Model Selection Testing for Diffusion Processes with Applications to Interest Rate and Exchange Rate Models

Hwan-sik Choi*
Texas A&M University

Minsoo Jeong†
Indiana University

Joon Y. Park‡
Indiana University

Sep, 2009

Abstract

A model selection test for nonnested misspecified diffusion models is developed based on the Kullback-Leibler information criterion. A new asymptotic framework accounts for the high significance of diffusion functions relative to drift functions for high frequency data. The test examines the hypothesis that two competing models are equivalent. Our approach distinguishes the roles of diffusion and drift functions and shows the equivalence of models must be understood differently depending on the sampling frequencies. When the sampling frequency is high, it is of primary importance for a model to have a diffusion function close to the true diffusion function, and we compare drift functions when the models can not be distinguished by the diffusion functions. As the sampling frequencies become higher, the diffusion functions are more important, and the information for ranking the drift functions is weaker. The drift functions are useful only when we sample data for long enough. Our new asymptotics deals with the different rates of information in the diffusion and drift functions by considering both the sampling interval Δ and the sampling span T, and we show the sampling span must increase at a relative speed faster than Δ^{-2} (or $\Delta^2 T \to \infty$) to ensure sufficient information to be collected for distinguishing two models by their drift functions. The limiting distribution of the test statistic is normal, and we compare different asymptotic approximations to the sampling distribution of the test statistic using subsampling, and nonparametric block bootstrap methods, as well as the standard normal approximation for the test statistics standardized by the heteroskedasticity autocorrelation consistent variance estimators. We apply our test to spot interest rate models and exchange rate models. We find that many popular models are observationally equivalent.

*hwansik.choi@tamu.edu. 3035 Allen, 4228 TAMU, Department of Economics, Texas A&M University, College Station, TX 77843-4228, USA.

†jeongm@indiana.edu. Wylie Hall, Department of Economics, Indiana Univesrity, Bloomington, IN 47405, USA.

‡joon@indiana.edu. Wylie Hall, Department of Economics, Indiana Univesrity, Bloomington, IN 47405, USA.

JEL: C22, C52

Keywords: nonnested, misspecification, diffusion, model selection, infill asymptotics, likelihood ratio, spot rate, exchange rate.

1 Introduction

We propose a model selection test for two nonnested misspecified diffusion models by comparing the likelihoods based on the Kullback-Leibler information criterion (KLIC, Kullback and Leibler (1951)) under an asymptotic environment that is useful to account for the dominant nature of diffusion functions for high frequency data. Our testing framework examines the null hypothesis of the equivalence of two models in the log-likelihood ratio criterion. The test is directional; a model with a higher likelihood is preferred to the other, and a failure to reject the null suggests the observational equivalence of candidate models.

We show the crucial role of sampling frequencies in distinguishing diffusion models. For high frequency data, having a good diffusion function specification is more important than having a good drift function. The asymmetric importance of diffusion and drift functions is more prominent as the sampling frequency becomes higher. This phenomenon can not be explained by the traditional model selection methods under a fixed sampling frequency such as Vuong (1989), Rivers and Vuong (2002), and Choi and Kiefer (2008). Although our approach is developed on a similar ground to the previous work, we separate the roles of diffusion and drift functions by considering sampling frequencies explicitly showing the different relative informational value of diffusion and drift functions. Specifically, we analyze the log-likelihood ratio criterion with the asymptotics under which data are measured at increasing frequencies (infill) and over increasing periods (long span). While the long span part of the asymptotics is useful in distinguishing two diffusion models, the infill part poses a challenge in distinguishing the information from diffusion functions and drift functions, because the information from drift functions does not accumulate by the infill.

As sampling frequencies increase, diffusion functions become the primary component in the log-likelihood ratio criterion for the model comparison, and two diffusion models are equivalent if their diffusion functions have an equal divergence from the true diffusion function. This implies that the likelihood ratio criterion selects the diffusion model with a better diffusion function regardless of the degree of misspecification in drift functions. Drift functions are useful only if the models are equivalent in terms of the diffusion functions. For example, suppose (μ_0, σ_0) represents the true drift function μ_0 and diffusion function σ_0. Let (μ_i^*, σ_i^*) be the "closest" members of diffusion models $i = 1, 2$ in terms of the likelihood criterion. The closest members are defined from the probability limits of maximum likelihood estimators. For high frequency data, a superior σ_i^* is always preferred, and a good μ_i^* is valued when σ_1^* and σ_2^* are equivalent under the likelihood ratio criterion. See Figure 1.

Figure 1 is about here.

When the diffusion functions are equivalent, the usefulness of the drift functions depends on the sampling frequency, and comparing the two models with the drift functions can be difficult for high frequency data. Since the diffusion functions have no information on the distinguishability of the models in this case, the primary model selection factor in the likelihood ratio becomes a noise. If a model has a better drift function than the other, the information about the superior drift function must accumulate fast enough to dominate this noise for a valid inference. This "signal-to-noise" ratio becomes lower as the sampling frequency is higher (or the sampling interval Δ is shorter) relative to the sampling span T. We show that if the sampling frequency is too high, it is not possible to distinguish the two models by their drift functions when they have equivalent diffusion functions. The relative maximum data frequency (or relative minimum sampling span) allowed for a valid model comparison is given by $\Delta\sqrt{T} \to \infty$. When the sampling frequency is higher than this bound, the drift functions are not meaningful; only the diffusion functions matter.

This result also indicates that, when it is difficult to distinguish two models with diffusion functions, it may help to sample infrequently to improve the signal-to-noise ratio. In practice, as we are less flexible in choosing a sampling span generally, an infrequent sampling implies fewer observations, and it is not clear if the infrequent sampling would help distinguish the models. But we advise to try lower frequencies when the test fails to reject at a higher frequency.

We also show the choice of an approximation method for transition densities affects the model selection criterion when the diffusion functions are equivalent. Therefore, the models are ranked by both drift functions and approximation methods in this case, and the choice of an approximation method should be considered as an integrated part of modeling.

A model selection test is different from nested or nonnested specification tests. Specification tests have been widely used to check the adequacy of a model in explaining data. Although the evaluation and the measurement of specification errors are important to choose or distinguish different models, since all models are approximations, these methods may only lead to the conclusion that we can not detect that a model is misspecified with a given sample size. If more data were available, the model would be rejected naturally. Moreover, specification tests do not give a good indicator for choosing the best model among many models, when multiple models could not be rejected making them observationally indistinguishable. When models are thought to be misspecified, defining a measure of misspecification and comparing the models for possible superiority can be a sensible alternative approach. This type of approach is called the model selection (Davidson and MacKinnon (1981)).

For the model selection approach, testing equivalence of two models with the likelihood criterion as a misspecification measure is studied in Vuong (1989) with i.i.d. data in discrete-time (fixed Δ) sampling environments. Rivers and Vuong (2002) and Choi and Kiefer (2008) considered more general situations using a large class of divergence measures for stationary data with unknown serial correlation. They used the heteroskedasticity autocorrelation consistent (HAC) variance estimator to construct a robust test. Choi and Kiefer (2008) used the fixed-b asymptotic approximation proposed by Kiefer and Vogelsang (2002) and Kiefer and Vogelsang (2005). The validity of these approaches critically depends on the condition that candidate models are nonnested and misspecified. See Section 6 in Rivers and Vuong (2002) for further discussion. Our new approach extends the model selection tests to diffusion models, and gives important new insights on the distinguishability of models and its relationship with sampling frequencies in the continuous-time framework.

Chen and Scott (1993) applied the test of Vuong (1989) to compare nonnested affine interest rate term structure models, and Aït-Sahalia and Kimmel (2008) used Vuong's test with the likelihoods obtained from the closed-form approximation of Aït-Sahalia (2002). In both the papers, the likelihoods should be i.i.d. for the validity of Vuong's test, and the sampling frequencies are expected to be low. Although we cover univariate diffusion processes only in this paper, our test does not require the likelihoods to be i.i.d. and uses a practical asymptotic framework especially relevant to financial data, for which high frequency measurements are available over a reasonably long span that may make drift functions useful. More importantly, the new asymptotics deals with the different roles of drift and diffusion functions directly, which was not considered in previous literature.

We apply our test to select a spot interest rate model and a foreign exchange rate model, and it is shown that many popular models are difficult to distinguish in practice.

2 Model Selection for Diffusion Models

2.1 Model

Consider a time-homogeneous stationary Itô diffusion process X_t and the standard Brownian motion W_t defined on a probability space $(\Omega, \mathfrak{F}, \mathbb{P})$. Let X_t be a weak solution to a stochastic differential equation (SDE)

$$dX_t = \mu_0(X_t)dt + \sigma_0(X_t)dW_t, \tag{2.1}$$

which satisfies the conditions in Karatzas and Shreve (1991) to admit a weak solution. We observe n samples $\{X_{i\Delta}\}_{i=0}^n$ from the diffusion process X_t measured from time zero to $T = n\Delta$ at a non-random time interval Δ.

Let $\mathcal{D}_X \subset \mathbb{R}$ be the range of X_t such that $\{X_t | t \geq 0\} \subset \mathcal{D}_X$, $\mathbb{P}$-a.s. Consider a parametric diffusion model $M(\theta)$ for X_t that solves a SDE

$$M(\theta) : dX_t = \mu(X_t; \theta)dt + \sigma(X_t; \theta)dW_t, \tag{2.2}$$

where $\mu(\cdot; \cdot)$ and $\sigma(\cdot; \cdot)$ are known functions with an unknown parameter vector θ in a compact set $\Theta \subset \mathbb{R}^k$.

Definition 2.1 *A diffusion model $M(\theta)$ is misspecified if*

$$\mathbb{P}\{(\mu_0(X_t), \sigma_0(X_t)) = (\mu(X_t;\theta), \sigma(X_t;\theta))\} < 1, \tag{2.3}$$

for all $\theta \in \Theta$.

Let $p_0(t,x,y)$ and $p(t,x,y;\theta)$ be the transition densities from $X_0 = x$ to $X_t = y$ of the true process X_t and the process that solves a misspecified diffusion model $M(\theta)$ respectively. Let the maximum likelihood (ML) estimator

$$\hat{\theta} = \underset{\theta\in\Theta}{\operatorname{argmax}} \sum_{i=1}^{n} \log p(\Delta, X_{(i-1)\Delta}, X_{i\Delta}; \theta) \tag{2.4}$$

converge in probability to an interior point θ^* of Θ as $\Delta \to 0$ and $T \to \infty$. The probability limit θ^* is the pseudo-true value, and we call the transition density $p(t,x,y;\theta^*)$ the pseudo-true transition density. The process implied by $M(\theta^*)$ is said to be the pseudo-true process. In practice, the true transition density $p_0(t,x,y)$ is not known, and the transition density $p(t,x,y;\theta)$ implied by $M(\theta)$ may not be available in closed-form except for a few classes of diffusions. When we do not have the closed-form transition density $p(t,x,y;\theta)$, we must obtain an approximate transition density from a method such as the Euler or Milstein approximations, or the simulated likelihood method of Brandt and Santa-Clara (2002), or the Hermite expansion of Aït-Sahalia (2002). We show that the choice of an approximation method has important implications in our model selection criterion, and models equivalent in the true likelihoods may not be equivalent in approximate likelihoods. Our asymptotic results can be used as long as the approximate transition densities are "close" enough to the true transition density satisfying the regularity conditions introduced later.

Suppose we have two misspecified diffusion models $M_1(\theta_1)$ and $M_2(\theta_2)$ with transition densities $p_j(t,x,y;\theta_j)$ and the log-likelihood functions

$$\mathcal{L}_j(\theta_j) = \mathcal{L}_j^{n,\Delta}(\theta_j) = \sum_{i=1}^{n} \log p_j(\Delta, X_{(i-1)\Delta}, X_{i\Delta}; \theta_j) \tag{2.5}$$

of data $\{X_{i\Delta}\}_{i=0}^n$ conditional on $X_0 = x_0$ for $j = 1, 2$, respectively. We suppress the dependency of the log-likelihoods on n and Δ (or equivalently T and Δ) for notational simplicity. Our model selection approach compares two models by the log-likelihood ratio

$$\mathcal{L}_1(\theta_1^*) - \mathcal{L}_2(\theta_2^*) = \sum_{i=1}^{n} \log \frac{p_1(\Delta, X_{(i-1)\Delta}, X_{i\Delta}; \theta_1^*)}{p_2(\Delta, X_{(i-1)\Delta}, X_{i\Delta}; \theta_2^*)}, \tag{2.6}$$

where θ_j^* $(j = 1, 2)$ are pseudo-true values. The comparison with the log-likelihood ratio is based on the KLIC between two probability measures as in Vuong (1989). In general, the KLIC(P,Q)

from a probability measure P to an equivalent probability measure Q, is defined by the expectation with respect to the measure P,

$$\mathrm{KLIC}(P,Q) = \mathbf{E}_P(\log(dP/dQ)),$$

and it is infinite if P and Q are not equivalent. See Csiszár (1967a,b, 1975) for more general divergence measures.

Although our test is based on the likelihood ratio, the asymptotic behaviors of our test statistics are different under our new framework. We test for the null hypothesis that two models are equivalent in the sense that the scaled log-likelihood ratio

$$\frac{1}{T}(\mathcal{L}_1(\theta_1^*) - \mathcal{L}_2(\theta_2^*)) \tag{2.7}$$

converges to zero in probability as $T \to \infty, \Delta \to 0$. Let

$$H_0: \operatorname*{plim}_{T\to\infty,\Delta\to 0} \frac{1}{T}(\mathcal{L}_1(\theta_1^*) - \mathcal{L}_2(\theta_2^*)) = 0, \tag{2.8}$$

$$H_1: \operatorname*{plim}_{T\to\infty,\Delta\to 0} \frac{1}{T}(\mathcal{L}_1(\theta_1^*) - \mathcal{L}_2(\theta_2^*)) > 0, \tag{2.9}$$

$$H_2: \operatorname*{plim}_{T\to\infty,\Delta\to 0} \frac{1}{T}(\mathcal{L}_1(\theta_1^*) - \mathcal{L}_2(\theta_2^*)) < 0. \tag{2.10}$$

The null hypothesis H_0 is tested against the alternative $H_1 \cup H_2$. Our test is directional; when $M_1(\theta_1)$ is preferred, the scaled log-likelihood ratio converges to a positive number (H_1).

2.2 Asymptotics

We derive the asymptotic results for the log-likelihood ratio $\mathcal{L}_1(\hat\theta_1) - \mathcal{L}_2(\hat\theta_2)$, where $\hat\theta_j$ ($j = 1,2$) are the ML estimators, with $T \to \infty, \Delta \to 0$. We first define the functions that describe the behavior of the transition densities during a small time interval t.

Definition 2.2 *The derivatives of log transition densities*

$$\ell_j(t,x,y) = \log p_j(t,x,y;\theta_j^*) \tag{2.11}$$

are

$$\ell_{jy}(t,x,y) = \partial \log p_j(t,x,y;\theta_j^*)/\partial y, \tag{2.12}$$

$$\ell_{jt}(t,x,y) = \partial \log p_j(t,x,y;\theta_j^*)/\partial t, \tag{2.13}$$

for $j=1,2$. Other derivatives, $\ell_{jyy}(t,x,x), \ell_{jyyy}(t,x,x), \ell_{jyyyy}(t,x,x), \ell_{jyyt}(t,x,x)$, are defined similarly. For $j=1,2$, define

$$A_j(x)=\lim_{t\to 0}\{\ell_{jy}(t,x,x)+t\ell_{jyt}(t,x,x)\},\ B_j(x)=\lim_{t\to 0} t\ell_{jyyy}(t,x,x),\ C_j(x)=\lim_{t\to 0} t\ell_{jyyyy}(t,x,x),$$
$$D_j(x)=\lim_{t\to 0}\left\{2\ell_{jt}(t,x,x)+t\ell_{jtt}(t,x,x)+(2t)^{-1}\right\},\ E_j(x)=\lim_{t\to 0}\left\{\ell_{jyy}(t,x,x)+t\ell_{jyyt}(t,x,x)\right\},$$

for all $x\in\mathcal{D}_X$.

For simplicity, we assume that drift and diffusion parameters are separable. We write $\theta_j = (\alpha_j,\beta_j)$, where α_j and β_j are parameter vectors for drift functions $\mu_j(\cdot;\alpha_j)$ and diffusion functions $\sigma_j(\cdot;\beta_j)$ for $j=1,2$, respectively, and denote $\theta_j^*=(\alpha_j^*,\beta_j^*)$.

Let $\mu_0=\mu_0(X_t)$, $\sigma_0=\sigma_0(X_t)$ and denote $\mu_j=\mu_j(X_t;\alpha_j^*)$, $\sigma_j=\sigma_j(X_t;\beta_j^*)$, $A_j=A_j(X_t)$, and their derivatives $\sigma_{j\beta}=\partial\sigma_j(X_t;\beta_j^*)/\partial\beta$, $A_{j\beta}=\partial A_j(X_t)/\partial\beta$ for $j=1,2$. Also $B_j,\dots,E_j$ $(j=1,2)$ are defined likewise.

We show the results in two cases. When models have different diffusion functions (denoted as Case 1), their superiority is determined by the diffusion functions. When models have identical diffusion functions (denoted as Case 2), they must be compared by the drift functions because the pseudo-true values of diffusion function parameters are the same under our asymptotics regardless of the drift specifications. The main results are in the following theorem. All proofs are in Appendix.

Theorem 2.3 (Asymptotic expansions) *(a) Case 1: $\sigma_1(\,\cdot\,;\beta_1^*)\neq\sigma_2(\,\cdot\,;\beta_2^*)$.*
If $\Delta^2T\to\infty$, $\Delta^3T\to 0$ as $T\to\infty$, $\Delta\to 0$, and Assumptions B.1-B.4 are satisfied, then

$$\begin{aligned}\mathcal{L}_1(\hat\theta_1)-\mathcal{L}_2(\hat\theta_2)=&-\frac{1}{\Delta}\int_0^T\left[\log\left(\frac{\sigma_1}{\sigma_2}\right)+\frac{1}{2}\left(\frac{1}{\sigma_1^2}-\frac{1}{\sigma_2^2}\right)\sigma_0^2\right]dt\\&+\int_0^T\left[(A_1-A_2)\mu_0+\frac{(B_1-B_2)\mu_0\sigma_0^2}{2}+\frac{(C_1-C_2)\sigma_0^4}{8}\right.\\&\left.+\frac{D_1-D_2}{2}+\frac{(E_1-E_2)\sigma_0^2}{2}-\frac{1}{2}\left(\frac{1}{\sigma_1^2}-\frac{1}{\sigma_2^2}\right)\mu_0^2\right]dt+O_p\left(\sqrt{\frac{T}{\Delta}}\right).\end{aligned}$$

(b) Case 2: $\sigma_1(\,\cdot\,;\beta_1^*)=\sigma_2(\,\cdot\,;\beta_2^*)$ and $\mu_1(\,\cdot\,;\alpha_1^*)\neq\mu_2(\,\cdot\,;\alpha_2^*)$.

If $\Delta T^2 \to 0$ as $T \to \infty$, $\Delta \to 0$, and Assumptions B.1-B.4 are satisfied, then

$$\begin{aligned}
&\mathcal{L}_1(\hat{\theta}_1) - \mathcal{L}_2(\hat{\theta}_2) \\
&\quad = \int_0^T \left[(A_1 - A_2)\mu_0 + \frac{D_1 - D_2}{2} + \frac{(E_1 - E_2)\sigma_0^2}{2} \right] dt \\
&\quad + \int_0^T (A_1 - A_2)\sigma_0 dW_t \\
&\quad - \int_0^T \frac{\sigma'_{1\beta}(\sigma_0^2 - \sigma_1^2)}{\sigma_1^3} dt \left(\int_0^T \frac{(\sigma_1 \sigma_{1\beta\beta'} - 3\sigma_{1\beta}\sigma'_{1\beta})(\sigma_0^2 - \sigma_1^2) - 2\sigma_1^2 \sigma_{1\beta}\sigma'_{1\beta}}{\sigma_1^4} dt \right)^{-1} \times \\
&\qquad \int_0^T \left[(A_{1\beta} - A_{2\beta})\mu_0 + \frac{D_{1\beta} - D_{2\beta}}{2} + \frac{(E_{1\beta} - E_{2\beta})\sigma_0^2}{2} \right] dt + o_p(\sqrt{T}).
\end{aligned}$$

For Case 1, the diffusion functions of the competing models are the most important factor in the likelihood ratio, and we have the second order term related to both the drift functions and the approximation methods of transition densities. The leading term of Case 2 depends on the drift functions and the approximation methods of transition densities as well as the sampling error of the estimators of diffusion function parameters. See Appendix for the examples of the explicit expansions for the Euler and Milstein approximations as well as Aït-Sahalia's Hermite expansion approximation.

In Case 1, the diffusion functions $\sigma_1(\beta_1^*)$ and $\sigma_2(\beta_2^*)$ of models are compared at their pseudo-true values β_1^* and β_2^* as the leading criterion. The diffusion functions dominate as data frequencies become higher. Moreover, β_1^* and β_2^* do not depend on drift function specifications, which gives some insights to the fact that estimated diffusion parameters do not usually depend much on drift function specifications for high frequency data. Therefore, a model with a good diffusion function is preferred no matter how bad its drift function is.

If we have

$$\begin{aligned}
\operatorname*{plim}_{T\to\infty,\Delta\to 0} \frac{\Delta}{T}(\mathcal{L}_1(\hat{\theta}_1) - \mathcal{L}_2(\hat{\theta}_2)) &= -\mathbf{E}\left[\log\left(\frac{\sigma_1}{\sigma_2}\right) + \frac{1}{2}\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right)\sigma_0^2 \right] \qquad (2.14) \\
&= 0, \qquad (2.15)
\end{aligned}$$

the diffusion functions are said to be equivalent. When the models have different but equivalent diffusion functions in our criterion, the second order asymptotic component becomes important. The second order term is related to drift functions, and having a good drift function pays off only then. The relative size of the leading and second order terms depends on the sampling frequency. For high frequency data, the informative second order term can be too small compared to the leading order term which does not have any information on the superiority of a model when the diffusion functions are equivalent. We can easily miss a good drift specification in this situation.

This indicates that a failure to reject the null hypothesis may be from having a sampling frequency too high to consider the drift functions.

Another implication of the theorem is that the error from using the approximate transition densities rather than the true transition densities can not be ignored even asymptotically. When the models are equivalent in terms of the diffusion functions, the superiority of a model depends on both the drift specification and the approximation methods, because $A_j, \ldots, E_j$ depend on the choice of the approximate likelihoods. Since the choice of the approximate transition densities leads to a new model selection criterion, the null hypothesis must imply that the models are equivalent under the particular approximation method actually used for estimation.

In Case 2, since the competing models have the identical diffusion functions, the model selection is based on the drift functions only. This is because the pseudo-true values β_1^* and β_2^* depend on the diffusion functions only. The evaluation and comparison of the models crucially depend on the sampling time span. High frequency data would not help estimate the drift parameters nor evaluate their ranking. Therefore when the sampling span is short, it is quite challenging to distinguish these models. Moreover, the approximation methods for transition densities are always important since the functions that depend on the approximation methods appear in the primary selection criterion.

For both Case 1 and 2, our model selection criterion using approximate transition densities becomes equivalent to the one using the true transition densities if the approximations are sufficiently accurate. The following corollary gives a sufficient condition that an approximate transition density should satisfy to make our testing procedure with approximate transition densities asymptotically equivalent to the test using the true transition density.

Corollary 2.4 (Equivalence condition) *Let A_j^0, B_j^0, C_j^0, D_j^0 and E_j^0 be defined as in Definition 2.2 using the true transition density of $M_j(\theta_j)$. If*

$$\mathbf{E}\left[(A_j - A_j^0)\mu_0 + \frac{(B_j - B_j^0)\mu_0\sigma_0^2}{2} + \frac{(C_j - C_j^0)\sigma_0^4}{8} + \frac{D_j - D_j^0}{2} + \frac{(E_j - E_j^0)\sigma_0^2}{2}\right] = 0, \qquad (2.16)$$

then the approximate likelihood $\mathcal{L}_j(\theta_j)$ gives the asymptotically equivalent model selection criterion to the true likelihood $\mathcal{L}_j^0(\theta_j)$, which implies

$$\frac{1}{T}\big(\mathcal{L}_j(\hat{\theta}_j) - \mathcal{L}_j^0(\hat{\theta}_j^0)\big) \xrightarrow{p} 0$$

as $T \to \infty$ and $\Delta \to 0$ for both Case 1 and Case 2, where $\hat{\theta}_j$ and $\hat{\theta}_j^0$ are the ML esitmators from $\mathcal{L}_j$ and $\mathcal{L}_j^0$, respectively.

For example, the Euler or the Hermite expansion approximations define the same model selection criterion as the true transition density for the Ornstein-Uhlenbeck process (see Appendix for the proof).

The following theorem gives the asymptotic distribution under the null hypothesis H_0 that competing models are equivalent.

Theorem 2.5 (Asymptotic null distribution) *(a) Case 1:* $\sigma_1(\,\cdot\,;\beta_1^*) \neq \sigma_2(\,\cdot\,;\beta_2^*)$.
If $\Delta^2 T \to \infty,\ \Delta^3 T \to 0$ *as* $T \to \infty$, $\Delta \to 0$, *and Assumptions B.1-B.4 are satisfied, then under the null* H_0 *in (2.8),*

$$\frac{\Delta}{\sqrt{T}}\big(\mathcal{L}_1(\hat{\theta}_1) - \mathcal{L}_2(\hat{\theta}_2)\big) \xrightarrow{d} \mathbf{N}\big(0, \mathbf{E}\mathcal{G}_1^2\big),$$

where

$$\mathcal{G}_1(x) = \sigma_0(x)s_0(x)\int_{x_0}^{x}\left[\log\left(\frac{\sigma_1(v)}{\sigma_2(v)}\right) + \frac{1}{2}\bigg(\frac{1}{\sigma_1^2(v)} - \frac{1}{\sigma_2^2(v)}\bigg)\sigma_0^2(v)\right]m_0(v)dv,$$

and the derivative of scale function $s_0(x)$ *and the speed density* $m_0(x)$ *are given as follows:*

$$\begin{aligned} s_0(x) &= \exp\bigg(-\int_{x_0}^{x}\frac{2\mu_0(v)}{\sigma_0^2(v)}\,dv\bigg), \\ m_0(x) &= \frac{1}{\sigma_0^2(x)s_0(x)} \end{aligned} \tag{2.17}$$

for $x_0 \in \mathcal{D}_X$.
(b) Case 2: $\sigma_1(\,\cdot\,;\beta_1^*) = \sigma_2(\,\cdot\,;\beta_2^*)$ *and* $\mu_1(\,\cdot\,;\alpha_1^*) \neq \mu_2(\,\cdot\,;\alpha_2^*)$.
If $\Delta T^2 \to 0$ *as* $T \to \infty$, $\Delta \to 0$, *and Assumptions B.1-B.4 are satisfied, then under the null* H_0 *in (2.8),*

$$\frac{1}{\sqrt{T}}\big(\mathcal{L}_1(\hat{\theta}_1) - \mathcal{L}_2(\hat{\theta}_2)\big) \xrightarrow{d} \mathbf{N}\big(0, C'\Sigma C\big),$$

where

$$\begin{aligned} \Sigma &= \mathbf{E}\big[\big(\mathcal{G}_a, \mathcal{G}_b'\big)'\big(\mathcal{G}_a, \mathcal{G}_b'\big)\big], \\ C &= \Bigg(1,\, \mathbf{E}\bigg[(A_{1\beta} - A_{2\beta})\mu_0 + \frac{D_{1\beta} - D_{2\beta}}{2} + \frac{(E_{1\beta} - E_{2\beta})\sigma_0^2}{2}\bigg]' \times \\ &\qquad \bigg[\mathbf{E}\frac{(\sigma_1\sigma_{1\beta\beta'} - 3\sigma_{1\beta}\sigma_{1\beta}')(\sigma_0^2 - \sigma_1^2) - 2\sigma_1^2\sigma_{1\beta}\sigma_{1\beta}'}{\sigma_1^4}\bigg]^{-1}\Bigg)', \end{aligned}$$

with

$$\mathcal{G}_a(x) = \sigma_0(x)s_0(x)\int_{x_0}^{x}\Big[\big(A_1(v)-A_2(v)\big)\mu_0(v) + \frac{D_1(v)-D_2(v)}{2} + \frac{\big(E_1(v)-E_2(v)\big)\sigma_0^2(v)}{2}\Big]m_0(v)dv$$
$$- \big(A_1(x)-A_2(x)\big)\sigma_0(x),$$
$$\mathcal{G}_b(x) = \sigma_0(x)s_0(x)\int_{x_0}^{x}\frac{\sigma_{1\beta}(v)\big(\sigma_0^2(v)-\sigma_1^2(v)\big)}{\sigma_1^3(v)}m_0(v)dv$$

for $x_0 \in \mathcal{D}_X$ *in (2.17).*

We also derive the asymptotic local power curves.

Theorem 2.6 (Local asymptotic power) *(a) Case 1:* $\sigma_1(\,\cdot\,;\beta_1^*) \neq \sigma_2(\,\cdot\,;\beta_2^*)$.
Let $\Delta^2 T \to \infty,\ \Delta^3 T \to 0$ *as* $T\to\infty$, $\Delta\to 0$, *and Assumptions B.1-B.4 are satisfied. Let the local alternative be*

$$\mathbf{E}\Big[\log\Big(\frac{\sigma_1}{\sigma_2}\Big) + \frac{1}{2}\Big(\frac{1}{\sigma_1^2}-\frac{1}{\sigma_2^2}\Big)\sigma_0^2\Big] = -\frac{\delta_1}{\sqrt{T}} \qquad \textit{and}$$
$$\mathbf{E}\Big[(A_1-A_2)\mu_0 + \frac{(B_1-B_2)\mu_0\sigma_0^2}{2} + \frac{(C_1-C_2)\sigma_0^4}{8} + \frac{D_1-D_2}{2} + \frac{(E_1-E_2)\sigma_0^2}{2} - \frac{1}{2}\Big(\frac{1}{\sigma_1^2}-\frac{1}{\sigma_2^2}\Big)\mu_0^2\Big] = \frac{\delta_2}{\Delta\sqrt{T}}$$

for some $\delta_1, \delta_2 \in \mathbb{R}$. *Then we have*

$$\frac{\Delta}{\sqrt{T}}\big(\mathcal{L}_1(\hat\theta_1) - \mathcal{L}_2(\hat\theta_2)\big) \xrightarrow{d} \mathbf{N}\big(\delta, \mathbf{E}\mathcal{G}_1^2\big),$$

where $\delta = \delta_1+\delta_2$ *and* $\mathcal{G}_1$ *is defined in Theorem 2.5. The asymptotic power function* $\mathcal{P}_\alpha(\delta)$ *for the two-sided test with level* α *is given by*

$$\mathcal{P}_\alpha(\delta) = 2 - \Phi\Big(z_{\alpha/2} - \frac{\delta}{\sqrt{\mathbf{E}\mathcal{G}_1^2}}\Big) - \Phi\Big(z_{\alpha/2} + \frac{\delta}{\sqrt{\mathbf{E}\mathcal{G}_1^2}}\Big),$$

where Φ *and* $z_{\alpha/2}$ *are the cdf and the* $(1-\alpha/2)$ *quantile of the standard normal distribution respectively.*
(b) Case 2: $\sigma_1(\,\cdot\,;\beta_1^*) = \sigma_2(\,\cdot\,;\beta_2^*)$ *and* $\mu_1(\,\cdot\,;\alpha_1^*) \neq \mu_2(\,\cdot\,;\alpha_2^*)$.
Let $\Delta T^2 \to 0$ *as* $T\to\infty$, $\Delta\to 0$, *and Assumptions B.1-B.4 are satisfied. Let the local alternative be*

$$\mathbf{E}\Big[(A_1-A_2)\mu_0 + \frac{D_1-D_2}{2} + \frac{(E_1-E_2)\sigma_0^2}{2}\Big] = \frac{\delta}{\sqrt{T}}.$$

Then we have

$$\frac{1}{\sqrt{T}}\big(\mathcal{L}_1(\hat\theta_1) - \mathcal{L}_2(\hat\theta_2)\big) \xrightarrow{d} \mathbf{N}\big(\delta, C'\Sigma C\big),$$

where C and Σ are defined in Theorem 2.5, and the asymptotic power function $\mathcal{P}_\alpha(\delta)$ for the two-sided test with level α is

$$\mathcal{P}_\alpha(\delta) = 2 - \Phi\left(z_{\alpha/2} - \frac{\delta}{\sqrt{C'\Sigma C}}\right) - \Phi\left(z_{\alpha/2} + \frac{\delta}{\sqrt{C'\Sigma C}}\right).$$

In Case 1, as discussed earlier, when the diffusion functions are equivalent, we must compare drift functions. The information on drift functions accumulates only if we sample data longer, and the test has low, or potentially no power. A large number of observations would not help in this case. To have power against a superior drift function, we must require $\Delta\sqrt{T} \to \infty$. The long sampling span T ensures the signal from the second order term dominates the leading noise term asymptotically.

In practice, we would reject the null easily if high frequency data are available with a moderate sampling span, since diffusion functions of models are not generally equivalent. Even if the highest frequency data available can not reject the null, we can hope that the models may be ranked by the drift functions using an infrequent sampling. But the infrequent sampling usually reduces the number of observations and can deteriorate the accuracy of approximate transition densities. We also suggest a direct comparison using an alternative selection criterion, which involves drift functions only, may be preferred to the log-likelihood ratios in this condition.

In Case 2, only T matters. The likelihood ratio does not carry much information to distinguish models when T is short, even if the number of observations is large. Also since diffusion functions are not useful, estimating their parameters only adds sampling errors in the criterion. Therefore we may try to compare the drift functions directly to improve the power of the test.

2.3 Test Statistics and Sampling Distributions

We consider both asymptotically pivotal and non-pivotal test statistics for comparison, and use different asymptotic approximations to their sampling distributions.

- Log-likelihood ratio (non-pivotal): we define the test statistic

$$\tau_{T,\Delta} = \frac{T^{1/2}}{\kappa(T,\Delta)}(\mathcal{L}_1(\hat\theta_1) - \mathcal{L}_2(\hat\theta_2)), \tag{2.18}$$

  where $\hat\theta_j$ $(j = 1, 2)$ are the ML estimators, and $\kappa(T,\Delta)$ is the scaling factor. As shown in the previous section, the proper scaling factor depends on whether the competing models have the same diffusion functions or not. We have $\kappa(T,\Delta) = T/\Delta = n$ for Case 1, and $\kappa(T,\Delta) = T$ for Case 2.

- Log-likelihood ratio normalized by the HAC variance estimator: let

$$\hat{u}_i = \log \frac{p_1(\Delta, X_{(i-1)\Delta}, X_{i\Delta}; \hat{\theta}_1)}{p_2(\Delta, X_{(i-1)\Delta}, X_{i\Delta}; \hat{\theta}_2)}, \tag{2.19}$$

and

$$\bar{u} = n^{-1} \sum_{i=1}^{n} \hat{u}_i. \tag{2.20}$$

Then

$$\mathcal{L}_1(\hat{\theta}_1) - \mathcal{L}_2(\hat{\theta}_2) = \sum_{i=1}^{n} \hat{u}_i, \tag{2.21}$$

and the test statistic t_n for both Case 1 and Case 2 is given by

$$t_n = \frac{n^{-1/2}(\mathcal{L}_1(\hat{\theta}_1) - \mathcal{L}_2(\hat{\theta}_2))}{s_{HAC}}, \tag{2.22}$$

where

$$s_{HAC}^2 = \sum_{l=1-n}^{n-1} k\left(\frac{l\Delta}{M}\right) \hat{\gamma}(l\Delta) \tag{2.23}$$

$k(x)$ is a kernel function, $0 < M \leq T$ is a bandwidth parameter such that $M/T \to b \in (0,1]$ as $T \to \infty$, and

$$\hat{\gamma}(l\Delta) = \frac{1}{n} \sum_{i=|l|+1}^{n} (\hat{u}_i - \bar{u})(\hat{u}_{i-|l|} - \bar{u}). \tag{2.24}$$

We are using the fixed-b approach of Kiefer and Vogelsang (2005), and the limiting distribution of t_n is derived under a high level assumption given below.

The normalized statistic t_n becomes the same form for both Case 1 and Case 2. Although the form also coincides with the test statistics of Choi and Kiefer (2008), which is developed with a fixed sampling interval, the implications are different; the pseudo-true values, the order of $\mathcal{L}_1(\theta_1^*) - \mathcal{L}_2(\theta_2^*)$, and the long-run variance of the numerator of the test statistics are different from the case with a fixed sampling interval, and they also depend on whether the test is in Case 1 or Case 2 in our framework. The normalized test statistics have the advantage of the convenience of having the same form regardless of Case 1, Case 2, or the case of a fixed sampling interval. The fundamental difference of our approach from the conventional methods is in the notion of the equivalence of models. Our approach shows the equivalence should be understood differently when sampling frequencies are high.

Definition 2.7 (Kiefer and Vogelsang (2005), p. 1141) *Let $\widetilde{W}_i(r)$ be an i-dimensional Brownian bridge $\widetilde{W}_i(r) = W_i(r) - rW_i(1)$, and $W_i(r)$ an i-dimensional standard Brownian motion. With a kernel function $k(x)$, let the $(i \times i)$ random matrix $\mathcal{Q}_i(b)$ be defined as follows.*
(a) if $k(x)$ is twice continuously differentiable everywhere,

$$\mathcal{Q}_i(b) = -\int_0^1 \int_0^1 \frac{1}{b^2} k''\left(\frac{r-s}{b}\right) \widetilde{W}_i(r)\widetilde{W}_i(s)' drds, \tag{2.25}$$

(b) if $k(x)$ is continuous, $k(x) = 0$ for $|x| \geq 1$, and $k(x)$ is twice continuously differentiable everywhere except for $|x| = 1$,

$$\mathcal{Q}_i(b) = -\int\int_{|r-s|<b} \frac{1}{b^2} k''\left(\frac{r-s}{b}\right) \widetilde{W}_i(r)\widetilde{W}_i(s)' drds \tag{2.26}$$

$$+ \frac{k'_-(1)}{b} \int_0^{1-b} \left(\widetilde{W}_i(r+b)\widetilde{W}_i(r)' + \widetilde{W}_i(r)\widetilde{W}_i(r+b)'\right) dr, \tag{2.27}$$

where $k'_-(1) = \lim_{h\to 0}[(k(1) - k(1-h))/h]$, i.e., $k'_-(1)$ is the derivative of $k(x)$ from the left at $x = 1$,
(c) if $k(x)$ is the Bartlett kernel,

$$\mathcal{Q}_i(b) = \frac{2}{b} \int_0^1 \widetilde{W}_i(r)\widetilde{W}_i(r)' dr \tag{2.28}$$

$$- \frac{1}{b} \int_0^{1-b} \left(\widetilde{W}_i(r+b)\widetilde{W}_i(r)' + \widetilde{W}_i(r)\widetilde{W}_i(r+b)'\right) dr. \tag{2.29}$$

Assumption 2.8 *Let*

$$u_i = \log \frac{p_1(\Delta, X_{(i-1)\Delta}, X_{i\Delta}; \theta_1^*)}{p_2(\Delta, X_{(i-1)\Delta}, X_{i\Delta}; \theta_2^*)}, \tag{2.30}$$

and define the partial sum process

$$g_{[rn]} = \frac{T^{1/2}}{\kappa(T, \Delta)} \sum_{i=1}^{[rn]} u_i, \tag{2.31}$$

where $\kappa(T, \Delta) = T/\Delta = n$ for Case 1, and $\kappa(T, \Delta) = T$ for Case 2.
(a) The partial sum satisfies the functional central limit theorem, i.e.

$$g_{[rn]} \Rightarrow \omega W(r), \tag{2.32}$$

where the long-run variances $\omega^2 = \mathbf{E}\mathcal{G}_1^2$ for Case 1 and $\omega^2 = C'\Sigma C$ for Case 2 (see Theorem 2.5 for their definitions), and $W(r)$ is the standard Brownian motion defined on $C[0, 1]$, and

(b)

$$\frac{T^{1/2}}{\kappa(T,\Delta)} \sum_{i=1}^{[rn]} (\hat{u}_i - u_i) \xrightarrow{p} 0, \tag{2.33}$$

uniformly in $r \in [0,1]$ *as* $T \to \infty,\ \Delta \to 0$.

Theorem 2.9 *Let* $M/T \to b \in (0,1]$. *If the assumptions in Theorem 2.5 and Assumption 2.8 are satisfied, then we have, as* $T \to \infty,\ \Delta \to 0$,
(a) for Case 1,

$$t_n = \frac{\frac{\Delta}{\sqrt{T}}(\mathcal{L}_1(\hat{\theta}_1) - \mathcal{L}_2(\hat{\theta}_2))}{\sqrt{\Delta \sum_{l=1-n}^{n-1} k\left(\frac{l\Delta}{M}\right) \hat{\gamma}(l\Delta)}} \tag{2.34}$$

$$= \frac{n^{-1/2}(\mathcal{L}_1(\hat{\theta}_1) - \mathcal{L}_2(\hat{\theta}_2))}{s_{HAC}} \Rightarrow \frac{W(1)}{\sqrt{\mathcal{Q}_1(b)}}, \tag{2.35}$$

and
(b) for Case 2,

$$t_n = \frac{\frac{1}{\sqrt{T}}(\mathcal{L}_1(\hat{\theta}_1) - \mathcal{L}_2(\hat{\theta}_2))}{\sqrt{\Delta^{-1} \sum_{l=1-n}^{n-1} k\left(\frac{l\Delta}{M}\right) \hat{\gamma}(l\Delta)}} \tag{2.36}$$

$$= \frac{n^{-1/2}(\mathcal{L}_1(\hat{\theta}_1) - \mathcal{L}_2(\hat{\theta}_2))}{s_{HAC}} \Rightarrow \frac{W(1)}{\sqrt{\mathcal{Q}_1(b)}}, \tag{2.37}$$

where the random function $\mathcal{Q}_1(b)$, *defined in Definition 2.7, is independent with* $W(1)$, *but depends on the kernel function* $k(x)$.

The critical values of the fixed-b asymptotic approximations must be obtained from simulations. Table 1 in Kiefer and Vogelsang (2005) tabulates them for popular kernel functions.

We compare the performance of the approximations to the sampling distributions of the above test statistics using the subsampling (Politis, Romano, and Wolf (1999)) and the nonparametric block bootstrap methods for $\tau_{T,\Delta}$ and t_n. For the asymptotically pivotal test t_n, we also consider the standard normal and the fixed-b approximations. In the all examples in this paper, we use the Bartlett kernel and try three different bandwidths for comparison. We explain the subsampling and the block bootstrap methods in detail.

2.3.1 Subsampling method

We calculate the subsample test statistics $\tau^s_{T,\Delta}, t^s_n$ $(s = 1, \ldots, n_s)$ of the same form as $\tau_{T,\Delta}, t_n$ from n_s blocks of subsamples of size S. Let $\hat{C}_n$ and $\hat{V}_n$ be the mean and the variance of subsample statistics (i.e. $\hat{C}_n = n_s^{-1}\sum_{s=1}^{n_s}\tau_{T,\Delta}^s$ and $\hat{V}_n = (n_s-1)^{-1}\sum_{s=1}^{n_s}(\tau_{T,\Delta}^s - \hat{C}_n)^2$ for $\tau_{T,\Delta}$). We use two different asymptotic approximations to the sampling distributions of $\tau_{T,\Delta}, t_n$ to get critical values.

1. Use $\mathbf{N}(\hat{C}_n, \hat{V}_n)$.

2. Use the empirical distributions of $\tau_{T,\Delta}^s, t_n^s$.

2.3.2 Block bootstrap method

We use the following algorithm.

1. Randomly select n_B blocks of consecutive observations of the length $l = n/n_B$ to construct the bootstrap samples $\{X_{i\Delta}^*\}_{i=1}^n$.

2. Calculate the bootstrap estimators $\tilde{\theta}_1, \tilde{\theta}_2$, the log-likelihood ratio $(\mathcal{L}_1^*(\tilde{\theta}_1) - \mathcal{L}_2^*(\tilde{\theta}_2))$, the statistic $\tau_{T,\Delta}^*$, and s_{HAC}^{2*} from $\{X_{i\Delta}^*\}_{i=1}^n$.

3. Calculate the bootstrap test statistics,

$$\tau_{T,\Delta}^b = \tau_{T,\Delta}^* - \tau_{T,\Delta}, \tag{2.38}$$

$$t_n^b = \frac{n^{-1/2}((\mathcal{L}_1^*(\tilde{\theta}_1) - \mathcal{L}_2^*(\tilde{\theta}_2)) - (\mathcal{L}_1(\hat{\theta}_1) - \mathcal{L}_2(\hat{\theta}_2))}{s_{HAC}^*}. \tag{2.39}$$

4. Repeat $1 \sim 3$ for $b = 1, \ldots, B$.

5. Instead of using $\tau_{T,\Delta}^b$ and t_n^b directly to calculate critical values, we use the empirical distributions of $\tau_{T,\Delta}^b - B^{-1}\sum_{b=1}^B \tau_{T,\Delta}^b$, and $t_n^b - B^{-1}\sum_{b=1}^B t_n^b$. The centering of the bootstrap statistics by their sample means is asymptotically negligible, but we found it help get more accurate sizes for our simulation examples. But we believe using the usual empirical distribution is also a reasonable choice. We use the equal-tailed percentile or percentile-t bootstrap method.

3 Finite Sample Properties

3.1 Four Examples

We consider 4 examples shown in Table 1 for Case 1 and 2. For all examples, competing models M_1 and M_2 are nonnested and misspecified, and the true processes M_0 are chosen to make M_1 and M_2 equivalent to satisfy the null hypothesis when the models use the Milstein approximate transition densities. Consequently, we must use the Milstein approximation for estimation. The models in Case 1 have different diffusion functions. For Case 2, the models have identical diffusion functions.

Table 1 is about here.

3.2 Size of the Tests

Table 2 shows the sizes of our test statistics (two-sided, 5% level).

Table 2 is about here.

For each example, we generated $T = 5$ and $T = 40$ years of data sampled at the daily frequency ($\Delta = 1/250$). We calculate the actual sizes of the tests from $1,000$ simulation iterations, and compare the performance of $\tau_{T,\Delta}$ and t_n.The statistics t_n uses the Bartlett kernel with three different bandwidths. The fixed-b approach does not provide an optimal choice of bandwidths. Although any bandwidth parameter should work as long as $M/T \to b$ is fixed, we choose the optimal bandwidth ($M = cT^{1/3}$) of Andrews (1991) as an example. Note that the bandwidth parameter does not have the optimality property under the fixed-b approach. We choose $b = M/T = 0.0855$ for $T = 5$ and $b = 0.0213$ for $T = 40$ by setting $c = 0.25$ in Andrews's method. Then we try doubling ($b = 0.171$ for $T = 5$, and $b = 0.0427$ for $T = 40$) and tripling ($b = 0.256$ for $T = 5$, and $b = 0.0641$ for $T = 40$) the bandwidth for comparison. Let $t_n(M)$ be the statistics t_n with a bandwidth parameter M (years). The chosen bandwidths for $T = 5$ are $M = 0.43$, 0.85, 1.28 years, and for $T = 40$, $M = 0.85$, 1.7, 2.56 years.

We approximate the sampling distributions of the test statistics with the four methods described earlier. Specifically, "Sub N" implies the subsampling approximation by the fitted normal distribution with the sample mean and variance of the subsample statistics, and "Sub Emp" means that we have used the empirical distribution of the subsample statistics directly.

The subsample method uses $n_s = 199$ sub-blocks of consecutive observations of the size equal to $S = T^{0.4} = 1.90$ years (38%) for $T = 5$ and $S = T^{0.7} = 13.23$ years (33%) for $T = 40$. The block bootstrap method is based on $B = 399$ bootstrap repetitions, and each repetition uses the bootstrap samples constructed from $n_B = 25$ blocks of equal size (or a block length equal to 1/25 of the total number of observations). The choice of the sub-block size for both the bootstrap and the subsampling can affect the results. In unreported experiments, we found that the effect of choosing a different sub-block size is larger for the subsampling method than for the bootstrap method. Although the effect is mild for the bootstrap method, it is not negligible. The optimal choice of a block size is a difficult problem and we do not consider it here. See, for example, Politis, Romano, and Wolf (1999) and Lahiri (2003) for further discussion on this topic.

Overall, the pivotal tests t_n perform better than $\tau_{T,\Delta}$. The test statistic $\tau_{T,\Delta}$ shows serious over-rejections in Case 1, and under-rejections in Case 2. In Case 1, the block bootstrap seems to work better than or similar to the subsampling. The bootstrap method is better for the non-pivotal statistic $\tau_{T,\Delta}$. The results from the standard normal approximations can be good (Example 1, $T = 40$) or bad (Example 1,2, $T = 5$ and Example 2, $T = 40$), they depend on the HAC bandwidths closely. For Case 2, the block bootstrap is better (especially in Example 4) than the subsampling method for $T = 5$, but the subsampling is comparable to the bootstrap for $T = 40$.

The standard normal approximation shows good performance when the HAC bandwidth is chosen properly, but performs poorly in general. The fixed-b approach performs well in Example 1 and 3. It is better when the bandwidth parameter is large. The bootstrap method seems to be similar to the fixed-b method in Example 1 and 3. In Example 2 and 4, the bootstrap is better than the fixed-b approach except for $T = 5$ in Example 2.

The result can improve as the sampling period is longer (Example 1 and 3), but it may not improve much (Example 2 and 4). This shows that we may need a very long period of data to get reliable approximations to the sampling distributions of the test statistics.

In summary, approximating the sampling distributions of the test statistics can be difficult. The asymptotically pivotal tests is more reliable than the non-pivotal test. The block bootstrap looks like a better choice for short sampling periods, and the subsampling improves as the sampling period becomes longer. The standard normal and the fixed-b approximations can be a better or worse choice than the bootstrap or the subsampling depending on the examples. The fixed-b approximation is better than the standard normal approximation.

3.3 Power of the Tests

We calculate size-corrected power using the modified true processes given in Table 3.

Table 3 is about here.

Size corrected critical values (equal tailed, 5% level) are from the 2.5% and 97.5% quantiles of $5,000$ repetitions of simulations under the null hypothesis. Table 4 shows the power of the tests from $5,000$ simulation iterations.

Table 4 is about here.

In most examples, the pivotal tests t_n have better power than the non-pivotal test $\tau_{T,\Delta}$. In Case 1, the models are distinguished by the diffusion functions and the power improves significantly when sampling periods increase from $T = 5$ to $T = 40$, and the number of observations is from $1,250$ to $10,000$. In Case 2, the models are compared by the drift functions only, the tests have low power when the sampling period is short ($T = 5$). Especially, Example 3 has little power with $T = 5$ years of data even if M_1 is the correct model. For $T = 40$ ($n = 10,000$), the power improves. It shows how difficult to distinguish models in Case 2, and the number of observations may not be a good indicator of the power of the test. Diffusion models with the same diffusion functions are essentially very similar.

4 Applications

4.1 Spot Interest Rates

Many popular spot rate models such as Vasicek (Vasicek (1977)), CIR (Cox, Ingersoll Jr, and Ross (1985)), DK (Duffie and Kan (1996)), CKLS (Chan, Karolyi, Longstaff, and Sanders (1992)), and AG (Ahn and Gao (1999)) models perform unsatisfactorily in practice. They are shown to be rejected in specification tests (see, for example, Aït-Sahalia (1996) and Hong and Li (2005)). Especially, Hong and Li (2005) found that all the models they considered indicate serious misspecification. Our approach provides a formal statistical testing for possible superiority among them.

Since the spot interest rates are not observed in practice, they are usually approximated with securities with short maturities. We use the annualized 1-month Eurodollar deposit (London) bid rates from 1/1/1971 to 12/31/2007 (source: Federal Reserve Statistical Release, `http://www.federalreserve.gov/releases/h15/data/Business_day/H15_ED_M1.txt`). We do not consider the "weekend" effects; Fridays are followed by Mondays. The sampling interval Δ is approximately set to be $T/n = 0.00393$, where $T = 37$ years is the sampling period, and $n = 9,411$ is the total number of observations.

We could have used the daily over-night federal funds rates or the 1-week Eurodollar rates, but securities with very short maturities tend to have the market microstructure effects, and need to be filtered (see Hamilton (1996) for a filtering and Aït-Sahalia (1996) for a discussion). Also, Duffee (1996) advocates the use of Eurodollar rates because of the idiosyncratic variations of the Treasury bill yields.

Figure 2 shows the sample path of the annualized daily 1-month Eurodollar deposit rates. The mean and the standard deviation are 0.06798 and 0.03556 respectively.

FIGURE 2 IS ABOUT HERE.

The candidate models are the AG (Ahn and Gao) model (Inverse Feller process)

$$M_1 : dr_t = \kappa(\mu - r_t)r_t dt + \sigma r_t^{3/2}\ dW_t, \tag{4.1}$$

and the CIR (Feller's square root process) model

$$M_2 : dr_t = \kappa(\mu - r_t)dt + \sigma\sqrt{r_t}\ dW_t. \tag{4.2}$$

We choose these models since they performed similarly in the misspecification test of Hong and Li (2005). The estimated models from the MLE with the Milstein approximate transition densities are

$$M_1 : dr_t = 3.365\ (0.0947 - r_t) r_t dt + 1.9246\ r_t^{3/2}\ dW_t\ \text{(AG)}, \tag{4.3}$$

$$M_2 : dr_t = 0.673\ (0.0671 - r_t) dt + 0.1495 \sqrt{r_t}\ dW_t\ \text{(CIR)}, \tag{4.4}$$

with estimated log-likelihoods $l_1 = 46,630.07$ and $l_2 = 43,950.82$ respectively.

Table 5 shows the results of the model selection tests. As in the Monte Carlo study in the previous section, we used four test statistics $\tau_{T,\Delta}$, $t_n(0.83)$, $t_n(1.67)$, and $t_n(2.50)$. The statistic $\tau_{T,\Delta}$ is the non-pivotal test based on the log-likelihood ratio, and $t_n(M)$ is using the HAC variance estimator with a bandwidth parameter M. The sampling distributions of the test statistics are approximated by the subsampling, the block bootstrap, and the standard normal distribution as described in the previous section. For the subsampling, we used 199 blocks of the subsamples of size $T^{0.7} = 12.5$ years, and for the block bootstrap, we used 25 blocks (1.48 years for each block) and 799 bootstrap repetitions to get the bootstrap critical values. "AG" or "CIR" represents the superiority of the respective model; "0" means failing to reject.

Table 5 is about here.

We have some evidence that AG model is better when the sampling distributions of test statistics are approximated by the standard normal distribution or the fixed-b approach, but they are not rejected when the subsampling or the bootstrap method is used with the asymptotically pivotal statistics. We note that $\tau_{T,\Delta}$ rejects except for the bootstrap method, but due to the unstable size performance from our simulation study, we weigh the results from the HAC statistics more. Overall, the information from the data is not enough to tell the superiority of the AG model.

4.2 Exchange Rates (Euro-US Dollar)

We use the daily (measured at 2:15 pm (C.E.T.), business days only) spot Euro-US dollar exchange rates from 1999.1 $\sim$ 2008.12 ($T = 10$ years, $n = 2,560$, $\Delta = 0.003906$) from European Central Bank's (ECB) Statistical Data Warehouse.

Figure 3 shows the sample path of the exchange rates. The mean and the standard deviation are 1.155 and 0.1934 respectively.

Figure 3 is about here.

We choose the AG and CIR models again to compare. The estimated models using the MLE with the Milstein approximate transition densities are

$$M_1 : dX_t = 0.0727\ (1.475 - X_t)X_t dt + 0.09925\ X_t^{3/2} dW_t \text{ (AG)} \tag{4.5}$$

$$M_2 : dX_t = 0.108\ (1.352 - X_t)dt + 0.1107\sqrt{X_t} dW_t \text{ (CIR)} \tag{4.6}$$

We used the same setup for the subsample and the block bootstrap as for the spot rate model selection (5.01 years of the sub-block size, 0.4 years for the bootstrap block size). The HAC tests are $t_n(0.54)$, $t_n(1.08)$, $t_n(1.62)$. The results are in Table 5. Although the CIR model has a higher likelihood value, we do not find significant evidence of the superiority compared to the AG model.

5 Conclusion

We proposed a likelihood ratio based model selection test for high frequency data from a diffusion process. Generally, the model selection test can rank two competing models by the proximity of the diffusion functions of the models to the true diffusion functions. Our results show that more observations may help to select a better model, but what we really need is a long enough sampling period to evaluate and compare the models. This insight becomes critical when the candidate models have the same diffusion functions (i.e. Case 2 in our paper); the number of observations does not matter, and only the sampling period is useful.

Our Monte Carlo study shows that the accurate approximation to the sampling distributions of the test statistics is challenging. We advise to use pivotal statistics with resampling methods such as the block bootstrap or the subsampling method.

The applications of our tests to the spot rate and exchange rate models show that Ahn and Gao's model and the CIR model are difficult to distinguish, although Ahn and Gao's model performs slightly better for the spot rates and the CIR model is better for the exchange rates.

Our new asymptotics gives new insights for high frequency data sampled over a reasonably long horizon. Although the new asymptotics has clear advantages, the standard procedures developed under a fixed sampling interval or a fixed sampling period must be carefully studied again. For example, the asymptotic properties of the block bootstrap methods need to be investigated under our new asymptotics.

We conclude with future research topics related to this paper. We are working on the extension of our results to nonstationary processes and multivariate diffusions. Stationary multivariate diffusions would have similar results to the current paper, but studying nonstationary processes is expected to be difficult because the asymptotic behavior of estimators and test statistics largely depends on true processes.

References

Ahn, D.-H., and B. Gao (1999): "A Parametric Nonlinear Model of Term Structure Dynamics," *The Review of Financial Studies*, 12(4), 721–762.

Aït-Sahalia, Y. (1996): "Testing Continuous-Time Models of the Spot Interest Rate," *Review of Financial Studies*, 9(2), 385–426.

Aït-Sahalia, Y. (2002): "Maximum Likelihood Estimation of Discretely Sampled Diffusions: A Closed-Form Approximation Approach," *Econometrica*, 70(1), 223–262.

Aït-Sahalia, Y., and R. Kimmel (2008): "Estimating Affine Multifactor Term Structure Models Using Closed-Form Likelihood Expansions," *Working Paper, Princeton University and Ohio State University.*

Andrews, D. (1991): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59(3), 817–858.

Berman, S. (1964): "Limiting distribution of the maximum of a diffusion process," *The Annals of Mathematical Statistics*, 35(1), 319–329.

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(2), 161–210.

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

Chen, R., and L. Scott (1993): "Maximum likelihood estimation for a multifactor equilibrium model of the term structure of interest rates," *The Journal of Fixed Income*, 3(3), 14–31.

Choi, H.-s., and N. M. Kiefer (2008): "Improving robust model selection test for dynamic models with an application to comparing predictive accuracy," *Working paper, Texas A&M University and Cornell University.*

Cline, D. B., and M. Jeong (2009): "Limit Theorems for the Integral of Diffusion Processes," *Working Paper, Texas A&M University.*

Cox, J., J. Ingersoll Jr, and S. Ross (1985): "A Theory of the Term Structure of Interest Rates.," *Econometrica*, 53(2), 385–407.

Csiszár, I. (1967a): "Information Type Measures of Difference of Probability Distributions and Indirect Observations," *Studia Scientiarum Mathematicarum Hungarica*, 2, 299–318.

——— (1967b): "On Topological Properties of f-Divergence," *Studia Scientiarum Mathematicarum Hungarica*, 2, 329–339.

——— (1975): "I-Divergence Geometry of Probability Distributions and Minimization Problems," *The Annals of Probability*, 3, 146–158.

Davidson, R., and J. G. MacKinnon (1981): "Several tests for model specification in the presence of alternative hypotheses," *Econometrica*, 49, 781–793.

Davis, R. (1982): "Maximum and minimum of one-dimensional diffusions," *Stoch. Process. Appl*, 13, 1–9.

Duffee, G. (1996): "Idiosyncratic variation of Treasury bill yields," *Journal of Finance*, 51(2), 527–551.

Duffie, D., and R. Kan (1996): "A yield-factor model of interest rates," *Mathematical Finance*, 6(4), 379–406.

Hamilton, J. (1996): "The Daily Market for Federal Funds," *The Journal of Political Economy*, 104(1), 26–56.

Hong, Y., and H. Li (2005): "Nonparametric Specification Testing for Continuous-Time Models with Applications to Term Structure of Interest Rates," *Review of Financial Studies*, 18(1), 37–84.

Jeong, M., and J. Y. Park (2009): "Asymptotics for the maximum likelihood estimators of diffusion models," *Working Paper, Indiana University.*

Karatzas, I., and S. Shreve (1991): *Brownian Motion and Stochastic Calculus*. Springer, 2nd edn.

Kiefer, N. M., and T. J. Vogelsang (2002): "Heteroskedasticity-Autocorrelation Robust Standard Errors Using the Bartlett Kernel Without Truncation," *Econometrica*, 70, 2093–2095.

Kiefer, N. M., and T. J. Vogelsang (2005): "A New Asymptotic Theory for Heteroskedasticity-Autocorrelation Robust Tests," *Econometric Theory*, 21, 1130–1164.

Kullback, S., and R. Leibler (1951): "On information and sufficiency," *Annals of Mathematical Statistics*, 22(1), 79–86.

Lahiri, S. (2003): *Resampling methods for dependent data*, Springer series in statistics. Springer.

Politis, D., J. Romano, and M. Wolf (1999): *Subsampling*. Springer Verlag.

Rivers, D., and Q. Vuong (2002): "Model selection tests for nonlinear dynamic models," *The Econometrics Journal*, 5(1), 1–39.

Stone, C. (1963): "Limit theorems for random walks, birth and death processes, and diffusion processes," *Illinois J. Math*, 7, 638–660.

van Zanten, H. (2000): "A multivariate central limit theorem for continuous local martingales," *Statistics and Probability Letters*, 50(3), 229–235.

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

Vuong, Q. (1989): "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses," *Econometrica*, 57(2), 307–333.

Appendices

A Examples

We give the explicit expansions in Theorem 2.3 for the Euler and Milstein approximations as well as Aït-Sahalia's Hermite expansion approximation in the following examples.

Example A.1 (Case 1) *(a) Euler approximation:*

$$\begin{aligned} A_j &= \frac{\mu_j}{\sigma_j^2}, \qquad D_j = -\frac{\mu_j^2}{\sigma_j^2}, \\ B_j &= 0, \qquad C_j = 0, \qquad E_j = 0 \end{aligned} \tag{A.1}$$

for $j = 1, 2$, thus

$$\begin{aligned} \mathcal{L}_1(\hat{\theta}_1) - \mathcal{L}_2(\hat{\theta}_2) = &-\frac{1}{\Delta}\int_0^T \left[\log\left(\frac{\sigma_1}{\sigma_2}\right) + \frac{1}{2}\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right)\sigma_0^2\right] dt \\ &- \frac{1}{2}\int_0^T \left(\frac{\big(\mu_0 - \mu_1\big)^2}{\sigma_1^2} - \frac{\big(\mu_0 - \mu_2\big)^2}{\sigma_2^2}\right) dt + O_p\left(\sqrt{\frac{T}{\Delta}}\right). \end{aligned}$$

(b) Milstein approximation:

$$\begin{aligned} A_j &= \frac{2\mu_j - 3\sigma_j\sigma_j^{\cdot}}{2\sigma_j^2}, \qquad B_j = \frac{3\sigma_j^{\cdot}}{\sigma_j^3}, \qquad C_j = -\frac{15\sigma_j^{\cdot 2}}{\sigma_j^4}, \\ D_j &= -\frac{4\mu_j^2 - 12\mu_j\sigma_j\sigma_j^{\cdot} + 5\sigma_j^2\sigma_j^{\cdot 2}}{4\sigma_j^2}, \qquad E_j = -\frac{\sigma_j^{\cdot}(6\mu_j - 7\sigma_j\sigma_j^{\cdot})}{2\sigma_j^3} \end{aligned}$$

for $j = 1, 2$, where $\sigma_j^{\cdot} = \sigma_j^{\cdot}(X_t, \beta_j^*)$ and $\sigma_j^{\cdot}(x, \beta_j^*) = \partial\sigma_j(x, \beta_j^*)/\partial x$, thus

$$\begin{aligned} &\mathcal{L}_1(\hat{\theta}_1) - \mathcal{L}_2(\hat{\theta}_2) \\ &\quad = -\frac{1}{\Delta}\int_0^T \left[\log\left(\frac{\sigma_1}{\sigma_2}\right) + \frac{1}{2}\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right)\sigma_0^2\right] dt \\ &\qquad - \int_0^T \left[\frac{1}{2}\left(\frac{\big(\mu_0 - \mu_1\big)^2}{\sigma_1^2} - \frac{\big(\mu_0 - \mu_2\big)^2}{\sigma_2^2}\right) - \frac{3}{2}\left(\frac{\sigma_1^{\cdot}(\sigma_1^2 - \sigma_0^2)(\mu_1 - \mu_0)}{\sigma_1^3} - \frac{\sigma_2^{\cdot}(\sigma_2^2 - \sigma_0^2)(\mu_2 - \mu_0)}{\sigma_2^3}\right)\right. \\ &\qquad\qquad \left. + \frac{5}{8}\left(\frac{\sigma_1^{\cdot 2}(\sigma_1^2 - \sigma_0^2)^2}{\sigma_1^4} - \frac{\sigma_2^{\cdot 2}(\sigma_2^2 - \sigma_0^2)^2}{\sigma_2^4}\right) - \frac{1}{2}\left(\frac{\sigma_1^{\cdot 2}}{\sigma_1^2} - \frac{\sigma_2^{\cdot 2}}{\sigma_2^2}\right)\sigma_0^2\right] dt + O_p\left(\sqrt{\frac{T}{\Delta}}\right). \end{aligned}$$

(c) Hermite expansion approximation (Aït-Sahalia (2002) with $K = 1$):

$$A_j = \frac{\mu_j}{\sigma_j^2} - \frac{3\dot{\sigma_j}}{2\sigma_j}, \qquad B_j = \frac{3\dot{\sigma_j}}{\sigma_j^3}, \qquad C_j = \frac{4\ddot{\sigma_j}}{\sigma_j^3} - \frac{11\dot{\sigma_j}^2}{\sigma_j^4},$$
$$D_j = -\dot{\mu_j} - \left(\frac{\mu_j}{\sigma_j} - \frac{\dot{\sigma_j}}{2}\right)^2 + \frac{\mu_j \dot{\sigma_j}}{\sigma_j} + \frac{\sigma_j \ddot{\sigma_j}}{2}, \qquad E_j = \frac{\dot{\mu_j}}{\sigma_j^2} - \frac{2\mu_j \dot{\sigma_j}}{\sigma_j^3} + \frac{3\dot{\sigma_j}^2}{2\sigma_j^2} - \frac{3\ddot{\sigma_j}}{2\sigma_j}$$

for $j = 1, 2$, where $\dot{\mu_j} = \dot{\mu_j}(X_t, \alpha_j^*)$ and $\dot{\mu_j}(x, \alpha_j^*) = \partial \mu_j(x, \alpha_j^*)/\partial x$, $\dot{\sigma_j} = \dot{\sigma_j}(X_t, \beta_j^*)$ and $\dot{\sigma_j}(x, \beta_j^*) = \partial \sigma_j(x, \beta_j^*)/\partial x$, and $\ddot{\sigma_j} = \ddot{\sigma_j}(X_t, \beta_j^*)$ and $\ddot{\sigma_j}(x, \beta_j^*) = \partial^2 \sigma_j(x, \beta_j^*)/\partial x^2$, thus

$$\begin{aligned}
&\mathcal{L}_1(\hat{\theta}_1) - \mathcal{L}_2(\hat{\theta}_2) \\
&\quad = -\frac{1}{\Delta}\int_0^T \left[\log\left(\frac{\sigma_1}{\sigma_2}\right) + \frac{1}{2}\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right)\sigma_0^2\right] dt \\
&\qquad - \int_0^T \left[\frac{1}{2}\left(\frac{\big(\mu_0 - \mu_1\big)^2}{\sigma_1^2} - \frac{\big(\mu_0 - \mu_2\big)^2}{\sigma_2^2}\right) + \frac{3}{2}\left(\frac{\dot{\sigma_1}}{\sigma_1} - \frac{\dot{\sigma_2}}{\sigma_2}\right)\mu_0 - \frac{3}{2}\left(\frac{\dot{\sigma_1}}{\sigma_1^3} - \frac{\dot{\sigma_2}}{\sigma_2^3}\right)\mu_0\sigma_0^2\right. \\
&\qquad\qquad - \left(\frac{\sigma_0^2}{\sigma_1^2} - 1\right)\left(\frac{\dot{\mu_1}}{2} - \frac{\mu_1\dot{\sigma_1}}{\sigma_1}\right) + \left(\frac{\sigma_0^2}{\sigma_2^2} - 1\right)\left(\frac{\dot{\mu_2}}{2} - \frac{\mu_2\dot{\sigma_2}}{\sigma_2}\right) - \frac{2}{3}\left(\frac{\dot{\sigma_1}^2}{\sigma_1^4} - \frac{\dot{\sigma_2}^2}{\sigma_2^4}\right)\sigma_0^4 \\
&\qquad\qquad + \frac{\dot{\sigma_1}^2}{8}\left(\frac{3\sigma_0^2}{\sigma_1^2} - 1\right)^2 - \frac{\dot{\sigma_2}^2}{8}\left(\frac{3\sigma_0^2}{\sigma_2^2} - 1\right)^2 - \frac{13}{48}\left(\frac{\sigma_1\ddot{\sigma_1}}{\sigma_1^4} - \frac{\sigma_2\ddot{\sigma_2}}{\sigma_2^4}\right)\sigma_0^4 \\
&\qquad\qquad \left. - \frac{\sigma_1\ddot{\sigma_1}}{2}\left(\frac{\sigma_0^2}{\sigma_1^2} - 2\right)^2 + \frac{\sigma_2\ddot{\sigma_2}}{2}\left(\frac{\sigma_0^2}{\sigma_2^2} - 2\right)^2\right] dt + O_p\left(\sqrt{\frac{T}{\Delta}}\right).
\end{aligned}$$

Example A.2 (Case 2) *(a) Euler approximation: $A_j, \ldots, E_j$ are in (A.1), and*

$$A_{j\beta} = -\frac{2\mu_j\sigma_{j\beta}}{\sigma_j^3}, \quad D_{j\beta} = \frac{2\mu_j^2\sigma_{j\beta}}{\sigma_j^3}, \quad E_{j\beta} = 0$$

for $j = 1, 2$, thus

$$\begin{aligned}
&\mathcal{L}_1(\hat{\theta}_1) - \mathcal{L}_2(\hat{\theta}_2) \\
&\quad = -\frac{1}{2}\int_0^T \frac{(\mu_0 - \mu_1)^2 - (\mu_0 - \mu_2)^2}{\sigma_1^2} dt + \int_0^T \frac{\sigma_0(\mu_1 - \mu_2)}{\sigma_1^2} dW_t \\
&\qquad + \int_0^T \frac{\sigma_{1\beta}'(\sigma_0^2 - \sigma_1^2)}{\sigma_1^3} dt \left(\int_0^T \frac{(\sigma_1\sigma_{1\beta\beta'} - 3\sigma_{1\beta}\sigma_{1\beta}')(\sigma_0^2 - \sigma_1^2) - 2\sigma_1^2\sigma_{1\beta}\sigma_{1\beta}'}{\sigma_1^4} dt\right)^{-1} \times \\
&\qquad \int_0^T \frac{2\sigma_{1\beta}\big((\mu_0 - \mu_1)^2 - (\mu_0 - \mu_2)^2\big)}{\sigma_1^3} dt + o_p(\sqrt{T}).
\end{aligned}$$

(b) Milstein approximation:

$$A_{j\beta} = -\frac{2\mu_j\sigma_{j\beta}}{\sigma_j^3} + \frac{3\sigma_j\dot{\sigma_j}\sigma_{1\beta}}{2\sigma_j^3} - \frac{3\dot{\sigma_{j\beta}}}{2\sigma_j}, \quad D_{j\beta} = \frac{2\mu_j^2\sigma_{j\beta}}{\sigma_j^3} - \frac{5\dot{\sigma_j}\dot{\sigma_{j\beta}}}{2} - \frac{3\mu_j\dot{\sigma_j}\sigma_{j\beta}}{\sigma_j^2} + \frac{3\mu_j\dot{\sigma_{j\beta}}}{\sigma_j},$$

$$E_{j\beta} = \frac{9\mu_j\dot{\sigma_j}\sigma_{j\beta}}{\sigma_j^4} - \frac{3\mu_j\dot{\sigma_{j\beta}}}{\sigma_j^3} - \frac{7\dot{\sigma_j}^2\sigma_{j\beta}}{\sigma_j^3} + \frac{7\dot{\sigma_j}\dot{\sigma_{j\beta}}}{\sigma_j^2}$$

for $j = 1, 2$, *where* $\dot{\sigma_{j\beta}} = \dot{\sigma_{j\beta}}(X_t, \beta_j^*)$ *and* $\dot{\sigma_{j\beta}}(x, \beta_j^*) = \partial\sigma_{j\beta}(x, \beta_j^*)/\partial x$, *thus*

$$\begin{aligned}
&\mathcal{L}_1(\hat{\theta}_1) - \mathcal{L}_2(\hat{\theta}_2) \\
&\quad = -\frac{1}{2}\int_0^T \left(\frac{(\mu_0-\mu_1)^2 - (\mu_0-\mu_2)^2}{\sigma_1^2} - \frac{3\dot{\sigma_1}(\sigma_1^2-\sigma^2)(\mu_1-\mu_2)}{\sigma_1^3}\right)dt \\
&\qquad + \int_0^T \frac{\sigma_0(\mu_1-\mu_2)}{\sigma_1^2}dW_t \\
&\qquad + \int_0^T \frac{\sigma_{1\beta}'(\sigma_0^2-\sigma_1^2)}{\sigma_1^3}dt\left(\int_0^T \frac{(\sigma_1\sigma_{1\beta\beta'} - 3\sigma_{1\beta}\sigma_{1\beta}')(\sigma_0^2-\sigma_1^2) - 2\sigma_1^2\sigma_{1\beta}\sigma_{1\beta}'}{\sigma_1^4}dt\right)^{-1}\times \\
&\qquad\quad \int_0^T 2\left[\frac{\sigma_{1\beta}\big((\mu_0-\mu_1)^2-(\mu_0-\mu_2)^2\big)}{\sigma_1^3} + \left(\frac{3\dot{\sigma_{1\beta}}}{2\sigma_1} + \frac{9\dot{\sigma_1}\sigma_{1\beta} - 3\sigma_1\dot{\sigma_{1\beta}}}{2\sigma_1^4}\sigma_0^2\right)(\mu_1-\mu_2)\right]dt \\
&\qquad + o_p(\sqrt{T}).
\end{aligned}$$

(c) Hermite expansion approximation (Aït-Sahalia (2002) with $K = 1$*):*

$$\begin{aligned}
A_{j\beta} &= -\frac{2\mu_j\sigma_{j\beta}}{\sigma_j^3} + \frac{3\sigma_j\dot{\sigma_j}\sigma_{1\beta}}{2\sigma_j^3} - \frac{3\dot{\sigma_{j\beta}}}{2\sigma_j}, \\
D_{j\beta} &= \frac{2\sigma_{j\beta}\mu_j^2}{\sigma_j^3} - \frac{2\sigma_{j\beta}\dot{\sigma_j}\mu_j}{\sigma_j^2} + \frac{2\dot{\sigma_{j\beta}}\mu_j}{\sigma_j} - \frac{1}{2}\dot{\sigma_j}\dot{\sigma_{j\beta}} + \frac{\sigma_{j\beta}\ddot{\sigma_j}}{2} + \frac{\sigma_j\ddot{\sigma_{j\beta}}}{2}, \\
E_{j\beta} &= -\frac{3\sigma_{j\beta}\dot{\sigma_j}^2}{\sigma_j^3} + \frac{6\mu_j\sigma_{j\beta}\dot{\sigma_j}}{\sigma_j^4} + \frac{3\dot{\sigma_{j\beta}}\dot{\sigma_j}}{\sigma_j^2} - \frac{2\sigma_{j\beta}\dot{\mu_j}}{\sigma_j^3} - \frac{2\mu_j\dot{\sigma_{j\beta}}}{\sigma_j^3} + \frac{3\sigma_{j\beta}\ddot{\sigma_j}}{2\sigma_j^2} - \frac{3\ddot{\sigma_{j\beta}}}{2\sigma_j}
\end{aligned}$$

for $j = 1, 2$, *where* $\dot{\sigma_{j\beta}} = \dot{\sigma_{j\beta}}(X_t, \beta_j^*)$ *and* $\dot{\sigma_{j\beta}}(x, \beta_j^*) = \partial\sigma_{j\beta}(x, \beta_j^*)/\partial x$, *and* $\ddot{\sigma_{j\beta}} = \ddot{\sigma_{j\beta}}(X_t, \beta_j^*)$ *and*

$\ddot{\sigma_{j\beta}}(x, \beta_j^*) = \partial^2 \sigma_{j\beta}(x, \beta_j^*)/\partial x^2$, *thus*

$$
\begin{aligned}
&\mathcal{L}_1(\hat{\theta}_1) - \mathcal{L}_2(\hat{\theta}_2) \\
&\quad = -\frac{1}{2}\int_0^T \left(\frac{(\mu_0 - \mu_1)^2 - (\mu_0 - \mu_2)^2}{\sigma_1^2} - \left(\frac{\sigma_0^2}{\sigma_1^2} - 1\right)\left(1 - \frac{2\dot{\sigma_1}}{\sigma_1}\right)(\mu_1 - \mu_2) \right) dt \\
&\quad\quad + \int_0^T \frac{\sigma_0(\mu_1 - \mu_2)}{\sigma_1^2} dW_t \\
&\quad\quad + \int_0^T \frac{\sigma_{1\beta}'(\sigma_0^2 - \sigma_1^2)}{\sigma_1^3} dt \left(\int_0^T \frac{(\sigma_1 \sigma_{1\beta\beta'} - 3\sigma_{1\beta}\sigma_{1\beta}')(\sigma_0^2 - \sigma_1^2) - 2\sigma_1^2 \sigma_{1\beta}\sigma_{1\beta}'}{\sigma_1^4} dt \right)^{-1} \times \\
&\quad\quad\quad \int_0^T 2\Bigg[\frac{\sigma_{1\beta}}{\sigma_1^3}\left((\mu_0 - \mu_1)^2 - (\mu_0 - \mu_2)^2\right) - \frac{\sigma_0^2 \sigma_{1\beta}}{\sigma_1^3}(\dot{\mu_1} - \dot{\mu_2}) \\
&\quad\quad\quad\quad + \left(\frac{3\sigma_0^2 \sigma_{1\beta}\dot{\sigma_1}}{\sigma_1^4} - \frac{\sigma_0^2 \dot{\sigma_{1\beta}}}{\sigma_1^3} - \frac{2\sigma_{1\beta}\dot{\sigma_1}}{\sigma_1^2} + \frac{\dot{\sigma_{1\beta}}}{\sigma_1} \right)(\mu_1 - \mu_2) \Bigg] dt + o_p(\sqrt{T}).
\end{aligned}
$$

For the Ornstein-Uhlenbeck process, it is shown that the model selection criterion using the Euler or the Hermite expansion approximation is equivalent to using the true transition density.

Example A.3 *(Ornstein-Uhlenbeck) If Model j is the Ornstein-Uhlenbeck process,*

$$dX_t = \kappa(\eta - X_t)dt + \gamma dW_t,$$

we have

$$
\begin{aligned}
&A_j^0(x) = \frac{\kappa(\eta - x)}{\gamma^2}, \quad B_j^0(x) = 0, \quad C_j^0(x) = 0, \\
&D_j^0(x) = -\frac{\kappa^2(\eta - x)^2}{\gamma^2} + \kappa, \quad E_j^0(x) = -\frac{\kappa}{\gamma^2}, \\
&A_{j\beta}^0(x) = -\frac{2\kappa(\eta - x)}{\gamma^3}, \quad D_{j\beta}^0(x) = \frac{2\kappa^2(\eta - x)^2}{\gamma^3}, \quad E_{j\beta}^0(x) = \frac{2\kappa}{\gamma^3}
\end{aligned}
$$

for the exact transition density. We obtain the same functions as above if we use the Hermite expansion approximation of Aït-Sahalia (2002). Also for the Euler approximation, the difference of the Euler approximated log-likelihood $\mathcal{L}_j$ and the true log-likelihood $\mathcal{L}_j^0$ is

$$\frac{1}{T}\left(\mathcal{L}_j(\hat{\theta}_j) - \mathcal{L}_j^0(\hat{\theta}_j^0)\right) = \frac{1}{T}\int_0^T \frac{\kappa^*\left(\gamma^{*2} - \sigma_0^2\right)}{2\gamma^{*2}} dt + O_p(T^{-1/2}) \xrightarrow{p} 0,$$

where κ^* and γ^* are pseudo-true values of κ and γ, since $\mathbf{E}\left(\gamma^{*2} - \sigma_0^2\right) = 0$ from Lemma B.11.

B Proofs

Hereafter we use a shorthand notation for the functional argument such as $fg(x) \doteq f(x)g(x)$. We denote $\ell^o(\Delta, x, y) = \ell(\Delta, x, y) + \log\sqrt{\Delta}$ and $\ell^*(\Delta, x, y) = \Delta\ell^o(\Delta, x, y)$ in the following. (That is, $\ell^o(\Delta, x, y) = \log[\sqrt{\Delta}p(\Delta, x, y)]$ and $\ell^*(\Delta, x, y) = \log\big([\sqrt{\Delta}p(\Delta, x, y)]^{\Delta}\big)$.) We define $f(0, x, x) = \lim_{\Delta\to 0} f(\Delta, x, x)$. Also, for notational convenience, hereafter we denote the derivative w.r.t. the parameters at $y = x$ and $\Delta = 0$ as $\ell^*_{\alpha y\Delta}(0, x, x) = (\ell^*_{y\Delta})_\alpha(0, x, x)$, for example.

B.1 Assumptions

Assumption B.1 (Differentiability) *For $j = 1, 2$, we assume that $\ell^*_i(t, x, y; \theta)$, $\mu_i(x; \alpha)$ and $\sigma_i(x; \beta)$ are infinitely differentiable in $t \geq 0$, $x, y \in \mathcal{D}$ and $\theta_i \in \Theta_i$, and denoting each derivative of ℓ^*_i, μ_i and σ_i as $f(t, x, y; \theta)$, there exist g which is locally bounded and increasing at a polynomial rate of order k on both boundaries, and for all $x \in \mathcal{D}$*

$$|f(t, x, y; \theta)| \leq g(x)$$

for all small $t \geq 0$, and all y close to x.

Note that Assumption B.1 guarantees the existence of the following limits, which we denote as

$$\lim_{\Delta\to 0} \ell^*_{jy\Delta}(\Delta, x, x) = A_j(x) \qquad \lim_{\Delta\to 0} \ell^*_{jyyy}(\Delta, x, x) = B_j(x)$$
$$\lim_{\Delta\to 0} \ell^*_{jyyyy}(\Delta, x, x) = C_j(x) \qquad \lim_{\Delta\to 0} \ell^*_{j\Delta\Delta}(\Delta, x, x) = D_j(x)$$
$$\lim_{\Delta\to 0} \ell^*_{jyy\Delta}(\Delta, x, x) = E_j(x).$$

Assumption B.2 (Extremal Bounds) *$X_t \in \mathcal{D}_X$ is stationary and there exists $p > 0$ such that*

$$T^{-p} \sup_{t\in[0,T]} \big|X_t\big| \xrightarrow{p} 0$$

as $T \to \infty$. When $\mathcal{D}_X = (0, \infty)$, additionally assume

$$T^{-p} \sup_{t\in[0,T]} X_t^{-1} \xrightarrow{p} 0.$$

For the properties of the extremal process of diffusion models, one can refer to Berman (1964), Davis (1982), and Stone (1963), or also to Cline and Jeong (2009).

Assumption B.3 (Likelihood Expansion) *For $j = 1, 2$, ℓ^*_j satisfies* $\lim_{\Delta\to 0} \ell^*_j(\Delta, x, x) = 0$,

$\lim_{\Delta\to 0} \ell^*_{jy}(\Delta, x, x) = 0$, *and*

$$\lim_{\Delta\to 0} \ell^*_{j\Delta}(\Delta, x, x) = -\log\big(\sigma_j(x)\big) + c$$
$$\lim_{\Delta\to 0} \ell^*_{jyy}(\Delta, x, x) = -1/\sigma_j^2(x)$$

for some constant c which does not depend on the parameters. Moreover, $B_j(x)$ and $C_j(x)$ do not depend on $\mu_j(x)$.

It can be shown that the true transition density of diffusion processes satisfy this condition as in Jeong and Park (2009). It is also not difficult to check these conditions are all satisfied by the Euler and the Milstein approximated transition densities, and also the closed-form approximation proposed by Aït-Sahalia (2002).

Assumption B.4 (Identification) *Denoting $f_{j\alpha\alpha}$ and $f_{j\beta\beta}$ as the following,*

$$f_{j\alpha\alpha}(x;\theta_j) = \mu_0(x)A_{j\alpha\alpha'}(x;\theta_j) + \frac{\sigma_0^2(x)}{2}E_{j\alpha\alpha'}(x;\theta_j) + \frac{1}{2}D_{j\alpha\alpha'}(x;\theta_j)$$
$$f_{j\beta\beta}(x;\theta_j) = \left(\frac{\sigma_{j\beta}\sigma'_{j\beta}}{\sigma_j^2} - \frac{\sigma_{j\beta\beta'}}{\sigma_j}\right)(x;\theta_j) - \left(\frac{3\sigma_{j\beta}\sigma'_{j\beta}}{\sigma_j^4} - \frac{\sigma_{j\beta\beta'}}{\sigma_j^3}\right)(x;\theta_j)\sigma_0^2(x)$$

for $j = 1, 2$, $\mathbf{E}f_{j\alpha\alpha}(X_t;\theta_j^*)$ and $\mathbf{E}f_{j\beta\beta}(X_t;\theta_j^*)$ are positive definite.

This is a condition to make the Hessian positive definite.

B.2 Proof of the Main Theorem

B.2.1 Useful Lemmas

For Lemmas B.8-B.10, we let them hold for both Case 1 and Case 2, that is, either when $\Delta^2 T \to \infty$ and $\Delta^3 T \to 0$, or when $\Delta T^2 \to 0$. We also introduce functional operators,

$$\mathcal{A}f(t,x,y) = f_t(t,x,y) + \mu_0(y)f_y(t,x,y) + \frac{1}{2}\sigma_0^2(y)f_{y^2}(t,x,y)$$
$$\mathcal{B}f(t,x,y) = \sigma_0(y)f_y(t,x,y)$$

for our derivation.

Lemma B.5 *Let $X_t \in \mathcal{D}_X$ be a positive recurrent process with a derivative of scale density $s_0(x)$ and a speed density $m_0(x)$. For f such that $\mathbf{E}f(X_t) = 0$ and $\mathbf{E}\mathcal{G}_f^2(X_t) < \infty$,*

$$\frac{1}{\sqrt{T}}\int_0^T f(X_t)dt \xrightarrow{d} \mathbf{N}\big(0, \mathbf{E}\mathcal{G}_f^2(X_t)\big)$$

as $T \to \infty$, where $\mathcal{G}_f(x) = \sigma_0(x)s_0(x)\int_{x_0}^x f(v)m_0(v)dv$ for $x_0 \in \mathcal{D}_X$ as chosen in (2.17).

Proof. Note that from Itô's lemma

$$g(X_T) - g(X_0) = \int_0^T \left(\mu_0 g^{\cdot} + \frac{\sigma_0^2 g^{\cdot\cdot}}{2}\right)(X_t)dt + \int_0^T \sigma_0 g^{\cdot}(X_t)dW_t.$$

Letting $\big(\mu_0 g^{\cdot} + \sigma_0^2 g^{\cdot\cdot}/2\big)(x) = f(x)$, we have

$$g^{\cdot}(x) = s_0(x)\int_{x_0}^x f(v)m_0(v)dv$$

as a solution of the partial differential equation. Thus we can write

$$\int_0^T f(X_t)dt = -\int_0^T \mathcal{G}_f(X_t)dW_t + g(X_T) - g(X_0), \tag{B.1}$$

where $\mathcal{G}_f(x) = \sigma_0(x)s_0(x)\int_{x_0}^x f(v)m_0(v)dv$ and $g(x) = \int_{x_0}^x \mathcal{G}_f(v)/\sigma_0(v)dv$. If $\mathbf{E}f(X_t) = 0$,

$$\frac{1}{\sqrt{T}}\int_0^T f(X_t)dt = -\frac{1}{\sqrt{T}}\int_0^T \mathcal{G}_f(X_t)dW_t + \frac{g(X_T)}{\sqrt{T}} - \frac{g(X_0)}{\sqrt{T}} \xrightarrow{d} \mathbf{N}\big(0, \mathbf{E}\mathcal{G}_f^2(X_t)\big)$$

as $T \to \infty$, and if $\mathbf{E}f(X_t) \neq 0$, we have $\mathbf{E}\mathcal{G}_f^2(X_t) = \infty$ since otherwise it gives a contradiction. ■

Lemma B.6 *Let $X_t \in \mathcal{D}_X$ be a positive recurrent process with a speed density $m_0(x)$ and $\mathbf{E}f(X_t) < \infty$. If*

$$\int_{\mathcal{D}_X} f(x)m_0(x)dx < \infty,$$

then

$$\int_{\mathcal{D}_X} \left(\mu_0 f^{\cdot} + \frac{\sigma_0^2 f^{\cdot\cdot}}{2}\right)(x)m_0(x)dx = 0.$$

Proof. From Itô's lemma,

$$f(X_t) - f(X_s) = \int_s^t \left(\mu_0 f^{\cdot} + \frac{\sigma_0^2 f^{\cdot\cdot}}{2}\right)(X_r)dr + \int_s^t \sigma_0 f^{\cdot}(X_r)dW_r$$

for any $t > s$. Taking expectation on both sides,

$$\mathbf{E}f(X_t) - \mathbf{E}f(X_s) = \int_s^t \mathbf{E}\left(\mu_0 f^{\cdot} + \frac{\sigma_0^2 f^{\cdot\cdot}}{2}\right)(X_r)dr = 0$$

so we should have

$$\mathbf{E}\bigg(\mu_0 f^{\cdot} + \frac{\sigma_0^2 f^{\cdot\cdot}}{2}\bigg)(X_r) = \int_{\mathcal{D}_X} \bigg(\mu_0 f^{\cdot} + \frac{\sigma_0^2 f^{\cdot\cdot}}{2}\bigg)(x)\frac{m_0(x)}{M(\mathcal{D}_X)}dx = 0,$$

where $M(\mathcal{D}_X) = \int_{\mathcal{D}_X} m_0(x)dx$. ■

Lemma B.7 *For f satisfying the conditions in Assumption B.1 and i.i.d. mean zero random variables $\mathcal{W}_{i,\Delta}$ with $\mathbf{E}\mathcal{W}_{i,\Delta}^2 = c\Delta^k$ for some c and k,*

$$\sum_{i=1}^n f(X_{(i-1)\Delta})\mathcal{W}_{i,\Delta} = O_p(\Delta^{(k-1)/2}T^{r+1/2})$$

as $T \to \infty$.

Proof. Since we can view it as a continuous martingale representation whose quadratic variation is given by the conditional variance process

$$c^2 \sum_{i=1}^n f^2(X_{(i-1)\Delta})\Delta^k = O_p(\Delta^{k-1}T^{2r+1}),$$

the stated result easily follows from this. ■

Lemma B.8 *(a) If the following repeated integrations only consist of the time (dt) integration,*

$$\sum_{i=1}^n g(X_{(i-1)\Delta}) \int_{(i-1)\Delta}^{i\Delta} \cdots \int_{(i-1)\Delta}^{s} f(r, X_{(i-1)\Delta}, X_r)dr \cdots dt = O_p(\Delta^{k-1}T)$$

as $T \to \infty$ and $\Delta \to 0$, where k is the number of the repeated integrations.
(b) Otherwise,

$$\sum_{i=1}^n g(X_{(i-1)\Delta}) \int_{(i-1)\Delta}^{i\Delta} \cdots \int_{(i-1)\Delta}^{s} f(r, X_{(i-1)\Delta}, X_r)dr \cdots dW_t = O_p(\Delta^{(2k_1+k_2-1)/2}\sqrt{T})$$

as $T \to \infty$ and $\Delta \to 0$, where k_1 is the number of integrals w.r.t. the time (dt), and k_2 is the number of integral w.r.t. the Brownian motion (dW_t). Here, though we did not write in the exact form not to make it too complicated, in the expression for the repeated integration, the integral can be with respect to either time, or to the Brownian motion, with any combinations of the two, with a condition that it has at least one dW_t term.

Proof. To prove this lemma, we utilize the result of Lemma 3 in Jeong and Park (2009), which states the order of the same integrals are $O_p(\Delta^{k-1}T^{2r+1})$ for (a), and $O_p(\Delta^{(2k_1+k_2-1)/2}T^{2r+1/2})$

for (b). Note that this Lemma 3 can be proven under Assumptions B.1 and B.2. For case (a), note that we can apply Itô's lemma

$$f(r,X_{(i-1)\Delta},X_r) = f(0,X_{(i-1)\Delta},X_{(i-1)\Delta}) \\ + \int_{(i-1)\Delta}^{r} \mathcal{A}f(t-(i-1)\Delta,X_{(i-1)\Delta},X_t)dt + \int_{(i-1)\Delta}^{r} \mathcal{B}f(t-(i-1)\Delta,X_{(i-1)\Delta},X_t)dW_t$$

repeatedly to get higher order terms. If we apply Itô's lemma $q \geq 4/3r - 1/3$ steps, then the biggest order term becomes $O_p(\Delta^{k-1+q/2}T^{2r+1/2})$ from the specified lemma and Lemma B.7, thus we obtain the stated result. Similarly for the case (b), we can apply Itô's lemma repeatedly $q \geq 4/3r$ steps, then the biggest order term becomes $O_p(\Delta^{(2k_1+k_2-1)/2+q/2}T^{2r+1/2})$, thus we obtain the stated result. ■

Lemma B.9 *For 6 times differentiable f,*

$$\Delta\sum_{i=1}^{n} f(X_{(i-1)\Delta}) = \int_0^T f(X_t)dt + O_p(\Delta\sqrt{T})$$

as $T\to\infty$ and $\Delta\to 0$.

Proof. From Itô's lemma,

$$\begin{aligned}\sum_{i=1}^{n} f(X_{(i-1)\Delta})\Delta &= \int_0^T f(X_t)dt - \sum_{i=1}^{n}\int_{(i-1)\Delta}^{i\Delta}\int_{(i-1)\Delta}^{t}\left(\mu_0 f^{\cdot} + \frac{\sigma_0^2 f^{\cdot\cdot}}{2}\right)(X_s)dsdt \\ &\quad - \sum_{i=1}^{n}\int_{(i-1)\Delta}^{i\Delta}\int_{(i-1)\Delta}^{t}\sigma_0 f^{\cdot}(X_s)dW_sdt \\ &= \int_0^T f(X_t)dt + A_{T,\Delta} + B_{T,\Delta},\end{aligned}$$

and $B_{T,\Delta} = O_p(\Delta\sqrt{T})$ from Lemma B.8. For $A_{T,\Delta}$, we can do the same expansions three steps further using Itô's lemma and we get

$$A_{T,\Delta} = \Delta\int_0^T f_1(X_t)dt + 2\Delta^2\int_0^T f_2(X_t)dt + R_{T,\Delta},$$

where $f_1(x) = \big(\mu_0 f^{\cdot} + \sigma_0^2 f^{\cdot\cdot}/2\big)(x)$, $f_2(x) = \big(\mu_0 f_1^{\cdot} + \sigma_0^2 f_1^{\cdot\cdot}/2\big)(x)$, and $R_{T,\Delta}$ is a remainder which consists of the terms appearing in Lemma B.8. It is straightforward to show each term and $R_{T,\Delta}$ are $O_p(\Delta\sqrt{T})$ using Lemma B.6, B.5 and B.8. ■

Lemma B.10 *For twice differentiable f,*

$$\sum_{i=1}^{n} f(X_{(i-1)\Delta})(W_{i\Delta} - W_{(i-1)\Delta}) = \int_0^T f(X_t)dW_t + O_p(\sqrt{\Delta T})$$

as $T \to \infty$ and $\Delta \to 0$.

Proof. Using Itô's lemma,

$$\begin{aligned}\sum_{i=1}^{n} f(X_{(i-1)\Delta})(W_{i\Delta} - W_{(i-1)\Delta}) &= \int_0^T f(X_t)dW_t - \sum_{i=1}^{n}\int_{(i-1)\Delta}^{i\Delta}\int_{(i-1)\Delta}^{t} \sigma_0 f^{\cdot}(X_s)dW_s dW_t \\ &\quad - \sum_{i=1}^{n}\int_{(i-1)\Delta}^{i\Delta}\int_{(i-1)\Delta}^{t}\left(\mu_0 f^{\cdot} + \frac{\sigma_0^2 f^{\cdot\cdot}}{2}\right)(X_s)dsdW_t \\ &= \int_0^T f(X_t)dW_t + O_p(\sqrt{\Delta T})\end{aligned}$$

from Lemma B.8. ■

Lemma B.11 *Denoting $f_{j\alpha}$ and $f_{j\beta}$ as*

$$\begin{aligned}f_{j\alpha}(x;\theta_j) &= \mu_0(x)A_{j\alpha}(x;\theta_j) + \frac{\sigma_0^2(x)}{2}E_{j\alpha}(x;\theta_j) + \frac{1}{2}D_{j\alpha}(x;\theta_j) \\ f_{j\beta}(x;\beta_j) &= -\frac{\sigma_{j\beta}}{\sigma_j}(x;\beta_j) + \sigma_0^2(x)\frac{\sigma_{j\beta}}{\sigma_j^3}(x;\beta_j)\end{aligned}$$

for $j = 1, 2$, $\mathbf{E} f_{1\alpha}(X_t;\theta_1^*) = 0$ and $\mathbf{E} f_{1\beta}(X_t;\beta_1^*) = 0$.

Proof. By applying Itô's lemma subsequently, we have

$$\begin{aligned}\mathcal{S}_{1\alpha}(\theta_1) &= \frac{1}{\Delta}\sum_{i=1}^{n}\ell_{1\alpha}^*(\Delta, x_i, y_i) \\ &= \frac{1}{\Delta}\sum_{i=1}^{n}\ell_{1\alpha}^*(0, x_i, x_i) + \sum_{i=1}^{n}\mathcal{A}\ell_{1\alpha}^*(0, x_i, x_i) + \frac{1}{\Delta}\sum_{i=1}^{n}\mathcal{B}\ell_{1\alpha}^*(0, x_i, x_i)\mathcal{W}_{1i} \\ &\quad + \frac{\Delta}{2}\sum_{i=1}^{n}\mathcal{A}^2\ell_{1\alpha}^*(0, x_i, x_i) + \frac{1}{\Delta}\sum_{i=1}^{n}\mathcal{B}\mathcal{A}\ell_{1\alpha}^*(0, x_i, x_i)\mathcal{W}_{2i} + \frac{1}{\Delta}\sum_{i=1}^{n}\mathcal{A}\mathcal{B}\ell_{1\alpha}^*(0, x_i, x_i)\mathcal{W}_{3i} \\ &\quad + \frac{1}{\Delta}\sum_{i=1}^{n}\mathcal{B}^2\ell_{1\alpha}^*(0, x_i, x_i)\mathcal{W}_{4i} + \frac{1}{\Delta}\sum_{i=1}^{n}\mathcal{B}^3\ell_{1\alpha}^*(0, x_i, x_i)\mathcal{W}_{5i} + O_p(\sqrt{\Delta T}) + O_p(\Delta T),\end{aligned}$$

where

$$\mathcal{W}_{1i} = W_{i\Delta} - W_{(i-1)\Delta}, \quad \mathcal{W}_{2i} = \int_{(i-1)\Delta}^{i\Delta} \int_{(i-1)\Delta}^{s} dW_r ds, \quad \mathcal{W}_{3i} = \int_{(i-1)\Delta}^{i\Delta} \int_{(i-1)\Delta}^{s} dr dW_s,$$
$$\mathcal{W}_{4i} = \int_{(i-1)\Delta}^{i\Delta} \int_{(i-1)\Delta}^{s} dW_r dW_s, \quad \mathcal{W}_{5i} = \int_{(i-1)\Delta}^{i\Delta} \int_{(i-1)\Delta}^{s} \int_{(i-1)\Delta}^{r} dW_u dW_r dW_s, \tag{B.2}$$

from Lemma B.8. From Assumption B.3 note that we have $\bar{\ell}_\alpha(0, x, x) = 0$, $\mathcal{A}\bar{\ell}_\alpha(0, x, x) = 0$, $\mathcal{B}\bar{\ell}_\alpha(0, x, x) = 0$, $\mathcal{B}^2\bar{\ell}_\alpha(0, x, x) = 0$ and $\mathcal{B}^3\bar{\ell}_\alpha(0, x, x) = 0$. Also note that $\mathcal{W}_{2i} = \frac{\Delta}{2}(W_{i\Delta} - W_{(i-1)\Delta}) + \frac{\Delta}{2\sqrt{3}}(Z_{i\Delta} - Z_{(i-1)\Delta})$ and $\mathcal{W}_{3i} = \frac{\Delta}{2}(W_{i\Delta} - W_{(i-1)\Delta}) - \frac{\Delta}{2\sqrt{3}}(Z_{i\Delta} - Z_{(i-1)\Delta})$, where Z is a standard Brownian motion independent of W. Thus, so that this score becomes a martingale in the limit, we should have

$$\begin{aligned}
&\mathbf{E}\big(\mathcal{A}^2\ell^*_{1\alpha}(0, X_t, X_t; \theta)\big) \\
&\quad = \mathbf{E}\bigg(\mu_0(X_t)\ell^*_{1\alpha y\Delta}(0, X_t, X_t; \theta) + \frac{\sigma_0^2(X_t)}{2}\ell^*_{1\alpha yy\Delta}(0, X_t, X_t; \theta) + \frac{1}{2}\ell^*_{1\alpha\Delta\Delta}(0, X_t, X_t; \theta)\bigg) = 0
\end{aligned}$$

at $\theta = \theta_1^*$ as the all other remainder terms become of smaller orders. Similarly for $\mathcal{S}_{1\beta}(\theta_1^*)$. ■

Lemma B.12 *Denoting ι_k as a $k \times 1$ one vector, let $w_{1\alpha} = T^{1/2}\iota_{k_{1\alpha}}$, where $k_{1\alpha}$ is the number of drift term parameters, and $w_{1\beta} = T^{1/2}\Delta^{-1/2}\iota_{k_{1\beta}}$, where $k_{1\beta}$ is the number of diffusion term parameters. Define $w_1 = Diag\big((w'_{1\alpha}, w'_{1\beta})'\big)$ as a diagonal matrix. Then*

$$\sup_{\theta \in \mathcal{N}_1} \left| w_1^{-1}\big(\mathcal{H}_1(\theta) - \mathcal{H}_1(\theta_1^*)\big)w_1^{-1\prime} \right| = o_p(1)$$

as $T \to \infty$ and $\Delta \to 0$, where $\mathcal{N}_1 = \{\theta : w_1'(\theta - \theta_1^*) \leq 1\}$. The same hold for Model 2.

Proof. With the same step as in (B.4), note that we have the expansion for the Hessian as

$$\begin{aligned}
\mathcal{H}_{1\alpha\alpha'}(\theta) &= \int_0^T \left[A_{1\alpha\alpha}(\theta)\mu_0 + \frac{1}{2}E_{1\alpha\alpha}(\theta)\sigma_0^2 + \frac{1}{2}D_{1\alpha\alpha}(\theta)\right](X_t)dt + O_p(\sqrt{T}) \\
\mathcal{H}_{1\alpha\beta'}(\theta) &= \int_0^T \left[A_{1\alpha\beta}(\theta)\mu_0 + \frac{1}{2}E_{1\alpha\beta}(\theta)\sigma_0^2 + \frac{1}{2}D_{1\alpha\beta}(\theta)\right](X_t)dt + O_p(\sqrt{T}) \\
\mathcal{H}_{1\beta\beta'}(\theta) &= \frac{1}{\Delta}\int_0^T \left[\frac{\sigma_{j\beta}\sigma'_{j\beta}}{\sigma_j^2}(\theta) - \frac{\sigma_{j\beta\beta'}}{\sigma_j}(\theta) - \left(\frac{3\sigma_{j\beta}\sigma'_{j\beta}}{\sigma_j^4} - \frac{\sigma_{j\beta\beta'}}{\sigma_j^3}\right)(\theta)\sigma_0^2\right](X_t)dt + O_p(\Delta^{-1/2}\sqrt{T})
\end{aligned}$$

with the notational exploitation for the functional arguments. We have

$$
\begin{aligned}
&\mathcal{H}_{1\alpha\alpha'}(\theta) - \mathcal{H}_{1\alpha\alpha'}(\theta_1^*) \\
&\quad = \big((\theta - \theta_1^*) \otimes I_{k_{\theta_1}}\big)' \int_0^T \left[\frac{\partial}{\partial\theta} A_{1\alpha\alpha}(\tilde{\theta})\mu_0 + \frac{1}{2}\frac{\partial}{\partial\theta} E_{1\alpha\alpha}(\tilde{\theta})\sigma_0^2 + \frac{1}{2}\frac{\partial}{\partial\theta} D_{1\alpha\alpha}(\tilde{\theta})\right](X_t)dt + O_p(\sqrt{T}) \\
&\quad = O_p(T||\theta - \theta_1^*||) + O_p(\sqrt{T})
\end{aligned}
$$

uniformly in $\tilde{\theta} \in \Theta$, where $\tilde{\theta}$ lies in the line segment of θ and θ_1^*, and similarly for $\mathcal{H}_{1\alpha\beta'}(\theta) - \mathcal{H}_{1\alpha\beta'}(\theta_1^*)$ and $\mathcal{H}_{1\beta\beta'}(\theta) - \mathcal{H}_{1\beta\beta'}(\theta_1^*)$. Thus

$$
\begin{pmatrix} T^{-1}\big(\mathcal{H}_{1\alpha\alpha'}(\theta) - \mathcal{H}_{1\alpha\alpha'}(\theta_1^*)\big) & \sqrt{\Delta}T^{-1}\big(\mathcal{H}_{1\alpha\beta'}(\theta) - \mathcal{H}_{1\alpha\beta'}(\theta_1^*)\big) \\ \sqrt{\Delta}T^{-1}\big(\mathcal{H}_{1\beta\alpha'}(\theta) - \mathcal{H}_{1\beta\alpha'}(\theta_1^*)\big) & \Delta T^{-1}\big(\mathcal{H}_{1\beta\beta'}(\theta) - \mathcal{H}_{1\beta\beta'}(\theta_1^*)\big) \end{pmatrix} \xrightarrow{p} 0
$$

uniformly in $\theta \in \Theta$. ■

B.2.2 Proof of Theorem 2.3

Hereafter we use the convention $f(0, x) = \lim_{\Delta \to 0} f(\Delta, x)$ for notational convenience. We also let $x_i = X_{(i-1)\Delta}$ and $y_i = X_{i\Delta}$.

Proof of (a).

Part 1. (Estimator Asymptotics) We have

$$
w_1^{-1}\mathcal{S}_1(\hat{\theta}_1) = w_1^{-1}\mathcal{S}_1(\theta_1^*) + w_1^{-1}\mathcal{H}_1(\theta_1^*)(\hat{\theta}_1 - \theta_1^*) + o_p\big(w_1'(\hat{\theta}_1 - \theta_1^*)\big)
$$

since from Lemma B.12

$$
\sup_{\theta \in \mathcal{N}_1} \left| w_1^{-1}\big(\mathcal{H}_1(\theta) - \mathcal{H}_1(\theta_1^*)\big)w_1^{-1\prime} \right| = o_p(1)
$$

where $\mathcal{N}_1 = \{\theta : w_1'(\theta - \theta_1^*) \leq 1\}$. Thus

$$
-w_1^{-1}\mathcal{H}_1(\theta_1^*)(\hat{\theta}_1 - \theta_1^*) = w_1^{-1}\mathcal{S}_1(\theta_1^*) + o_p\big(w_1'(\hat{\theta}_1 - \theta_1^*)\big)
$$

wpa1, and we get

$$
\hat{\theta}_1 - \theta_1^* = -\mathcal{H}_1^{-1}(\theta_1^*)\mathcal{S}_1(\theta_1^*) + o_p(\hat{\theta}_1 - \theta_1^*). \tag{B.3}
$$

We let $x_i = X_{(i-1)\Delta}$ and $y_i = X_{i\Delta}$ hereafter for the simplicity. From Assumption B.3 and Lemma

B.8,

$$\begin{aligned}
\mathcal{S}_{1\alpha}(\theta_1^*) &= \frac{1}{\Delta}\sum_{i=1}^n \ell_{1\alpha}^*(\Delta, x_i, y_i) \qquad \text{(B.4)}\\
&= \frac{1}{\Delta}\sum_{i=1}^n \ell_{1\alpha}^*(0, x_i, x_i) + \sum_{i=1}^n \mathcal{A}\ell_{1\alpha}^*(0, x_i, x_i) + \frac{1}{\Delta}\sum_{i=1}^n \mathcal{B}\ell_{1\alpha}^*(0, x_i, x_i)\mathcal{W}_{1i}\\
&\quad + \frac{\Delta}{2}\sum_{i=1}^n \mathcal{A}^2\ell_{1\alpha}^*(0, x_i, x_i) + \frac{1}{\Delta}\sum_{i=1}^n \mathcal{B}\mathcal{A}\ell_{1\alpha}^*(0, x_i, x_i)\mathcal{W}_{2i} + \frac{1}{\Delta}\sum_{i=1}^n \mathcal{A}\mathcal{B}\ell_{1\alpha}^*(0, x_i, x_i)\mathcal{W}_{3i}\\
&\quad + \frac{1}{\Delta}\sum_{i=1}^n \mathcal{B}^2\ell_{1\alpha}^*(0, x_i, x_i)\mathcal{W}_{4i} + \frac{1}{\Delta}\sum_{i=1}^n \mathcal{B}^3\ell_{1\alpha}^*(0, x_i, x_i)\mathcal{W}_{5i} + O_p(\Delta T),
\end{aligned}$$

where $\mathcal{W}_{ki}$'s are defined in (B.2), and from Lemma B.11,

$$\begin{aligned}
\mathcal{S}_{1\alpha}(\theta_1^*) &= \Delta\sum_{i=1}^n \left(\mu_0(x_i)\ell_{1\alpha y\Delta}^*(0, x_i, x_i) + \frac{\sigma_0^2(x_i)}{2}\ell_{1\alpha yy\Delta}^*(0, x_i, x_i) + \frac{1}{2}\ell_{1\alpha\Delta\Delta}^*(0, x_i, x_i)\right) + O_p(\Delta T)\\
&= O_p(\Delta T).
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathcal{S}_{1\beta}(\theta_1^*) &= O_p(T) \qquad \text{(B.5)}\\
\mathcal{H}_{1\alpha\alpha}(\theta_1^*) &= O_p(T) \qquad \text{(B.6)}\\
\mathcal{H}_{1\beta\beta}(\theta_1^*) &= O_p(\Delta^{-1}T) \qquad \text{(B.7)}\\
\mathcal{H}_{1\alpha\beta}(\theta_1^*) &= O_p(T). \qquad \text{(B.8)}
\end{aligned}$$

Also, from Assumption B.4, we can check that

$$\begin{aligned}
\mathcal{H}_{1\beta\beta}^{-1}(\theta_1^*) &= O_p(\Delta T^{-1})\\
\mathcal{H}_{1\alpha\beta}^{-1}(\theta_1^*) &= O_p(T^{-1}),
\end{aligned}$$

thus

$$\hat{\theta}_1 - \theta_1^* = O_p(\Delta).$$

Part 2. (Likelihood Asymptotics) We have

$$\begin{aligned}
\mathcal{L}_1(\hat{\theta}_1) - \mathcal{L}_2(\hat{\theta}_2) &= \mathcal{L}_1(\theta_1^*) - \mathcal{L}_2(\theta_2^*) + (\hat{\theta}_1 - \theta_1^*)'\mathcal{S}_1(\theta_1^*) - (\hat{\theta}_2 - \theta_2^*)'\mathcal{S}_2(\theta_2^*)\\
&\quad + \frac{1}{2}(\hat{\theta}_1 - \theta_1^*)'\mathcal{H}_1(\theta_1^*)(\hat{\theta}_1 - \theta_1^*) - \frac{1}{2}(\hat{\theta}_2 - \theta_2^*)'\mathcal{H}_2(\theta_2^*)(\hat{\theta}_2 - \theta_2^*) + o_p(\Delta T)
\end{aligned}$$

since from Lemma B.12

$$\sup_{\theta\in\mathcal{N}_1}\left|w_1^{-1}\big(\mathcal{H}_1(\theta)-\mathcal{H}_1(\theta_1^*)\big)w_1^{-1\prime}\right| = o_p(1)$$

where $\mathcal{N}_1 = \{\theta : w_1'(\theta-\theta_1^*) \le 1\}$, so

$$(\hat\theta_1-\theta_1^*)'\big(\mathcal{H}_1(\tilde\theta_1)-\mathcal{H}_1(\theta_1^*)\big)(\hat\theta_1-\theta_1^*) = o_p(\Delta T).$$

We also have

$$\begin{aligned}
&(\hat\theta_1-\theta_1^*)'\mathcal{S}_1(\theta_1^*) = O_p(\Delta T)\\
&(\hat\theta_1-\theta_1^*)'\mathcal{H}_1(\theta_1^*)(\hat\theta_1-\theta_1^*) = O_p(\Delta T),
\end{aligned}$$

thus

$$\mathcal{L}_1(\hat\theta_1)-\mathcal{L}_2(\hat\theta_2) = \mathcal{L}_1(\theta_1^*)-\mathcal{L}_2(\theta_2^*)+O_p(\Delta T).$$

Hereafter, let us suppress the parameter arguments, and let $\ell_1^o(\Delta,x,y) = \ell_1(\Delta,x,y)+\log\sqrt{\Delta}$. Note that

$$\mathcal{L}_1(\theta_1^*)-\mathcal{L}_2(\theta_2^*) = \sum_{i=1}^n \ell_1(\Delta,x_i,y_i)-\sum_{i=1}^n \ell_2(\Delta,x_i,y_i) = \sum_{i=1}^n \ell_1^o(\Delta,x_i,y_i)-\sum_{i=1}^n \ell_2^o(\Delta,x_i,y_i).$$

From Lemma B.8,

$$\begin{aligned}
\sum_{i=1}^n \ell_1^o(\Delta,x_i,y_i) &= \frac{1}{\Delta}\sum_{i=1}^n \ell_1^*(0,x_i,x_i)+\sum_{i=1}^n \mathcal{A}\ell_1^*(0,x_i,x_i)+\frac{1}{\Delta}\sum_{i=1}^n \mathcal{B}\ell_1^*(0,x_i,x_i)\mathcal{W}_{1i}\\
&\quad+\frac{\Delta}{2}\sum_{i=1}^n \mathcal{A}^2\ell_1^*(0,x_i,x_i)+\frac{1}{\Delta}\sum_{i=1}^n \mathcal{B}\mathcal{A}\ell_1^*(0,x_i,x_i)\mathcal{W}_{2i}+\frac{1}{\Delta}\sum_{i=1}^n \mathcal{A}\mathcal{B}\ell_1^*(0,x_i,x_i)\mathcal{W}_{3i}\\
&\quad+\frac{1}{\Delta}\sum_{i=1}^n \mathcal{B}^2\ell_1^*(0,x_i,x_i)\mathcal{W}_{4i}+\frac{1}{\Delta}\sum_{i=1}^n \mathcal{B}^3\ell_1^*(0,x_i,x_i)\mathcal{W}_{5i}+O_p(\Delta T),
\end{aligned}$$

where $\mathcal{W}_{ki}$'s are defined in (B.2), so from Lemma B.8, the stated result follows by applying the operators to each function and simplifying the equation. ■

Proof of (b).

Part 1. (Estimator Asymptotics)
First note that $\beta_1^* = \beta_2^*$ when $\sigma_1 = \sigma_2$ from Lemma B.11 and Assumption B.3. From (B.3), (B.4),

(B.6), (B.5), (B.7) and (B.8), we can check that

$$\hat{\theta}_1 - \theta_1^* = O_p(T^{-1/2}). \tag{B.9}$$

From Lemma B.11 and Assumption B.3

$$\begin{aligned}
\mathcal{S}_{1\beta}(\theta_1^*) &= \frac{1}{\Delta}\sum_{i=1}^{n} \ell_{1\beta}^*(\Delta, x_i, y_i) \\
&= \frac{1}{\Delta}\sum_{i=1}^{n} \ell_{1\beta}^*(0, x_i, x_i) + \sum_{i=1}^{n} \mathcal{A}\ell_{1\beta}^*(0, x_i, x_i) + \frac{1}{\Delta}\sum_{i=1}^{n} \mathcal{B}\ell_{1\beta}^*(0, x_i, x_i)\mathcal{W}_{1i} \\
&\quad + \frac{\Delta}{2}\sum_{i=1}^{n} \mathcal{A}^2\ell_{1\beta}^*(0, x_i, x_i) + \frac{1}{\Delta}\sum_{i=1}^{n} \mathcal{B}\mathcal{A}\ell_{1\beta}^*(0, x_i, x_i)\mathcal{W}_{2i} + \frac{1}{\Delta}\sum_{i=1}^{n} \mathcal{A}\mathcal{B}\ell_{1\beta}^*(0, x_i, x_i)\mathcal{W}_{3i} \\
&\quad + \frac{1}{\Delta}\sum_{i=1}^{n} \mathcal{B}^2\ell_{1\beta}^*(0, x_i, x_i)\mathcal{W}_{4i} + \frac{1}{\Delta}\sum_{i=1}^{n} \mathcal{B}^3\ell_{1\beta}^*(0, x_i, x_i)\mathcal{W}_{5i} + o_p(T), \\
&= \dot{S}_{T,\Delta} + \ddot{S}_{1T,\Delta} + o_p(T),
\end{aligned} \tag{B.10}$$

where $\mathcal{W}_{ki}$'s are defined in (B.2), $\dot{S}_{T,\Delta} = O_p(\sqrt{T}/\Delta)$ and does not depend on μ_1, and

$$\ddot{S}_{1T,\Delta} = \frac{1}{2}\int_0^T \mathcal{A}^2\ell_{1\beta}^*(0, X_t, X_t)dt.$$

It is because we can set

$$\dot{S}_{T,\Delta} = \sum_{i=1}^{n} \mathcal{A}\ell_{1\beta}^*(0, x_i, x_i) = \frac{1}{\Delta}\int_0^T \mathcal{A}\ell_{1\beta}^*(0, X_t, X_t)dt + o_p(\sqrt{T}/\Delta)$$

and we have

$$\frac{\Delta}{2}\sum_{i=1}^{n} \mathcal{A}^2\ell_{1\beta}^*(0, x_i, x_i) = \ddot{S}_{T,\Delta} + o_p(T).$$

Similarly from Assumption B.3

$$
\begin{aligned}
\mathcal{H}_{1\beta\beta}(\theta_1^*) &= \frac{1}{\Delta}\sum_{i=1}^{n}\ell_{1\beta\beta}^*(\Delta, x_i, y_i) \\
&= \frac{1}{\Delta}\sum_{i=1}^{n}\ell_{1\beta\beta}^*(0, x_i, x_i) + \sum_{i=1}^{n}\mathcal{A}\ell_{1\beta\beta}^*(0, x_i, x_i) + \frac{1}{\Delta}\sum_{i=1}^{n}\mathcal{B}\ell_{1\beta\beta}^*(0, x_i, x_i)\mathcal{W}_{1i} \\
&\quad + \frac{\Delta}{2}\sum_{i=1}^{n}\mathcal{A}^2\ell_{1\beta\beta}^*(0, x_i, x_i) + \frac{1}{\Delta}\sum_{i=1}^{n}\mathcal{B}\mathcal{A}\ell_{1\beta\beta}^*(0, x_i, x_i)\mathcal{W}_{2i} + \frac{1}{\Delta}\sum_{i=1}^{n}\mathcal{A}\mathcal{B}\ell_{1\beta\beta}^*(0, x_i, x_i)\mathcal{W}_{3i} \\
&\quad + \frac{1}{\Delta}\sum_{i=1}^{n}\mathcal{B}^2\ell_{1\beta\beta}^*(0, x_i, x_i)\mathcal{W}_{4i} + \frac{1}{\Delta}\sum_{i=1}^{n}\mathcal{B}^3\ell_{1\beta\beta}^*(0, x_i, x_i)\mathcal{W}_{5i} + o_p(T), \\
&= \dot{H}_{T,\Delta} + \ddot{H}_{1T,\Delta} + o_p(T),
\end{aligned} \tag{B.11}
$$

where $\dot{H}_{T,\Delta} = O_p(T/\Delta)$ and does not depend on μ_1, and $\ddot{H}_{1T,\Delta} = O_p(T)$. Note that

$$
\dot{H}_{T,\Delta} = \frac{1}{\Delta}\int_0^T \mathcal{A}\ell_{1\beta\beta'}^*(0, X_t, X_t)dt + O_p(\sqrt{T}).
$$

Now from

$$
\mathcal{S}_1(\hat{\theta}_1) = \mathcal{S}_1(\theta_1^*) + \mathcal{H}_1(\theta_1^*)(\hat{\theta}_1 - \theta_1^*) + \big(\mathcal{H}_1(\tilde{\theta}_1) - \mathcal{H}_1(\theta_1^*)\big)(\hat{\theta}_1 - \theta_1^*),
$$

we have

$$
-\mathcal{H}_{1\beta\beta}(\theta_1^*)(\hat{\beta}_1 - \beta_1^*) = \mathcal{S}_{1\beta}(\theta_1^*) + \big(\mathcal{H}_{1\beta\beta}(\tilde{\theta}_1) - \mathcal{H}_{1\beta\beta}(\theta_1^*)\big)(\hat{\beta}_1 - \beta_1^*) + O_p(\sqrt{T})
$$

wpa1 from (B.6), (B.7), (B.8) and (B.9) with the same steps as in the proof of Lemma B.12. Denoting a differential operator w.r.t. any single element of α_1 as ∂_α,

$$
0 = \partial_\alpha\mathcal{S}_{1\beta}(\theta_1^*) + \mathcal{H}_{1\beta\beta}(\tilde{\theta}_1)\partial_\alpha\hat{\beta}_1 + \partial_\alpha\mathcal{H}_{1\beta\beta}(\tilde{\theta}_1)(\hat{\beta}_1 - \beta_1^*) + O_p(\sqrt{T}),
$$

and to avoid any contradiction, we should have $\partial_\alpha\hat{\beta}_1 = O_p(\Delta)$ from (B.9), (B.10) and (B.11). That is, $\hat{\beta}_1 - \beta_1^*$ does not depend on α_1 up to order Δ. Now note that

$$
\mathcal{H}_{1\beta\beta}(\tilde{\theta}_1) - \mathcal{H}_{1\beta\beta}(\theta_1^*) = \frac{\big((\tilde{\beta}_1 - \beta_1^*) \otimes \iota_{k_{\beta_1}}\big)'}{\Delta}\int_0^T \frac{\partial}{\partial\beta_1}\mathcal{A}\ell_{1\beta\beta'}^*(0, X_t, X_t; \check{\beta}_1)dt + O_p(T) \tag{B.12}
$$

since $\ell^*_{1\beta\beta\Delta}$ and $\ell^*_{1\beta\beta yy}$ are functions of only β_1, thus

$$
\begin{aligned}
\hat{\beta}_1 - \beta_1^* &= -\mathcal{H}_{1\beta\beta}(\theta_1^*)^{-1}\mathcal{S}_{1\beta}(\theta_1^*) - \mathcal{H}_{1\beta\beta}(\theta_1^*)^{-1}\big(\mathcal{H}_{1\beta\beta}(\tilde{\theta}_1) - \mathcal{H}_{1\beta\beta}(\theta_1^*)\big)(\hat{\beta}_1 - \beta_1^*) + O_p(\Delta T^{-1/2}) \\
&= -\mathcal{H}_{1\beta\beta}(\theta_1^*)^{-1}\mathcal{S}_{1\beta}(\theta_1^*) + Q_{T,\Delta} + O_p(\Delta T^{-1/2})
\end{aligned}
$$

with $Q_{T,\Delta} = O_p(T^{-1/2})$ and does not depend on α_1. From (B.10) and (B.11),

$$
\hat{\beta}_1 - \beta_1^* = -\dot{H}_{T,\Delta}^{-1}\dot{S}_{T,\Delta} + Q_{T,\Delta} - \dot{H}_{T,\Delta}^{-1}\ddot{S}_{T,\Delta} + o_p(\Delta),
$$

thus

$$
\hat{\beta}_1 - \beta_1^* = \dot{\beta}_{T,\Delta} + \ddot{\beta}_{1T,\Delta} + o_p(\Delta), \tag{B.13}
$$

where

$$
\begin{aligned}
\dot{\beta}_{T,\Delta} &= -\dot{H}_{T,\Delta}^{-1}\dot{S}_{T,\Delta} + Q_{T,\Delta} \\
&= -\bigg(\int_0^T \mathcal{A}\ell^*_{1\beta\beta'}(0, X_t, X_t)dt\bigg)^{-1} \int_0^T \mathcal{A}\ell^*_{1\beta}(0, X_t, X_t)dt + o_p(T^{-1/2})
\end{aligned}
$$

which does not depend on μ_1, and

$$
\ddot{\beta}_{1T,\Delta} = -\frac{\Delta}{2}\bigg(\int_0^T \mathcal{A}\ell^*_{1\beta\beta'}(0, X_t, X_t)dt\bigg)^{-1} \int_0^T \mathcal{A}^2\ell^*_{1\beta}(0, X_t, X_t)dt.
$$

Part 2. (Likelihood Asymptotics) We have

$$
\begin{aligned}
\mathcal{L}_1(\hat{\theta}_1) - \mathcal{L}_2(\hat{\theta}_2) = \mathcal{L}_1(\theta_1^*) - \mathcal{L}_2(\theta_2^*) &+ (\hat{\theta}_1 - \theta_1^*)'\mathcal{S}_1(\theta_1^*) - (\hat{\theta}_2 - \theta_2^*)'\mathcal{S}_2(\theta_2^*) \\
&+ \frac{1}{2}(\hat{\theta}_1 - \theta_1^*)'\mathcal{H}_1(\theta_1^*)(\hat{\theta}_1 - \theta_1^*) - \frac{1}{2}(\hat{\theta}_2 - \theta_2^*)'\mathcal{H}_2(\theta_2^*)(\hat{\theta}_2 - \theta_2^*) + O_p(1).
\end{aligned}
$$

This is because

$$
\begin{aligned}
(\hat{\alpha}_1 - \alpha_1^*)'\big(\mathcal{H}_{1\alpha\alpha}(\tilde{\theta}_1) - \mathcal{H}_{1\alpha\alpha}(\theta_1^*)\big)(\hat{\alpha}_1 - \alpha_1^*) = o_p(1) \\
(\hat{\alpha}_1 - \alpha_1^*)'\big(\mathcal{H}_{1\alpha\beta}(\tilde{\theta}_1) - \mathcal{H}_{1\alpha\beta}(\theta_1^*)\big)(\hat{\beta}_1 - \beta_1^*) = o_p(1)
\end{aligned}
$$

from the same step as in the proof of Lemma B.12, and

$$
(\hat{\beta}_1 - \beta_1^*)'\big(\mathcal{H}_{1\beta\beta}(\tilde{\theta}_1) - \mathcal{H}_{1\beta\beta}(\theta_1^*)\big)(\hat{\beta}_1 - \beta_1^*) - (\hat{\beta}_2 - \beta_2^*)'\big(\mathcal{H}_{2\beta\beta}(\tilde{\theta}_2) - \mathcal{H}_{2\beta\beta}(\theta_2^*)\big)(\hat{\beta}_2 - \beta_2^*) = O_p(1)
$$

from (B.12) and (B.13). For the Hessian terms above, we have

$$(\hat{\alpha}_1 - \alpha_1^*)'\mathcal{H}_{1\alpha\alpha}(\theta_1^*)(\hat{\alpha}_1 - \alpha_1^*) = O_p(1)$$
$$(\hat{\alpha}_1 - \alpha_1^*)'\mathcal{H}_{1\alpha\beta}(\theta_1^*)(\hat{\beta}_1 - \beta_1^*) = O_p(1)$$

and

$$\begin{aligned}
(\hat{\beta}_1 - \beta_1^*)'\mathcal{H}_{1\beta\beta}(\theta_1^*)(\hat{\beta}_1 - \beta_1^*) &- (\hat{\beta}_2 - \beta_2^*)'\mathcal{H}_{2\beta\beta}(\theta_2^*)(\hat{\beta}_2 - \beta_2^*) \\
&= \dot{\beta}_{T,\Delta}\dot{H}_{T,\Delta}(\ddot{\beta}_{1T,\Delta} - \ddot{\beta}_{2T,\Delta}) + (\ddot{\beta}_{1T,\Delta} - \ddot{\beta}_{2T,\Delta})\dot{H}_{T,\Delta}\dot{\beta}_{T,\Delta} + O_p(1) \\
&= \int_0^T \mathcal{A}\ell_{1\beta}^*(0, X_t, X_t)dt\left(\int_0^T \mathcal{A}\ell_{1\beta\beta'}^*(0, X_t, X_t)dt\right)^{-1} \times \\
&\quad \int_0^T \left[\mathcal{A}^2\ell_{1\beta}^*(0, X_t, X_t) - \mathcal{A}^2\ell_{2\beta}^*(0, X_t, X_t)\right]dt + o_p(\sqrt{T})
\end{aligned}$$

from (B.11) and (B.13). Also note that from (B.4) and (B.9),

$$(\hat{\theta}_1 - \theta_1^*)'\mathcal{S}_1(\theta_1^*) = (\hat{\beta}_1 - \beta_1^*)'\mathcal{S}_{1\beta}(\theta_1^*) + O_p(1)$$

thus from (B.10) and (B.13),

$$\begin{aligned}
(\hat{\theta}_1 - \theta_1^*)'\mathcal{S}_1(\theta_1^*) &- (\hat{\theta}_2 - \theta_2^*)'\mathcal{S}_2(\theta_2^*) \\
&= \dot{\beta}_{T,\Delta}\big(\ddot{S}_{1T,\Delta} - \ddot{S}_{2T,\Delta}\big) + \big(\ddot{\beta}_{1T,\Delta} - \ddot{\beta}_{2T,\Delta}\big)\dot{S}_{T,\Delta} + o_p(\sqrt{T}) \\
&= -\int_0^T \mathcal{A}\ell_{1\beta}^*(0, X_t, X_t)dt\left(\int_0^T \mathcal{A}\ell_{1\beta\beta'}^*(0, X_t, X_t)dt\right)^{-1} \times \\
&\quad \int_0^T \left[\mathcal{A}^2\ell_{1\beta}^*(0, X_t, X_t) - \mathcal{A}^2\ell_{2\beta}^*(0, X_t, X_t)\right]dt + o_p(\sqrt{T}).
\end{aligned}$$

For the asymptotics of $\mathcal{L}_1(\theta_1^*) - \mathcal{L}_2(\theta_2^*)$, rest of the steps are similar to the proof of Theorem 2.3. We get the stated result by applying the operators to each function and simplifying the equation. ■

Remark B.13 *The conditions $\Delta^3 T \to 0$ in Case 1 and $\Delta T^2 \to 0$ in Case 2 are not necessary and they are only technical conditions to make the proof simpler. We can make the order arbitrarily small if necessary, but only with higher order expansions in (B.4), for example. This also applies to the other theorems and corollaries.*

B.2.3 Proof of Corollary 2.4

Proof. It is straightforward from Theorem 2.3 using $\mathcal{L}_j(\hat{\theta}_j) - \mathcal{L}_j^0(\hat{\theta}_j^0)$ for $\mathcal{L}_1(\hat{\theta}_1) - \mathcal{L}_2(\hat{\theta}_2)$. ■

B.2.4 Proof of Theorem 2.5

Proof. Proof for (a) is straightforward by applying Lemma B.5 to Theorem 2.3. For (b), note that we can write

$$\mathcal{L}_1(\hat{\theta}_1) - \mathcal{L}_2(\hat{\theta}_2) = -\int_0^T \mathcal{G}_a(X_t)dW_t$$
$$+\frac{1}{2}\int_0^T \mathcal{G}_b(X_t)dW_t\left(\int_0^T \mathcal{A}\ell^*_{1\beta\beta'}(0,X_t,X_t)dt\right)^{-1}\int_0^T \left[\mathcal{A}^2\ell^*_{1\beta} - \mathcal{A}^2\ell^*_{2\beta}\right](0,X_t,X_t)dt + o_p(\sqrt{T})$$

from (B.1). Since

$$\frac{1}{\sqrt{T}}\left(\int_0^T \mathcal{G}_a(X_t)dW_t, \int_0^T \mathcal{G}_b'(X_t)dW_t\right)' \xrightarrow{d} \mathbf{N}(0,\Sigma)$$

from Theorem 4.1 of van Zanten (2000) and

$$\left(1, \frac{1}{2}\int_0^T \left[\mathcal{A}^2\ell^*_{1\beta} - \mathcal{A}^2\ell^*_{2\beta}\right]'(0,X_t,X_t)dt\left[\int_0^T \mathcal{A}\ell^*_{1\beta\beta'}(0,X_t,X_t)dt\right]^{-1}\right)' \to_{a.s.} C$$

as $T \to \infty$, the stated result easily follows by applying the operators to each function and simplifying the equation. ■

B.2.5 Proof of Corollary 2.6

Proof. The proof easily follows from Theorem 2.3 with the same steps as in the proof of Theorem 2.5, and the proof is omitted. ■

B.2.6 Proof of Theorem 2.9

Proof. We rewrite the partial sum $g_{[rn]}$ as

$$g_{[rn]} = \frac{T^{1/2}}{\kappa(T,\Delta)}\sum_{i=1}^{[rn]} u_i = \begin{cases} \frac{1}{\sqrt{n}}\sum_{i=1}^{[rn]} \Delta^{1/2}u_i & \text{(for Case 1)} \\ \frac{1}{\sqrt{n}}\sum_{i=1}^{[rn]} \Delta^{-1/2}u_i & \text{(for Case 2)} \end{cases}, \tag{B.14}$$

and let $v_i = \Delta^{1/2}u_i$ for Case 1 and $v_i = \Delta^{-1/2}u_i$ for Case 2. Then we define the sample autocovariance function

$$\hat{\gamma}_v(l\Delta) = \frac{1}{n}\sum_{i=|l|+1}^{n} (\hat{v}_i - \bar{v})(\hat{v}_{i-|l|} - \bar{v}), \tag{B.15}$$

where $\hat{v}_i = \Delta^{1/2}\hat{u}_i$ for Case 1, $\hat{v}_i = \Delta^{-1/2}\hat{u}_i$ for Case 2, and $\bar{v} = n^{-1}\sum_{i=1}^{n}\hat{v}_i$. Note that we have the relationship,

$$\hat{\gamma}_v(l\Delta) = \begin{cases} \Delta\hat{\gamma}(l\Delta) & \text{(for Case 1)} \\ \Delta^{-1}\hat{\gamma}(l\Delta) & \text{(for Case 2)} \end{cases}. \tag{B.16}$$

The long-run variance estimator $\hat{\Omega}$ of the numerator $g_{[1n]}$ of t_n is given by

$$\hat{\Omega} = \sum_{l=1-n}^{n-1} k\left(\frac{l\Delta}{M}\right)\hat{\gamma}_v(l\Delta) = \begin{cases} \Delta\sum_{l=1-n}^{n-1} k\left(\frac{l\Delta}{M}\right)\hat{\gamma}(l\Delta) & \text{(for Case 1)} \\ \Delta^{-1}\sum_{l=1-n}^{n-1} k\left(\frac{l\Delta}{M}\right)\hat{\gamma}(l\Delta) & \text{(for Case 2)} \end{cases}. \tag{B.17}$$

Since $g_{[rn]} = n^{-1/2}\sum_{i=1}^{[rn]} v_i$ satisfies the functional central limit theorem (Assumption 2.8), the limiting distribution of $\hat{\Omega}$ is given by Theorem 1 in Kiefer and Vogelsang (2005) replacing $g_{[rT]}$ and Assumption 2 in Kiefer and Vogelsang (2005) with $g_{[rn]}$ and our Assumption 2.8 respectively. Although v_i depends on Δ, we use the high level assumption on v_i to apply the results of Kiefer and Vogelsang (2005). Therefore we have the test statistics t_n given by

$$t_n = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{[rn]}\hat{v}_i}{\sqrt{\hat{\Omega}}} = \begin{cases} \frac{\Delta}{\sqrt{T}}\sum_{i=1}^{n}\hat{u}_i \Big/ \sqrt{\Delta\sum_{l=1-n}^{n-1} k\left(\frac{l\Delta}{M}\right)\hat{\gamma}(l\Delta)} & \text{(for Case 1)} \\ \frac{1}{\sqrt{T}}\sum_{i=1}^{n}\hat{u}_i \Big/ \sqrt{\Delta^{-1}\sum_{l=1-n}^{n-1} k\left(\frac{l\Delta}{M}\right)\hat{\gamma}(l\Delta)} & \text{(for Case 2)} \end{cases} \tag{B.18}$$

$$= \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{u}_i}{\sqrt{\sum_{l=1-n}^{n-1} k\left(\frac{l\Delta}{M}\right)\hat{\gamma}(l\Delta)}} \text{ (for Case 1 and Case 2).} \tag{B.19}$$

■

C Tables

Case 1	Example 1 (Only the diffusion functions are misspecified)
	$M_0 : dX_t = 0.2\ (0.07 - X_t)dt + 0.07\sqrt{X_t}\ dW_t$ (CIR)
	$M_1 : dX_t = \kappa(\mu - X_t)dt + \sigma X_t^{0.9}\ dW_t$
	$M_2 : dX_t = \kappa(\mu - X_t)dt + \sigma X_t^{0.0785}\ dW_t$
	Example 2 (Both the drift and diffusion functions are misspecified)
	$M_0 : dX_t = 0.2\ (0.07 - X_t)dt + 0.07\sqrt{X_t}\ dW_t$ (CIR)
	$M_1 : dX_t = \kappa(0.0726 - X_t)X_t dt + \sigma X_t^{1.5}\ dW_t$ (Ahn and Gao)
	$M_2 : dX_t = \kappa(0.05 - X_t)X_t dt + \sigma X_t^{-0.648}\ dW_t$
Case 2	Example 3 (Only the drift functions are misspecified)
	$M_0 : dX_t = -X_t dt + 0.04\ dW_t$ (Vasicek)
	$M_1 : dX_t = \kappa(0.01 - X_t)dt + \sigma dW_t$
	$M_2 : dX_t = \kappa(-0.01 - X_t)dt + \sigma dW_t$.
	Example 4
	$M_0 : dX_t = 0.2\ (0.07 - X_t)dt + 0.07\sqrt{X_t}\ dW_t$ (CIR)
	$M_1 : dX_t = \kappa(0.0780 - X_t)X_t dt + \sigma X_t^{1.5}\ dW_t$ (Ahn and Gao)
	$M_2 : dX_t = \kappa(0.0888 - X_t)dt + \sigma X_t^{1.5}\ dW_t$ (CKLS,CEV $\rho = 1.5$)

Table 1: Four examples for the Monte Carlo experiments of our model selection tests. M_0 is the true process, M_j for $j = 1, 2$, are competing nonnested misspecified models. In Case 1, models have different diffusion functions, and In Case 2, they have the identical diffusion functions.

	$T = 5$				$T = 40$			
	$\tau_{T,\Delta}$	$t_n(0.43)$	$t_n(0.85)$	$t_n(1.28)$	$\tau_{T,\Delta}$	$t_n(0.85)$	$t_n(1.70)$	$t_n(2.56)$
Case 1								
Example 1								
Sub N	18.9	4.4	3.0	2.2	14.2	6.5	6.2	5.6
Sub Emp	18.4	5.9	4.7	3.4	14.3	7.7	8.2	7.4
Bootstrap	7.0	6.9	6.0	6.0	7.2	7.6	7.8	6.7
N(0,1)	N/A	9.0	11.8	14.2	N/A	7.2	7.1	7.7
Fixed-b	N/A	5.4	4.9	4.4	N/A	6.1	4.9	4.8
Example 2								
Sub N	25.9	5.7	5.2	3.4	31.7	12.9	10.7	9.9
Sub Emp	26.1	8.1	7.1	5.4	31.5	14.1	11.5	10.3
Bootstrap	12.5	8.7	7.9	6.9	22.0	18.2	15.0	12.9
N(0,1)	N/A	11.5	14.0	15.7	N/A	29.9	23.8	21.0
Fixed-b	N/A	7.3	5.1	5.2	N/A	28.6	20.7	16.4
Case 2								
Example 3								
Sub N	1.5	7.4	6.7	5.2	2.7	3.7	2.8	2.9
Sub Emp	3.2	11.6	10.2	8.7	4.0	5.0	4.6	4.9
Bootstrap	0.4	5.5	7.3	7.3	5.1	8.2	8.4	8.3
N(0,1)	N/A	7.2	12.8	18.1	N/A	5.5	8.8	10.3
Fixed-b	N/A	4.1	5.6	5.7	N/A	4.6	7.1	8.1
Example 4								
Sub N	1.9	5.2	3.9	3.0	5.4	6.6	7.9	8.5
Sub Emp	2.7	7.9	6.9	5.7	6.5	11.2	10.8	11.0
Bootstrap	0.6	4.8	4.6	4.5	1.7	5.6	7.1	8.1
N(0,1)	N/A	1.4	3.2	5.6	N/A	2.1	4.7	7.2
Fixed-b	N/A	0.7	0.9	1.3	N/A	1.6	3.5	5.2

Table 2: Size of the tests. The rejection rates (%) are from the Monte Carlo experiments for the two sided, 5% level tests with daily observations. The number of simulation iterations is $1,000$. The statistic $\tau_{T,\Delta}$ is the non-pivotal log-likelihood ratio statistics. $t_n(M)$ is the pivotal statistic with the HAC variance estimator using the Bartlett kernel with a bandwidth of M years. The subsampling method is based on 199 blocks of equal size $S = T^{0.4}$ for $T = 5$ and $S = T^{0.7}$ for $T = 40$. "Sub N" implies the subsampling approximations by fitting the normal distribution with the sample mean and variance of the subsample statistics, and "Sub Emp" means that we have used the empirical distribution of the subsample statistics directly. The block bootstrap ("bootstrap") is based on 399 bootstrap repetitions with a block length $l = T/25$. "Fixed-b" uses the critical values from the fixed-b asymptotic approximations.

Case 1	Example 1: M_1 is preferred
	$M_0 : dX_t = 0.2\ (0.07 - X_t)dt + 0.07\ X_t^{\mathit{0.6}}\ dW_t$
	Example 2: M_2 is preferred
	$M_0 : dX_t = 0.2\ (0.07 - X_t)dt + 0.07\ X_t^{\mathit{0.45}}\ dW_t$
Case 2	Example 3: M_1 is preferred
	$M_0 : dX_t = (\mathit{0.01} - X_t)dt + 0.04\ dW_t$
	Example 4: M_1 is preferred
	$M_0 : dX_t = 0.2\ (\mathit{0.05} - X_t)dt + 0.07\sqrt{X_t}\ dW_t$

Table 3: Specifications of true processes for power simulations. The italic numbers show the modifications of the DGPs in the size simulations.

	$T = 5$				$T = 40$			
	$\tau_{T,\Delta}$	$t_n(0.43)$	$t_n(0.85)$	$t_n(1.28)$	$\tau_{T,\Delta}$	$t_n(0.85)$	$t_n(1.7)$	$t_n(2.56)$
Case 1								
Example 1	5.98	11.78	10.38	10.20	62.90	89.10	86.94	83.74
Example 2	10.52	7.92	7.14	6.58	19.26	19.32	20.30	19.96
Case 2								
Example 3	4.66	5.16	5.68	5.20	19.34	31.38	32.70	32.18
Example 4	2.54	5.94	6.76	6.78	19.62	21.52	21.96	22.10

Table 4: Size corrected power of the tests. Rejection rates (%) are from the Monte Carlo experiments for the size corrected power of the two sided, 5% level tests based on $5,000$ simulation iterations.. Size corrected critical values are from the 2.5% and 97.5% quantiles of $5,000$ repetitions of simulations under the null hypotheses.

Spot rates ($T = 37$ years, $n = 9,411$)				
Statistic	$\tau_{T,\Delta}$	$t_n(0.83)$	$t_n(1.67)$	$t_n(2.50)$
	27.620	3.542	2.737	2.437
Sub N	AG	0	0	0
Sub Emp	AG	0	0	0
Bootstrap	0	0	0	0
$N(0,1)$	N/A	AG	AG	AG
Fixed-b	N/A	AG	AG	AG

Exchange rates ($T = 10$ years, $n = 2,588$)				
Statistic	$\tau_{T,\Delta}$	$t_n(0.54)$	$t_n(1.08)$	$t_n(1.62)$
	-1.051	-0.837	-0.754	-0.775
Sub N	0	0	0	0
Sub Emp	CIR	0	0	0
Bootstrap	0	0	0	0
$N(0,1)$	N/A	0	0	0
Fixed-b	N/A	0	0	0

Table 5: Two candidate models (Ahn and Gao (AG) and CIR models) are compared for the spot rate and the exchange rate data. Four different statistics are used. $\tau_{T,\Delta}$ is non-pivotal, $t_n(M)$ is using the HAC variance estimator with a bandwidth parameter M years. The sampling distributions of the test statistics are approximated by the subsampling (199 blocks with size $T^{0.7}$ which is 12.5 and 5.01 years for the spot and the exchange rates respectively) with the fitted normal distribution ("Sub N") or the empirical distribution ("Sub Emp"), the block bootstrap ("Bootstrap") with 25 equal-sized blocks and 399 bootstrap repititions, and the standard normal distribution $N(0,1)$ (for asymptotically pivotal statistics only). "AG" or "CIR" represents the superiority of the respective model; "0" represents failing to reject.

D Figures

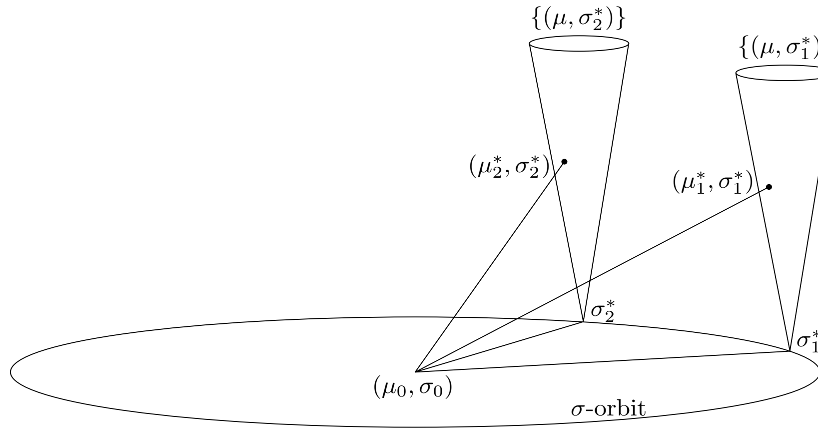

Figure 1: Two diffusion models (μ_1^*, σ_1^*) and (μ_2^*, σ_2^*) at their pseudo-true parameter values are considered to be equivalent when their diffusion functions σ_1^* and σ_2^* have equal divergence from the true diffusion function σ_0, i.e. they are on the same σ-orbit. If the models are equivalent, they can be distinguished further by the divergence of their drift functions measured by the vertical elevation from the σ-orbit. The cones are diffusions with the same diffusion functions. The null hypothesis of our model selection test is that the models are equivalent in terms of both the diffusion functions (on the same σ-orbit) and the drift functions (of equal elevation).

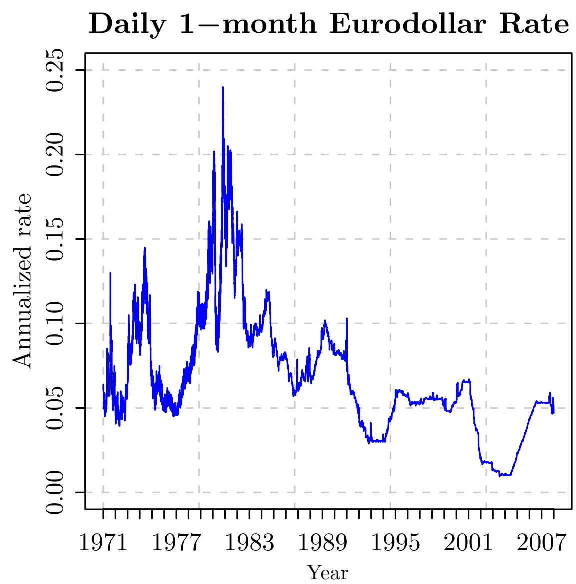

Figure 2: Annualized daily 1-month Eurodollar deposit rates from 01/01/1971 ∼ 12/31/2007 (business days only)

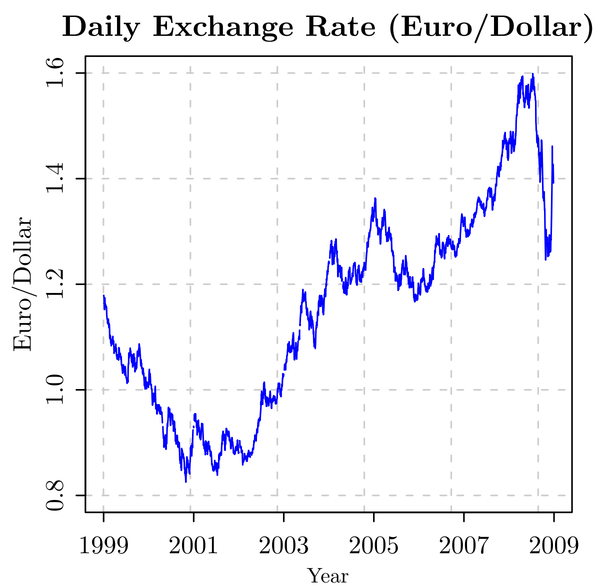

Figure 3: Daily Euro/Dollar exchange rates from 01/01/1999 ∼ 12/31/2008 (business days only)