Delta+O(\Delta^2). \qquad (A.10)
\end{aligned}$$

For $1 \le i < j \le k$, combining the Markov property of $\{X_t, t \ge 0\}$ with (A.2), (A.3), and (A.6), we have

$$\begin{aligned}
&E\{(X_{t+i\Delta}-x_0)(X_{t+j\Delta}-x_0)|X_t=x_0\}\\
&\quad = E[(X_{t+i\Delta}-x_0)E\{(X_{t+j\Delta}-x_0)|X_{t+i\Delta}\}|X_t=x_0]\\
&\qquad\qquad \text{(Markovian property)}\\
&\quad = E[(X_{t+i\Delta}-x_0)\{(X_{t+i\Delta}-x_0)+\mu(X_{t+i\Delta})(j-i)\Delta\\
&\qquad +O(\Delta^2)\}|X_t=x_0]\\
&\quad = E\{(X_{t+i\Delta}-x_0)^2+(X_{t+i\Delta}-x_0)\mu(X_{t+i\Delta})(j-i)\Delta\\
&\qquad +(X_{t+i\Delta}-x_0)O(\Delta^2)|X_t=x_0\}\\
&\quad = \sigma^2(x_0)i\Delta+O(\Delta^2). \qquad (A.11)
\end{aligned}$$

We also obtain, according to (A.2), that

$$\begin{aligned}
&E\{(X_{t+i\Delta}-x_0)|X_t=x_0\}E\{(X_{t+j\Delta}-x_0)|X_t=x_0\}\\
&\quad = \{\mu(x_0)i\Delta+O(\Delta^2)\}\{\mu(x_0)j\Delta+O(\Delta^2)\}=O(\Delta^2). \qquad (A.12)
\end{aligned}$$

The expression (21) follows readily from the combination of (A.9), (A.10), (A.11), and (A.12).

We now consider the conditional variance of the squared diffusion estimator. In the same vein, from equations (17) and (20), we have

$$\begin{aligned}
&\sigma^2_{2,\Delta}(x_0)\\
&\quad = \Delta^{-2}\Bigg[\sum_{1\le j\le k} a^2_{k,j}\operatorname{var}\{(X_{t+j\Delta}-X_t)^2|X_t=x_0\}+2\sum_{1\le i<j\le k}\sum a_{k,i}a_{k,j}\\
&\qquad \times\operatorname{cov}\{((X_{t+i\Delta}-x_0)^2,(X_{t+j\Delta}-x_0)^2)|X_t=x_0\}\Bigg]. \qquad (A.13)
\end{aligned}$$

For $j \ge 1$, (A.3) and (A.5) imply that

$$\begin{aligned}
&\operatorname{var}\{(X_{t+j\Delta}-X_t)^2|X_t=x_0\}\\
&\quad = E\{(X_{t+j\Delta}-X_t)^4|X_t=x_0\}-[E\{(X_{t+j\Delta}-X_t)^2|X_t=x_0\}]^2\\
&\quad = 2\sigma^4(x_0)(j\Delta)^2+O(\Delta^3). \qquad (A.14)
\end{aligned}$$

For $1 \le i < j \le k$, combining the Markov property of $\{X_t, t \ge 0\}$ with (A.5), (A.7), and (A.8), we have

$$\begin{aligned}
&E\{(X_{t+i\Delta}-x_0)^2(X_{t+j\Delta}-x_0)^2|X_t=x_0\}\\
&\quad = E[(X_{t+i\Delta}-x_0)^2E\{(X_{t+j\Delta}-x_0)^2|X_{t+i\Delta}\}|X_t=x_0]\\
&\qquad\qquad \text{(Markovian property)}\\
&\quad = E[(X_{t+i\Delta}-x_0)^2\{(X_{t+i\Delta}-x_0)^2+(2(X_{t+i\Delta}-x_0)\mu(X_{t+i\Delta})\\
&\qquad +\sigma^2(X_{t+i\Delta}))(j-i)\Delta+O(\Delta^3)\}|X_t=x_0]\\
&\quad = E\{(X_{t+i\Delta}-x_0)^4+2(X_{t+i\Delta}-x_0)^3\mu(X_{t+i\Delta})(j-i)\Delta\\
&\qquad +(X_{t+i\Delta}-x_0)^2\sigma^2(X_{t+i\Delta})(j-i)\Delta+O(\Delta^3)|X_t=x_0\}\\
&\quad = 3\sigma^4(x_0)(i\Delta)^2+O(\Delta^3)+\sigma^4(x_0)(i\Delta)(j-i)\Delta+O(\Delta^3)\\
&\quad = 2\sigma^4(x_0)(i\Delta)^2+\sigma^4(x_0)ij\Delta^2+O(\Delta^3). \qquad (A.15)
\end{aligned}$$

We also obtain from (A.3) that

$$\begin{aligned}
&E\{(X_{t+i\Delta}-x_0)^2|X_t=x_0\}E\{(X_{t+j\Delta}-x_0)^2|X_t=x_0\}\\
&\quad = \{\sigma^2(x_0)i\Delta+O(\Delta^2)\}\{\sigma^2(x_0)j\Delta+O(\Delta^2)\}\\
&\quad = \sigma^4(x_0)ij\Delta^2+O(\Delta^3). \qquad (A.16)
\end{aligned}$$

The equality (22) follows directly from the combination of (A.13), (A.14), (A.15), and (A.16).

A.3 Proof of Theorem 3

The proofs in this section are based on some combinatorial relations. Let $\gamma = \lim_{n\to\infty}\{\sum_{k=1}^{n} k^{-1} - \log(n)\} \approx .577216$ be the Euler's constant and $\psi(z) = \Gamma'(z)/\Gamma(z)$ be the Psi function, where $\Gamma(z) = \int_0^\infty u^{z-1}e^{-u}\,du$ for $z > 0$. First, we consider part (a). With the aid of Mathematica, we obtain the identities

$$\sum_{j=1}^{k}\binom{k}{j}^2\frac{(j+2)}{(j+1)^2} = \frac{(2k+1)!}{\{(k+1)!\}^2} + \frac{4^{k+1}\Gamma(3/2+k)}{(k+1)^3\pi^{1/2}k!} - \frac{2k^2+4k+3}{(k+1)^2}, \tag{A.17}$$

$$\sum_{j=1}^{k}\binom{k}{j}^2\frac{(j+3)}{(j+1)^2} = \frac{(2k+1)!}{\{(k+1)!\}^2} + \frac{2^{2k+3}\Gamma(3/2+k)}{(k+1)^3\pi^{1/2}k!} - \frac{3k^2+6k+5}{(k+1)^2}, \tag{A.18}$$

and

$$\sum_{j=2}^{k}\left\{\sum_{i=1}^{j-1}(-1)^{i+1}\binom{k}{i}\right\}(-1)^{j+1}\binom{k}{j}\Big/ j = \frac{1+\gamma k}{k} - \frac{1}{k}\binom{2k}{k} + \psi(k+1). \tag{A.19}$$

Consequently, putting $a_{k,j} = (-1)^{j+1}\binom{k}{j}/j$ and simplifying the right sides of (A.17) and (A.18), we have

$$\sum_{j=1}^{k} j a_{k,j}^2 > \sum_{j=1}^{k}\binom{k}{j}^2\frac{(j+2)}{(j+1)^2} = \frac{(2k+1)(k+3)}{(k+1)^3}\binom{2k}{k} - \frac{2k^2+4k+3}{(k+1)^2} \tag{A.20}$$

and

$$\sum_{j=1}^{k} j a_{k,j}^2 \le \sum_{j=1}^{k}\binom{k}{j}^2\frac{(j+3)}{(j+1)^2} = \frac{(2k+1)(k+5)}{(k+1)^3}\binom{2k}{k} - \frac{3k^2+6k+5}{(k+1)^2}. \tag{A.21}$$

Applying (A.19) and the identity $\psi(n) = \sum_{j=1}^{n-1} j^{-1} - \gamma$, which holds for any integer $n \ge 2$, we deduce

$$\sum\sum_{1\le i<j\le k} i a_{k,i}a_{k,j} = \frac{1}{k} + \sum_{j=1}^{k}\frac{1}{j} - \frac{1}{k}\binom{2k}{k}. \tag{A.22}$$

Hence (21), (A.9), and (A.22), together with inequalities (A.20) and (A.21), ensure that $V_1(k)$ has a lower bound

$$\frac{k^2-3k-2}{k(k+1)^3}\binom{2k}{k} + \frac{2}{k} + 2\sum_{j=1}^{k}\frac{1}{j} - \frac{2k^2+4k+3}{(k+1)^2} \tag{A.23}$$

and an upper bound

$$\frac{5k^2-k-2}{k(k+1)^3}\binom{2k}{k} + \frac{2}{k} + 2\sum_{j=1}^{k}\frac{1}{j} - \frac{3k^2+6k+5}{(k+1)^2}. \tag{A.24}$$

The conclusion follows from applying Stirling's formula $n! = (2\pi n)^{1/2}(n/e)^n\exp\{\theta/(12n)\}$ for some $0 < \theta < 1$ to the first dominating terms of (A.23) and (A.24).

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

$$\sum_{j=1}^{k} j^2 a_{k,j}^2 = \binom{2k}{k} - 1. \tag{A.25}$$

Again with the aid of Mathematica, we obtain the identity that for $k > 1$ and $2 \le j \le k$,

$$\sum_{i=1}^{j-1}(-1)^{i+1}\binom{k}{i} i = \frac{(-1)^j j\Gamma(k)}{\Gamma(j)\Gamma(k-j+1)} - \frac{(-1)^j\Gamma(k-1)}{\Gamma(j)\Gamma(k-j)}, \tag{A.26}$$

which implies that

$$\sum\sum_{1\le i<j\le k} i^2 a_{k,i}a_{k,j} = \frac{1}{k-1}\sum_{j=2}^{k}\binom{k-1}{j}\binom{k}{j} - \sum_{j=2}^{k}\binom{k-1}{k-j}\binom{k}{j} = -\frac{\binom{2k-1}{k}(k-2)+1}{k-1}. \tag{A.27}$$

The conclusion (b) follows from (22), (A.13), (A.25), (A.27) and Stirling's formula.

A.4 Proof of Theorem 4

It suffices to consider only Part (1); similar treatments apply to Part (2). We denote a generic constant by C. Let $\mathbf{X} = ((X^*_{i\Delta} - x_0)^j)_{i=1,\ldots,n-k;\,j=0,\ldots,q}$, $\mathbf{y} = (Y^*_\Delta, \ldots, Y^*_{(n-k)\Delta})^T$, $\mathbf{W} = \operatorname{diag}\{K_h(X^*_{i\Delta} - x_0), i = 1, \ldots, n-k\}$, and $\mathbf{m} = (E(Y^*_\Delta|X^*_\Delta), \ldots, E(Y^*_{(n-k)\Delta}|X^*_{(n-k)\Delta}))^T$. Denote $S_n = \mathbf{X}^T\mathbf{W}\mathbf{X}$ and $T_n = \mathbf{X}^T\mathbf{W}\mathbf{y}$. Then by (19), we can write $\hat{\boldsymbol{\beta}}(x_0) = S_n^{-1}T_n$ and thus

$$\hat{\boldsymbol{\beta}}(x_0) - \boldsymbol{\beta}(x_0) = S_n^{-1}\mathbf{X}^T\mathbf{W}\{\mathbf{m} - \mathbf{X}\boldsymbol{\beta}(x_0)\} + S_n^{-1}\mathbf{X}^T\mathbf{W}(\mathbf{y}-\mathbf{m}), \equiv \mathbf{b} + \mathbf{t}.$$

We first establish the asymptotic behavior of the bias vector $\mathbf{b} = (b_0, b_1, \ldots, b_q)^T$. Set $Z_{n,\ell} = K_h(X^*_{\ell\Delta} - x_0)(X^*_{\ell\Delta} - x_0)^j$ and $S_{n,j} = \sum_{\ell=1}^{n-k} Z_{n,\ell}$; then $S_n = (S_{n,i+j-2})_{i,\,j=1,\,\ldots,\,q+1}$. A Taylor expansion leads to the expression

$$\mathbf{b} = S_n^{-1}\{\beta_{q+1}(S_{n,q+1}, \ldots, S_{n,2q+1})^T + \beta_{q+2}(S_{n,q+2}, \ldots, S_{n,2q+2})^T + o_P(nh^{q+2})H\mathbf{1}\}, \tag{A.28}$$

with a $(q+1)\times(q+1)$ matrix $H = \operatorname{diag}(1, h, \ldots, h^q)$ and a $(q+1)\times 1$ vector $\mathbf{1} = (1, \ldots, 1)^T$. To derive the asymptotic form of $\mathbf{b}$, we need only apply the expression

$$S_{n,j} = nh^j\{p(x_0)\mu_j + hp'(x_0)\mu_{j+1} + O_P(a_n)\}, \tag{A.29}$$

where $a_n = h^2 + (nh)^{-1/2}$. Equation (A.29) can be obtained via procedures similar to those of Fan and Gijbels (1996, thm. 3.1). However, to verify the term $O_P(a_n)$ in our current context, we need to do the variance calculation for $S_{n,j}$, which is different than that of Fan and Gijbels. To this end, using the assumption on the transition density, we first obtain

$$|\operatorname{cov}(Z_{n,1}, Z_{n,\ell+1})| \le Ch^{2j}\{1 + o(1)\}. \tag{A.30}$$

Recall for a bounded real-valued Borel measurable function g, the transition probability operator $\mathcal{T}^\ell$ of the process $\{X^*_{i\Delta}, i = 1, \ldots, n-k\}$ is defined by

$$(\mathcal{T}^\ell g)(x) = E\{g(X^*_{(\ell+1)\Delta})|X^*_\Delta = x\}.$$

By the G_2 condition of Rosenblatt (1970), there exists a constant $\rho \in (0,1)$ for $\mathcal{T}$, such that for $g(\cdot) = K_h(\cdot - x_0)(\cdot - x_0)^j - E\{K_h(\cdot - x_0)(\cdot - x_0)^j\}$, we have

$$\begin{aligned}|\operatorname{cov}(Z_{n,1}, Z_{n,\ell+1})| &= |E\{g(X^*_\Delta)\mathcal{T}^\ell g(X^*_\Delta)\}| \\ &\le \|g(X^*_\Delta)\|_2\|\mathcal{T}^\ell g(X^*_\Delta)\|_2 \\ &\le \|g(X^*_\Delta)\|_2^2|\mathcal{T}^\ell|_2 \\ &\le Ch^{2j-1}\rho^\ell,\end{aligned} \tag{A.31}$$

where $|\mathcal{T}^\ell|_2 = \sup_{g:g\neq E(g)} \frac{\|\mathcal{T}^\ell g - E(g)\|_2}{\|g-E(g)\|_2}$, and E stands for expectation with respect to the stationary density $p(\cdot)$. Now select an integer d_n so that $d_n \to \infty$ and $d_n h \to 0$ (e.g., $d_n = h^{-1/2}$); then (A.30) and (A.31) give

$$\sum_{\ell=1}^{n-k-1} |\mathrm{cov}(Z_{n,1}, Z_{n,\ell+1})| = \left(\sum_{\ell=1}^{d_n} + \sum_{\ell=d_n+1}^{n-k-1}\right)|\mathrm{cov}(Z_{n,1}, Z_{n,\ell+1})| = o(h^{2j-1}). \quad \text{(A.32)}$$

This, along with the stationarity assumption, yield

$$\mathrm{var}(S_{n,j}) = (n-k)\mathrm{var}(Z_{n,1}) + 2\sum_{\ell=1}^{n-k-1}(n-k-\ell)\mathrm{cov}(Z_{n,1}, Z_{n,\ell+1}) = nh^{2j-1}\left[p(x_0)\nu_{2j} + o(1) + 2h^{-(2j-1)} \times \sum_{\ell=1}^{n-k-1}\left(1-\frac{\ell}{n-k}\right)\mathrm{cov}(Z_{n,1}, Z_{n,\ell+1})\right],$$

from whence (A.29) is obtained.

The asymptotic bias expression in (23) then results from the decomposition

$$\hat{\mu}_{1,\Delta}(x_0) - \mu(x_0) = \{\hat{\mu}_{1,\Delta}(x_0) - E(Y^*_{i\Delta}|X^*_{i\Delta} = x_0)\} - \{E(Y^*_{i\Delta}|X^*_{i\Delta} = x_0) - \mu(x_0)\}.$$

On the right side, we see that $\hat{\mu}_{1,\Delta}(x_0) - E(Y^*_{i\Delta}|X^*_{i\Delta} = x_0) = b_0$; by (13), we see that $E(Y^*_{i\Delta}|X^*_{i\Delta} = x_0) - \mu(x_0) = (-1)^{k+1} \times \frac{\mathcal{L}^{k+1} f_1(x_0, t_0+i\Delta)}{(k+1)}\Delta^k + O(\Delta^{k+1})$. This completes the proof of (23).

Next, consider the asymptotic variance of $\hat{\mu}_{1,\Delta}(x_0)$. By (A.29), $\mathbf{t} = p^{-1}(x_0)H^{-1}S^{-1}\mathbf{u}\{1+o_P(1)\}$, where $\mathbf{u} = n^{-1}H^{-1}\mathbf{X}^T\mathbf{W}(\mathbf{y}-\mathbf{m})$. For any constant vector $\mathbf{c}$, define

$$Q_n = \mathbf{c}^T\mathbf{u} = \frac{1}{n}\sum_{i=1}^{n-k}\{Y^*_{i\Delta} - E(Y^*_{i\Delta}|X^*_{i\Delta})\}C_h(X^*_{i\Delta} - x_0),$$

where $C(x) = \sum_{j=0}^{p} c_j\, x^j K(x)$, and $C_h(x) = C(x/h)/h$. Set $v_{n,\ell} = \{Y^*_{i\Delta} - E(Y^*_{i\Delta}|X^*_{i\Delta})\}C_h(X^*_{i\Delta} - x_0)$. Then direct calculations give that

$$\mathrm{var}(v_{n,1}) = (h\Delta)^{-1}\sigma_1^2(x_0;k)p(x_0)\mathbf{c}^T S^*\mathbf{c}\{1+o(1)\}. \quad \text{(A.33)}$$

Similar procedures to those used in (A.30)–(A.32) lead to

$$\sum_{\ell=1}^{n-k-1}|\mathrm{cov}(v_{n,1}, v_{n,\ell+1})| \le d_n h^2\Delta^{-2} + h\Delta^{-2}\sum_{\ell=d_n+1}^{n-k-1}\rho^\ell = o(h\Delta^{-2}),$$

which, combined with (A.33) and the assumption on h, imply that $\mathrm{var}(\mathbf{u}) = (nh\Delta)^{-1}\sigma_1^2(x_0;k)p(x_0)S^*\{1+o(1)\}$ and, therefore, (25).

[Received November 2000. Revised February 2002.]

REFERENCES

Aït-Sahalia, Y. (1996), "Nonparametric Pricing of Interest Rate Derivative Securities," *Econometrica*, 64, 527–560.

Allen, D. M. (1974), "The Relationship Between Variable and Data Augmentation and a Method of Prediction," *Technometrics*, 16, 125–127.

Arfi, M. (1995), "Non-Parametric Drift Estimation from Ergodic Samples," *Journal of Nonparametric Statistics*, 5, 381–389.

——— (1998), "Non-Parametric Variance Estimation from Ergodic Samples," *Scandinavia Journal of Statistics*, 25, 225–234.

Banon, G. (1978), "Nonparametric Identification for Diffusion Processes," *SIAM Journal of Control and Optimization*, 16, 380–395.

Banon, G., and Nguyen, H. T. (1981), "Recursive Estimation in Diffusion Models," *SIAM Journal of Control and Optimization*, 19, 676–685.

Chan, K. C., Karolyi, A. G., Longstaff, F. A., and Sanders, A. B. (1992), "An Empirical Comparison of Alternative Models of the Short-Term Interest Rate," *Journal of Finance*, 47, 1209–1227.

Chapman, D. A., and Pearson, N. D. (2000), "Is the Short Rate Drift Actually Nonlinear?," *Journal of Finance*, 55, 355–388.

Cox, J. C., Ingersoll, J. E., and Ross, S. A. (1980), "An Analysis of Variable Rate Loan Contracts," *Journal of Finance*, 35, 389–403.

——— (1985), "A Theory of the Term Structure of Interest Rates," *Econometrica*, 53, 385–407.

Denker, M., and Keller, G. (1983), "On U Statistics and V.Mises's Statistics for Weakly Dependent Processes," *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 64, 505–522.

Fan, J. (1992), "Design-Adaptive Nonparametric Regression," *Journal of the American Statistical Association*, 87, 998–1004.

Fan, J., and Gijbels, I. (1995), "Data-Driven Bandwidth Selection in Local Polynomial Fitting: Variable Bandwidth and Spatial Adaptation," *Journal of the Royal Statistical Society*, Ser. B, 57, 371–394.

——— (1996), *Local Polynomial Modeling and Its Applications*, London: Chapman and Hall.

Fan, J., and Yao, Q. W. (1998), "Efficient Estimation of Conditional Variance Functions in Stochastic Regression," *Biometrika*, 85, 645–660.

Fan, J., Zhang, C. M., and Zhang, J. (2001), "Generalized Likelihood Ratio Statistics and Wilks Phenomenon," *The Annals of Statistics*, 29, 153–193.

Gallant, A. R., and Long, J. R. (1997), "Estimating Stochastic Differential Equations Efficiently by Minimum Chi-Squared," *Biometrika*, 84, 125–141.

Hansen, L. P. (1982), "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054.

Jiang, G. J., and Knight, J. L. (1997), "A Nonparametric Approach to the Estimation of Diffusion Processes, With an Application to a Short-Term Interest Rate Model," *Econometric Theory*, 13, 615–645.

Kloeden, P. E., and Platen, E. (1992), *Numerical Solution of Stochastic Differential Equations*, Berlin: Springer-Verlag.

Kloeden, P. E., Platen, E., Schurz, H., and Sørensen, M. (1996), "On Effects of Discretization on Estimators of Drift Parameters for Diffusion Processes," *Journal of Applied Probability*, 33, 1061–1076.

Øksendal, B. (1985), *Stochastic Differential Equations: An Introduction With Applications*, New York: Springer-Verlag.

Osborne, M. F. M. (1959), "Brownian Motion in the Stock Market," *Operations Research*, 7, 145–173.

Pham, D. T. (1981), "Nonparametric Estimation of the Drift Coefficient in the Diffusion Equation," *Mathematische Operationsforschung und Statistik Series Statistics*, 12, 61–73.

Prakasa Rao, B. L. S. (1985), "Estimation of the Drift for Diffusion Process," *Statistics*, 16, 263–275.

Rosenblatt, M. (1970), "Density Estimates and Markov Sequences," in *Nonparametric Techniques in Statistical Inferences*, ed. M. Puri, London: Cambridge University Press, pp. 199–210.

——— (1971), *Markov Processes, Structure and Asymptotic Behavior*, New York: Springer-Verlag.

Ruppert, D. (1997), "Empirical-Bias Bandwidths for Local Polynomial Nonparametric Regression and Density Estimation," *Journal of the American Statistical Association*, 92, 1049–1062.

Ruppert, D., Sheather, S. J., and Wand, M. P. (1995), "An Effective Bandwidth Selector for Local Least Squares Regression," *Journal of the American Statistical Association*, 90, 1257–1270.

Stanton, R. (1997), "A Nonparametric Model of Term Structure Dynamics and the Market Price of Interest Rate Risk," *Journal of Finance*, 52, 1973–2002.

Stone, M. (1974), "Cross-Validatory Choice and Assessment of Statistical Predictions" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 36, 111–147.

Vasicek, O. A. (1977), "An Equilibrium Characterization of the Term Structure," *Journal of Financial Economics*, 5, 177–188.

Wahba, G. (1977), "A Survey of Some Smoothing Problems and the Method of Generalized Cross-validation for Solving them," in *Applications of Statistics*, ed. P. R. Krishnaiah, Amsterdam: North-Holland, pp. 507–523.

Wong, E. (1971), *Stochastic Processes in Information and Dynamical Systems*, New York: McGraw-Hill.