Nonparametric Specification Testing for Continuous-Time Models with Applications to Term Structure of Interest Rates

Yongmiao Hong
Cornell University and Tsinghua University

Haitao Li
Cornell University

We develop a nonparametric specification test for continuous-time models using the transition density. Using a data transform and correcting for the boundary bias of kernel estimators, our test is robust to serial dependence in data and provides excellent finite sample performance. Besides univariate diffusion models, our test is applicable to a wide variety of continuous-time and discrete-time dynamic models, including time-inhomogeneous diffusion, GARCH, stochastic volatility, regime-switching, jump-diffusion, and multivariate diffusion models. A class of separate inference procedures is also proposed to help gauge possible sources of model misspecification. We strongly reject a variety of univariate diffusion models for daily Eurodollar spot rates and some popular multivariate affine term structure models for monthly U.S. Treasury yields.

Continuous-time models have been widely used in finance to capture the dynamics of important economic variables, such as interest rates, exchange rates, and stock prices. The well-known option pricing model of Black and Scholes (1973) and the term structure model of Cox, Ingersoll, and Ross (1985, CIR), for example, assume that the underlying state variables follow diffusion processes. Economic theories usually do not suggest a functional form for continuous-time models, and researchers often use convenient specifications for deriving closed-form solutions for various security prices.

We would like to thank seminar participants at Cornell University, Hong Kong Baptist University, Humboldt University, Princeton University, the 2002 American Finance Association Meeting, the 2nd World Congress of Bachelier Finance Society, the 2002 Western Finance Association Meeting, the 2002 North American Econometric Society Summer Meeting, the 2002 International Conference on Nonparametric Statistics, the 2002 European Econometric Society Meeting, the 2002 European Financial Association Meeting, the 2002 American Statistical Association Meeting, and Andrew Ang, Alexei Egorov, Wolfgang Härdle, Bob Jarrow, Bob Kimmel, Ken Singleton (co-editor), George Tauchen, Kevin Wang, Qingshi Wang, and an anonymous referee for their comments. We also thank Yacine Ait-Sahalia and Greg Duffee for providing us with the Eurodollar interest rates and U.S. Treasury yields data, respectively. Support from the National Science Foundation under Grant SES-0111769 and computational assistance from the Parker Center at the Johnson Graduate School of Management and Doug Elias are greatly appreciated. Address correspondence to: Yongmiao Hong, Department of Economics and Department of Statistical Science, Cornell University, 424 Uris Hall, Ithaca, NY 14850, or e-mail: yh20@cornell.edu.

The Review of Financial Studies Vol. 18, No. 1
doi:10.1093/rfs/hhh006 Advance Access publication March 26, 2004

The last decade has seen the development of a large and still growing literature on the estimation of continuous-time models.[1] Motivated by Lo's (1988) finding that estimating the discretized version of a continuous-time model can result in inconsistent parameter estimates, many econometric methods have been developed to estimate continuous-time models using discretely sampled data.[2] However, there is relatively little effort on specification analysis for continuous-time models. Model misspecification generally yields inconsistent estimators of model parameters and their variance–covariance matrix, which could lead to misleading conclusions in inference and hypothesis testing. Moreover, a misspecified model can yield large errors in pricing, hedging, and risk management. It is therefore important to develop reliable specification tests for continuous-time models.

In a pioneering work, Ait-Sahalia (1996) develops probably the first nonparametric test for diffusion models. Observing that the drift and diffusion functions completely characterize the marginal density of a diffusion model, Ait-Sahalia (1996) compares a model marginal density estimator with a nonparametric counterpart using discretely sampled data. The test makes no restrictive assumption on the diffusion process and can detect a wide range of alternatives — an appealing property not shared by parametric approaches [e.g., Conley, Hansen, Luttmer, and Scheinkman (1997)]. In an application to daily Eurodollar interest rates, Ait-Sahalia (1996) rejects all existing univariate linear drift models using asymptotic theory and finds that "the principal source of rejection of existing models is the strong nonlinearity of the drift." Stanton (1997), using nonparametric kernel regression, also finds a significant nonlinear drift in spot rate data.[3]

Subsequent studies have pointed out the limitations of the nonparametric methods used by Ait-Sahalia (1996) and Stanton (1997) and questioned the findings of nonlinear drift. Pritsker (1998) shows that

[1] Sundaresan (2001) states that "perhaps the most significant development in the continuous-time field during the last decade has been the innovations in econometric theory and in the estimation techniques for models in continuous time." For other reviews of the recent literature, see, e.g., Tauchen (1997) and Campbell, Lo, and MacKinlay (1997).

[2] Available estimation procedures include, among many others, the nonparametric methods of Ait-Sahalia (1996), Stanton (1997), and Jiang and Knight (1997), the simulated method of moments of Duffie and Singleton (1993), the efficient method of moments (EMM) of Gallant and Tauchen (1996), the generalized method of moments of Hansen and Scheinkman (1995), the maximum likelihood method of Lo (1988) (numerically solving the forward Kolmogorov equation) and Ait-Sahalia (2002a,b) (closed-form approximation using Hermite polynomials), the simulated maximum likelihood method of Pedersen (1995) and Brandt and Santa-Clara (2002), the empirical characteristic function approach of Singleton (2001) and Jiang and Knight (2002), and the Monte Carlo Markov Chain method of Eraker (2001).

[3] The main finding of Ait-Sahalia (1996) is that the drift is zero when the interest rate is between 4 and 16% and mean-reverting occurs at both extremes, which he called "locally unit root behavior." Thus, nonlinearity does not refer exclusively to what happens below 4% and above 16% (where confidence intervals are wide due to the scarcity of data) but to the overall shape including the lack of a visible drift between 4 and 16%. A similar explanation also applies to Stanton's (1997) results.

Ait-Sahalia's (1996) test has poor finite sample performance because of persistent dependence in interest rate data and slow convergence of the nonparametric density estimator. Using an empirically relevant Vasicek (1977) model, Pritsker finds that the test tends to overreject the null hypothesis, and it needs 2755 years of daily data for the asymptotic theory to work adequately. Chapman and Pearson (2000) also show that the nonparametric methods used by Ait-Sahalia (1996) and Stanton (1997) produce biased estimates near the boundaries of data, which could produce a spurious nonlinear drift. The findings of Pritsker (1998) and Chapman and Pearson (2000) thus cast serious doubts on the applicability of nonparametric methods in finance, since persistent dependence is a stylized fact for interest rates and many other high-frequency financial data.

We develop an omnibus nonparametric specification test for continuous-time models based on the transition density, which, unlike the marginal density used by Ait-Sahalia (1996), captures the full dynamics of a continuous-time process. Our basic idea is that if a model is correctly specified, then the probability integral transform of data via the model transition density should be i.i.d. $U[0, 1]$. The probability integral transform can be called the "generalized residuals" of the continuous-time model. We shall test the i.i.d. $U[0, 1]$ hypothesis for the model generalized residuals by comparing a kernel estimator of the joint density of the generalized residuals with the product of two $U[0, 1]$ densities.[4] Our approach has several advantages.

First, our test significantly improves the size and power performance of the marginal density-based test, thanks to the use of the transition density and the probability integral transform. The marginal density-based test is computationally convenient and can detect many alternatives. However, it can easily miss the alternatives that have the same marginal density as the null model. In contrast, our transition density-based test can effectively pick them up. The probability integral transform helps achieve robustness of our test to persistent dependence in data. Because there is no serial dependence in the generalized residuals under correct model specification, nonparametric density estimators are expected to perform well in finite samples. Also, we use a modified kernel to alleviate the notorious "boundary bias" of kernel estimators. Simulations involving univariate and multivariate models show that our test has reasonable size and good power against a variety of alternatives in finite samples even for persistently dependent data.

Second, as we impose regularity conditions on the transition density rather than the stochastic differential equation of the underlying process,

[4] While the transition density has no closed form for most continuous-time models, many methods exist in the literature to provide accurate approximations of the transition density [e.g., Ait-Sahalia (2002a,b), Ait-Sahalia and Kimmel (2002), Duffie, Pedersen and Singleton (2003)].

our test is generally applicable: besides the univariate time-homogeneous diffusion models considered in Ait-Sahalia (1996), a wide variety of continuous-time and discrete-time dynamic models, such as time-inhomogeneous diffusion, GARCH, stochastic volatility, regime-switching, jump-diffusion, and multivariate diffusion models, are also covered. Many financial models have non-nested specifications with different estimation methods, so it has been challenging to formally compare their relative goodness of fit [see, e.g., Dai and Singleton (2000) for a discussion of affine models]. Our approach provides a unified framework under which the relative performance of different models can be compared by a metric measuring the departures of their generalized residuals from i.i.d. $U[0, 1]$.

Third, the model generalized residuals provide valuable information about sources of model misspecification. Intuitively, the i.i.d. property characterizes correct dynamic specification, and the $U[0, 1]$ property characterizes correct specification of the stationary distribution. To fully utilize the rich information in the generalized residuals, we also develop a class of rigorous separate inference procedures based on the autocorrelations in the powers of the model generalized residuals. These procedures complement Gallant and Tauchen's (1996) popular EMM-based individual t-tests and can be used to gauge how well a model captures various dynamic aspects of the underlying process.

To highlight our approach, we apply our tests to evaluate a variety of popular univariate spot rate models and multivariate term structure models. Using the same daily Eurodollar interest rate data, we reexamine the spot rate models considered in Ait-Sahalia (1996). While Ait-Sahalia (1996) rejects all linear drift models using asymptotic critical values, one would not reject Chan, Karolyi, Longstaff, and Sanders's (1992, CKLS) model and Ait-Sahalia's (1996) nonlinear drift model using the empirical critical values obtained in Pritsker (1998) for Ait-Sahalia's (1996) test. In contrast, our omnibus test firmly rejects all univariate diffusion models, and we find that nonlinear drift does not significantly improve goodness of fit. Our omnibus test also overwhelmingly rejects the three-factor completely and essentially affine models of Dai and Singleton (2000) and Duffee (2002), using monthly U.S. Treasury yields over the last 50 years. Affine models that are flexible in capturing the conditional variance and correlation of state variables and market prices of risk have the best performance and they fit the middle and long end of the yield curve better than the short end.

In Section 1, we introduce our omnibus test and separate inference procedures. In Section 2, we study finite sample size and the power of the omnibus test via simulation. In Sections 3 and 4, we evaluate a variety of popular univariate diffusion models for daily Eurodollar interest rates and multivariate affine term structure models for monthly U.S. Treasury yields, respectively. Section 5 concludes the discussion. The appendix gives the asymptotic theory. A GAUSS code for implementing our omnibus test is available from yh20@cornell.edu upon request.

1. Approach and Test Statistics

We now develop an omnibus nonparametric specification test for continuous-time models using the transition density. As our test is most closely related to Ait-Sahalia's (1996) marginal density-based test, we first follow Ait-Sahalia (1996) and consider univariate diffusion processes for comparison. In later sections, we will consider more general processes.

1.1 Dynamic probability integral transform

Suppose a state variable X_t follows a continuous-time (possibly time-inhomogeneous) diffusion:

$$dX_t = \mu_0(X_t, t)dt + \sigma_0(X_t, t)dW_t,$$

where $\mu_0(X_t, t)$ and $\sigma_0(X_t, t)$ are the true drift and diffusion functions, and W_t is a standard Brownian motion. Often it is assumed that $\mu_0(X_t, t)$ and $\sigma_0(X_t, t)$ belong to a certain parametric family:

$$\mu_0 \in \mathcal{M}_\mu \equiv \{\mu(\cdot, \cdot, \theta), \theta \in \Theta\} \quad \text{and} \quad \sigma_0 \in \mathcal{M}_\sigma \equiv \{\sigma(\cdot, \cdot, \theta), \theta \in \Theta\},$$

where Θ is a finite-dimensional parameter space. We say that models $\mathcal{M}_\mu$ and $\mathcal{M}_\sigma$ are correctly specified for drift $\mu_0(X_t, t)$ and diffusion $\sigma_0(X_t, t)$, respectively, if

$$\begin{aligned}\mathbb{H}_0 : P[\mu(X_t, t, \theta_0) = \mu_0(X_t, t), \sigma(X_t, t, \theta_0) = \sigma_0(X_t, t)] = 1 \\ \text{for some } \theta_0 \in \Theta. \qquad (1)\end{aligned}$$

The alternative hypothesis is that there exists no parameter value $\theta \in \Theta$ such that $\mu(X_t, t, \theta)$ and $\sigma(X_t, t, \theta)$ coincide with $\mu_0(X_t, t)$ and $\sigma_0(X_t, t)$, respectively, that is,

$$\begin{aligned}\mathbb{H}_A : P[\mu(X_t, t, \theta) = \mu_0(X_t, t), \sigma(X_t, t, \theta) = \sigma_0(X_t, t)] < 1 \\ \text{for all } \theta \in \Theta. \qquad (2)\end{aligned}$$

We will test whether a continuous-time model is correctly specified using $\{X_{\tau\Delta}\}_{\tau=1}^n$, a discrete sample of $\{X_t\}$ observed over a time span T at interval Δ, with the sample size $n \equiv T/\Delta$.

Assuming that X_t is a strictly stationary time-homogenous diffusion, Ait-Sahalia (1996) observes that any pair of drift model $\mu(X_t, \theta)$ and diffusion model $\sigma(X_t, \theta)$ completely characterizes the model marginal density

$$\pi(x, \theta) = \frac{\xi(\theta)}{\sigma^2(x, \theta)} \exp\left[\int_{x_0}^{x} \frac{2\mu(y, \theta)}{\sigma^2(y, \theta)} dy\right],$$

where $\xi(\theta)$ is a standardization factor and x_0 is the lower bound of the support of X_t. Ait-Sahalia (1996) proposes a novel test by comparing a kernel-based marginal density estimator $\hat{\pi}_0(\cdot)$ for $\{X_{\tau\Delta}\}$ with a parametric counterpart $\pi(\cdot,\hat{\theta})$, where $\hat{\theta}$ is a minimum-distance estimator for θ. The test is easy to implement and has power against many alternatives, but it will miss the alternatives that have the same marginal density as the null model.

Unlike the marginal density, the transition density of X_t can capture its full dynamics. Let $p_0(x,\ t\,|\,y,\ s)$ be the transition density of X_t; that is, the conditional density of $X_t = x$ given $X_s = y$, $s < t$. For a given pair of drift model $\mu(X_t, t, \theta)$ and diffusion model $\sigma(X_t, t, \theta)$, a family of transition densities $\{p(x,t\,|\,y,s,\theta)\}$ is characterized. When (and only when) $\mathbb{H}_0$ in Equation (1) holds, there exists some $\theta_0 \in \Theta$ such that $p(x,t\,|\,y,s,\theta_0) = p_0(x,\ t\,|\,y,\ s)$ almost everywhere for all $t > s$. Hence, the hypotheses of interest $\mathbb{H}_0$ in Equation (1) versus $\mathbb{H}_A$ in Equation (2) can be equivalently written as:

$$\mathbb{H}_0 : p(x,t\,|\,y,s,\theta_0) = p_0(x,t\,|\,y,s) \quad \text{almost everywhere for some } \theta_0 \in \Theta \qquad (3)$$

versus the alternative hypothesis

$$\mathbb{H}_A : p(x,t\,|\,y,s,\theta) \neq p_0(x,t\,|\,y,s) \quad \text{for some } t > s \text{ and for all } \theta \in \Theta. \qquad (4)$$

A natural approach to testing $\mathbb{H}_0$ in Equation (3) versus $\mathbb{H}_A$ in Equation (4) would be to follow Ait-Sahalia (1996) and compare a model transition density estimator $p(x,t\,|\,y,s,\hat{\theta})$ with a nonparametric counterpart, say $\hat{p}_0(x,t\,|\,y,s)$.[5] From Pritsker's (1998) analysis, however, we expect that the size performance of such a nonparametric test could be even worse than the marginal density-based test, because $\hat{p}_0(x,t\,|\,y,s)$ converges more slowly than the marginal density estimator $\hat{\pi}_0(x)$ due to the well-known "curse of dimensionality."[6] Furthermore, the finite sample distribution of the resulting test statistic is expected to be sensitive to dependent persistence in data.

Instead of comparing $p(x,t\,|\,y,s,\hat{\theta})$ and $\hat{p}_0(x,t\,|\,y,s)$ directly, we first transform the sample $\{X_{\tau\Delta}\}_{\tau=1}^{n}$ via the following dynamic probability integral transform:

$$Z_\tau(\theta) \equiv \int_{-\infty}^{X_{\tau\Delta}} p[x, \tau\Delta | X_{(\tau-1)\Delta}, (\tau-1)\Delta, \theta]dx, \quad \tau = 1, \ldots, n. \qquad (5)$$

[5] In addition to the marginal density-based test, Ait-Sahalia (1996) also develops a nonparametric test based on some functional of the transition density as implied by the forward and backward Kolmogorov equations under stationarity. This test, however, cannot capture the full dynamics of X_t.

[6] Under certain regularity conditions, the optimal convergence rates of $\hat{\pi}_0(x)$ and $\hat{p}_0(x,t\,|\,y,s)$ are $O(n^{-2/5})$ and $O(n^{-1/3})$, respectively. See Robinson (1983) for relevant discussion in the time series context.

Under $\mathbb{H}_0$ in Equation (3), there exists some $\theta_0 \in \Theta$ such that

$$p[x, \tau\Delta | X_{(\tau-1)\Delta}, (\tau-1)\Delta, \theta_0] = p_0[x, \tau\Delta | X_{(\tau-1)\Delta}, (\tau-1)\Delta]$$
$$\text{almost surely for all } \Delta > 0,$$

and the series $\{Z_\tau \equiv Z_\tau(\theta_0)\}_{\tau=1}^n$ is i.i.d. $U[0, 1]$. This is first proven, in a simpler context, by Rosenblatt (1952), and is used to evaluate out-of-sample density forecasts by Diebold, Gunther, and Tay (1998) and Hong (2001) in discrete-time contexts. We call $\{Z_\tau(\theta)\}$ the "generalized residuals" of model $p(x, t | y, s, \theta)$. Intuitively, the i.i.d. property characterizes correct specification of the model dynamics, and the $U[0, 1]$ property characterizes correct specification of the model marginal distribution.

Thus, we can test $\mathbb{H}_0$ versus $\mathbb{H}_A$ by checking whether the generalized residual series $\{Z_\tau(\theta)\}$ is i.i.d. $U[0, 1]$ for some $\theta = \theta_0$. In general, it is difficult to compute $\{Z_\tau(\theta)\}$ because the transition density of most continuous-time models has no closed-form. However, we can accurately approximate the model transition density by using the simulation methods of Pedersen (1995) and Brandt and Santa-Clara (2002), the Hermite expansion approach of Ait-Sahalia (2002a,b), or, for affine diffusions, the closed-form approximation of Duffie, Pedersen, and Singleton (2003) and the empirical characteristic function approach of Singleton (2001) and Jiang and Knight (2002).

1.2 Nonparametric omnibus test

It is nontrivial to test the joint hypothesis of i.i.d. $U[0, 1]$ for $\{Z_\tau\}_{\tau=1}^n$. One may suggest using the well-known Kolmogorov–Smirnov test, which unfortunately checks $U[0, 1]$ under the i.i.d. assumption rather than tests i.i.d. and $U[0, 1]$ jointly. It would easily miss the non-i.i.d. alternatives with uniform marginal distribution. Moreover, the Kolmogorov–Smirnov test cannot be used directly because it does not take into account the impact of parameter estimation uncertainty on the asymptotic distribution of the test statistic.

We propose to test i.i.d. $U[0, 1]$ by comparing a kernel estimator $\hat{g}_j(z_1, z_2)$ for the joint density $g_j(z_1, z_2)$ of $\{Z_\tau, Z_{\tau-j}\}$ with unity, the product of two $U[0, 1]$ densities. Our approach has at least three advantages. First, as there is no serial dependence in $\{Z_\tau\}$ under $\mathbb{H}_0$ in Equation (3), nonparametric joint density estimators are expected to perform much better in finite samples. In particular, the finite sample distribution of the resulting test will be robust to dependent persistence in data. Second, there is no asymptotic bias for nonparametric density estimators under $\mathbb{H}_0$ in Equation (3), because the conditional density of Z_τ given $\{Z_{\tau-1}, Z_{\tau-2}, \dots\}$ is a constant. Third, our test can be applied to time-inhomogeneous continuous-time processes because $\{Z_\tau\}$ is always i.i.d. $U[0, 1]$ under correct model specification.[7]

Our kernel estimator of the joint density $g_j(z_1, z_2)$ is, for any integer, $j>0$,

$$\hat{g}_j(z_1, z_2) \equiv (n-j)^{-1} \sum_{\tau=j+1}^{n} K_h(z_1, \hat{Z}_\tau) K_h(z_2, \hat{Z}_{\tau-j}), \tag{6}$$

where $\hat{Z}_\tau = Z_\tau(\hat{\theta})$, $\hat{\theta}$ is *any* $\sqrt{n}$-consistent estimator for θ_0, and $K_h(z_1, z_2)$ is a boundary-modified kernel defined below. For $x \in [0, 1]$, we define

$$K_h(x,y) \equiv \begin{cases} h^{-1}k\left(\frac{x-y}{h}\right) \Big/ \int_{-(x/h)}^{1} k(u)du, & \text{if } x \in [0,h), \\ h^{-1}k\left(\frac{x-y}{h}\right), & \text{if } x \in [h, 1-h], \\ h^{-1}k\left(\frac{x-y}{h}\right) \Big/ \int_{-1}^{(1-x)/h} k(u)du, & \text{if } x \in (1-h, 1], \end{cases} \tag{7}$$

where the kernel $k(\cdot)$ is a prespecified symmetric probability density, and $h \equiv h(n)$ is a bandwidth such that $h \to 0$, $nh \to \infty$ as $n \to \infty$. One example of $k(\cdot)$ is the quartic kernel

$$k(u) = \tfrac{15}{16}(1-u^2)^2 \mathbf{1}(|u| \le 1),$$

where $\mathbf{1}(\cdot)$ is the indicator function. We will use this kernel in our simulation study and empirical applications. In practice, the choice of h is more important than the choice of $k(\cdot)$. Like Scott (1992), we choose $h = \hat{S}_Z n^{-\frac{1}{6}}$, where $\hat{S}_Z$ is the sample standard deviation of $\{\hat{Z}_\tau\}_{\tau=1}^{n}$. This simple bandwidth rule attains the optimal rate for bivariate density estimation.

We use the modified kernel in Equation (7) because the standard kernel density estimator produces biased estimates near the boundaries of data due to asymmetric coverage of the data in the boundary regions. The denominators of $K_h(x,y)$ for $x \in [0,h) \cup (1-h, 1]$ account for the asymmetric coverage and ensure that the kernel density estimator is asymptotically unbiased uniformly over the entire support $[0,1]$. Our modified kernel has advantages over some alternative solutions to the boundary bias problem. One popular solution is to simply ignore the data in the boundary regions and use only the data in the interior region. Such trimming is simple, but in the present context, it may lead to the loss of a significant amount of information. If $h = sn^{-\frac{1}{5}}$, where $s^2 = \mathrm{var}(Z_t)$, for example, then about 23, 20, and 10% of a U [0, 1] sample will fall into

[7] Egorov, Li, and Xu (2003) extend Ait-Sahalia's (2002a) Hermite expansion approach to obtain accurate closed-form approximation for the transition density of time-inhomogeneous diffusion models.

the boundary regions when $n = 100$, 500, and 5000 respectively.[8] For financial time series $\{X_{\tau\Delta}\}$, one may be particularly interested in its tail distribution, which is exactly contained in (and only in) the boundary regions of $\{Z_\tau\}$!

Alternatively, one can follow Chapman and Pearson (2000) to use the jackknife kernel to eliminate the boundary bias.[9] The jackknife kernel, however, has the undesired property that it may generate negative density estimates in the boundary regions. It also gives a relatively large variance for the kernel estimates in the boundary regions, adversely affecting the power of the test in finite samples. In contrast, our modified kernel in Equation (7) always gives nonnegative density estimates with a smaller variance in the boundary regions than a jackknife kernel.

Similar to that in Ait-Sahalia (1996), our test is based on a quadratic form between $\hat{g}_j(z_1, z_2)$ and 1, the product of two $U[0, 1]$ densities:

$$\hat{M}(j) \equiv \int_0^1 \int_0^1 [\hat{g}_j(z_1, z_2) - 1]^2 dz_1 dz_2. \tag{8}$$

Our test statistic is a properly centered and scaled version of $\hat{M}(j)$:

$$\hat{Q}(j) \equiv \left[(n-j)h\hat{M}(j) - A_h^0\right] / V_0^{1/2}, \tag{9}$$

where the nonstochastic centering and scale factors

$$A_h^0 \equiv \left[(h^{-1} - 2)\int_{-1}^1 k^2(u)du + 2\int_0^1 \int_{-1}^b k_b^2(u)dudb\right]^2 - 1, \tag{10}$$

$$V_0 \equiv 2\left[\int_{-1}^1 \left[\int_{-1}^1 k(u+v)k(v)dv\right]^2 du\right]^2, \tag{11}$$

and $k_b(\cdot) \equiv k(\cdot)/\int_{-1}^b k(v)dv$. The modification of $k(\cdot)$ in the boundary regions does not affect the asymptotic variance V_0, but it affects the centering constant A_h^0, and this effect does not vanish even when $n \to \infty$.

Under correct model specification, we can show (see Theorem 1 in the appendix) that as $n \to \infty$,

$$\hat{Q}(j) \to N(0, 1) \quad \text{in distribution.}$$

The first lag $j = 1$ is often the most informative and important, but other lags may also reveal useful information on model misspecification. We

[8] The generalized residual series $\{Z_\tau\}$ is uniformly distributed over [0, 1] under $\mathbb{H}_0$ regardless of the distribution of the original series $\{X_{\tau\Delta}\}$. Under $\mathbb{H}_A$, the proportion of the original observations $\{X_{\tau\Delta}\}_{\tau=1}^n$ that fall into the boundary regions of $\{Z_\tau\}$ can be either smaller or larger than those under $\mathbb{H}_0$, depending on the model and the true data generating process X_t.

[9] See Härdle (1990) for further discussion on the jackknife kernel.

have, under correct model specification,

$$\mathrm{cov}[\hat{Q}(i), \hat{Q}(j)] \to 0 \quad \text{in probability for } i \neq j$$

as $n \to \infty$. This implies that $\hat{Q}(i)$ and $\hat{Q}(j)$ are asymptotically independent whenever $i \neq j$ (cf. Theorem 2 in the appendix). As a result, we can simultaneously use multiple statistics $\{\hat{Q}(j)\}$ with different lags to examine at which lag(s) the i.i.d. $U[0, 1]$ property is violated.

Under model misspecification, we can show that as $n \to \infty$,

$$\hat{Q}(j) \to \infty \quad \text{in probability}$$

whenever $\{Z_\tau, Z_{\tau-j}\}$ are not independent or $U[0, 1]$ (see Theorem 3 in the appendix).[10]

We now summarize our omnibus test procedure: (1) estimate the continuous-time model using any method that yields a $\sqrt{n}$-consistent estimator $\hat{\theta}$;[11] (2) compute the model generalized residuals $\{\hat{Z}_\tau = Z_\tau(\hat{\theta})\}_{\tau=1}^n$, where $Z_\tau(\theta)$ is given in Equation (5); (3) compute the boundary-modified kernel joint density estimator $\hat{g}_j(z_1, z_2)$ in Equation (6) for a prespecified lag j; (4) compute the test statistic $\hat{Q}(j)$ in Equation (9), and compare it with the upper-tailed N(0,1) critical value C_α at level α (e.g., $C_{0.05} = 1.645$). If $\hat{Q}(j) > C_\alpha$, reject $\mathbb{H}_0$ at level α. Note that upper-tailed N(0,1) critical values are suitable because for sufficiently large n, negative values of $\hat{Q}(j)$ occur only under $\mathbb{H}_0$.[12]

1.3 Separate inference

When a model is rejected using $\hat{Q}(j)$, it would be interesting to explore possible sources of the rejection. The model generalized residuals $\{Z_\tau(\theta)\}$ contain rich information on model misspecification. In an out-of-sample density forecast context, Diebold, Gunther, and Tay (1998) illustrate how to use the histogram of $\{Z_\tau\}$ and autocorrelogram in the powers of $\{Z_\tau\}$ to reveal sources of model misspecification. Similarly, we can compare the kernel-based marginal density estimator of the generalized residuals with

[10] In both our simulation study and empirical application below, we find that the power of $\hat{Q}(j)$ is more or less robust to lag order j. However, from a theoretical perspective, the power of $\hat{Q}(j)$ could be sensitive to the choice of j. To avoid this, we can consider a portmanteau test statistic $W(p) = p^{-1/2} \sum_{j=1}^{p} \hat{Q}(j)$, which, for any fixed integer $p > 0$, converges to $N(0, 1)$ under $\mathbb{H}_0$ given Theorems 1 and 2. The power of $W(p)$ still depends on the choice of p, but such dependence is expected to be less severe than the dependence of $\hat{Q}(j)$ on the choice of j.

[11] As an important feature of our test, we only require that the parameter estimator $\hat{\theta}$ be $\sqrt{n}$-consistent. We need not use asymptotically most efficient estimator. The sampling variation in $\hat{\theta}$ has no impact on the asymptotic distribution of $\hat{Q}(j)$. This delivers a convenient and generally applicable procedure in practice, because asymptotically most efficient estimators such as MLE or approximated MLE may be difficult to obtain in practice. One could choose a suboptimal, but convenient, estimator in implementing our procedure.

[12] We also develop an omnibus test based on the Hellinger metric, which is a quadratic form between $\sqrt{\hat{g}_j(z_1, z_2)}$ and $\sqrt{1 \cdot 1} = 1$. The Hellinger metric is more robust to imprecise parameter estimates and outliers in data. Our simulation shows that the Hellinger metric test has less (more) accurate size than the $\hat{Q}(j)$ test when $n < 1000$ ($n \geq 1000$). These results are available from the authors upon request.

$U[0, 1]$ to check how well a model fits the stationary distribution of the underlying process (e.g., whether the model can explain skewness and kurtosis). We can also examine autocorrelations in the powers of Z_τ, which are very informative about how well a model fits various dynamic aspects of the underlying process.

Although intuitive and convenient, these graphical methods ignore the impact of parameter estimation uncertainty in $\hat{\theta}$ on the asymptotic distribution of evaluation statistics, which generally exists even when $n \to \infty$. Here, we provide a class of rigorous separate inference procedures that address this issue:

$$M(m,l) \equiv \left[\sum_{j=1}^{n-1} w^2(j/p)(n-j)\hat{\rho}^2_{ml}(j) - \sum_{j=1}^{n-1} w^2(j/p) \right] \Big/ \left[2\sum_{j=1}^{n-2} w^4(j/p) \right]^{1/2}, \quad (12)$$

where $\hat{\rho}_{ml}(j)$ is the sample cross-correlation between $\hat{Z}^m_\tau$ and $\hat{Z}^l_{\tau-|j|}$, and $w(\cdot)$ is a weighting function of lag order j. We assume that $w(\cdot)$ is symmetric about zero and continuous on $\mathbb{R}$ except for a finite number of points. An example is the Bartlett kernel $w(z) = (1 - |z|)\mathbf{1}(|z| \leq 1)$. If $w(\cdot)$ has bounded support, p is a lag truncation order; if $w(\cdot)$ has unbounded support, all $n-1$ lags in the sample $\{X_{\tau\Delta}\}_{\tau=1}^n$ are used. Usually $w(\cdot)$ discounts higher order lags. This will give better power than equal weighting when $|\rho_{ml}(j)|$ decays to zero as lag order j increases. This is typically the case for most financial markets, where the recent events tend to have bigger impact than the remote past events.

The tests $M(m, l)$ are an extension of Hong's (1996) spectral density tests for the adequacy of discrete-time linear dynamic models with exogenous regressors. Extending the proof of Hong (1996), we can show that for each given pair of positive integers (m, l),

$$M(m,l) \to N(0,1) \quad \text{in distribution}$$

under correct model specification, provided the lag truncation order $p \equiv p(n) \to \infty$, $p/n \to 0$. Moreover, parameter estimation uncertainty in $\hat{\theta}$ has no impact on the asymptotic distribution of $M(m, l)$. Although the moments of the generalized residuals $\{Z_\tau\}$ are not the same as that of the original data $\{X_{\tau\Delta}\}$, they are highly correlated. In particular, the choice of $(m, l) = (1, 1), (2, 2), (3, 3), (4, 4)$ is very sensitive to autocorrelations in level, volatility, skewness, and kurtosis of $\{X_{\tau\Delta}\}$, respectively (cf. Diebold, Gunther, and Tay 1998). Furthermore, the choice of $(m, l) = (1, 2)$ and $(2, 1)$ is sensitive to ARCH-in-Mean and "leverage" effects, respectively. Different choices of order (m, l) can thus examine various dynamic aspects of the underlying process. Like $\hat{Q}(j)$, upper-tailed N(0, 1) critical values are suitable for $M(m, l)$.

Our omnibus and separate inference tests complement Gallant and Tauchen's (1996) popular EMM-based tests, which are also based on a nonparametric approach. While we use the model transition density directly, Gallant and Tauchen (1996) examine the simulation-based expectation of an auxiliary semi-nonparametric score function under the model distribution, which takes the value of zero under correct model specification. Like our tests, the EMM tests are applicable to stationary continuous/discrete-time univariate and multivariate models. In addition to the minimum chi-square test for generic model misspecification, the EMM approach also provides a spectrum of constructive individual t-statistics, which, similar to our separate inference tests $M(m, l)$, are informative in revealing possible sources of model misspecification.

In empirical financial studies, it is difficult to formally compare the relative performance of non-nested models using most existing tests, including EMM [e.g., Dai and Singleton (2000)]. In contrast, our nonparametric approach makes it possible to compare the performance of non-nested models via a metric measuring the distance of the model generalized residuals from i.i.d. $U[0, 1]$. As the transition density can capture the full dynamics of $\{X_t\}$, our omnibus test has power against any model misspecification when there is only one observable component in X_t, as is the case of univariate diffusion models. In contrast, as pointed out by Tauchen (1997, Section 4.3), the EMM minimum chi-square test may lack power against some alternatives, because the semi-nonparametric score may have zero expectation under the distribution of a misspecified model.

2. Finite Sample Performance

We now study the finite sample performance of the $\hat{Q}(j)$ test for both univariate and multivariate continuous-time models. For univariate models, we adopt the same simulation design as Pritsker (1998), who has conducted a simulation study on Ait-Sahalia's (1996) test. For multivariate models, we focus on affine diffusions given their importance in the existing asset pricing literature [see, e.g., Duffie, Pan, and Singleton (2000)]. We choose the simulation design of Ait-Sahalia and Kimmel (2002), who study the finite sample performance of parameter estimators for multivariate affine diffusion models using Ait-Sahalia's (2002b) closed-form likelihood expansion.

2.1 Univariate models

2.1.1. Size of the $\hat{Q}(j)$ test. To examine the size of $\hat{Q}(j)$ for univariate models, we simulate data from the Vasicek (1977) model:

$$dX_t = \kappa(\alpha - X_t)dt + \sigma dW_t, \tag{13}$$

where α is the long run mean and κ is the speed of mean reversion. The smaller κ is, the stronger the serial dependence in $\{X_t\}$, and consequently, the slower the convergence to the long run mean. We are particularly interested in the impact of dependent persistence in $\{X_t\}$ on the size of $\hat{Q}(j)$. Given that the finite sample performance of $\hat{Q}(j)$ may depend on both the marginal density and dependent persistence of $\{X_t\}$, we follow Pritsker (1998) to keep the marginal density unchanged while varying the speed of mean reversion. This is achieved by changing κ and σ^2 in the same proportion. In this way, we can focus on the impact of dependent persistence. We consider both low and high levels of dependent persistence and adopt the same parameter values as Pritsker (1998): $(\kappa, \alpha, \sigma^2) = (0.85837, 0.089102, 0.002185)$ and $(0.214592, 0.089102, 0.000546)$ for the low and high persistent dependence cases, respectively.

For each parameterization, we simulate 1000 data sets of a random sample $\{X_{\tau\Delta}\}_{\tau=1}^{n}$ at daily frequency for $n = 250$, 500, 1000, 2500, and 5500, respectively. These sample sizes correspond to about 1–22 years of daily data. For each data set, we estimate the model parameters $\theta = (\kappa, \alpha, \sigma^2)'$ via the maximum likelihood estimation (MLE) method and compute our $\hat{Q}(j)$ statistic using the generalized residuals of the estimated Vasicek model. We consider the empirical rejection rates using the asymptotic critical values (1.28 and 1.65) at the 10 and 5% levels, respectively.

Figures 1*a*–*d* report the empirical sizes of $\hat{Q}(j)$, $j = 1, \ldots, 20$, for $n = 250$, 500, 1000, 2500, and 5500. Figures 1*a* and *c* (*b* and *d*) show the rejection rates of $\hat{Q}(j)$ at the 10 and 5% levels under a correct Vasicek model with low (high) persistence of dependence. Overall, $\hat{Q}(j)$ has reasonable sizes for sample sizes as small as $n = 250$ (i.e., about one year of daily data), at both the 10 and 5% levels. The most striking difference from Ait-Sahalia's (1996) test is that the impact of dependent persistence on the size of $\hat{Q}(j)$ is minimal: the sizes of $\hat{Q}(j)$ are virtually the same in both the low and high persistent cases. In contrast, under the same simulation setting, Pritsker (1998) finds that Ait-Sahalia's (1996) test shows strong overrejection under a correct Vasicek model even when $n = 5500$, and it becomes worse when dependence becomes stronger.

2.1.2. Power of the $\hat{Q}(j)$ test. To investigate the power of $\hat{Q}(j)$ for univariate models, we simulate data from four popular univariate diffusion processes and test the null hypothesis that the data is generated from a Vasicek model. The four processes are:

1. The CIR model:

$$dX_t = \kappa(\alpha - X_t)dt + \sigma\sqrt{X_t}dW_t, \tag{14}$$

where $(\kappa, \alpha, \sigma^2) = (0.89218, 0.090495, 0.032742)$.

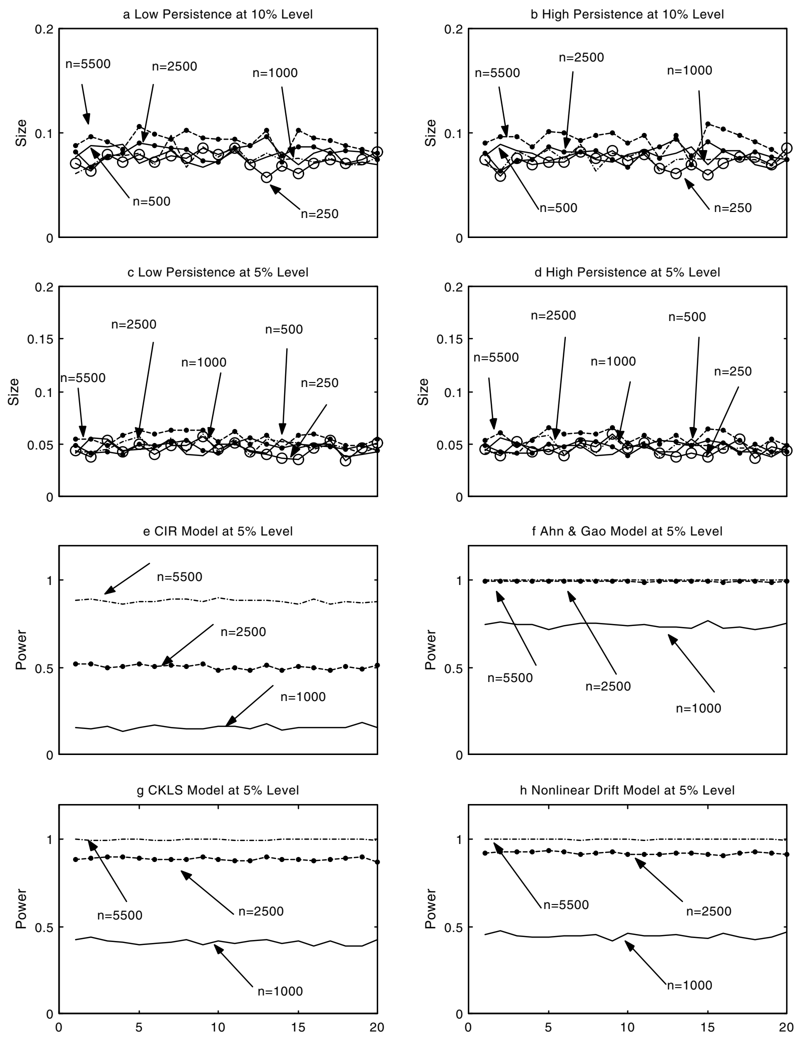

Figure 1
The finite sample size and power performance of Q-test for univariate diffusions
To examine the size of $\hat{Q}(j)$, we simulate 1000 data sets from the Vasicek model, $dX_t = \kappa\,(\alpha - X_t)\,dt + \sigma\,dW_t$, at daily frequency for $n = 250$, 500, 1000, 2500, and 5500, respectively. We choose $(\kappa, \alpha, \sigma^2) = (0.85837, 0.089102, 0.002185)$ and $(0.214592, 0.089102, 0.000546)$ for the low and high persistent dependence cases, respectively. For each data set, we estimate $\theta = (\kappa, \alpha, \sigma^2)'$ via MLE and compute $\hat{Q}(j)$ statistic using the generalized residuals of the estimated Vasicek model. We consider the empirical rejection rates using the asymptotic critical values (1.28 and 1.65) at the 10 and 5% levels, respectively. Figures 1*a* and *c* (*b* and *d*) report the rejection rates of $\hat{Q}(j)$ at the 10 and 5% levels under a correct Vasicek model with low (high) persistence of dependence. To examine the power of $\hat{Q}(j)$, we simulate 500 data sets from each of the four alternatives: the CIR, Ahn and Gao, CKLS, and Ait-Sahalia's nonlinear drift model, at daily frequency for $n = 1000$, 2500, and 5500, respectively. For each data set, we estimate a Vasicek model and use its generalized residuals to compute $\hat{Q}(j)$. Figures 1*e*–*h* report the power of $\hat{Q}(j)$ at the 5% level using asymptotic critical values against the CIR, Ahn and Gao, CKLS, and Ait-Sahalia's nonlinear drift model, respectively.

2. Ahn and Gao's (1999) Inverse-Feller model:

$$dX_t = X_t[\kappa - (\sigma^2 - \kappa\alpha)X_t]dt + \sigma X_t^{3/2} dW_t, \tag{15}$$

where $(\kappa, \alpha, \sigma^2) = (0.181, 15.157, 0.032742)$.

3. CKLS model:

$$dX_t = \kappa(\alpha - X_t)dt + \sigma X_t^{\rho} dW_t, \tag{16}$$

where $(\kappa, \alpha, \sigma^2, \rho) = (0.0972, 0.0808, 0.52186, 1.46)$.

4. Ait-Sahalia's (1996) nonlinear drift model:

$$dX_t = (\alpha_{-1}X_t^{-1} + \alpha_0 + \alpha_1 X_t + \alpha_2 X_t^2)dt + \sigma X_t^{\rho} dW_t, \tag{17}$$

where $(\alpha_{-1}, \alpha_0, \alpha_1, \alpha_2, \sigma^2, \rho) = (0.00107, -0.0517, 0.877, -4.604, 0.64754, 1.50)$.

The parameter values for the CIR model are taken from Pritsker (1998). For the other three models, which Pritsker (1998) does not consider, the parameter values are taken from Ait-Sahalia's (1999) estimates. For each of these four alternatives, we generate 500 realizations of a random sample $\{X_{\tau\Delta}\}_{\tau=1}^{n}$, where $n = 1000$, 2500, and 5500, respectively. For the CIR and Ahn and Gao's models, we simulate data from the model transition density, which has a closed-form. For the CKLS and Ait-Sahalia's nonlinear drift models whose transition density has no closed form, we simulate data using the convenient Milstein scheme. To reduce discretization bias, we simulate five observations each day and sample the data at daily frequency.

For each data set, we estimate a Vasicek model and use its generalized residuals to compute $\hat{Q}(j)$. Figures 1*e*–*h* report the power of $\hat{Q}(j)$ at the 5% level using asymptotic critical values for $n = 1000$, 2500, and 5500.[13] The test has overall good power in detecting misspecification of the Vasicek model against the four alternatives. When $n = 5500$, $\hat{Q}(j)$ rejects the CIR model with a power of about 90%. For comparison, Pritsker (1998), under the same simulation setting, reports that the size-corrected power of Ait-Sahalia's (1996) test at the 5% level in detecting the Vasicek model against the CIR alternative is about 38% when $n = 5500$. Moreover, $\hat{Q}(j)$ has virtually unit power against the other three alternatives when $n = 5500$. It appears that the transition density-based test is indeed more powerful than the marginal density-based test.

2.2 Multivariate models

For multivariate models, we cannot directly use the probability integral transform $\{Z_\tau(\theta)\}$ defined in Equation (5) for univariate models. Instead,

[13] The results using empirical critical values are very similar and are available from the authors upon request.

we have to conduct the probability integral transform for each individual observable component of X_t, conditioning on an appropriate information set. If the transformed series is not i.i.d. $U[0, 1]$ for some observable components of X_t, there then exists evidence of model misspecification.

Suppose we have a set of discrete-time observations of an N-dimensional continuous-time process X_t, $\{X_{i,\tau\Delta}\}_{\tau=1}^{n}$, $i = 1, \ldots, N$. Following Diebold, Hahn, and Tay (1999), we can partition the model-implied joint transition density of the N state variables $(X_{1,\tau\Delta}, \ldots, X_{N,\tau\Delta})$ at time $\tau\Delta$ into the products of N conditional densities,

$$p(X_{1,\tau\Delta}, \ldots, X_{N,\tau\Delta,\tau}\Delta | I_{(\tau-1)\Delta}, (\tau-1)\Delta, \theta)$$
$$= \prod_{i=1}^{N} p(X_{i,\tau\Delta,\tau}\Delta | X_{i-1,\tau\Delta}, \ldots, X_{1,\tau\Delta}, I_{(\tau-1)\Delta}, (\tau-1)\Delta, \theta),$$

where the conditional density $p(X_{i,\tau\Delta}|X_{i-1,\tau\Delta}, \ldots, X_{1,\tau\Delta}, I_{(\tau-1)\Delta}, (\tau-1)\Delta, \theta)$ of the i-th component $X_{i,\tau\Delta}$ depends on not only the past information $I_{(\tau-1)\Delta}$ but also other contemporaneous variables $\{X_{l,\tau\Delta}\}_{l=1}^{i-1}$. We then transform $X_{i,\tau\Delta}$ via its corresponding model-implied transition density

$$Z_{i,\tau}^{(1)}(\theta) \equiv \int_{-\infty}^{X_{i,\tau\Delta}} p(x, t\Delta | X_{i-1,\tau\Delta}, \ldots, X_{1,\tau\Delta}, I_{(\tau-1)\Delta}, (\tau-1)\Delta, \theta) dx,$$
$$i = 1, \ldots, N. \tag{18}$$

This approach produces N generalized residual samples, $\{Z_{i,\tau}^{(1)}(\theta)\}_{\tau=1}^{n}, i = 1, \ldots, N$, which we can use to evaluate the performance of a given multivariate model in capturing the dynamics of X_t. For each i, the series $\{Z_{i,\tau}^{(1)}(\theta)\}_{\tau=1}^{n}$ should be i.i.d. $U[0, 1]$ when $\theta = \theta_0$ under correct model specification.

There are $N!$ ways of factoring the joint transition density of X_t and it is possible that some particular sequence of conditioning partition may lead to low or no power for the test. While we only consider one of such $N!$ possibilities here, in practice there may be some natural ordering given the kind of questions to be addressed. For example, in term structure modeling, the transition density of long term bond yields could be conditioned on contemporaneous and past short term yields if one is interested in inferring the dynamics of long bond yields from short bond yields. This is especially appropriate when one is interested in the impact on long bond yields of a sudden policy change by the Fed on the spot rates.

Examination of generalized residuals for each observable component in X_t will reveal useful information about how well a multivariate model can capture the dynamics of each component in X_t. To assess the overall performance of the model in capturing the joint dynamics of X_t, we can combine the N individual generalized residuals $\{Z_{i,\tau}^{(1)}(\theta)\}_{\tau=1}^{n}$ in a suitable way to generate a new sequence, which we may call the *combined* generalized residuals of a multivariate model:

$$Z^{(2)}(\theta) \equiv \left[Z^{(1)}_{1,1}(\theta),\ldots,Z^{(1)}_{N,1}(\theta),Z^{(1)}_{1,2}(\theta),\ldots,Z^{(1)}_{N,2}(\theta),\ldots,Z^{(1)}_{1,n}(\theta),\ldots,Z^{(1)}_{N,n}(\theta)\right]', \tag{19}$$

The series $\{Z^{(2)}_\tau(\theta)\}^{nN}_{\tau=1}$ is also i.i.d. $U[0, 1]$ for $\theta=\theta_0$ under correct model specification and this property can be used to check the overall performance of the multivariate model.

2.2.1. Size of the $\hat{Q}(j)$ test. To examine the size of $\hat{Q}(j)$ for multivariate models, we simulate data from a three-factor Vasicek model:

$$d\begin{bmatrix} X_{1t} \\ X_{2t} \\ X_{3t} \end{bmatrix} = \begin{bmatrix} \kappa_{11} & \kappa_{12} & \kappa_{13} \\ \kappa_{21} & \kappa_{22} & \kappa_{23} \\ \kappa_{31} & \kappa_{32} & \kappa_{33} \end{bmatrix} \begin{bmatrix} \alpha_1 - X_{1t} \\ \alpha_2 - X_{2t} \\ \alpha_3 - X_{3t} \end{bmatrix} dt + \begin{bmatrix} \sigma_{11} & 0 & 0 \\ 0 & \sigma_{22} & 0 \\ 0 & 0 & \sigma_{33} \end{bmatrix} d\begin{bmatrix} W_{1t} \\ W_{2t} \\ W_{3t} \end{bmatrix}. \tag{20}$$

Following Ait-Sahalia and Kimmel (2002), we set $(\kappa_{11}, \kappa_{21}, \kappa_{22}, \kappa_{31}, \kappa_{32}, \kappa_{33}, \sigma_{11}, \sigma_{22}, \sigma_{33})' = (0.50, -0.20, 1.00, 0.10, 0.20, 2.00, 1.00, 1.00, 1.00)'$, and $(\kappa_{12}, \kappa_{13}, \kappa_{23}, \alpha_1, \alpha_2, \alpha_3)' = (0, 0, 0, 0, 0, 0)'$. We simulate 1000 data sets of the random sample $\{X_{\tau\Delta}\}^n_{\tau=1}$ at the monthly frequency for $n=250$, 500, and 1000, respectively. These sample sizes correspond to about 20–100 years of monthly data.[14] For each data set, we use MLE to estimate all model parameters in Equation (20), with no restrictions on the κ matrix and intercept coefficients. We then compute the $\hat{Q}(j)$ statistic using the generalized residuals of the estimated three-factor Vasicek model.[15] The empirical rejection rates using the asymptotic critical values (1.28 and 1.65) at the 10 and 5% levels are considered.

Figure 2 reports the empirical sizes of $\hat{Q}(j)$, $j=1, \ldots, 20$, for $n=250$, 500, and 1000, respectively. Figures 2*a–d* show the rejection rates of $\hat{Q}(j)$ at the 10% level for three individual and combined generalized residuals, while Figures 2*e–h* report the rejection rates at the 5% level. Overall, $\hat{Q}(j)$ has good sizes at both the 10 and 5% levels for sample sizes as small as $n=250$ (i.e., about 20 years of monthly data). Our results show that the good finite sample size performance of $\hat{Q}(j)$ in the univariate models carries over to the multivariate models as well.

[14] We simulate 100 observations each month and sample the data at the monthly frequency. We choose the monthly sampling frequency to match our empirical application to multivariate models below, where we use monthly data of U.S. Treasury yields.

[15] Both the likelihood estimation and the probability integral transforms are done in closed-form, because the transition density of X_t in (20) is multivariate Gaussian. See Ait-Sahalia (2002b) for more details.

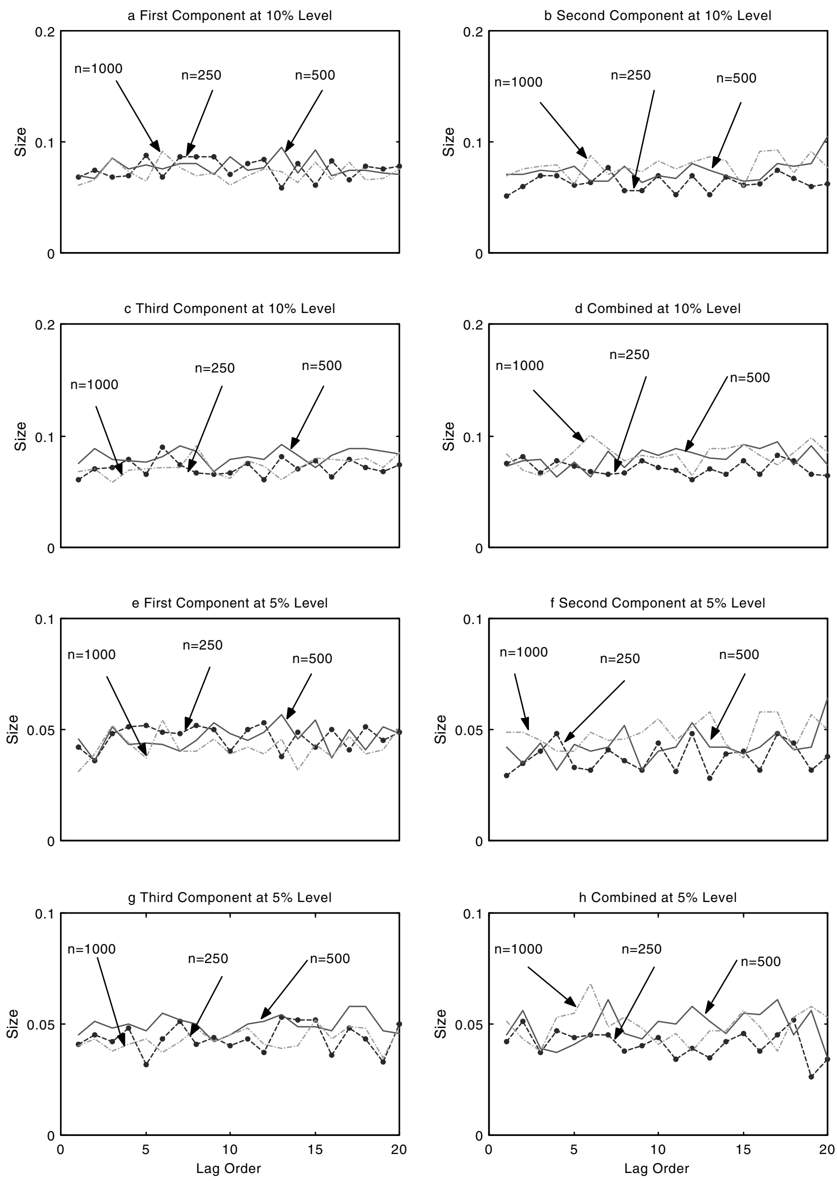

Figure 2

The finite sample size performance of Q-test for multivariate affine diffusions

To examine the size of $\hat{Q}(j)$, we simulate 1000 data sets from a three-factor Vasicek model at monthly frequency for $n = 250$, 500, and 1000, respectively. For each data set, we estimate model parameters via MLE and compute $\hat{Q}(j)$ statistic using the generalized residuals of the estimated Vasicek model. We consider the empirical rejection rates using the asymptotic critical values (1.28 and 1.65) at the 10 and 5% levels, respectively. Figures 2*a–d* report the rejection rates of $\hat{Q}(j)$ at the 10% level for three individual and combined generalized residuals, respectively, while Figures 2*e–h* report the rejection rates of $\hat{Q}(j)$ at the 5% level.

2.2.2. Power of the $\hat{Q}(j)$ test. To investigate the power of $\hat{Q}(j)$ for multivariate models, we simulate data from three other canonical affine diffusions and test the null hypothesis that the data is generated from the Vasicek model in Equation (20). We choose the same parameter values as Ait-Sahalia and Kimmel (2002) in each of the following three canonical affine diffusions:

1. $\mathbf{A}_1$ (3):

$$d\begin{bmatrix} X_{1t} \\ X_{2t} \\ X_{3t} \end{bmatrix} = \begin{bmatrix} \kappa_{11} & 0 & 0 \\ \kappa_{21} & \kappa_{22} & \kappa_{23} \\ \kappa_{31} & \kappa_{32} & \kappa_{33} \end{bmatrix} \begin{bmatrix} \alpha_1 - X_{1t} \\ -X_{2t} \\ -X_{3t} \end{bmatrix} dt + \begin{bmatrix} X_{1t}^{1/2} & 0 & 0 \\ 0 & (1+\beta_{21}X_{1t})^{1/2} & 0 \\ 0 & 0 & (1+\beta_{31}X_{1t})^{1/2} \end{bmatrix} d\begin{bmatrix} W_{1t} \\ W_{2t} \\ W_{3t} \end{bmatrix}, \tag{21}$$

where $(\kappa_{11}, \kappa_{22}, \kappa_{32}, \kappa_{33}, \alpha_1) = (0.50, 2.00, -0.10, 5.00, 2.00)$ and $(\kappa_{21}, \kappa_{23}, \kappa_{31}, \beta_{21}, \beta_{31}) = (0, 0, 0, 0, 0)$.

2. $\mathbf{A}_2$ (3):

$$d\begin{bmatrix} X_{1t} \\ X_{2t} \\ X_{3t} \end{bmatrix} = \begin{bmatrix} \kappa_{11} & \kappa_{12} & 0 \\ \kappa_{21} & \kappa_{22} & 0 \\ \kappa_{31} & \kappa_{32} & \kappa_{33} \end{bmatrix} \begin{bmatrix} \alpha_1 - X_{1t} \\ \alpha_2 - X_{2t} \\ -X_{3t} \end{bmatrix} dt + \begin{bmatrix} X_{1t}^{1/2} & 0 & 0 \\ 0 & X_{2t}^{1/2} & 0 \\ 0 & 0 & (1+\beta_{31}X_{1t}+\beta_{32}X_{2t})^{1/2} \end{bmatrix} d\begin{bmatrix} W_{1t} \\ W_{2t} \\ W_{3t} \end{bmatrix}, \tag{22}$$

where $(\kappa_{11}, \kappa_{22}, \kappa_{33}, \alpha_1, \alpha_2) = (0.50, 2.00, 5.00, 2.00, 1.00)$ and $(\kappa_{12}, \kappa_{21}, \kappa_{31}, \kappa_{32}, \beta_{31}, \beta_{32}) = (0, 0, 0, 0, 0, 0)$.

3. $\mathbf{A}_3$ (3):

$$d\begin{bmatrix} X_{1t} \\ X_{2t} \\ X_{3t} \end{bmatrix} = \begin{bmatrix} \kappa_{11} & \kappa_{12} & \kappa_{13} \\ \kappa_{21} & \kappa_{22} & \kappa_{23} \\ \kappa_{31} & \kappa_{32} & \kappa_{33} \end{bmatrix} \begin{bmatrix} \alpha_1 - X_{1t} \\ \alpha_2 - X_{2t} \\ \alpha_3 - X_{3t} \end{bmatrix} dt + \begin{bmatrix} X_{1t}^{1/2} & 0 & 0 \\ 0 & X_{2t}^{1/2} & 0 \\ 0 & 0 & X_{3t}^{1/2} \end{bmatrix} d\begin{bmatrix} W_{1t} \\ W_{2t} \\ W_{3t} \end{bmatrix}, \tag{23}$$

where $(\kappa_{11}, \kappa_{22}, \kappa_{33}, \alpha_1, \alpha_2, \alpha_3) = (0.50, 2.00, 1.00, 2.00, 1.00, 1.00)$ and $(\kappa_{12}, \kappa_{13}, \kappa_{21}, \kappa_{23}, \kappa_{31}, \kappa_{32}) = (0, 0, 0, 0, 0, 0)$.

For each of alternatives (21)–(23), we generate 500 data sets of the random sample $\{X_{\tau\Delta}\}_{\tau=1}^{n}$, for $n=250$, 500, and 1000, respectively at the monthly sample frequency. For each data set, we estimate the Vasicek model in Equation (20) via MLE and use its generalized residuals to compute $\hat{Q}(j)$. Again, we make no restrictions on the κ matrix and intercept coefficients in our estimation.

Figures 3*a*–*d* report the power of $\hat{Q}(j)$ under the $\mathbf{A}_1$ (3) alternative at the 5% level using asymptotic critical values for three individual and combined generalized residuals, respectively, for $n=250$, 500, and 1000. The test based on the generalized residuals of the first component X_{1t} has excellent power in detecting misspecification in this component, which is non-Gaussian under $\mathbf{A}_1$ (3). The power increases with sample size n and approaches unity when $n=1000$. One possible reason for such excellent performance is that the transition density of X_{1t}, which is noncentral Chi-square, deviates more significantly from the Gaussian distribution at the monthly sample interval than at daily sample interval. Interestingly, the powers of the test based on the generalized residuals of the other two (Gaussian) individual components are close to 5%, which shows that these individual tests do not overreject correctly specified individual components. On the other hand, the overall $\hat{Q}(j)$ test based on the combined generalized residuals has power against the Vasicek model under the $\mathbf{A}_1$ (3) alternative and its power increases with the sample size n. But the power of the overall test is smaller than that of the test based on the first component, apparently due to the fact that two-thirds of the combined generalized residuals are approximately i.i.d. $U[0, 1]$. Interestingly, the power of the test based on the combined generalized residuals is significantly higher at lag orders that are multiples of 3 (the dimension of X_t), displaying a clear periodic pattern. This is apparently due to the way in which the combined generalized residuals are constructed: at a lag order j that is multiples of 3, the generalized residuals $(\hat{Z}_{\tau}^{(2)}, \hat{Z}_{\tau-j}^{(2)})$ correspond to the same individual component in X_t, thus giving stronger serial dependence when the observations are from the misspecified first component X_{1t}. Consequently, it is easier to reject the three-factor Vasicek model at these lags.

Finally, we observe that the power patterns of the $\hat{Q}(j)$ test against the Vasicek model under both the $\mathbf{A}_2$ (3) alternative (Figures 3*e*–*h*) and the $\mathbf{A}_3$ (3) alternative (Figures 3*i*–*l*) are very similar to those under the $\mathbf{A}_1$ (3) alternative.

Overall, our simulation study shows that with the use of the probability integral transform and the boundary-modified kernel density estimator, our nonparametric $\hat{Q}(j)$ test performs rather well for both univariate and multivariate models, even for highly persistently dependent data with sample sizes often encountered in empirical finance.

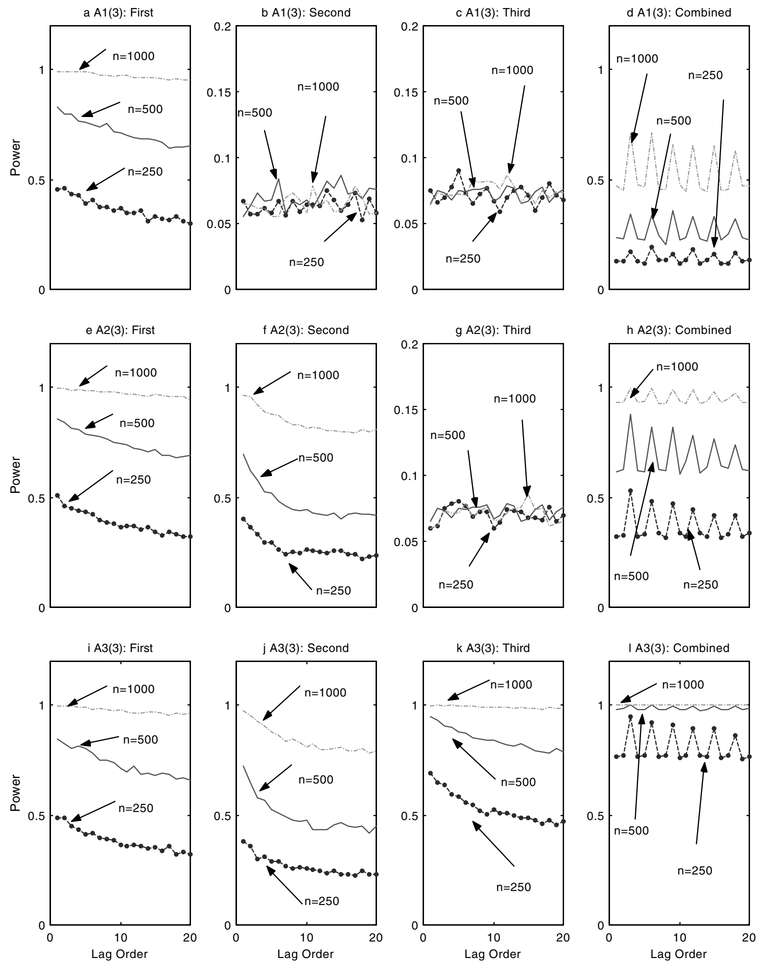

Figure 3
The finite sample power performance of Q-test for multivariate affine diffusions
To examine the power of $\hat{Q}(j)$, we simulate 500 data sets from three canonical trivariate affine diffusions: $\mathbf{A}_1$ (3), $\mathbf{A}_2$ (3), and $\mathbf{A}_3$ (3), at monthly frequency for $n = 250$, 500, and 1000, respectively. For each data set, we estimate a three-factor Vasicek model and use its generalized residuals to compute $\hat{Q}(j)$. Figures 3*a–d* report the power of $\hat{Q}(j)$ at the 5% level using asymptotic critical values against $\mathbf{A}_1$ (3) for three individual and combined generalized residuals, respectively. Figures 3*e–h* (i–l) report the power of $\hat{Q}(j)$ at the 5% level using asymptotic critical values against $\mathbf{A}_2$ (3) [$\mathbf{A}_3$ (3)] for three individual and combined generalized residuals, respectively.

3. Applications to Spot Interest Rate Models

We first use our tests to reexamine the spot rate models considered in Ait-Sahalia (1996), using the same data set: daily Eurodollar rates from June 1, 1973 to February 25, 1995, with a total of 5505 observations. Ait-Sahalia (1996) provides detailed summary statistics for the data.

We consider five popular models: the Vasicek, CIR, Ahn and Gao, CKLS, and Ait-Sahalia's nonlinear drift models, as given in Equations (13)–(17). For each model, we estimate parameters via MLE. For the Vasicek, CIR, and Ahn and Gao's models, the model likelihood function has a closed-form. For the CKLS and Ait-Sahalia's nonlinear drift models, we use Ait-Sahalia's (2002a) Hermite expansion to obtain a closed-form approximation for the model likelihood. Table 1 gives parameter estimates.

Figure 4*a* shows that the $\hat{Q}(j)$ statistics for lag order j from 1 to 20 for the five models range from 349.81 to 1574.02. Compared to upper-tailed $N(0, 1)$ critical values (e.g., 2.33 at the 1% level), these huge $\hat{Q}(j)$ statistics are overwhelming evidence that all five models are severely misspecified. The Vasicek model performs the worst, with the $\hat{Q}(j)$ values around 1400 for all lags from 1 to 20. This is probably due to its restrictive assumption of constant volatility. The CIR model dramatically reduces the $\hat{Q}(j)$ values to about 620, and the goodness of fit is further improved, in their order, by Ahn and Gao's model, Ait-Sahalia's nonlinear drift model, and the CKLS model. The latter performs the best, with the $\hat{Q}(j)$ values around 400. This suggests that level effect is important for modeling interest rate dynamics, but in contrast to the findings by Ait-Sahalia (1996) and Stanton (1997), nonlinear drift does not improve goodness of fit.

Although some models perform relatively better than others, the extremely large $\hat{Q}(j)$ statistics indicate that none of the five univariate diffusion models adequately captures the interest rate dynamics. There is a

Table 1
Parameter estimates for univariate diffusion models of spot interest rate

Parameters	Vasicek	CIR	Ahn & Gao	CKLS	Nonlinear drift
α_{-1}	0.0	0.0	0.0	0.0	0.0001
					(0.0033)
α_0	0.13	0.096	0.0	0.04	−0.02
	(0.034)	(0.033)		(0.0196)	(0.0138)
α_1	−1.59	−1.27	0.94	−0.62	1.47
	(0.380)	(0.474)	(0.281)	(0.3131)	(0.826)
α_2	0.0	0.0	−12.60	0.0	−15.41
			(5.04)		(7.269)
σ	0.064	0.19	2.17	1.48	1.50
	(0.0006)	(0.0018)	(0.020)	(0.080)	(0.080)
ρ	0.0	0.5	1.5	1.35	1.36
				(0.0214)	(0.021)
Log-likelihood	22503.6	23605.1	24364.1	24385.7	24388.5

This table reports parameter estimates for five univariate diffusion models (13)–(17) of spot interest rate using daily Eurodollar interest rates of Ait-Sahalia (1996) from June 1, 1973 to February 25, 1995. For convenience, we write each model as a special case of Ait-Sahalia's nonlinear drift model. Therefore, we have, Vasicek: $dX_t = [\alpha_0 + \alpha_1 X_t]dt + \sigma dW_t$; CIR: $dX_t = [\alpha_0 + \alpha_1 X_t]dt + \sigma\sqrt{X_t}dW_t$; Ahn and Gao: $dX_t = [\alpha_1 X_t + \alpha_2 X_t^2]dt + \sigma X_t^{\rho} dW_t$; CKLS: $dX_t = [\alpha_0 + \alpha_1 X_t]dt + \sigma X_t^{\rho} dW_t$; and the nonlinear drift model: $dX_t = [\alpha_{-1} X_t^{-1} + \alpha_0 + \alpha_1 X_t + \alpha_2 X_t^2]dt + \sigma X_t^{\rho} dW_t$. Parameter estimates are obtained using the maximum likelihood method: the model likelihood function is used if available; otherwise, the Hermite approximation of the model likelihood function is used. Standard errors are given in the parentheses.

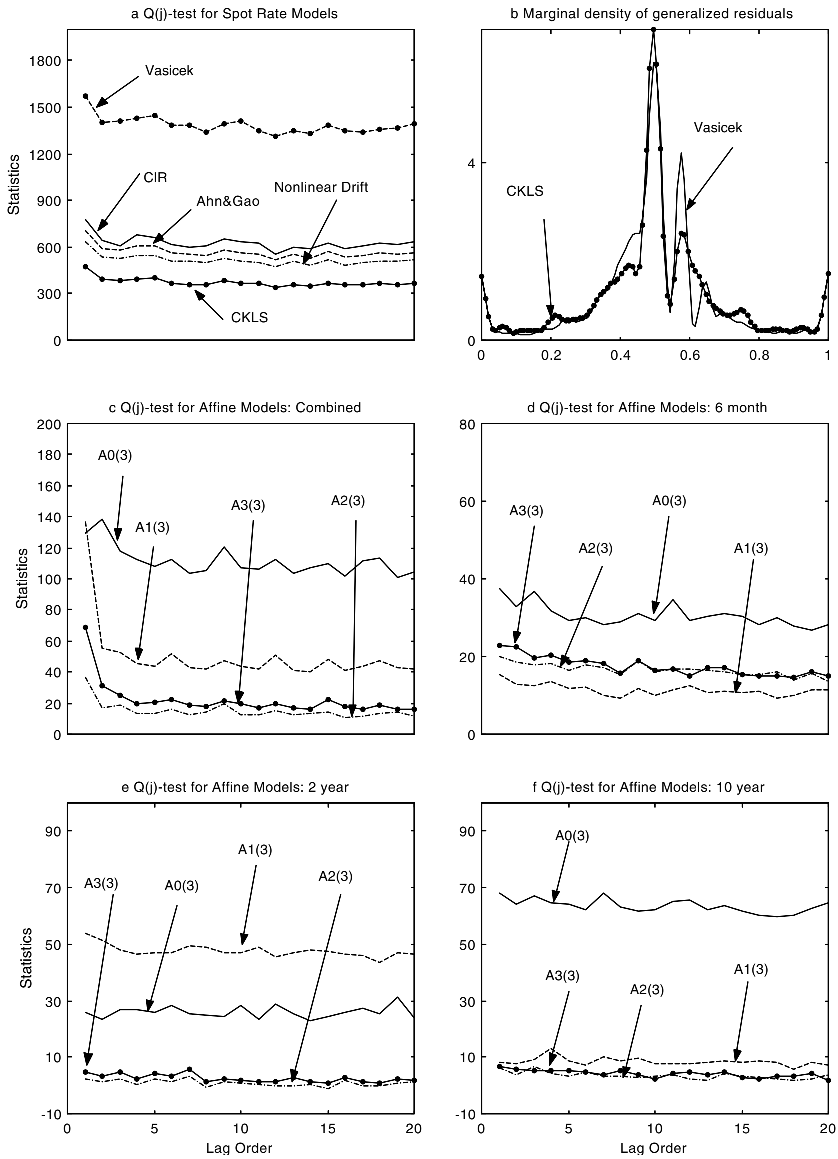

Figure 4

The empirical performance of univariate spot rate models and multivariate affine term structure models

Figure 4*a* reports the $\hat{Q}(j)$ statistics for lag order j from 1 to 20 for the Vasicek, CIR, Ahn and Gao, CKLS, and Ait-Sahalia's nonlinear drift models estimated using daily Eurodollar rates from June 1, 1973 to February 25, 1995. Figure 4*b* plots the kernel density estimators of the generalized residuals for the Vasicek and CKLS models (all other models give similar results). Figures 4*c*–*f* report the $\hat{Q}(j)$ statistics for the combined and three individual generalized residuals, respectively for four completely affine models, $\mathbf{A}_0$ (3), $\mathbf{A}_1$ (3), $\mathbf{A}_2$ (3), and $\mathbf{A}_3$ (3), estimated using monthly 6-month, 2- and 10-year zero-coupon Treasury yields from January 1952 to December 1998. The essentially affine models have $\hat{Q}(j)$ statistics that are similar to that of their completely affine counterparts.

long way to go before obtaining an adequate specification from any of these diffusion models. Our findings demonstrate the power of our tests: they overwhelmingly reject all diffusion models, including the CKLS and the nonlinear drift model, which Ait-Sahalia's (1996) marginal density-based test fails to reject using the empirical critical values in Pritsker (1998).

Next, we explore possible reasons for the rejection of the univariate diffusion models by separately examining the i.i.d. and $U[0, 1]$ properties of their generalized residuals $\{\hat{Z}_\tau\}_{\tau=1}^n$. In Figure 4*b*, we plot the kernel density estimators of the generalized residuals $\{\hat{Z}_\tau\}_{\tau=1}^n$ for the Vasicek and CKLS models (all other models give similar results). These density estimates all have peaks near 0.5, which indicates that too many observations fall into the area near the mean of the interest rate level than predicted by each model. In other words, the model-implied probability of the interest rates around the mean is always smaller than that implied by the true data generating process. Obviously, all five models cannot adequately capture the kurtosis of Eurodollar interest rates. This could be due to underestimating the speed of mean reversion (if any) or overestimating the magnitude of volatility.

To further examine how well the univariate diffusion models capture various dynamic aspects of Eurodollar interest rates, we report, in Table 2, various separate inference statistics $M(m, l)$ defined in Equation (12) for each model. We find that all five models fail to capture some of the most important features of $\{\hat{Z}_\tau\}_{\tau=1}^n$. The large $M(1, 1)$ statistics indicate that all univariate diffusion models fail to satisfactorily capture the conditional mean dynamics of $\{\hat{Z}_\tau\}_{\tau=1}^n$. In particular, the nonlinear drift model actually underperforms the linear drift models. The $M(2, 2)$ statistics show that the univariate diffusion models perform rather poorly in capturing the conditional variance of $\{\hat{Z}_\tau\}_{\tau=1}^n$, although introducing level effect substantially reduces the $M(2, 2)$ statistics from 1500 for the Vasicek model to about 700 for the CKLS model. The $M(3, 3)$ and $M(4, 4)$ statistics also show that the univariate diffusion models perform poorly in

Table 2
Separate inference statistics for spot rate models

Model	$M(1,1)$	$M(1,2)$	$M(2,1)$	$M(2,2)$	$M(3,3)$	$M(4,4)$
Vasicek	40.74	12.34	50.20	1540.37	115.74	997.45
CIR	45.64	4.76	28.23	985.28	139.16	719.24
Ahn & Gao	76.75	4.48	6.57	750.31	177.56	503.83
CKLS	72.78	4.94	7.55	682.35	170.41	485.62
Nonlinear drift	70.68	3.14	9.10	686.09	170.50	487.32

This table reports the separate inference statistics $M(m, l)$ in (12) for the five spot rate models. The asymptotically normal statistic $M(m, l)$ can be used to test whether the cross-correlation between the m-th and l-th moments of $\{Z_{\tau\Delta}\}$ is significantly different from zero. The choice of $(m, l) = (1, 1)$, $(2, 2)$, $(3, 3)$, and $(4, 4)$ is sensitive to autocorrelations in mean, variance, skewness, and kurtosis of $\{X_{\tau\Delta}\}$, respectively. We only report results for lag truncation order $p = 20$; the results for $p = 10$ and 30 are rather similar.

capturing the conditional skewness and kurtosis of $\{\hat{Z}_\tau\}_{\tau=1}^n$. The univariate diffusion models, especially the CKLS and Ait-Sahalia's non-linear drift models, however, provide reasonable description of ARCH-in-Mean and "leverage" effects as measured by $M(1, 2)$ and $M(2, 1)$.

Overall, we find that the five univariate diffusion models are quite restrictive in capturing both the unconditional distribution and the dependence structure of the model generalized residuals $\{\hat{Z}_\tau\}_{\tau=1}^n$. Of course, this may not be surprising given the overwhelming evidence of stochastic volatility [e.g., Andersen and Lund (1997)] and jumps [e.g., Johannes (2004) and Das (2002)] in interest rates. We also examine the performance of various discrete-time models with GARCH, regime-switching, and jump components to capture volatility clustering, mean shifts, and infrequent large movements in interest rate data. These models provide some improvements over the univariate diffusion models. For example, regime switching and jumps help capture the heavy tail or excess kurtosis of Eurodollar interest rates, and GARCH significantly improves the modeling of interest rate volatility. Despite these improvements, however, even the most sophisticated model that includes GARCH, regime switching, and jumps together still cannot adequately capture the spot rate dynamics. This model has a $\hat{Q}(j)$ statistic as large as 400.

4. Applications to Affine Term Structure Models

It is important to bear in mind that the Eurodollar rates considered above are encumbered with all kinds of "money-market" distortions that are largely absent from existing theoretical term structure models. To further illustrate the power of our tests and to better understand term structure dynamics, we next study some popular multivariate affine term structure models using yields on zero coupon bonds. These models provide a nice balance between richness of model specification and tractability. Assuming that the state variables follow affine diffusions, affine models can generate rich term structure dynamics while still allowing closed-form pricing for a wide variety of fixed-income securities [e.g., Duffie, Pan, and Singleton (2001), and Chacko and Das (2002)]. As a result, affine models have become probably the most widely studied term structure models in the literature.[16]

4.1 Affine term structure models

In affine models, it is assumed that the spot rate R_t is an affine function of N latent state variables $X_t = [X_{1,t}, \ldots, X_N,t]'$:

$$R_t = \delta_0 + \delta' X_t, \tag{24}$$

[16] For excellent reviews of the literature on affine models, see Dai and Singleton (2002) and Piazzesi (2002).

where δ_0 is a scalar and δ is an $N \times 1$ vector. In the absence of arbitrage opportunities, the time t-price of a zero-coupon bond that matures at $t+\tau_m$ ($\tau_m > 0$) equals

$$P(t, \tau_m) = E_t^Q\left[\exp\left(-\int_t^{t+\tau_m} R_s ds\right)\right],$$

where the expectation E_t^Q is taken under the risk-neutral measure Q. Thus, the whole yield curve is determined by X_t, which is assumed to follow an affine diffusion under the risk-neutral measure:

$$dX_t = \tilde{\kappa}\left(\tilde{\theta} - X_t\right)dt + \Sigma S_t d\tilde{W}_t, \tag{25}$$

where $\tilde{W}_t$ is an $N \times 1$ independent standard Brownian motion under measure Q, $\tilde{\kappa}$ and Σ are $N \times N$ matrices, and $\tilde{\theta}$ is an $N \times 1$ vector. The matrix S_t is diagonal with the (i, i)-th element

$$S_{t(ii)} \equiv [\alpha_i + \beta_i' X(t)]^{1/2}, \quad i = 1, \ldots, N, \tag{26}$$

where α_i is a scalar and β_i is an $N \times 1$ vector.

Under assumptions (24)–(26), the yields of zero coupon bonds, $Y(X_t, \tau_m) \equiv -\frac{1}{\tau_m}\log P(X_t, \tau_m)$, are an affine function of the state variables:

$$Y(X_t, \tau_m) = \frac{1}{\tau_m}[-A(\tau_m) + B(\tau_m)' X_t],$$

where the scaler function $A(\cdot)$ and the $N \times 1$ vector-valued function $B(\cdot)$ either have a closed-form or can be easily solved via numerical methods.

Completely affine models assume that the market prices of risk

$$\Lambda_t = S_t \lambda_1, \tag{27}$$

where λ_1 is an $N \times 1$ parameter vector. This implies that the compensation for risk is a fixed multiple of the variance of risk, a restriction that makes it difficult to replicate some stylized facts of historical excess bond returns. As a result, completely affine models provide poor forecasts of future bond yields and forecast errors are large when the slope of the term structure is steep. Duffee (2002) extends completely affine models to essentially affine models by assuming

$$\Lambda_t = S_t \lambda_1 + S_t^- \lambda_2 X_t, \tag{28}$$

where S_t^- is an $N \times N$ diagonal matrix with the (i, i)-th element

$$S_{t(ii)}^- = \begin{cases} (\alpha_i + \beta_i' X_t)^{-1/2}, & \text{if } \inf(\alpha_i + \beta_i' X_t) > 0, \\ 0, & \text{otherwise,} \end{cases} \quad i = 1, \ldots, N,$$

and λ_2 is an $N \times N$ matrix.

Under both specifications of Λ_t in Equations (27) and (28), X_t is also affine under the physical measure:

$$dX_t = \tilde{\kappa}(\tilde{\theta} - X_t)dt + \Sigma S_t \Lambda_t dt + \Sigma S_t dW_t,$$

where W_t is an $N \times 1$ standard Brownian motion under the physical measure.

Dai and Singleton (2000) greatly simplify the econometric analysis of affine models by providing a systematic scheme that classifies all admissible N-factor affine models into $N+1$ subfamilies, denoted as $\mathbf{A}_m$ (N), where $m \in \{0, 1, \ldots, N\}$ is the number of state variables that affect the instantaneous variance of X_t. They also introduce a canonical representation for $\mathbf{A}_m$ (N), which has the most flexible specification within each subfamily, as it either nests or is equivalent to (under an invariant transform) all the models in $\mathbf{A}_m$ (N).

We follow Dai and Singleton (2000) and Duffee (2002) to consider the canonical forms of the three-factor completely affine models $\mathbf{A}_m$ (3), $m = 0, 1, 2, 3$, and essentially affine models $\mathbf{E}_m$ (3), $m = 0, 1, 2$.[17] As the transition density of an affine model generally has no closed-form, MLE is infeasible. Following Duffee (2002), we estimate model parameters via Quasi-MLE, which is rather convenient for affine models, because the conditional mean and variance of X_t have a closed-form [see Duffee (2002) for details].[18]

4.2 Dynamic probability integral transform for affine models

We now discuss how to apply our tests to multivariate affine term structure models. The key is how to compute suitable generalized residuals for these models.

Suppose we have a time series of observations on the yields of N zero-coupon bonds with different maturities, $\{Y_{i,\tau\Delta}\}_{\tau=1}^n, i = 1, \ldots, N$.[19] Assuming that the yields are observed without error, given a parameter estimator $\hat{\theta}$, we can solve for the underlying state variables $\{X_{i,\tau\Delta}\}_{\tau=1}^n, i = 1, \ldots, N$. To examine whether the model transition density $p(X_{\tau\Delta}, \tau\Delta | I_{(\tau-1)\Delta}, (\tau-1)\Delta, \theta)$ of $X_{\tau\Delta}$ given $I_{(\tau-1)\Delta} \equiv \{X_{(\tau-1)\Delta}, \ldots, X_\Delta\}$ under the physical measure completely captures the dynamics of X_t, we can test whether the

[17] In the canonical representation, Σ is normalized to the identity matrix and the state vector X_t is ordered so that the first m elements of X_t affect the instantaneous variance of X_t. Setting $\alpha_i = 0$ for $i = 1, \ldots, m$, and $\alpha_i = 1$ for $i = m+1, \ldots N$, we have $S_{t(ii)} = X_{it}^{1/2}$ and $S^-_{t(ii)} = 0$ for $i = 1, \ldots, m$, and $S_{t(ii)} = (1 + \beta_i' X_t)^{1/2}$ and $S^-_{t(ii)} = (1 + \beta_i' X_t)^{-1/2}$ for $i = m+1, \ldots, N$, where $\beta_i = (\beta_{i1}, \ldots, \beta_{im}, 0, \ldots, 0)'$.

[18] We could also use other methods, such as the EMM method of Gallant and Tauchen (1996), the approximated MLE of Ait-Sahalia and Kimmel (2002) and Duffie, Pedersen, and Singleton (2003), the simulated MLE of Brandt and Santa-Clara (2002), and the empirical characteristic function method of Singleton (2001) and Jiang and Knight (2002).

[19] We arrange the yields so that the i-th bond has shorter maturity than the $(i+1)$-th bond.

probability integral transforms of $\{Y_{i,\tau\Delta}\}_{\tau=1}^{n}$, $i=1, \ldots, N$, with respect to the model transition density is i.i.d. $U[0, 1]$.

There are different ways to conduct the probability integral transform for affine models. Similar to our simulation study, we partition the joint density of the N different yields $(Y_{1,\tau\Delta}, \ldots, Y_{N,\tau\Delta})$ at time $\tau\Delta$ under the physical measure into the products of N conditional densities,

$$\begin{aligned} &p(Y_{1,\tau\Delta}, \ldots, Y_{N,\tau\Delta}, \tau\Delta | I_{(\tau-1)\Delta}, (\tau-1)\Delta, \hat{\theta}) \\ &= \prod_{i=1}^{N} p(Y_{i,\tau\Delta}, \tau\Delta | Y_{i-1,\tau\Delta}, \ldots, Y_{1,\tau\Delta}, I_{(\tau-1)\Delta}, (\tau-1)\Delta, \hat{\theta}), \end{aligned}$$

where the conditional density $p(Y_{1,\tau\Delta}, \ldots, Y_{N,\tau\Delta}, \tau\Delta | I_{(\tau-1)\Delta}, (\tau-1)\Delta, \hat{\theta})$ of $Y_{i,\tau\Delta}$ depends on not only the past information $I_{(\tau-1)\Delta}$ but also $\{Y_{l,\tau\Delta}\}_{l=1}^{i-1}$, the contemporaneous yields with shorter maturities.[20] We then transform the yield $Y_{i,\tau\Delta}$ via its corresponding model-implied transition density

$$\begin{aligned} Z_{i,\tau}^{(1)}(\hat{\theta}) \equiv \int_0^{Y_{i,\tau\Delta}} &p(Y_{1,\tau\Delta}, \ldots, Y_{N,\tau\Delta}, \tau\Delta | I_{(\tau-1)\Delta}, (\tau-1)\Delta, \hat{\theta}) \\ &i = 1, \ldots, N. \end{aligned} \tag{29}$$

This approach produces N generalized residual samples, $\{Z_{i,\tau}^{(1)}(\hat{\theta})\}_{\tau=1}^{n}$, $i=1, \ldots, N$. For each i, the sample $\{Z_{i,\tau}^{(1)}(\hat{\theta})\}_{\tau=1}^{n}$ is approximately i.i.d. $U[0, 1]$ under correct model specification.[21]

We can also combine the N generalized residuals $\{Z_{i,\tau}^{(1)}(\hat{\theta})\}_{\tau=1}^{n}$ in Equation (29) to obtain the *combined* generalized residuals of an affine model:

$$\begin{aligned} &Z^{(2)}(\hat{\theta}) \\ &\equiv \left[Z_{1,1}^{(1)}(\hat{\theta}), \ldots, Z_{N,1}^{(1)}(\hat{\theta}), Z_{1,2}^{(1)}(\hat{\theta}), \ldots, Z_{N,2}^{(1)}(\hat{\theta}), \ldots, Z_{1,n}^{(1)}(\hat{\theta}), \ldots, Z_{N,n}^{(1)}(\hat{\theta})\right]'. \end{aligned} \tag{30}$$

The combined generalized residuals $\{Z_{\tau}^{(2)}(\hat{\theta})\}_{\tau=1}^{nN}$ in Equation (30) is approximately i.i.d. $U[0, 1]$ under correct model specification and this property can be used to check the overall performance of an affine model. In contrast, each individual sample of generalized residuals $\{Z_{i,\tau}(\hat{\theta})\}_{\tau=1}^{n}$ in Equation (29) can be used to check the adequacy of an

[20] As noted earlier, there are $N!$ ways of factoring the joint transition density of yields with different maturities. In our application, we let the transition density of the yields of long term bonds depend on the contemporaneous yields of shorter maturity bonds, because the short-end of the yield curve is generally more sensitive to various economic shocks and is more volatile.

[21] We can also compute the probability integral transform of the yields $Y_{i,\tau\Delta}$, $i=1, \ldots, N$, with respect to $p(y, \tau\Delta | I_{(\tau-1)\Delta}, (\tau-1)\Delta, \hat{\theta})$, the model-implied conditional density of $Y_{i,\tau\Delta}$ given only $I_{(\tau-1)\Delta}$ under the physical measure. This approach, however, ignores the information on the joint distribution among bond yields with different maturities. Results (not reported here) show that such generalized residuals are further away from i.i.d. $U[0, 1]$ than $\{Z_{i,\tau}^{(1)}(\hat{\theta})\}$. This indicates that the joint transition density of yields with different maturities provides additional information on term structure dynamics.

affine model in capturing the dynamics of each specific yield, as has been illustrated in our simulation study.

Because the transition density has no closed form for most affine models, we use the simulation methods of Pedersen (1995) and Brandt and Santa-Clara (2002) to obtain an approximation for the transition density. This method is applicable to not only affine diffusions, but also other general multivariate diffusions. We could use other approximation methods mentioned earlier.

4.3 Empirical results

We now evaluate the performance of affine models in capturing the joint dynamics of U.S. Treasury yields during the second half of the last century. We use the same data as Duffee (2002): monthly yields on zero-coupon bonds with 6-month, 2- and 10-year maturities from January 1952 to December 1998. The zero-coupon bond yields are interpolated from coupon bond prices using the method of McCulloch and Kwon (1993), whose sample is extended by Bliss (1997) beyond February 1991. Assuming that the three yields are observed without error, we use them to estimate model parameters via Quasi-MLE. For most models, our estimates are very close to those of Duffee (2002), who includes three other yields (observed with measurement error). The results on model performance obtained from both Duffee's (2002) estimates and ours are qualitatively the same, although our estimates generally provide better model performance in terms of the $\hat{Q}(j)$ criterion. Parameter estimates for the canonical forms of seven completely and essentially affine models are reported in Tables 3 and 4. The only restrictions on model parameters are those required by the canonical form.

Based on parameter estimates in Tables 3 and 4, we calculate the generalized residuals $\{Z_{i,\tau}^{(1)}(\hat{\theta})\}_{\tau=1}^{n}$ in Equation (29) for the 6-month $(i=1)$, 2-year $(i=2)$ and 10-year $(i=3)$ yields, and the combined generalized residuals $\{Z_{\tau}^{(2)}(\hat{\theta})\}_{\tau=1}^{nN}$ in Equation (30). Dai and Singleton (2000) point out that it is difficult to formally assess the relative goodness of fit of non-nested classes of affine models using existing methods. One advantage of our approach is that the performance of non-nested models can be compared by a metric measuring the closeness of their generalized residuals to i.i.d. $U[0, 1]$.

We first examine the overall performance of each model via the $\hat{Q}(j)$ statistic based on the combined generalized residuals $\{Z_{\tau}^{(2)}(\hat{\theta})\}_{\tau=1}^{nN}$ in Equation (30), which is shown in Figure 4*c*.[22] Although some models perform relatively better than others, $\hat{Q}(j)$ overwhelmingly rejects all affine models at conventional significance levels. Among the seven affine

[22] For ease of exposition, we only report the $\hat{Q}(j)$ statistics $(j=1, \ldots, 20)$ for the completely affine models in Figures 4*c*–*f*. The essentially affine models have $\hat{Q}(j)$ statistics that are similar to that of their completely affine counterparts.

Table 3
Parameter estimates for multivariate affine term structure models completely affine models ($\lambda_2 = 0$)

	A_0 (3)		A_1 (3)		A_2 (3)		A_3 (3)	
	Est.	SE	Est.	SE	Est.	SE	Est.	SE
α_1	1.0		0.0		0.0		0.0	
α_2	1.0		1.0		0.0		0.0	
α_3	1.0		1.0		1.0		0.0	
δ_0	0.046	(0.015)	0.022	(0.012)	0.015	(0.0013)	−0.0052	(0.0086)
δ_1	0.0	(0.0)	0.00092	(0.00059)	0.00080	(0.00046)	0.0014	(0.00018)
δ_2	0.010	(0.002)	0.000093	(0.00018)	0.0014	(0.00015)	0.00056	(0.00064)
δ_3	0.00044	(0.00069)	0.0035	(0.00089)	0.0029	(0.00097)	0.0088	(0.0013)
$(K\theta)_1$	0.0		0.16	(0.052)	0.0	(0.0)	0.83	(0.067)
$(K\theta)_2$	0.0		−0.52		0.25	(0.37)	0.0	(0.0)
$(K\theta)_3$	0.0		0.27		−3.22		0.34	(0.012)
κ_{11}	−0.56	(0.14)	−0.030	(0.019)	−0.14	(0.098)	−0.55	(0.18)
κ_{12}	0.0		0.0		0.31	(0.097)	0.072	(0.078)
κ_{13}	0.0		0.0		0.0		0.0	(0.0)
κ_{21}	1.67	(0.15)	0.095	(0.0039)	0.17	(0.124)	0.0	(0.0)
κ_{22}	−1.64	(0.58)	−0.32	(0.056)	−0.48	(0.217)	−0.041	(0.029)
κ_{23}	0.0		−17.66	(0.035)	0.0		0.12	(0.020)
κ_{31}	0.13	(0.087)	−0.050	(0.036)	−0.87	(0.35)	1.84	(0.14)
κ_{32}	−0.18	(0.0)	0.018	(0.051)	3.38	(0.80)	0.0	(0.0)
κ_{33}	−0.0035	(0.17)	−1.86	(0.041)	−1.73	(0.12)	−2.02	(0.20)
β_{11}	0.0		1.0		1.0		1.0	
β_{12}	0.0		0.0		0.0		0.0	
β_{13}	0.0		0.0		0.0		0.0	
β_{21}	0.0		42.09	(0.062)	0.0		0.0	
β_{22}	0.0		0.0		1.0		1.0	
β_{23}	0.0		0.0		0.0		0.0	
β_{31}	0.0		0.32	(0.028)	0.0	(0.0)	0.0	
β_{32}	0.0		0.0		4.33	(0.071)	0.0	
β_{33}	0.0		0.0		0.0		1.0	
λ_{11}	−0.036	(0.00059)	−0.040	(0.024)	−0.029	(0.020)	−0.049	(0.016)
λ_{12}	−0.59	(0.0026)	−0.017	(0.011)	−0.054	(0.029)	−0.030	(0.0086)
λ_{13}	−0.16	(0.00049)	−0.11	(0.0030)	−0.11	(0.053)	−0.32	(0.020)
Likelihood	10.52		9.81		7.96		7.81	

This table reports the quasi-maximum likelihood estimates for three-factor completely affine models using monthly 6-month, 2- and 10-year zero coupon yields from January 1952 to December 1998. All models share the following specifications for the instantaneous interest rate, the physical dynamics of state variables and the market price of risk:

$$R_t = \delta_0 + \delta_1 X_{1t} + \delta_2 X_{2t} + \delta_3 X_{3t},$$

$$d\begin{pmatrix} X_{1t} \\ X_{2t} \\ X_{3t} \end{pmatrix} = \left[\begin{pmatrix} (K\theta)_1 \\ (K\theta)_2 \\ (K\theta)_3 \end{pmatrix} - \begin{pmatrix} \kappa_{11} & \kappa_{12} & \kappa_{13} \\ \kappa_{21} & \kappa_{22} & \kappa_{23} \\ \kappa_{31} & \kappa_{32} & \kappa_{33} \end{pmatrix}\begin{pmatrix} X_{1t} \\ X_{2t} \\ X_{3t} \end{pmatrix}\right] dt + S_t d\begin{pmatrix} W_{1t} \\ W_{2t} \\ W_{3t} \end{pmatrix},$$

$$S_{t(ii)} = \sqrt{\alpha_i + (\beta_{i1}\ \beta_{i2}\ \beta_{i3})X_t}, i = 1, 2, 3, \quad \text{and} \quad \Lambda_t = S_t\begin{pmatrix} \lambda_{11} \\ \lambda_{12} \\ \lambda_{13} \end{pmatrix} + S_t^{-}\begin{pmatrix} \lambda_{2(11)} & \lambda_{2(12)} & \lambda_{2(13)} \\ \lambda_{2(21)} & \lambda_{2(22)} & \lambda_{2(23)} \\ \lambda_{2(31)} & \lambda_{2(32)} & \lambda_{2(33)} \end{pmatrix}.$$

models, $\mathbf{A}_2$ (3) and $\mathbf{E}_2$ (3) have the best overall performance with their $\hat{Q}(1)$ statistics around 30, and $\mathbf{A}_3$ (3) performs worse, with its $\hat{Q}(1)$ statistics around 70. The models $\mathbf{A}_1$ (3) and $\mathbf{E}_1$ (3) have much worse performance with their $\hat{Q}(1)$ statistics close to 140, while $\mathbf{A}_0$ (3) and $\mathbf{E}_0$ (3) have the worst performance with most of their $\hat{Q}(j)$ statistics over 100.

Table 4
Parameter estimates for multivariate affine term structure models essentially affine models

	$\mathbf{E}_0$ (3)		$\mathbf{E}_1$ (3)		$\mathbf{E}_2$ (3)	
	Est.	SE	Est.	SE	Est.	SE
α_1	1.0		0.0		0.0	
α_2	1.0		1.0		0.0	
α_3	1.0		1.0		1.0	
δ_0	0.017	(0.00066)	0.023	(0.0041)	0.015	(0.0048)
δ_1	0.032	(0.0041)	0.0014	(0.00018)	0.00083	(0.00027)
δ_2	0.013	(0.0012)	0.000066	(0.00063)	0.0013	(0.0)
δ_3	0.02	(0.0098)	0.0035	(0.0018)	0.0024	(0.00043)
$(\mathrm{K}\theta)_1$	0.0		0.16	(0.039)	0.0	0.0
$(\mathrm{K}\theta)_2$	0.0		−0.52		0.25	(0.13)
$(\mathrm{K}\theta)_3$	0.0		0.27		−3.22	
κ_{11}	−2.69	(0.0168)	−0.030	(0.073)	−0.14	(0.032)
κ_{12}	0.0		0.0		0.31	(0.033)
κ_{13}	0.0		0.0		0.0	
κ_{21}	0.12	(0.0011)	0.095	(0.0069)	0.16	(0.045)
κ_{22}	−0.13	(0.0057)	−0.32	(0.13)	−0.48	(0.072)
κ_{23}	0.0		−17.66	(0.070)	0.0	
κ_{31}	−0.43	(0.0046)	−0.050	(0.0057)	−0.87	(0.33)
κ_{32}	−0.22	(0.052)	0.018	(0.097)	3.38	(0.25)
κ_{33}	−0.17	(0.056)	−1.86	(0.091)	−1.73	(0.57)
β_{11}	0.0		1.0		1.0	
β_{12}	0.0		0.0		0.0	
β_{13}	0.0		0.0		0.0	
β_{21}	0.0		42.09	(0.30)	0.0	
β_{22}	0.0		0.0		1.0	
β_{23}	0.0		0.0		0.0	
β_{31}	0.0		0.32	(0.29)	0.0	(0.0)
β_{32}	0.0		0.0		4.32	(0.14)
β_{33}	0.0		0.0		0.0	
λ_{11}	−0.32	(0.00055)	−0.040	(0.12)	−0.029	(0.0028)
λ_{12}	−0.86	(0.0020)	−0.017	(0.075)	−0.054	(0.010)
λ_{13}	−0.40	(0.0018)	−0.11	(0.020)	0.62	(0.17)
$\lambda_{2(11)}$	−0.56	(0.23)	0.0		0.0	
$\lambda_{2(12)}$	−0.46	(1.96)	0.0		0.0	
$\lambda_{2(13)}$	0.075	(0.053)	0.0		0.0	
$\lambda_{2(21)}$	0.20	(0.35)	3.49	(0.11)	0.0	
$\lambda_{2(22)}$	−5.37	(1.84)	−0.10	(0.037)	0.0	
$\lambda_{2(23)}$	−0.19	(0.16)	5.85	(0.022)	0.0	
$\lambda_{2(31)}$	−0.57	(0.19)	−0.064	(0.0069)	0.068	(0.036)
$\lambda_{2(32)}$	1.92	(1.06)	0.015	(0.078)	2.15	(0.12)
$\lambda_{2(33)}$	−0.19	(0.081)	−1.68	(0.081)	0.12	(0.041)
Likelihood	13.03		10.68		8.28	

This table reports the quasi-maximum likelihood estimates for three-factor essentially affine models using monthly 6-month, 2- and 10-year zero coupon yields from January 1952 to December 1998. All models share the following specifications for the instantaneous interest rate, the physical dynamics of state variables and the market price of risk:

$$R_t = \delta_0 + \delta_1 X_{1t} + \delta_2 X_{2t} + \delta_3 X_{3t},$$

$$d\begin{pmatrix} X_{1t} \\ X_{2t} \\ X_{3t} \end{pmatrix} = \left[\begin{pmatrix} (K\theta)_1 \\ (K\theta)_2 \\ (K\theta)_3 \end{pmatrix} - \begin{pmatrix} \kappa_{11} & \kappa_{12} & \kappa_{13} \\ \kappa_{21} & \kappa_{22} & \kappa_{23} \\ \kappa_{31} & \kappa_{32} & \kappa_{33} \end{pmatrix}\begin{pmatrix} X_{1t} \\ X_{2t} \\ X_{3t} \end{pmatrix}\right] dt + S_t d\begin{pmatrix} W_{1t} \\ W_{2t} \\ W_{3t} \end{pmatrix},$$

$$S_{t(ii)} = \sqrt{\alpha_i + (\beta_{i1}\ \beta_{i2}\ \beta_{i3})X_t},\ i = 1, 2, 3, \quad \text{and} \quad \Lambda_t = S_t\begin{pmatrix} \lambda_{11} \\ \lambda_{12} \\ \lambda_{13} \end{pmatrix} + S_t^{-}\begin{pmatrix} \lambda_{2(11)} & \lambda_{2(12)} & \lambda_{2(13)} \\ \lambda_{2(21)} & \lambda_{2(22)} & \lambda_{2(23)} \\ \lambda_{2(31)} & \lambda_{2(32)} & \lambda_{2(33)} \end{pmatrix}.$$

Except for $\mathbf{E}_2$ (3), the other two essentially affine models do not outperform their completely affine counterparts.

Our findings are consistent with the well-known trade-off between the flexibility in modeling the conditional variances of X_t on one hand, and the conditional correlations between the components of X_t and excess bond returns on the other hand. Both $\mathbf{A}_0$ (3) and $\mathbf{E}_0$ (3) have greater flexibility in modeling the conditional correlation of X_t and excess bond returns,[23] but they assume constant conditional variance for X_t. On the other hand, $\mathbf{A}_3$ (3), although allowing each state variable to follow a square root process, imposes strong restrictions on the correlations between the components of X_t and the market prices of risk.[24] Dai and Singleton (2000) conjecture that the models that are able to accommodate both time-varying volatilities of state variables and their time-varying correlations, such as $\mathbf{A}_1$ (3) and $\mathbf{A}_2$ (3), are more likely to perform better.

While Dai and Singleton (2000) show that $\mathbf{A}_1$ (3) outperforms $\mathbf{A}_2$ (3) in modeling the weekly U.S. swap rates during the past 15 years, we find that $\mathbf{A}_2$ (3) and $\mathbf{E}_2$ (3) perform best for monthly U.S. Treasury yields over the past 50 years. This difference is most likely due to the use of different data and sample periods. Our data covers a much longer period and includes late 1970s and early 1980s — a period with extremely high and persistent volatilities. Thus, modeling time-varying volatility might be more important for Treasury yields than swap rates. Our finding confirms Dai and Singleton's (2003) conjecture that "for other market of different sample periods, where conditional volatility is more pronounced in the data, the relative goodness-of-fit of models in the branches $\mathbf{A}_1$ (3) and $\mathbf{A}_2$ (3) may change." While Dai and Singleton (2000) only consider the completely affine models $\mathbf{A}_1$ (3) and $\mathbf{A}_2$ (3), we find that $\mathbf{A}_3$ (3) outperforms $\mathbf{A}_1$ (3), and $\mathbf{E}_2$ (3) outperforms all the completely affine models for Treasury yields.

Figures 4*d–f* report the $\hat{Q}(j)$ statistics for 6-month, 2-, and 10-year yields, respectively. Consistent with their ranking in overall performance, $\mathbf{A}_2$ (3), $\mathbf{E}_2$ (3), and $\mathbf{A}_3$ (3) perform well in modeling each individual yield. One interesting finding that is not obvious from the overall performance evaluation is that the above three models capture the 2- and 10-year yields much better than the 6-month yields: most $\hat{Q}(j)$ statistics for $\mathbf{E}_2$ (3) for the 2-year yields are not significant at the 5% level. Models $\mathbf{A}_1$ (3) and $\mathbf{E}_1$ (3) have similar performance to the above three models for the 6-month and 10-year yields, but they perform extremely poorly for the 2-year

[23] Gaussian models allow correlations of all the state variables with different signs and market prices of risk to depend on all the state variables in their essentially affine form.

[24] In the $\mathbf{A}_3$ (3) model, the conditional correlations among the components of X_t are equal to zero and the unconditional correlations among the components of X_t must be positive. In addition, the market prices of risk can only depend on the volatility of each risk factor and there is no corresponding essential affine version for this model.

yield. Models $\mathbf{A}_0$ (3) and $\mathbf{E}_0$ (3) have the worst performance for all individual yields.

We further investigate the reasons for the failure of affine models by separately examining the $U[0, 1]$ and i.i.d. properties of the generalized residuals of each individual yield. Figure 5 displays the kernel marginal density estimates of the combined generalized residuals in (30) and the individual residuals in Equation (29) for 6-month, 2- and 10-year yields, for models $\mathbf{A}_0$ (3), $\mathbf{A}_1$ (3), and $\mathbf{A}_2$ (3).[25] Consistent with the $\hat{Q}(j)$ statistics, we find that $\mathbf{A}_2$ (3) captures the marginal distributions of the 2- and 10-year yields very well: the marginal density estimates of the generalized residuals for these two yields are very close to $U[0, 1]$. However, $\mathbf{A}_2$ (3) fails to adequately capture the distribution of the 6-month yield: its marginal density exhibits a pronounced peak in the middle, suggesting that the model cannot adequately capture the center of the distribution of the 6-month yield. Both $\mathbf{A}_0$ (3) and $\mathbf{A}_1$ (3) capture the marginal distribution of all three yields very poorly: there are pronounced peaks at both ends of the distribution, suggesting that there are too many observations in the tails than predicted by the model. This indicates the failure of these two models in capturing the heavy tails (or large movements) of all three yields.

We also calculate $M(m, l)$ statistics in Equation (12) to examine the aspects of the dynamics of the generalized residuals that an affine model fails to capture. The $M(m, l)$ statistics reported in Table 5 show that all models fail to capture dependence in the conditional variance and kurtosis of the generalized residuals for all three yields. Both $M(2, 2)$ and $M(4, 4)$ are overwhelmingly significant for all yields under all affine models, although they become smaller for non-Gaussian models. All affine models have some difficulties in modeling the conditional mean [$M(1, 1)$] and skewness [$M(3, 3)$] of their generalized residuals, especially for the 6-month and 10-year yields. Nevertheless, all affine models seem to be able to adequately capture ARCH-in-Mean effects [$M(1, 2)$] and "leverage" effects [$M(2, 1)$] for all yields, except for some rare cases involving the 6-month and 2-year yields. All affine models fail to satisfactorily capture the dependence in conditional mean, variance, skewness, and kurtosis of the generalized residuals for the 6-month yield. Except for $M(2, 2)$ and $M(4, 4)$, models $\mathbf{A}_2$ (3), $\mathbf{E}_2$ (3), and $\mathbf{A}_3$ (3) provide fairly good description of the 2-year yield.

In summary, we find that some affine models [e.g., $\mathbf{A}_2$ (3) and $\mathbf{E}_2$ (3)] do a reasonably good job in modeling the 2- and 10-year yields, although they fail to adequately capture the conditional variance and kurtosis of the generalized residuals of these yields. However, these models fail to

[25] The marginal densities of $\mathbf{A}_3$ (3) and $\mathbf{E}_2$ (3) are very similar to that of $\mathbf{A}_2$ (3), while the marginal densities of $\mathbf{E}_0$ (3) and $\mathbf{E}_1$ (3) are similar to those of $\mathbf{A}_0$ (3) and $\mathbf{A}_1$ (3), respectively.

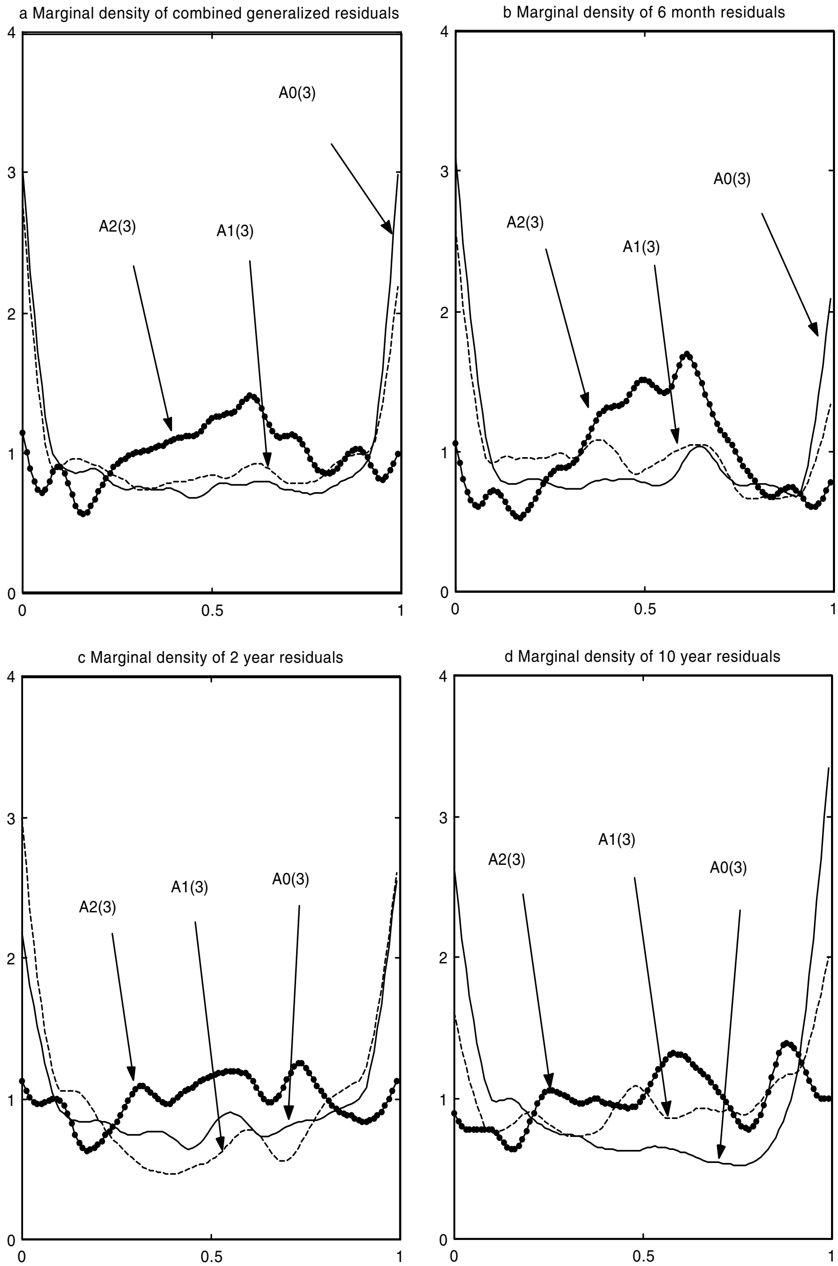

Figure 5
The nonparametric marginal density of the generalized residuals of combined, 6-month, 2- and 10-year yields under affine term structure models
Figures 5*a–d* plot the kernel density estimators of the combined and three individual generalized residuals, respectively for three completely affine models $\mathbf{A}_0$ (3), $\mathbf{A}_1$ (3), and $\mathbf{A}_2$ (3), respectively.

Table 5
Separate inference statistics for affine models

Model	Maturity	M (1,1)	M (1,2)	M (2,1)	M (2,2)	M (3,3)	M (4,4)
A_0 (3)	6 month	15.45	5.27	0.30	93.47	13.03	105.96
	2 year	0.015	−0.22	−0.25	44.34	0.045	47.90
	10 year	1.14	0.51	−0.57	20.92	1.02	21.12
A_1 (3)	6 month	10.92	2.40	3.52	43.77	9.78	53.60
	2 year	4.28	2.76	−1.37	5.24	2.20	6.04
	10 year	1.95	0.088	−0.49	23.07	2.54	19.73
A_2 (3)	6 month	14.34	0.49	2.77	67.83	10.80	69.01
	2 year	0.67	1.31	1.19	8.76	−0.13	8.63
	10 year	10.35	1.23	−1.12	22.19	13.49	18.63
A_3 (3)	6 month	18.74	0.73	1.91	78.46	13.03	80.49
	2 year	0.11	1.01	1.04	12.65	−0.81	12.50
	10 year	−0.51	1.14	2.13	19.43	−0.55	20.77
E_0 (3)	6 month	17.98	−0.25	21.71	206.54	16.69	239.44
	2 year	0.64	0.77	0.73	36.72	2.02	58.41
	10 year	0.17	−0.22	2.13	48.31	3.32	69.82
E_1 (3)	6 month	9.48	2.26	1.63	41.43	8.91	55.13
	2 year	3.32	1.65	−1.54	7.09	1.82	7.98
	10 year	1.46	0.37	−0.79	38.23	2.02	35.35
E_2 (3)	6 month	15.58	2.42	2.83	51.35	12.80	48.51
	2 year	0.32	0.99	0.22	7.20	−0.36	7.38
	10 year	9.85	1.20	−1.01	21.62	13.35	19.33

This table reports the separate inference statistics $M(m, l)$ in (12) for the seven three-factor completely and essentially affine models. The asymptotically statistic $M(m, l)$ can be used to test whether the cross-correlation between the m-th and l-th moments of $\{Z_{\tau\Delta}\}$ is significantly different from zero. The choice of $(m, l) = (1, 1), (2, 2), (3, 3)$, and $(4, 4)$ is very sensitive to autocorrelations in mean, variance, skewness, and kurtosis of $(X_{\tau\Delta}\}$, respectively. We only show results for lag truncation order $p = 20$; the results for $p = 10$ and 30 are similar.

satisfactorily capture the short end of the yield curve: they have difficulties in modeling the marginal distribution and the dynamics of the generalized residuals of the 6-month yield.

5. Conclusion

Nonparametric methods have enjoyed enormous success in many areas of econometrics and statistics. In the recent finance literature, there are some concerns that nonparametric approach, despite its many appealing features, might not be suitable for financial data, which are typically highly persistently dependent and thus render poor finite sample performance of nonparametric methods.

In this article, we have developed an omnibus nonparametric specification test that has good finite sample performance, and can be applied to a wide range of continuous-time and discrete-time, univariate and multivariate dynamic economic models. A class of separate inference procedures are supplemented to gauge possible sources of model misspecification. To highlight our approach, we have applied our tests to evaluate a variety of popular univariate spot rate diffusion models and multivariate affine term structure models, obtaining many interesting new empirical findings. Our study shows that contrary to the general perception in the literature, nonparametric methods can be a reliable and powerful tool for analyzing financial data.

Appendix

To illustrate the essence of our proof, we present a detailed proof for the case that X_t is a Markovian process. The case that X_t is non-Markovian can be proved similarly, but with more tedious algebra. Throughout, we use C to denote a generic bounded constant, $|\cdot|$ to denote the usual Euclidean norm, and $\frac{\partial}{\partial\theta}Z_\tau(\theta_0)$ to denote $\frac{\partial}{\partial\theta}Z_\tau(\theta)|_{\theta=\theta_0}$. All convergencies are taken as $n\to\infty$. We now provide regularity conditions.

Assumption 1. *Let $(\Omega,\mathcal{F},P)$ be a complete probability space.* (1) *$X_t\equiv X_t(\omega)$, where $\omega\in\Omega$ and $t\in[0,T]\subset\mathbb{R}^+$, is a continuous-time process with transition density $p_0(x,t\,|\,y,s)$, where $s<t$;* (2) *a discrete sample $\{X_{\tau\Delta}\}_{\tau=1}^n$ of X_t is observed, where Δ is a fixed sample interval, and $\{X_{\tau\Delta}\}$ is an α-mixing process with mixing coefficient $\alpha(\cdot)$ satisfying $\sum_{\tau=0}^\infty\alpha(\tau\Delta)^{(\nu-1)/\nu}\le C$ for some constant $\nu>1$.*

Assumption 2. *Let Θ be a finite-dimensional parameter space.* (1) *The model transition density $p(x,t\,|\,y,s,\theta)$ for the underlying process X_t is a measurable function of (x,y) for each $\theta\in\Theta$;* (2) *$p(x,t\,|\,y,s,\theta)$ is twice-continuously differentiable with respect to θ in a neighborhood Θ_0 of θ_0, with $\lim_{n\to\infty}\sum_{\tau=1}^n E|\frac{\partial}{\partial\theta}Z_\tau(\theta_0)|^4\le C$ and $\lim_{n\to\infty}\sum_{\tau=1}^n E\sup_{\theta\in\Theta_0}|\frac{\partial^2}{\partial\theta\partial\theta'}Z_\tau(\theta_0)|^2\le C$, where $Z_\tau(\theta)$ is defined in* (5).

Assumption 3. (1) *The function $G_{\tau-1}(z)\equiv E\{[\frac{\partial}{\partial\theta}Z_\tau(\theta)]_{\theta=\theta_0}|Z_\tau(\theta_0)=z,X_{(\tau-1)\Delta}\}$ is a measurable function of z and $X_{(\tau-1)\Delta}$;* (2) *with probability one, $G_{\tau-1}(z)$ is continuously differentiable with respect to $z\in[0,1]$ for each $\tau>0$, with $\lim_{n\to\infty}\sum_{\tau=1}^n E|G'_{\tau-1}[Z_\tau(\theta_0)]|^2\le C$.*

Assumption 4. *$\hat\theta\in\Theta$ is a parameter estimator such that $\sqrt{n}(\hat\theta-\theta^*)=O_P(1)$, where $\theta^*\equiv p\lim\hat\theta$ is an interior element in Θ and $\theta^*=\theta_0$ under $\mathbb{H}_0$.*

Assumption 5. *The kernel function k: $[-1,1]\to\mathbb{R}^+$ is a symmetric, bounded, and twice continuously differentiable probability density such that $\int_{-1}^1k(u)du=1$, $\int_{-1}^1uk(u)du=0$, and $\int_{-1}^1u^2k(u)du<\infty$.*

We first state the asymptotic distribution of the proposed test statistic under $\mathbb{H}_0$.

Theorem 1. *Suppose that Assumptions* 1–5 *hold and $h=cn^{-\delta}$ for $c\in(0,\infty)$ and $\delta\in(0,\frac{1}{5})$. Then for any integer $j>0$ such that $j=o(n^{1-\delta\,(5-2'\nu)})$ where ν is as in Assumption* 1, *we have $\hat Q(j)\to^d N(0,1)$ under $\mathbb{H}_0$.*

Theorem 2. *Put $\hat Q\equiv[\hat Q(j_1),\ldots,\hat Q(j_L)]'$, where $j_1,\ldots,j_L$ are L distinct positive integers, and L is a fixed integer. Then, under the same conditions of Theorem 1, $\hat Q\to^d N(O,I)$ under $\mathbb{H}_0$, where I is a $L\times L$ identity matrix. Consequently, $\hat Q(j_{c_1})$ and $\hat Q(j_{c_2})$ are asymptotically independent whenever $j_{c_1}\ne j_{c_2}$.*

Next, we consider the asymptotic power of our tests under the alternative $\mathbb{H}_A$.

Assumption 6. *For each integer $j>0$, the joint density $g_j(x,y)$ of the transformed random vector $\{Z_\tau,Z_{\tau-j}\}$, where $Z_\tau\equiv Z_\tau(\theta^*)$ and θ^* is as in Assumption* 4, *exists and is continuously differentiable on* $[0,1]^2$.

Theorem 3. *Suppose that Assumptions* 1–6 *hold and $h=cn^{-\delta}$ for $c\in(0,\infty)$ and $\delta\in(0,\frac{1}{5})$. Then we have $(nh)^{-1}\hat Q(j)\to^p V_0^{-1/2}\int_0^1\int_0^1[g_j(z_1,z_2)-1]^2dz_1dz_2$ $\mathbb{H}_A$* for any integer $j>0$ *such that $j=o(n^{1-\delta(5-2/\nu)})$ where ν is as in Assumption* 1. *Consequently, for any sequence of constants $\{C_n=o(nh)\}$, $P[\hat Q(j)>C_n]\to1$ whenever Z_τ and $Z_{\tau-j}$ are not independent nor* $U[0,1]$.

Proof of Theorem 1. Throughout, we put $w \equiv (z_1, z_2) \in \mathbb{I}^2$, where $\mathbb{I} \equiv [0, 1]^2$. Let $\tilde{g}_j(w)$ be defined in the same way as $\hat{g}_j(w)$ in (6) but with $\{Z_\tau\}$ replacing $\{\hat{Z}_\tau\}$, and let $\tilde{Q}(j)$ be defined in the same way as $\hat{Q}(j)$ in Equation (9) with $\tilde{g}_j(z)$ replacing $\hat{g}_j(z)$. We shall prove the following theorems. ■

Theorem A.1. $\hat{Q}(j) - \tilde{Q}(j) \to^p 0.$

Theorem A.2. $\tilde{Q}(j) \to^d N(0, 1).$

Proof of Theorem A.1. Let $\tilde{M}(j)$ be defined as $\hat{M}(j)$ in Equation (8) with $\{\tilde{g}_j(w)\}$ replacing $\{\hat{g}_j(w)\}$. We write

$$\hat{M}(j) - \tilde{M}(j) = \int_{\mathbb{I}^2} [\hat{g}_j(w) - \tilde{g}_j(w)]^2 dw + 2\int_{\mathbb{I}^2} [\tilde{g}(w) - 1][\hat{g}(w) - \tilde{g}(w)]dw$$
$$\equiv \hat{\Delta}_1(j) + 2\hat{\Delta}_2(j). \tag{A1}$$

We shall show Proposition A.1 and A.2 below. Throughout, put $n_j \equiv n - j$. ■

Proposition A.1. $n_j h \hat{\Delta}_1(j) \to^p 0.$

Proposition A.2. $n_j h \hat{\Delta}_2(j) \to^p 0.$

To show these propositions, we first state a useful lemma.

Lemma A.1. *Let K_h (z_1, z_2) be defined in* (7). *Then for* $m = 0,\ 1,\ 2$ *and* $\lambda \geq 1$, *we have* $\int_0^1 |\frac{\partial^m}{\partial^m z_2} K_h(z_1, z_2)|^\lambda dz_1 \leq Ch^{1-\lambda(m+1)}$ *for all* $z_2 \in [0, 1]$ *and* $\int_0^1 |\frac{\partial^m}{\partial^m z_2} K_h(z_1, z_2)|^\lambda dz_2 \leq Ch^{1-\lambda(m+1)}$ *for all* $z_1 \in [0, 1]$.

Proof of Lemma A.1. The results follow from change of variable and Assumption 5. ■

Proof of Proposition A.1. Put $\kappa_h(w, w') \equiv K_h\ (z_1,\ z'_{\rm i})\ K_h\ (z_2, z'_2) - 1$ and $W_{j\tau}\ (\theta) \equiv [Z_\tau\ (\theta), Z_{\tau-j}\ (\theta)]$. By a second order Taylor series expansion, we have

$$\hat{g}_j(w) - \tilde{g}_j(w) = (\hat{\theta} - \theta_0)' n_j^{-1} \sum_{\tau=j+1}^{n} \frac{\partial \kappa_h[w, W_{j\tau}(\theta_0)]}{\partial \theta}$$
$$+ \frac{1}{2}(\hat{\theta} - \theta_0)' n_j^{-1} \sum_{\tau=j+1}^{n} \frac{\partial^2 \kappa_h[w, W_{j\tau}(\bar{\theta})]}{\partial\theta\partial\theta'}(\hat{\theta} - \theta_0), \tag{A2}$$

where $\bar{\theta}$ lies between the segment of $\hat{\theta}$ and θ_0. It follows that

$$\hat{\Delta}_1(j) \leq 2|\hat{\theta} - \theta_0|^2 \int_{\mathbb{I}^2} \left| n_j^{-1} \sum_{\tau=j+1}^{n} \frac{\partial \kappa_h[w, W_{j\tau}(\theta_0)]}{\partial\theta} \right|^2 dw + |\hat{\theta} - \theta_0|^4 \int_{\mathbb{I}^2} \left| n_j^{-1} \sum_{\tau=j+1}^{n} \frac{\partial^2 \kappa_h[w, W_{j\tau}(\bar{\theta})]}{\partial\theta\partial\theta'} \right|^2 dw$$
$$\equiv 2|\hat{\theta} - \theta_0|^2 \hat{\Delta}_{11}(j) + |\hat{\theta} - \theta_0|^4 \hat{\Delta}_{12}. \tag{A3}$$

Put $\frac{\partial}{\partial\theta}\hat{\kappa}_h(w) \equiv n_j^{-1} \sum_{\tau=j+1}^{n} \frac{\partial}{\partial\theta}\kappa_h[w, W_{j\tau}(\theta_0)]$. Then

$$\hat{\Delta}_{11}(j) \leq 2\int_{\mathbb{I}^2} \left| E\frac{\partial\hat{\kappa}_h(w)}{\partial\theta} \right|^2 dw + 2\int_{\mathbb{I}^2} \left| \frac{\partial\hat{\kappa}_h(w)}{\partial\theta} - E\frac{\partial\hat{\kappa}_h(w)}{\partial\theta} \right|^2 dw \equiv 2\hat{D}_{1n}(j) + 2\hat{D}_{2n}(j). \tag{A4}$$

We now compute the order of magnitude for $\hat{D}_{1n}(j)$. Using the identity that

$$\frac{\partial \kappa_h[w, W_{j\tau}(\theta)]}{\partial\theta} = \frac{\partial K_h[z_1, Z_\tau(\theta)]}{\partial\theta} K_h[z_2, Z_{\tau-j}(\theta)] + K_h[z_1, Z_\tau(\theta)] \frac{\partial K_h[z_2, Z_{\tau-j}(\theta)]}{\partial\theta}, \tag{A5}$$

the law of iterated expectations, and $E\{K_h[z_1,\ Z_\tau(\theta_0)]|I_{\tau-1}\} = EK_h\ [z_1,\ Z_\tau(\theta_0)] = 1$ under $\mathbb{H}_0$, we have

$$\begin{aligned} E\left\{\frac{\partial \kappa_h[w, W_{j\tau}(\theta_0)]}{\partial\theta}\right\} &= E\left\{E\left[\frac{\partial K_h[z_1, Z_\tau(\theta_0)]}{\partial\theta}\middle| I_{\tau-1}\right] K_h\left[z_2, Z_{\tau-j}(\theta_0)\right]\right\} \\ &\quad + E\left\{\frac{\partial K_h[z_2, Z_{\tau-j}(\theta_0)]}{\partial\theta}\right\}. \end{aligned} \tag{A6}$$

Recall $G_{\tau-1}(z) \equiv E\{[\frac{\partial}{\partial\theta} Z_\tau(\theta)]_{\theta=\theta_0} | Z_\tau(\theta_0) = z, I_{\tau-1}\}$ in Assumption 3. Because

$$\frac{\partial K_h[z_1, Z_\tau(\theta)]}{\partial\theta} = \frac{\partial K_h[z_1, Z_\tau(\theta)]}{\partial Z_\tau(\theta)} \frac{\partial Z_\tau(\theta)}{\partial\theta} \tag{A7}$$

and $\frac{\partial}{\partial Z_\tau(\theta_0)} K_h[z_1, Z_\tau(\theta_0)]$ is a function of $Z_\tau(\theta_0)$, which is independent of $I_{\tau-1}$ under $\mathbb{H}_0$, we have

$$\begin{aligned} E\left\{\frac{\partial K_h[z_1, Z_\tau(\theta_0)]}{\partial\theta}\middle| I_{\tau-1}\right\} &= \int_0^1 \frac{\partial K_h(z_1, z)]}{\partial z} G_{\tau-1}(z)dz \\ &= [G_{\tau-1}(z)K_h(z_1, z)]_{z=0}^{z=1} - \int_0^1 K_h(z_1, z)G'_{\tau-1}(z)dz \\ &= -G'_{\tau-1}(z_1) + o(1), \end{aligned} \tag{A8}$$

where the first equality follows by iterated expectations and the i.i.d. U[0,1] property of $\{Z_\tau(\theta_0)\}$, the second by integration by part, and the last by change of variable $z = z_1 + hu$ and Assumption 3. For the last equality, we have used the fact that $G_{\tau-1}(0) = G_{\tau-1}(1) = 0$ for all integers $\tau > 0$. It follows from Equations (A6) and (A8) that

$$E\left\{\frac{\partial \kappa_h[w, W_{j\tau}(\theta_0)]}{\partial\theta}\right\} = -\{E[G_{\tau-1}(z_1)K_h(z_2, Z_{\tau-j}(\theta_0)] + E[G_{\tau-j-1}(z_2)]\}[1 + o(1)]. \tag{A9}$$

Hence, for the first term in Equation (A4), by change of variable and Assumption 3, we have

$$\hat{D}_{1n}(j) = \int_{\mathbb{I}^2} \left| n_j^{-1} \sum_{\tau=j+1}^{n} E \frac{\partial \kappa_h[w, W_{j\tau}(\theta_0)]}{\partial\theta} \right|^2 dw = O(1). \tag{A10}$$

Next, we consider the second term $\hat{D}_{2n}(j)$ in Equation (A4). Given the Markovian property of the diffusion process X_t, $\frac{\partial}{\partial\theta}\kappa_h[w, W_{j\tau}(\theta_0)]$, as given in Equation (A5), is a measurable function of at most $\{X_{\tau\Delta}, X_{(\tau-1)\Delta}, X_{(\tau-j)\Delta}, X_{(\tau-j-1)\Delta}\}$ and thus is an α-mixing process with α-mixing coefficient $\alpha_j(l\Delta) \le 1$ if $l \le j+1$ and $\alpha_j(l\Delta) = \alpha\ [(l-j-1)\Delta]$ if $l > j+1$ [cf. White (1984), Proposition 6.1.8, p. 153]. By the Cauchy–Schwarz inequality and a standard α-mixing inequality [Hall and Heyde (1980), Corollary A.2, p. 278], we have

$$\begin{aligned}
&\int_{\mathbb{I}^2} E\left|\frac{\partial\hat{\kappa}_h(w)}{\partial\theta} - E\frac{\partial\hat{\kappa}_h(w)}{\partial\theta}\right|^2 dw \\
&\leq 2n_j^{-1}\sum_{l=0}^{n-1}\sum_{\tau=l+1}^{n}\int_{\mathbb{I}^2} cov\left\{\frac{\partial\kappa_h[w, W_{j\tau}(\theta_0)]}{\partial\theta}, \frac{\partial\kappa_h[w, W_{j(\tau-l)}(\theta_0)]}{\partial\theta}\right\} dw \\
&\leq Cn_j^{-1}\left[\sum_{l=0}^{\infty}\alpha_j(l\Delta)^{\frac{\nu}{\nu-1}}\right] n_j^{-1}\sum_{\tau=l+1}^{n}\int_{\mathbb{I}^2}\left\{E\left|\frac{\partial\kappa_h[w, W_{j\tau}(\theta_0)]}{\partial\theta}\right|^{2\nu}\right\}^{1/\nu} dw \\
&= O(n_j^{-1}jh^{-6+2/\nu}),
\end{aligned} \tag{A11}$$

where we made use of the fact that $\sum_{l=0}^{\infty}\alpha_j(l\Delta)^{\frac{\nu}{\nu-1}} \leq (j+1) + \sum_{l=0}^{\infty}\alpha(l\Delta)^{\frac{\nu}{\nu-1}} \leq C(j+1)$ give Assumption 1, and the fact that by Jensen's inequality, the C_r-inequality, (A5), (A7), Lemma A.1, and Assumption 2, we have

$$n_j^{-1}\sum_{\tau=l+1}^{n}\int_{\mathbb{I}^2}\left\{E\left|\frac{\partial\kappa_h[w, W_{j\tau}(\theta_0)]}{\partial\theta}\right|^{2\nu}\right\}^{1/\nu} dw \leq 2C(\nu)h^{-6+2/\nu}\left[n_j^{-1}\sum_{\tau=j+1}^{n}E\left|\frac{\partial Z_\tau(\theta_0)}{\partial\theta}\right|^{2\nu}\right]^{1/\nu}.$$

It follows from (A11) and Markov's inequality that $\hat{D}_{2n}(j) = O_P(n_j^{-1}jh^{-6+2/\nu})$. This, Equations (A4), and (A10) imply

$$\hat{\Delta}_{11}(j) = O_P(1 + n_j^{-1}jh^{-6+2/\nu}). \tag{A12}$$

Next, we consider the second term $\hat{\Delta}_{12}$ in Equation (A3). Noting that

$$\begin{aligned}
\frac{\partial^2\kappa_h[w, W_{j\tau}(\theta)]}{\partial\theta\partial\theta'} = &\frac{\partial^2 K_h[z_1, Z_\tau(\theta)]}{\partial\theta\partial\theta'}K_h[z_2, Z_{\tau-j}(\theta)] \\
&+ K_h[z_1, Z_\tau(\theta)]\frac{\partial^2 K_h[z_2, Z_{\tau-j}(\theta)]}{\partial\theta\partial\theta'} \\
&+ 2\frac{\partial K_h[z_1, Z_\tau(\theta)]}{\partial\theta}\frac{\partial K_h[z_2, Z_{\tau-j}(\theta)]}{\partial\theta'},
\end{aligned} \tag{A13}$$

we write

$$\begin{aligned}
\frac{1}{8}\hat{\Delta}_{12}(j) \leq &\int_{\mathbb{I}^2}\left|n_j^{-1}\sum_{\tau=j+1}^{n}\frac{\partial^2 K_h[z_1, Z_\tau(\theta)]}{\partial\theta\partial\theta'}K_h[z_2, Z_{\tau-j}(\theta)]\right|^2 dw \\
&+ \int_{\mathbb{I}^2}\left|n_j^{-1}\sum_{\tau=j+1}^{n}K_h[z_1, Z_\tau(\theta)]\frac{\partial^2 K_h[z_2, Z_{\tau-j}(\theta)]}{\partial\theta\partial\theta'}\right|^2 dw \\
&+ \int_{\mathbb{I}^2}\left|n_j^{-1}\sum_{\tau=j+1}^{n}\frac{\partial K_h[z_1, Z_\tau(\theta)]}{\partial\theta}\frac{\partial K_h[z_2, Z_{\tau-j}(\theta)]}{\partial\theta'}\right|^2 dw \\
\equiv &\hat{D}_{3n}(j) + \hat{D}_{4n}(j) + \hat{D}_{5n}(j).
\end{aligned} \tag{A14}$$

For the first term in Equation (A14), by the Cauchy–Schwarz inequality, the identity that

$$\frac{\partial^2 K_h[z_1, Z_\tau(\theta)]}{\partial\theta\partial\theta'} = \frac{\partial^2 K_h[z_1, Z_\tau(\theta)]}{\partial^2 Z_\tau(\theta)}\frac{\partial Z_\tau(\theta)}{\partial\theta}\frac{\partial Z_\tau(\theta)}{\partial\theta'} + \frac{\partial K_h[z_1, Z_\tau(\theta)]}{\partial Z_\tau(\theta)}\frac{\partial^2 Z_\tau(\theta)}{\partial\theta\partial\theta'}, \tag{A15}$$

Lemma A.1, and Assumption 2, we have $\hat{D}_{3n}(j) = O_P(h^{-6})$ and $\hat{D}_{4n}(j) = O_P(h^{-6})$. For $\hat{D}_{5n}(j)$ in Equation (A14), we have $\hat{D}_{5n}(j) \leq \{n_j^{-1}\sum_{\tau=1}^{n}\int_0^1|\frac{\partial}{\partial\theta}K_h[z_1, Z_\tau(\theta)]|^2 dz_1\}^2 = O(h^{-6})$

by the Cauchy–Schwarz inequality, Equation (A7), Lemma A.1, and Assumption 2. It follows from (A14) and $\hat{\theta}-\theta_0 = O_P(n^{-1/2})$ that $\hat{\Delta}_{12}(j) = O_P(h^{-6})$. This, (A3), (A12), and Assumption 3 imply $\hat{\Delta}_1(j) = O_P(n_j^{-1} + n_j^{-2}jh^{-6+2/\nu} + n_j^{-2}h^{-6}) = o_P(n_j^{-1}h^{-1})$ given $h=ch^{-\delta}$ for $\delta \in (0, \frac{1}{5})$ and $j = o(n^{1-\delta\,(5-2/\nu)})$. The proof of Proposition A.1 is completed. ■

Proof of Proposition A.2. Using Equation (A2), we have

$$\begin{aligned}\hat{\Delta}_2(j) &= (\hat{\theta}-\theta_0)'\int_{\mathbb{I}^2}[\tilde{g}_j(w)-1]n_j^{-1}\sum_{\tau=j+1}^{n}\frac{\partial\kappa_h[w, W_{j\tau}(\theta_0)]}{\partial\theta}dw\\ &\quad+\frac{1}{2}(\hat{\theta}-\theta_0)'\int_{\mathbb{I}^2}[\tilde{g}_j(w)-1]n_j^{-1}\sum_{\tau=j+1}^{n}\frac{\partial^2\kappa_h[w, W_{j\tau}(\bar{\theta})]}{\partial\theta\partial\theta'}dw(\hat{\theta}-\theta_0)\\ &\equiv (\hat{\theta}-\theta_0)'\hat{\Delta}_{21}(j)+\frac{1}{2}(\hat{\theta}-\theta_0)'\hat{\Delta}_{22}(j)(\hat{\theta}-\theta_0).\end{aligned}\tag{A16}$$

We first consider $\hat{\Delta}_{21}(j)$. Recall the definition of $\frac{\partial}{\partial\theta}\hat{\kappa}(w)$ as used in Equation (A4). We have

$$\begin{aligned}\hat{\Delta}_{21}(j) &= \int_{\mathbb{I}^2}E\frac{\partial\hat{\kappa}_h(w)}{\partial\theta}[\tilde{g}_j(w)-1]dw+\int_{\mathbb{I}^2}\left[\frac{\partial\hat{\kappa}_h(w)}{\partial\theta}-E\frac{\partial\hat{\kappa}_h(w)}{\partial\theta}\right][\tilde{g}_j(w)-1]dw\\ &\equiv \hat{D}_{6n}(j)+\hat{D}_{7n}(j).\end{aligned}\tag{A17}$$

We write $\hat{D}_{6n}(j) = n_j^{-1}\sum_{\tau=j+1}^{n}D_{6n\tau}(j)$, where $D_{6n\tau}(j) \equiv \int_{\mathbb{I}^2}\kappa_h(w, W_{j\tau})E[\frac{\partial}{\partial\theta}\kappa_h(w, W_{j\tau})]dw$ is a j-dependent process with zero mean given $E\kappa_h[w, W_{j\tau}(\theta_0)] = 0$ under $\mathbb{H}_0$. Because $E\{\kappa_h[w, W_{j\tau}(\theta_0)]\,\kappa_h(w, W_{js}(\theta_0)]\} = 0$ unless $\tau = s,\ s \pm j$, we have $E|\hat{D}_{6n}(j)|^2 \le 3n_j^{-1}\sum_{\tau=j+1}^{n}E|D_{6n\tau}(j)|^2 = O(n_j^{-1})$ by Equation (A9), change of variables and Assumption 3. Thus, we have $\hat{D}_{6n}(j) = O_P(n_j^{-1/2})$.

For the second term in Equation (A17), by the Cauchy–Schwarz inequality, Equation (A11), Markov's inequality, and $\sup_{z\in\mathbb{I}^2}|\tilde{g}_j(w)-1| = O_P(n_j^{-1/2}h^{-1}/\ln(n_j))$ as follows from a standard uniform convergence argument for kernel density estimation with application of Bernstein's large deviation inequality, we have $\hat{D}_{7n}(j) = O_P(n_j^{-1}j^{1/2}h^{-4+1/\nu}/\ln(n_j))$. It follows from Equation (A17) that

$$\hat{\Delta}_{21}(j) = O_P(n_j^{-1/2}+n_j^{-1}j^{1/2}h^{-4+1/\nu}/\ln(n_j)).\tag{A18}$$

Next, we consider the second term $\hat{\Delta}_{22}(j)$ in Equation (A16). Using Equation (A13), we have

$$\begin{aligned}\hat{\Delta}_{22}(j) &= \int_{\mathbb{I}^2}[\tilde{g}_j(w)-1]\left\{n_j^{-1}\sum_{\tau=j+1}^{n}\frac{\partial^2K_h[z_1, Z_\tau(\bar{\theta})]}{\partial\theta\partial\theta'}K_h[z_2, Z_{\tau-j}(\bar{\theta})]\right\}dw\\ &\quad+\int_{\mathbb{I}^2}[\tilde{g}_j(w)-1]\left\{n_j^{-1}\sum_{\tau=j+1}^{n}K_h[z_1, Z_\tau(\bar{\theta})]\frac{\partial^2K_h[z_2, Z_{\tau-j}(\bar{\theta})]}{\partial\theta\partial\theta'}\right\}dw\\ &\quad+2\int_{\mathbb{I}^2}[\tilde{g}_j(w)-1]\left\{n_j^{-1}\sum_{\tau=j+1}^{n}\frac{\partial K_h[z_1, Z_\tau(\bar{\theta})]}{\partial\theta}\frac{\partial K_h[z_2, Z_{\tau-j}(\bar{\theta})]}{\partial\theta'}\right\}dw\\ &\equiv \hat{D}_{8n}(j)+\hat{D}_{9n}(j)+2\hat{D}_{10n}(j).\end{aligned}\tag{A19}$$

For the first term in Equation (A19), using Equation (A15), Lemma A.1, and Assumption 2, we have

$$\begin{aligned}\left|\hat{D}_{8n}(j)\right| &\le \sup_{w\in\mathbb{I}^2}|\tilde{g}_j(w)-1|n_j^{-1}\sum_{\tau=j+1}^{n}\int_0^1\left|\frac{\partial^2K_h[z_1, Z_\tau(\bar{\theta})]}{\partial\theta\partial\theta'}\right|dz_1\int_0^1 K_h[z_2, Z_{\tau-j}(\bar{\theta})]dz_2\\ &= O_P(n_j^{-1/2}h^{-3}/\ln(n_j)).\end{aligned}$$

Similarly, we can show $\hat{D}_{9n}(j) = O_P(n_j^{-1/2}h^{-3}/\ln(n_j))$. We also have $\hat{D}_{10n}(j) = O_P(n_j^{-1/2}h^{-3}/\ln(n_j))$ by Equation (A7), Lemma A.1, and Assumption A.2. It follows from Equation (A19) that $\hat{\Delta}_{22}(j) = O_P(n_j^{-1/2}h^{-3}/\ln(n_j))$. This, Equations (A16), (A18), Assumption 4, and the conditions on h and j imply $\hat{\Delta}_2(j) = o_P(n_j^{-1}h^{-1})$. ■

Proof of Theorem A.2. In the proof of Theorem A.2, we set $W_{j\tau} \equiv (Z_\tau, Z_{\tau-j})$ and $Z_\tau \equiv Z_\tau(\theta_0)$. Recalling $\tilde{g}_j(w) - 1 = n_j^{-1}\sum_{\tau=j+1}^{n} \kappa(w, W_{j\tau})$, we write

$$
\begin{aligned}
n_j\widetilde{M}(j) &= n_j^{-1}\sum_{\tau=j+1}^{n}\int_{\mathbb{I}^2}\kappa_h^2(w, W_{j\tau})dw + 2n_j^{-1}\sum_{\tau=j+2}^{n}\sum_{s=j+1}^{\tau-1}\int_{\mathbb{I}^2}\kappa_h(w, W_{j\tau})\kappa_h(w, W_{j\tau})dw \\
&\equiv \tilde{A}_n(j) + \tilde{B}_n(j). \qquad \text{(A20)}
\end{aligned}
$$

We shall show Propositions A.3 and A.4 below. ■

Proposition A.3. *$h\tilde{A}_n(j) - hA_h^0 = O_P(n_j^{-1/2}h^{-3/2})$, where A_h^0 is as in Equation* (10).

Proposition A.4. *$h\tilde{B}_n(j) \to^d N(0, V_0)$, where V_0 is as in Equation* (11).

We first state a useful lemma.

Lemma A.2. *Put $a_h(z_1, z_2) \equiv K_h(z_1, z_2) - 1$ and $b_h(z_1, z_2) \equiv \int_0^1 a_h(z, z_1)a_h(z, z_2)dz$. Then for any $h \equiv h(n) \in (0,\ 1)$ and any integer $n > 0$, we have:* (1) $\int_0^1 a_h(z, z_1)dz_1 = 0$ *and* $\int_0^1 |a_h(z_1, z)|dz_1 \le C$ *for all* $z \in [0, 1]$; (2) $\int_0^1\int_0^1 a_h^2(z_1, z_2)dw = (h^{-1} - 2)\ \int_{-1}^1 k^2(u)du + 2\int_0^1\int_{-b}^1 k_b^2(u)dudb - 1$; (3) $\int_0^1\int_0^1 a_h^4(z_1, z_2)dw = O(h^{-3})$; (4) $\int_0^1 b_h(z_1, z)dz_1 = \int_0^1 b_h(z, z_1)dz_1 = 0$ *for all* $z \in [0, 1]$; (5) $\int_0^1\int_0^1 b_h^2(z_1, z_2)dw = h^{-1}V_0^{1/2}[1 + o(1)]$;(6) $\int_0^1\int_0^1 b_h^4(z_1, z_2)dw = O(h^{-3})$.

Proof of Lemma A.2. The results follow by change of variable, Equation (7), and Assumption 5. ■

Proof of Proposition A.3. By the definition of $\kappa_h(w, w')$ and $a_h(z_1, z_2)$, we can write

$$
\kappa_h(w, W_{j\tau}) = a_h(z_1, Z_\tau)a_n(z_2, Z_{\tau-j}) + a_h(z_1, Z_\tau) + a_h(z_2, Z_{\tau-j}). \qquad \text{(A22)}
$$

Hence, by the Cauchy–Schwarz inequality, Lemma A.2 (6), and $\mathbb{H}_0$, we have $E[\int_0^1\int_0^1 \kappa_h^2(w, W_{j\tau})dz]^2 = O(h^{-6})$ for $j > 0$. Moreover, observing that $W_{j\tau}$ is a j-dependent process such that $W_{j\tau}$ and W_{js} are independent unless $\tau = s,\ s \pm j$, and using Chebyshev's inequality, we have $\tilde{A}_n(j) - E\tilde{A}_n(j) = O_P(n_j^{-1/2}h^{-3})$.

It remains to show $E\tilde{A}_n(j) = A_h^0$. Using the fact that Z_τ and $Z_{\tau-j}$ are independent for $j > 0$, the law of iterated expectations, and Lemmas A.2(1, 2), we have for $j > 0$, $E\tilde{A}_n(j) = [\int_0^1 Ea_h^2(z, Z_1)dz]^2 + 2\int_0^1 Ea_h^2(z, Z_1)dz = A_h^0$. ■

Proof of Proposition A.4. Using (A22) and the definition of $b_h(z_1, z_2)$, we obtain

$$
\begin{aligned}
n_j\widetilde{B}_n(j) = {} & 2n_j^{-1}\sum_{\tau=j+2}^{n}\sum_{s=j+1}^{\tau-1} b_h(Z_\tau, Z_s)b_n(Z_{\tau-j}, Z_{s-j}) \\
& + 2n_j^{-1}\sum_{\tau=j+2}^{n}\sum_{s=j+1}^{\tau-1}[b_h(Z_\tau, Z_s) + b_h(Z_{\tau-j}, Z_{s-j}) \\
& + b_h(Z_\tau, Z_s)\int_0^1 a_h(z_2, Z_{\tau-j})dz_2 + b_h(Z_\tau, Z_s)\int_0^1 a_h(z_2, Z_{s-j})dz_2 \\
& + b_h(Z_{\tau-j}, Z_{s-j})\int_0^1 a_h(z_1, Z_\tau)dz_1 + b_n(Z_{\tau-j}, Z_{s-j})\int_0^1 a_h(z_1, Z_s)dz_1 \\
& + \int_0^1 a_h(z_1, Z_\tau)dz_1\int_0^1 a_h(z_2, Z_{s-j})dz_2 + \int_0^1 a_h(z_1, Z_s)dz_1\int_0^1 a_h(z_2, Z_{\tau-j})dz_2] \\
\equiv {} & \tilde{U}_n(j) + 2\sum_{c=1}^{8}\tilde{B}_{cn}(j).
\end{aligned}
$$

■

Lemma A.3. $n_j h\tilde{B}_n(j) = n_j h\tilde{U}_n(j) + o_P(1)$.

Proof of Lemma A.3. We shall show $n_j h\tilde{B}_{cn}(j) \to^p 0$ for $c=1, \ldots, 8$. First, we consider $c=1$. By Lemma A.2(4) and $\mathbb{H}_0$, $\tilde{B}_{1n}(j)$ is a degenerate U-statistic because $E[b_h(Z_t, Z_s)|Z_s] = E[b_h(Z_t, Z_s)|Z_\tau] = 0$. Hence, we have $E[n_j\tilde{B}_{1n}(j)]^2 = n_j^{-2}\sum_{\tau=j+2}^{n}\sum_{s=j+1}^{\tau-1} Eb_h^2(Z_\tau, Z_s) = O(h^{-1})$, where the last equality follows from Lemma A.2(5). Therefore, $n_j h\tilde{B}_{1n}(j) = O_P(h^{1/2}) = o_P(1)$ given $h \to 0$. Similarly, we have $n_j h\tilde{B}_{2n}(j) = o_P(1)$.

Next, we consider $c=3$. Let $\mathcal{F}_\tau$ be the sigma field consisting of $\{Z_s,\ s \le \tau\}$. Thus, by Lemma A.2(2,4,5) and the law of iterated expectations (conditional on $\mathcal{F}_{\tau-1}$), we obtain $E\tilde{B}_{3n}(j) = 0$, and $E[n_j\tilde{B}_{3n}(j)]^2 = O(h^{-1})$. Hence, we have $n_j h\tilde{B}_{3n}(j) = O_P(h^{1/2}) = o_P(1)$. Similarly, $n_j h\tilde{B}_{cn}(j) \to^p 0$ for $c=4, 5, 6$.

Finally, we consider $c=7$ and 8. For $c=7$, we write $\tilde{B}_{7n}(j) = n_j^{-1}\sum_{\tau=j+2}^{n} B_{7n\tau}(j)$, where $B_{7n\tau}(j) \equiv \int_0^1 a_h(z_1, Z_\tau)dz_1 \sum_{s=j+1}^{\tau-1}\int_0^1 a_h(z_2, Z_{s-j})dz_2$. By Lemma A.2(1) and $\mathbb{H}_0$, we have $E[B_{7n\tau}(j)|\mathcal{F}_{\tau-1}] = 0$ and $E[n_j\tilde{B}_{7n\tau}(j)]^2 \le C(\tau-j)$.. It follows that $E[n_j\tilde{B}_{7n}(j)]^2 \le C$ and $n_j h\tilde{B}_{7n}(j) = o_P(1)$. For $c=8$, we write

$$n_j\tilde{B}_{8n}(j) = n_j^{-1}\sum_{\tau=j+2}^{n}\sum_{s=j+1}^{\tau-1}[\mathbf{1}(s=\tau-j) + \mathbf{1}(s\neq\tau-j)]\int_0^1 a_h(z_1, Z_s)dz_1 \times \int_0^1 a_h(z_2, Z_{\tau-j})dz_2,$$

where the first term (with $\mathbf{1}(s=\tau-j)$) is $O_P(1)$ by Markov's inequality and Lemma A.2(1), and the second term is $O_P(1)$ by using reasoning analogous to that for $n_j\tilde{B}_{7n}(j)$. It follows that $n_j h\tilde{B}_{8n}(j) = O_P(h) = o_P(1)$.

We now show $n_j h\tilde{U}_n(j) \to^d N(0, V_0)$ under $\mathbb{H}_0$. Note that $\tilde{U}_n(j)$ is a U-statistic of a j-dependent process $\{W_{j\tau}\}$ where j is allowed to grow as $n\to\infty$. We shall use Brown's (1971) martingale limit theorem. This approach has also been used in Hong and White (2005) to derive the limit distribution of nonparametric entropy measures of serial dependence. We first approximate $\tilde{U}_n(j)$ by a simpler U-statistic. ■

Lemma A.4. *$n_j h[\tilde{U}_n(j) - \tilde{U}_n^*(j)] \to^p 0$, where $\tilde{U}_n^*(j) \equiv n_j^{-1}\sum_{\tau=2j+2}^{n}\sum_{s=j+1}^{\tau-j-1} U_n(W_{j\tau}, W_{js})$ and $U_n(W_{j\tau}, W_{js}) \equiv 2b_h(Z_\tau, Z_{\tau-j})\, b_h(Z_s, Z_{s-j})$.*

Proof of Lemma A.4. Given the definitions of $\tilde{U}_n(j)$ and $\tilde{U}_n^*(j)$, we write

$$n_j\tilde{U}_n(j) - n_j\tilde{U}_n^*(j) = n_j^{-1}\sum_{\tau=j+2}^{n}\sum_{s=\max(j+1,\tau-j)}^{\tau-1} U_n(W_{j\tau}, W_{js}) \equiv n_j^{-1}\sum_{\tau=j+2}^{n} R_{n\tau}(j), \qquad \text{(A23)}$$

where $\tau - s \le j$ in the first equality. Because $E[U_n(W_{j\tau}, W_{js})|\mathcal{F}_{\tau-1}] = 0$ a.s. for $\tau > s$ under $\mathbb{H}_0$ by Lemma A.1(1), $\{R_{n\tau}(j), \mathcal{F}_{\tau-1}\}$ is an adapted martingale difference sequence (m.d.s.). It follows that $E[n_j\hat{U}_n(j) - n_j\tilde{U}_n^*(j)] = 0$ and $E[n_j\hat{U}_n(j) - n_j\tilde{U}_n^*(j)]^2 = n_j^{-2}\sum_{\tau=j+2}^{n} ER_{n\tau}^2(j)$, where

$$ER_{n\tau}^2(j) \le 2E\,[U_n^2(W_{j\tau}, W_{j(\tau-j)})] + 2E\left[\sum_{s=\max(j=1,\tau-j+1)}^{\tau-1} U_n(W_{j\tau}, W_{js})\right]^2. \qquad \text{(A24)}$$

The first term in Equation (A24) is the contribution from $s=\tau-j$ and the second from $s>\tau-j$. For the first term in Equation (A24), $W_{j\tau}$ and $W_{j(\tau-j)}$ are not independent, but we have

$$EU_n^2(W_{j\tau}, W_{j(\tau-j)})] \le 4Eb_h^4(Z_\tau, Z_s) = O(h^{-3}) \qquad \text{(A25)}$$

by the Cauchy–Schwarz inequality and Lemma A.2(6). For the second term in Equation (A24), $W_{j\tau}$ and W_{js} are independent given $s>\tau-j$. By Lemma A.2(4), we have $E[U_n(w, W_{js})|\mathcal{F}_{s-1}]=E[b_n(z_1, Z_s)|\mathcal{F}_{s-1}]b_h(z_2, Z_{s-j})=0$ a.s. for all $w\in[0,1]^2$. It follows by the law of iterated expectations, $\mathbb{H}_0$ and Lemma A.2(5) that

$$E\left[\sum_{s=\max(j+1,\tau-j+1)}^{\tau-1} U_n(W_{j\tau}, W_{js})\right]^2$$

$$= 4\sum_{s=\max(j+1,\tau-j+1)}^{\tau-1} E[b_h(Z_\tau, Z_s)b_h(Z_{\tau-j}, Z_{s-j})]^2 = O(jh^{-2}).$$

It follows from Equations (A23)–(A26) that $E[n_j\hat{U}_n(j)-n_j\tilde{U}_n(j)]^2 = O(n_j^{-1}h^{-3}+n_j^{-1}jh^{-2})$. Hence, $n_jh[\tilde{U}_n(j)-\tilde{U}_n^*(j)] = O_P(n_j^{-1/2}h^{-1/2}+n_j^{-1/2}j^{1/2}) = o_P(1)$ by Chebyshev's inequality, $h=cn^{-\delta}$ for $\delta\in(0,\frac{1}{5})$ and $j/n_j\to 0$.

We now consider the limit distribution of $\tilde{U}_n^*(j)$. ■

Lemma A.5. $n_jh\tilde{U}_n^*(j)\to^d N(0, V_0)$.

Proof of Lemma A.5. We write $n_jh\tilde{U}_n^*(j)=n_j^{-1}\sum_{\tau=j+1}^n U_{n\tau}^*(j)$, where $U_{n\tau}^*(j)\equiv h\sum_{s=j+1}^{\tau-j-1}U_n(W_{j\tau}, W_{js}) = 2h\sum_{s=j+1}^{\tau-j-1}b_h(Z_\tau, Z_s)b_h(Z_{\tau-j}, Z_{s-j})$. Because $\{U_{n\tau}^*(j), \mathcal{F}_{\tau-1}\}$ is an adapted m.d.s. by $\mathbb{H}_0$ and Lemma A.2(4), we use Brown's (1971) martingale theorem, which states that $V_n^{-1/2}(j)n_jh\tilde{U}_n^*(j)\to^d N(0,1)$ if

$$\text{var}^{-1}[n_jh\tilde{U}_n^*(j)]n_j^{-2}\sum_{\tau=2j+2}^{n} E\{U_{n\tau}^*(j)^2\mathbf{1}[|U_{n\tau}^*(j)|>\epsilon n_j\text{var}^{1/2}[n_jh\tilde{U}_n^*(j)]]\}\to 0 \quad \forall\epsilon>0, \qquad \text{(A27)}$$

$$\text{var}^{-1}[n_jh\tilde{U}_n^*(j)]n_j^{-2}\sum_{\tau=2j+2}^{n}E[U_{n\tau}^*(j)^2|\mathcal{F}_{\tau-1}] \xrightarrow{p} 0. \qquad \text{(A28)}$$

We first show $V_n(j)\equiv\text{var}[n_jh\tilde{U}_n^*(j)]\to V_0$. Noting that $W_{j\tau}$ and W_{js} are independent for $s<\tau-j$, we have $E[U_n(W_{j\tau}, W_{js_1})\,U_n(W_{j\tau}, W_{js_2})|W_{j\tau}]=0$ for any $s_1, s_2<\tau-j$ and $s_1\neq s_2$. It follows by the law of iterated expectations and $\mathbb{H}_0$ that $E[U_{n\tau}^*(j)^2] = h^2\sum_{s=j+1}^{\tau-j-1}EU_n^2(W_{j\tau}, W_{js}) = 4(\tau-2j-1)h^2[Eb_h^2(Z_1, Z_2)]^2$. Therefore, we have $V_n(j) = n_j^{-2}\sum_{\tau=j+2}^{n}E[U_{n\tau}^*(j)^2]\to V_0$ by Lemma A.2(4).

We now can verify condition Equation (A27) by showing $n_j^{-4}\sum_{\tau=2j+2}^{n}E[U_{n\tau}^*(j)^4]\to 0$. Because $E[U_n(w, W_{js})|\mathcal{F}_{s-1}]=0$ a.s. for all $w\in[0,1]^2$ under $\mathbb{H}_0$, we have $E[\sum_{s=j+1}^{n-j-1}U_n(w, W_{js})]^4\leq 4\{\sum_{s=j+1}^{n-j-1}[EU_n^4(w, W_{js})]^{1/2}\}^2$. It follows that $E[U_{n\tau}^*(j)^4]\leq 4\{\sum_{s=j+1}^{\tau-j-1}[EU_n^4(W_{j\tau}, W_{js})]^{1/2}\}^2 = O(\tau^2h^{-6})$ by the law of iterated expectations and Minkowski's inequality, where for $\tau>s+j$,

$$EU_n^4(W_{j\tau}, W_{js})\leq 4[Eb_n^4(Z_\tau, Z_s)]^2\leq Ch^{-6}[1+o(1)] \qquad \text{(A29)}$$

by independence between $W_{j\tau}$ and W_{js} and Lemma A.2(6). Thus, $n_j^{-4}\sum_{\tau=2j+2}^{n}E[U_{n\tau}^*(j)^4] = O(n_j^{-1}h^{-2})\to 0$.

Next, we verify (A28) by showing $E\{n_j^{-2}\sum_{\tau=2j+2}^{n}E[U_{n\tau}^*(j)^2|\mathcal{F}_{\tau-1}]-E[U_{n\tau}^*(j)^2]\}^2\to 0$. For notational simplicity, we put $E_\tau(\cdot)\equiv E(\cdot|\mathcal{F}_\tau)$. Then, we can write

$$E_{\tau-1}[U_{n\tau}^*(j)^2] = h^2\sum_{s=j+2}^{\tau-j-1}E_{\tau-1}[U_n^2(W_{j\tau}, W_{js})] + 2h^2\sum_{s_2=j+1}^{\tau-j-1}\sum_{s_1=j+1}^{s_2-1}E_{\tau-1}[U_n(W_{j\tau}, W_{js_1})U_n(W_{j\tau}, W_{js_2})]$$

$$= h^2\sum_{s=j+2}^{\tau-j-1}EU_n^2(W_{j\tau}, W_{js}) + h^2\sum_{c=1}^{4}Q_{cn\tau}(j),$$

where

$$Q_{1n\tau}(j) \equiv 2 \sum_{s_2=j+1}^{\tau-j-1} \sum_{s_1=j+1}^{s_2-1} E_{\tau-1}[U_n(W_{j\tau}, W_{js_1}) U_n(W_{j\tau}, W_{js_2})],$$

$$Q_{2n\tau}(j) \equiv \sum_{s=j+1}^{\tau-j-1} \{E_{\tau-1}[U_n^2(W_{j\tau}, W_{js}) - E_{\tau-1-j}[U_n^2(W_{j\tau}, W_{js})]\},$$

$$Q_{3n\tau}(j) \equiv \sum_{s=j+1}^{\tau-j-1} \{E_{\tau-1-j}[U_n^2(W_{j\tau}, W_{js}) - E_{s-1}[U_n^2(W_{j\tau}, W_{js})]\},$$

$$Q_{4n\tau}(j) \equiv \sum_{s=j+1}^{\tau-j-1} \{E_{s-1}[U_n^2(W_{j\tau}, W_{js}) - E[U_n^2(W_{j\tau}, W_{js})]\}.$$

Thus, noting $E[U_{n\tau}^*(j)^2] = h^2 \sum_{s=j+1}^{\tau-j-1} EU_n^2(W_{j\tau}, W_{js})$, we obtain

$$E\left[n_j^{-2} \sum_{\tau=2j+2}^{n} \{E_{\tau-1}[U_{n\tau}^*(j)^2] - E[U_{n\tau}^*(j)^2]\}\right]^2 \leq 8n_j^{-4}h^4 \sum_{c=1}^{4} E\left[\sum_{\tau=2j+2}^{n} Q_{cn\tau}(j)\right]^2. \quad \text{(A30)}$$

We first consider $c=1$. Given $U_n(W_{j\tau}, W_{js}) = 2b_h(Z_\tau, Z_s)b_h(Z_{\tau-j}, Z_{s-j})$, we write $Q_{1n\tau}(j) = 2\sum_{s_2=j+1}^{\tau-j-1} \sum_{s_1=j+1}^{s_2-1} q_h(Z_{s_1}, Z_{s_2}) b_h(Z_{\tau-j}, Z_{s_1-j}) b_h(Z_{\tau-j}, Z_{s_2-j})$, where $q_h(z_1, z_2) \equiv E[b_h(Z, z_1)\, b_h(Z, z_2)]$. By Lemma A.2(4), we have $E[q_h(Z, z)] = E[q_h(z, Z)] = 0$ for all $z \in \mathbb{I}$. Moreover, $Z_{\tau-j}$ is independent of (W_{js_1}, W_{js_2}) for $\tau > s_1 - j$, $\tau > s_2 - j$. Thus, conditional on $Z_{\tau-j}$, $Q_{1n\tau}(j)$ has a structure similar to that of $n_j h[\tilde{U}_n(j) - \tilde{U}_n^*(j)]$ in Lemma A.3. Following reasoning analogous to that for $E[n_j h\tilde{U}_n(j) - n_j h\tilde{U}_n^*(j)]^2$ in the proof of Lemma A.3, we have

$$EQ_{1n\tau}^2(j) = E\{E[Q_{1n\tau}^2|Z_{\tau-j}]\} \leq 2 \sum_{s_2=j+2}^{\tau-j-1} \sum_{s_1=j+1}^{s_2-1} E\{E_{\tau-1}[U_n(W_{j\tau}, W_{js_1}) U_n(W_{j\tau}, W_{js_2})]\}^2$$

$$= O(\tau h^{-6} + \tau^2 h^{-2}),$$

where we made use of the facts that (1) for $s_2 = s_1 + j$, W_{js1} and W_{js2} are not independent, but by the Cauchy–Schwarz inequality, Jensen's inequality, and Equation (A29), we have $E\{E_{\tau-1}[U_n(W_{j\tau}, W_{js_1})U_n(W_{j\tau}, W_{js_2})]\}^2 = O(h^{-6})$; and (2) for $s_2 \neq s_1 + j$, W_{js_1} and W_{js_2} are independent, and so

$$E\{E_{\tau-1}[U_n(W_{j\tau}, W_{js_1}) U_n(W_{j\tau}, W_{js_2})]\}^2 = \{E[U_n(W_{j\tau}, W_{js_1}) U_n(W_{j\tau}, W_{js_2})]\}^2 = O(h^{-2}).$$

It follows by Minkowski's inequality that

$$n_j^{-4} h^4 E\left[\sum_{\tau=j+2}^{n} Q_{1n\tau}(j)\right]^2 \leq n_j^{-4} h^4 \left\{\sum_{s=j+1}^{n-j-1} [EQ_{1n\tau}^2(j)]^{1/2}\right\}^2 = O(n_j^{-1} h^{-2} + h^2). \quad \text{(A31)}$$

Next, we bound $E[\sum_{\tau=2j+2}^{n} Q_{2n\tau}(j)]^2$, the second term in Equation (A30). Write $\sum_{\tau=2j+2}^{n} Q_{2n\tau}(j) = \sum_{s=j+1}^{n-j-1} \tilde{Q}_{2ns}(j)$, where $\tilde{Q}_{2ns}(j) \equiv \sum_{\tau=s+j+1}^{n} \{E_{\tau-1}[U_n^2(W_{j\tau}, W_{js})] - E_{\tau-j-1}[U_n^2(W_{j\tau}, W_{js})]\}$. Because the summand in $\tilde{Q}_{2ns}(j)$ is an m.d.s. with respect to $\mathcal{F}_{\tau-j-1}$, we have

$$E\tilde{Q}_{2ns}^2(j) = \sum_{\tau=s+j+1}^{n} E\{E_{\tau-1}[U_n^2(W_{j\tau}, W_{js})] - E_{\tau-j-1}[U_n^2(W_{j\tau}, W_{js})]\}^2 \leq C(n-s-j)h^{-6}[1+o(1)]$$

by Jensen's inequality and Equation (A29). It follows by Minkowski's inequality that

$$n_j^{-4} h^4 E\left[\sum_{\tau=2j+2}^{n} Q_{2n\tau}(j)\right]^2 \leq n_j^{-4} h^4 \left\{\sum_{s=j+1}^{n-j-1} [E\tilde{Q}_{2ns}^2(j)]^{1/2}\right\}^2 = O(n_j^{-1} h^{-2}). \quad \text{(A32)}$$

Similarly, because the summand in $Q_{3n\tau}(j)$ is an m.d.s. with respect to $\mathcal{F}_{s-1}$, and the summand in $Q_{4n\tau}(j)$ is an independent sequence with zero mean, we have $EQ_{3n\tau}^2 \leq C\tau h^{-6}[1+o(1)]$ and $EQ_{4n\tau}^2 \leq C\tau h^{-6}[1+o(1)]$. Hence, by Minkowski's inequality, we have

$$n_j^{-4}h^4 E\left[\sum_{\tau=j+2}^{n} Q_{cn\tau}(j)\right]^2 = O(n_j^{-1}h^{-2}), \quad c = 3, 4. \tag{A33}$$

It follows from Equations (A30)–(A33) that $E\{n_j^{-2}\sum_{\tau=2j+2}^{n} E[U_{n\tau}^*(j)^2|\mathcal{F}_{\tau-1}] - E[U_{n\tau}^*(j)^2]\}^2 = O(n_j^{-1}h^{-2}+h^2)$. Thus, condition (A28) holds given $h = cn^{-\delta}$ for $\delta \in (0, \frac{1}{5})$. Thus, we have $n_j h\tilde{U}_n^*(j) \to^d N(0, V_0)$ by Brown's theorem. ■

Proof of Theorem 2. We use the Cramer–Wold device (e.g., White 1984, Proposition 5.1, p.108). Let $\lambda \equiv (\lambda_1, \ldots, \lambda_L)'$ be a $L \times 1$ vector such that $\lambda'\lambda = 1$. Consider the statistic $\hat{Q}_\lambda \equiv \sum_{c=1}^{L}\lambda_c\hat{Q}(j_c)$. Following reasoning analogous to that for Theorem 1, it can be shown that under $\mathbb{H}_0$ we have $nh\hat{Q}_\lambda = \sum_{c=1}^{L}\lambda_c V_0^{-1/2}[nh\tilde{U}_n^*(j_c) - hA_h^0] + o_P(1)$ and $\sum_{c=1}^{L}\lambda_c V_0^{-1/2}[nh\tilde{U}_n^*(j_c) - hA_h^0] \to^d N(0, V_\lambda)$, where the asymptotic variance

$$\begin{aligned} V_\lambda &\equiv \lim_{n\to\infty}\operatorname{var}\left[\sum_{c=1}^{L}\lambda_c V_0^{-1/2}[nh\tilde{U}_n^*(j_c) - hA_h^0]\right] \\ &= V_0^{-1}\lim_{n\to\infty}\sum_{c=1}^{L}\lambda_c^2\operatorname{var}\left[nh\tilde{U}_n^*(j_c)\right] + 2V_0^{-1}\lim\sum_{c_2=2}^{L-1}\sum_{c_1=1}^{c_2-1}\lambda_{c_1}\lambda_{c_2}\operatorname{cov}\left[nh\tilde{U}_n^*(j_{c_1}), nh\tilde{U}_n^*(j_{c_2})\right] \\ &= \sum_{c=1}^{L}\lambda_c^2, \end{aligned}$$

given $\operatorname{var}[nh\tilde{U}_n^*(j_c)] \to V_0$ and $\operatorname{cov}[nh_n\tilde{U}_n^*(j_{c_1}), nh\tilde{U}_n^*(j_{c_2})] = 0$ whenever $j_{c_1} \neq j_{c_2}$. It follows by the Cramer–Wold device that $\hat{Q}_L \equiv [\hat{Q}(j_1), \ldots, \hat{Q}(j_L)]' \to^d N(0, I)$. ■

Proof of Theorem 3. Put $M_1(j) \equiv \int_{\mathbb{I}^2}[g_j(w)-1]^2 dw$. Then

$$\hat{M}(j) - M_1(j) = \int_{\mathbb{I}^2}[\hat{g}_j(w) - g_j(w)]^2 dw + 2\int_{\mathbb{I}^2}[\hat{g}_j(w) - g_j(w)][g_j(w)-1]dw. \tag{A34}$$

We now show $\hat{M}(j) - M_1(j) \to^p 0$. Note that $\int_{\mathbb{I}^2}[\hat{g}_j(w) - g_j(w)]^2 dw \leq 2\int_{\mathbb{I}^2}[\hat{g}_j(w) - \tilde{g}_j(w)]^2 dw + \int_{\mathbb{I}^2}[\tilde{g}_j(w) - g_j(w)]^2 dw$. For the first term, we have $\int_{\mathbb{I}^2}[\hat{g}_j(w) - \tilde{g}_j(w)]^2 dw \to^p 0$ following reasoning analogous to that of Theorem A.1. For the second term, using the decomposition that $\tilde{g}_j(w) - g_j(w) = [\hat{g}_j(w) - E\hat{g}_j(w)] + [E\hat{g}_j(w) - g_j(w)]$, the α-mixing condition in Assumption 1, change of variable, and a Taylor series expansion for the bias, we can show $\int_{\mathbb{I}^2}[\tilde{g}_j(w) - g_j(w)]^2 dw = O_P(n_j^{-1}h^{-2} + h^2)$, where the $O(h^2)$ term is the squared bias given Assumption 6. It follows that $\int_{\mathbb{I}^2}[\hat{g}_j(w) - g_j(w)]^2 dw \to^p 0$ given $h = cn^{-\delta}$ for $c \in (0, \infty)$ and $\delta \in (0, \frac{1}{5})$. We thus have $\hat{M}(j) - M_1(j) \to^p 0$ by the Cauchy–Schwarz inequality and Equation (A34). Moreover, given $(n_j h)^{-1}A_h^0 = O(n_j^{-1}h^{-3}) = o(1)$, we have $(n_j h)^{-1}\hat{Q}(j) = V_0^{-1/2}M_1(j) + o_P(1)$. It follows that $P[\hat{Q}(j) > C_n] \to 1$ for any $C_n = o(nh)$ whenever $M_1(j) > 0$, which holds when $\{Z_\tau, Z_{\tau-j}\}$ are not independent or U[0, 1]. ■

References

Ahn, D., and B. Gao, 1999, "A Parametric Nonlinear Model of Term Structure Dynamics," *Review of Financial Studies*, 12, 721–762.

Ait-Sahalia, Y., 1996, "Testing Continuous-Time Models of the Spot Interest Rate," *Review of Financial Studies*, 9, 385–426.

Ait-Sahalia, Y., 1999, "Transition Densities for Interest Rate and Other Nonlinear Diffusions," *Journal of Finance*, 54, 1361–1395.

Ait-Sahalia, Y., 2002a, "Maximum-Likelihood Estimation of Discretely Sampled Diffusions: A Closed-Form Approach," *Econometrica*, 70, 223–262.

Ait-Sahalia, Y., 2002b, "Closed-Form Expansion for Multivariate Diffusions," working paper, Princeton University.

Ait-Sahalia, Y., and R. Kimmel, 2002, "Estimating Affine Multifactor Term Structure Models Using Closed-Form Likelihood Expansions," working paper, Princeton University.

Andersen, T.G., and J. Lund, 1997, "Estimating Continuous-Time Stochastic Volatility Models of the Short-Term Interest Rate," *Journal of Econometrics*, 77, 343–378.

Black, F., and M. Scholes, 1973, "The Pricing of Options and Corporate Liabilities," *Journal of Political Economy*, 81, 637–654.

Bliss, R., 1997, "Testing Term Structure Estimation Methods," *Advances in Futures and Options Research*, 9, 197–231.

Brandt, M., and P. Santa-Clara, 2002, "Simulated Likelihood Estimation of Diffusions with an Application to Exchange Rate Dynamics in Incomplete Markets," *Journal of Financial Economics*, 63, 161–210.

Brown, B.M. 1971, "Martingale Limit Theorems," *Annals of Mathematical Statistics*, 42, 59–66.

Campbell, J., A. Lo, and C. MacKinlay, 1997, *The Econometrics of Financial Markets*, Princeton University Press, NJ.

Chacko, G., and S. Das, 2002, "Pricing Interest Rate Derivatives: A General Approach," *Review of Financial Studies*, 15, 195–241.

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

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

Conley, T.G., L.P. Hansen, E.G.J. Luttmer, and J.A. Scheinkman, 1997, "Short-Term Interest Rates as Subordinated Diffusions," *Review of Financial Studies*, 10, 525–578.

Cox, J.C., J.E. Ingersoll, and S.A. Ross, 1985, "A Theory of the Term Structure of Interest Rates," *Econometrica*, 53, 385–407.

Dai, Q., and K. Singleton, 2000, "Specification Analysis of Affine Term Structure Models," *Journal of Finance*, 55, 1943–1978.

Dai, Q., and K. Singleton, 2002, "Term Structure Dynamics in Theory and Reality," *Review of Financial Studies*, 16, 631–678.

Das, S. R., 2002, "The Surprise Element: Jumps in Interest Rates," *Journal of Econometrics*, 106, 27–65.

Diebold, F.X., T. Gunther, and A. Tay, 1998, "Evaluating Density Forecasts with Applications to Financial Risk Management," *International Economic Review*, 39, 863–883.

Diebold, F.X., J. Hahn, and A.S. Tay, 1999, "Multivariate Density Forecast Evaluation and Calibration in Financial Risk Management: High-Frequency Returns of Foreign Exchange," *Review of Economics and Statistics*, 81, 661–673.

Duffee, G., 2002, "Term Premia and Interest Rate Forecasts in Affine Models," *Journal of Finance*, 57, 405–443.

Duffie, D., J. Pan, and K. Singleton, 2000, "Transform Analysis and Asset Pricing for Affine Jump-Diffusions," *Econometrica*, 68, 1343–1376.

Duffie, D., L. Pedersen, and K. Singleton, 2003, "Modeling Credit Spreads on Sovereign Debt: A Case Study of Russian Bonds," *Journal of Finance*, 55, 119–159.

Duffie, D., and K. Singleton, 1993, "Simulated Moments Estimation of Markov Models of Asset Prices," *Econometrica*, 61, 929–952.

Egorov, A., H. Li, and Y. Xu, 2003, "Maximum Likelihood Estimation of Time-Inhomogeneous Diffusions," *Journal of Econometrics*, 114, 107–139.

Eraker, B., 2001, "MCMC Analysis of Diffusion Models with Applications to Finance," *Journal of Business and Economic Statistics*, 19–2, 177–191.

Gallant, A.R., and G. Tauchen, 1996, "Which Moments to Match?" *Econometric Theory*, 12, 657–681.

Hall, P., and C. Heyde, 1980, *Martingale Limit Theory and Its Application*, Academic Press, New York.

Hansen, L.P., and J.A. Scheinkman, 1995, "Back to the Future: Generating Moment Implications for Continuous Time Markov Processes," *Econometrica*, 63, 767–804.

Härdle, W., 1990, *Applied Nonparametric Regression*, Cambridge University Press, New York.

Hong, Y., 1996, "Consistent Testing for Serial Correlation of Unknown Form," *Econometrica*, 64, 837–864.

Hong, Y., 2001, *Evaluation of Out-of-Sample Density Forecasts with Applications to Stock Prices*, Department of Economics and Department of Statistical Science, Working Paper, Cornell University.

Hong, Y., and H. White, 2005, *Asymptotic Theory for Nonparametric Entropy Measures of Serial Dependence, Econometrica*, forthcoming.

Jiang, G., and J. Knight, 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.

Jiang, G., and J. Knight, 2002, "Estimation of Continuous-Time Processes via the Empirical Characteristic Function," *Journal of Business and Economic Statistics*, 20, 198–212.

Johannes, M., 2004, "The Statistical and Economic Role of Jumps in Interest Rates," *Journal of Finance*, 59, 227–260.

Lo, A.W., 1988, "Maximum Likelihood Estimation of Generalized Ito Processes with Discretely Sampled Data," *Econometric Theory*, 4, 231–247.

McCulloch, J., and H. Kwon, 1993, "U.S. Term Structure Data, 1947–1991," Ohio State University working paper 93–96.

Pedersen, A. R., 1995, "A New Approach to Maximum Likelihood Estimation for Stochastic Differential Equations Based on Discrete Observations," *Scandinavian Journal of Statistics*, 22, 55–71.

Piazzesi, M. 2002, "Affine Term Structure Models," working paper, University of California, Los Angeles.

Pritsker, M., 1998, "Nonparametric Density Estimation and Tests of Continuous Time Interest Rate Models," *Review of Financial Studies*, 11, 449–487.

Robinson, P., 1983, "Nonparametric Estimators for Time Series," *Journal of Time Series Analysis*, 4, 184–207.

Rosenblatt, M., 1952, "Remarks on a Multivariate Transformation," *Annals of Mathematical Statistics*, 23, 470–472.

Scott, D.W., 1992, *Multivariate Density Estimation, Theory, Practice, and Visualization*, John Wiley and Sons, New York.

Singleton, K., 2001, "Estimation of Affine Asset Pricing Models Using the Empirical Characteristic Function," *Journal of Econometrics*, 102, 111–141.

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

Sundaresan, S., 2001, "Continuous-Time Methods in Finance: A Review and an Assessment," *Journal of Finance*, 55, 1569–1622.

Tauchen, G., 1997, "New Minimum Chi-Square Methods in Empirical Finance," in D. Kreps and K. Wallis (eds.), *Advances in Econometrics: Seventh World Congress*, Cambridge University Press, Cambridge, UK.

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

White, H., 1984, *Asymptotic Theory for Econometricians*, Academic Press, San Diego.