Information Theory for Maximum Likelihood Estimation of Diffusion Models

Hwan-sik Choi
Binghamton University
Mar. 2014

Abstract

We develop an information theoretic framework for maximum likelihood estimation of diffusion models. Two information criteria that measure the divergence of a diffusion process from the true diffusion are defined. The maximum likelihood estimator (MLE) converges asymptotically to the limit that minimizes the criteria when sampling interval decreases as sampling span increases. When both drift and diffusion specifications are correct, the criteria become zero and the inverse curvatures of the criteria equal the asymptotic variance of the MLE. For misspecified models, the minimizer of the criteria defines pseudo-true parameters. Pseudo-true drift parameters depend on approximate transitions if used.

JEL: C13, C22, C40.
Keywords: Kullback-Leibler information criterion, fiber bundle, high frequency, infinitesimal generator, infill asymptotics, misspecification.
1 Introduction

We develop an information theoretic framework for maximum likelihood estimation of diffusion models in an asymptotic environment in which sampling interval shrinks to zero as sampling span increases to infinity. The new approach provides a unified framework within which we can analyze both correctly specified and misspecified diffusion models.

For discrete time models, White (1982) gives an information theoretic foundation for analyzing the maximum likelihood estimator (MLE). His theory is suitable for the models analyzed with the conventional large sample theory with fixed sampling interval. For diffusion models, estimation and inference under decreasing sampling interval were studied previously by, for example, Bandi and Phillips (2003), Jeong and Park (2013), and Choi, Jeong, and Park (2013). Bandi and Phillips (2003) study fully nonparametric estimation. Jeong and Park (2013) study the maximum likelihood estimation for correctly specified parametric models for both stationary and non-stationary processes. Choi, Jeong, and Park (2013) provide some asymptotic results for misspecified models, although their main focus is on model selection testing.

The theory developed in Jeong and Park (2013) and Choi, Jeong, and Park (2013) can explain the differential properties of the drift and diffusion parameter estimators observed in high frequency samples. They show that the rates of information accumulation for drift and diffusion parameters are different, and both sample size and sampling span are important for diffusion parameter estimation whereas sampling span only matters for drift parameter estimation. The present paper provides an information theoretic foundation for their new asymptotic theory.

In White’s theory, the Kullback-Leibler information criterion (KLIC, Kullback and Leibler (1951)) plays a key role in understanding the asymptotic properties of the MLE. He shows that the MLE converges to the limit that minimizes the KLIC from the true distribution and derives asymptotic properties of the MLE under model misspecification. He also shows that the usual inference procedure with the Likelihood Ratio, Lagrange Multiplier, and Wald test statistics is not valid for a misspecified model. Naturally, we are interested in the questions similar to those that White has addressed in his paper. Under the new asymptotic framework with decreasing sampling interval, do the MLEs of misspecified diffusion models converge to some limits asymptotically, do the limits have any meaning, are the usual inference procedures based on likelihood ratios valid, what is the consequence of partial misspecification, what is the impact of having to use an approximate
transition when the exact transition is unknown in closed-form? We provide an answer to each of these questions in this paper.

Based on the KLIC, we develop two-tier information criteria, which we refer to as the primary and secondary information criteria. The criteria measure the divergence of a diffusion process $M$ from the true diffusion $X$. The primary criterion depends on the diffusion function of $M$, and equals zero if and only if the diffusion function is the same as the diffusion function of $X$. The secondary criterion becomes zero if both drift and diffusion functions of $M$ are correct.

Our new information criteria play a similar role as the KLIC in White (1982). When sampling interval decreases as sampling span increases, the MLE asymptotically minimizes the two criteria, and the minimizer of the criteria is said to be the pseudo-true parameter. Because the primary criterion does not depend on the drift specification of $M$, the MLE for a diffusion function parameter is consistent to the same limit regardless of drift specifications. The secondary criterion depends on both diffusion and drift specification. When the exact transition is not known in closed-form and an approximate transition is used, the secondary criterion depends on the choice of an approximate transition as well. Therefore, the pseudo-true drift parameter depends on diffusion and drift function specifications and the choice of approximate transitions if used.

Under the partial misspecification where the diffusion function is correctly specified, the pseudo-true values of drift function parameters become invariant for a large class of approximate transitions. Thus, the MLE of drift parameters would be less sensitive to the choice of an approximate transition when the diffusion function is correctly specified. When the drift function is correctly specified but the diffusion function is misspecified, the drift estimator does not converge to the true value in general. However, we show that the Euler approximate transition has a robustness property in the sense that the MLE of the drift parameters converges asymptotically to the true value. For the Milstein approximate transition and the closed-form approximate transition of Aït-Sahalia (2002), the MLE of the drift parameter may not converge to the true parameter value.

We also study the asymptotic distribution of the MLE for both correctly specified and misspecified models. When a diffusion model is correctly specified, we show that the asymptotic variances of the MLE of diffusion and drift parameters are given by the inverse curvatures of the primary and secondary criteria respectively. For misspecified models, the asymptotic variance of the MLE depends on both expected Hessian matrix and covariance matrix of the score functions.
We develop the new information criteria in the next section, and give information theoretic interpretation of the asymptotic properties of the MLE in Section 3.

2 Divergence of diffusion processes

Let $W$ be the standard Brownian motion defined on a probability space $(\Omega, \mathcal{F}, P)$, and $X$ be a stationary diffusion on $\mathcal{D} \subset \mathbb{R}$ that solves the stochastic differential equation (SDE)

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

with a drift function $\mu_0(\cdot)$ and a diffusion function $\sigma_0(\cdot)$. We assume that SDEs in this paper satisfy the conditions in Karatzas and Shreve (1991) to admit a weak solution. Suppose we have a parametric diffusion model $M(\theta)$ for $X$ given by

$$M(\theta) : dX_t = \mu(X_t; \alpha) \, dt + \sigma(X_t; \beta) \, dW_t,$$

where $\mu(\cdot; \alpha)$ and $\sigma(\cdot; \beta) > 0$ are known functions with an unknown parameter vector $\theta = (\alpha, \beta)$ in a compact set $\Theta \subset \mathbb{R}^k$. A diffusion model $M(\theta)$ is misspecified if there exists no $\theta \in \Theta$ such that $\mu_0(\cdot) = \mu(\cdot; \alpha)$ and $\sigma_0(\cdot) = \sigma(\cdot; \beta)$ on $\mathcal{D}$.

Let $p_0(t, x, y)$ be the transition density from $X_0 = x$ to $X_t = y$ of the true process in (2.1) and $p(t, x, y; \theta)$ be the exact or approximate transition density of $M(\theta)$. We consider the KLIC from the transition density $p_0(t, x, y)$ to $p(t, x, y; \theta)$. The KLIC of $P$ to $Q$, the divergence measure

$$D_t(\theta) \equiv \mathbb{E} \left( \log \frac{p_0(t, X_0, X_t)}{p(t, X_0, X_t; \theta)} \right),$$

as a function of $t$ given $\theta$. The function $D_t(\theta)$ characterizes the divergence from the true process $X$ to $M(\theta)$ at various transition intervals. For instance, $D_t(\theta)$ for small $t$ would measure how close the model transition $p(t, x, y; \theta)$ is from the true transition $p_0(t, x, y)$ for observations $X_0$ and $X_t$ obtained over a small time interval $t$. In the limit of $t \to \infty$,
For $M(\theta) : dX_t = 0.2(\alpha - X_t)dt + \beta\sqrt{X_t}dW_t$, the two curves show $D_t(\theta)$ from the true diffusion $\theta = (\alpha, \beta) = (0.06, 0.1)$ to two diffusion processes $\theta = (0.02, 0.11)$ and $\theta = (0.08, 0.07)$. For small $t$, the process with $\theta = (0.02, 0.11)$ is closer to the true process, but $\theta = (0.08, 0.07)$ is closer for large $t$.

$D_\infty(\theta)$ would measure the divergence of the marginal densities of $X$ and $M(\theta)$.

Since $D_t(\theta)$ depends on the transition interval $t$ given $\theta$, the proximity of two processes $X$ and $M(\theta)$ in $D_t(\theta)$ should be understood separately for different $t$. In this sense, we may consider the combination of $M(\theta)$ and $t$ as a model if sampling occurs at a specific time interval only. Consequently, for the discrete time analysis with fixed sampling interval, it is sufficient to consider $D_t(\theta)$ at a specific $t$ as the only relevant criterion. In this paper, however, we are particularly interested in the asymptotic framework in which $t \to 0$ since the exact transition of a diffusion model is rarely known in closed-form and all approximate transitions would be valid only if $t \to 0$. This is also one of the reasons that diffusion models are mostly used for high frequency samples.

In Figure 1, we give two examples of the divergence measure $D_t(\theta)$ calculated with the CIR process (Cox, Ingersoll, and Ross (1985))

$$M(\theta) : dX_t = 0.2(\alpha - X_t)dt + \beta\sqrt{X_t}dW_t$$

with a two dimensional parameter vector $\theta = (\alpha, \beta)$. The true process $X$ is defined by $\theta = (\alpha, \beta) = (0.06, 0.1)$, and $D_t(\theta)$ from $X$ to the diffusion processes with $\theta = (0.02, 0.11)$
and $\theta = (0.08, 0.07)$ are shown in Figure 1. The process with $\theta = (0.02, 0.11)$ has the diffusion function parameter close to the true value $\beta = 0.1$ whereas $\theta = (0.08, 0.07)$ has the drift function parameter close to the true parameter $\alpha = 0.06$ compared to the other. The figure shows that the process with $\theta = (0.02, 0.11)$ is closer to $X$ in $D_t(\theta)$ at small $t$. As $t$ increases, the process $\theta = (0.08, 0.07)$ becomes closer to the true process.

To see the relative effect of $\alpha$ and $\beta$ on $D_t(\theta)$ more clearly as $t$ changes, we compare the contour plots of $D_t(\theta)$ for various $\theta$ in Figure 2 under quarterly ($t = 1/4$), monthly ($t = 1/12$), weekly ($t = 1/52$), and daily ($t = 1/250$) transition intervals. The true process $X$ is shown as the dot at the center of the plot. Each contour level curve represents the diffusion processes with equal $D_t(\theta)$ from $X$. From this figure, we can notice that the drift function parameter $\alpha$ (horizontal moves) and the diffusion function parameter $\beta$ (vertical moves) influence the divergence differently depending on the transition interval. Given $\beta$, a fixed amount of deviation of $\alpha$ from its true value 0.06 increases the divergence very much with quarterly transition interval, but the increase is quite small under daily transition interval. On the other hand, deviation of $\beta$ from its true value 0.1 has similar impact on $D_t(\theta)$ at all four transition intervals. This implies that the divergence measure $D_t(\theta)$ is more sensitive to diffusion function specification than drift function specification as transition interval decreases. This example motivates that it is crucial to study the behavior of $D_t(\theta)$ with small $t$ to understand the asymmetric impact of drift and diffusion specifications on $D_t(\theta)$ for high frequency data. From this motivation, we define the primary criterion as follows.

**Definition 2.1 (Primary Criterion).** The primary criterion is the limit

$$C^P(\theta) \equiv \lim_{t \to 0} D_t(\theta)$$

of the divergence function $D_t(\theta)$ as $t \to 0$.

We show that $C^P(\theta)$ plays a crucial role in analyzing the MLE under the asymptotic framework in which the sampling interval decreases to zero. To ensure that $C^P(\theta)$ is well defined, we develop our theory under the following assumptions. Let $l(t, x, y) = \log p(t, x, y)$ be a generic log transition density of $M(\theta)$ for any $\theta \in \Theta$ or of the true process $X$ with a drift function $\mu(\cdot)$ and a diffusion function $\sigma(\cdot)$. We consider the scaled log transition density

$$\tilde{l}(t, x, y) \equiv t(l(t, x, y) + \log \sqrt{t}), \text{ for } t > 0. \quad (2.3)$$
Figure 2: Contour plots of the divergence measure $D_t(\theta)$ defined in \cite{2.2} from the true process $X$ to $M(\theta) : dX_t = 0.2(\alpha - X_t)dt + \beta \sqrt{X_t}dW_t$. The true process is defined by $\theta = (\alpha, \beta) = (0.06, 0.1)$ and shown as the dot at the center. Each level represents the processes in $M(\theta)$ with equal divergence from $X$ under quarterly ($t = 1/4$), monthly ($t = 1/12$), weekly ($t = 1/52$), and daily ($t = 1/250$) transition intervals.
Although the transition $p(t, x, y)$ is not well defined at $t = 0$, we extend the domain of the scaled function $\tilde{l}(t, x, y)$ to $t \geq 0$ by defining its value at $t = 0$ as $\lim_{t \to 0} \tilde{l}(t, x, y)$ which is assumed to exist by the following assumption. The derivatives of $\tilde{l}(t, x, y)$ at $t = 0$ are defined similarly.

**Assumption 1.** $\tilde{l}(t, x, y), \mu(x)$ and $\sigma(x)$ are infinitely differentiable in $t \geq 0$, $x, y \in \mathcal{D}$ and $\theta \in \Theta$. Let $h(t, x, y)$ be one of these functions or their derivative functions. Then $h(t, x, y)$ is dominated by $g$ in the sense that

$$|h(t, x, y)| \leq g(x)g(y),$$

for all small $t \geq 0$, $y, x \in \mathcal{D}$, and $\theta$ in the interior of $\Theta$. Moreover, $g$ is locally bounded, increasing at a polynomial rate of order $p$ on the boundary of $\mathcal{D}$, and $\mathbb{E}g(X_t)^{2d} < \infty$, where $d$ is the dimension of $\theta$.

**Assumption 2.** There exists $q > 0$ such that

$$T^{-q} \sup_{t \in [0, T]} |X_t| \overset{p}{\to} 0$$

as $T \to \infty$. If $\mathcal{D} = (0, \infty)$, we further assume that

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

With the above technical conditions, we introduce the following notation for simplicity of exposition. Let $f(t, x, y)$ be a function that satisfies Assumption 1. We denote the functions

$$f_{/0}(x) \equiv f(0, x, x), \quad f_{/t}(x) \equiv \frac{\partial f(0, x, x)}{\partial t},$$

with the single argument $x$ as the function $f(t, x, y)$ and its derivative with respect to $t$ at $y = x$ and $t = 0$. Other higher order derivatives such as $f_{/y}(x), f_{/yy}(x)$, or $f_{/yt}(x)$ are defined similarly. Using this notation, for the scaled log transition density $\tilde{l}(t, x, y)$ in (2.3), we can write $\tilde{l}_{/0}(x) = \tilde{l}(0, x, x)$, $\tilde{l}_{/t}(x) = \partial \tilde{l}(0, x, x)/\partial t$, $\tilde{l}_{/y}(x) = \partial \tilde{l}(0, x, x)/\partial y$, and other derivatives similarly. We also assume the following for $\tilde{l}(t, x, y)$ and its derivatives at $y = x$ and $t = 0$. 

8
Assumption 3. We have the following limits for $\bar{l}(t, x, y)$.

\[
\begin{align*}
\bar{l}_{/0}(x) &= 0, & \bar{l}_{/t}(x) &= -\frac{1}{2} \log \sigma^2(x) - \frac{1}{2} \log(2\pi), \\
\bar{l}_{/y}(x) &= 0, & \bar{l}_{/yy}(x) &= -\frac{1}{\sigma^2(x)}.
\end{align*}
\]

Assumption 4. The functions $\bar{l}_{/yyy}(x)$ and $\bar{l}_{/yyyy}(x)$ depend on $\sigma(x)$ but not on $\mu(x)$.

Assumptions 3 and 4 can be shown to be satisfied for the exact transition density always. When the exact transition must be approximated because it is not available in closed-form, we can show that the Euler and Milstein approximate transitions and the closed-form approximate transition of Aït-Sahalia (2002) satisfy them also. See Lemma A.5 of Jeong and Park (2013) for details.

For expositional simplicity, we write $f \equiv f(X_0)$ for any function $f(\cdot)$ with a single argument. For example, we write $\bar{l}_{/0} = \bar{l}_{/0}(X_0)$, $\mu = \mu(X_0)$ and $\sigma = \sigma(X_0)$. Also, we suppress $\theta$ for the functions such as the log-likelihoods and the drift and diffusion functions of $M(\theta)$. With this notational convention, we derive the explicit form of $C^P$ in the following theorem. All proofs are in Appendix.

**Theorem 2.1.** Let

\[
\mathcal{A} = \frac{\partial}{\partial t} + \mu_0(y) \frac{\partial}{\partial y} + \frac{\sigma_0^2(y)}{2} \frac{\partial^2}{\partial y^2}
\]

be the infinitesimal generator of the true process $X$. Given $X_0 = x$, define

\[
f(t, x, y) \equiv \bar{l}_0(t, x, y) - \bar{l}(t, x, y),
\]

where $\bar{l}_0(t, x, y)$ and $\bar{l}(t, x, y)$ are the scaled log-likelihoods defined in (2.3) of the true process $X$ and the model $M(\theta)$ respectively. If Assumptions 3 and 4 are satisfied, then we have

\[
C^P(\theta) = \mathbb{E}(Af)/0 = \frac{1}{2} \mathbb{E} \left[ -\log \frac{\sigma_0^2}{\sigma^2} + \left( \frac{\sigma_0^2}{\sigma^2} - 1 \right) \right].
\]

Note that $(Af)/0 = (Af)/0(X_0)$, $\sigma_0^2 = \sigma_0^2(X_0)$ and $\sigma^2 = \sigma^2(X_0; \beta)$ from our notational convention. The theorem shows that the drift specification of $M(\theta)$ becomes irrelevant for $D_1(\theta)$ as $t \to 0$ because the primary criterion $C^P(\theta)$ does not depend on the drift function.
parameter $\alpha$. This result also supports our observation from Figure 2. We will denote $C_P(\theta)$ as $C_P(\beta)$ in the rest of this paper.

There is a convenient way to interpret the primary criterion. Suppose that $Y(\beta)$ is a random variable of which the distribution is the marginal distribution of the diffusion process

$$r(X_t; \beta) \equiv \frac{\sigma_0^2(X_t)}{\sigma^2(X_t; \beta)}.$$ 

Then we can write

$$C_P(\beta) = \frac{1}{2} \mathbb{E}\{(Y(\beta) - 1) - \log Y(\beta)\}.$$ 

Intuitively, $C_P(\beta)$ measures the expected error from approximating $\log Y$ with its linear approximation $Y - 1$ for given $\beta$. For a correctly specified diffusion function with $\sigma_0^2(\cdot) = \sigma^2(\cdot; \beta_0)$, we would have $C_P(\beta_0) = 0$ since $r(\cdot; \beta_0) = 1$.

For given $\beta = \beta^*$, let $M_{\beta^*}(\alpha) \subset M(\theta)$ be the sub-model defined on the subset $\Theta_{\beta^*} = \{(\alpha, \beta) \in \Theta | \beta = \beta^*\}$ of $\Theta$. Theorem 2.1 implies that the primary criterion $C_P(\beta)$ is invariant on $\Theta_{\beta}$. Therefore, if there exists $\beta^*$ that minimizes $C_P(\beta)$, all the diffusion processes in $M_{\beta^*}(\alpha)$ would minimize the primary criterion and be considered to be closest to $X$. We define the secondary divergence measure that depends on both $\alpha$ and $\beta$ to further distinguish the members of $M_{\beta^*}(\alpha)$.

**Definition 2.2 (Secondary Criterion).** Let

$$D_t'(\theta) \equiv \frac{\partial D_t(\theta)}{\partial t}$$

be the derivative of $D_t(\theta)$ at $t$. The secondary criterion is the limit

$$C_S(\theta) \equiv \lim_{t \to 0} D_t'(\theta)$$

as $t \to 0$.

We derive the explicit form of $C_S(\theta)$ to understand the role of diffusion and drift specifications in determining the secondary divergence of $M(\theta)$ from the true process $X$. Let $\mu_{/x}(x) \equiv \partial \mu(x)/\partial x$ and $\sigma_{/x}^2(x) \equiv \partial \sigma^2(x)/\partial x$, and denote $\mu_{/x} \equiv \mu_{/x}(X_0)$ and $\sigma_{/x}^2 \equiv \sigma_{/x}^2(X_0)$. Other functions and their derivatives without arguments are defined in the same manner.
Theorem 2.2. Suppose Assumptions 1-4 are satisfied. We have

\[ C^S(\theta) = \frac{1}{2} \mathbb{E}(A^2 f)/0 \]

\[ = \frac{1}{2} \mathbb{E}\left\{ \left( \mu_0 + \frac{\sigma^2_0}{2} \right) + \left( \sigma^2_0 \left( \frac{\mu_0}{x} + \frac{\sigma_0^2}{4} \right) \right) \left( -\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2} \right) \right\} \tag{2.7} \]

\[ + \sigma^2_0 \left( \mu_0 + \frac{\sigma^2_0}{2} \right) \left( \tilde{l}_{y/y} - \tilde{l}_{yy} \right) + \frac{\sigma_0^4}{4} \left( \tilde{l}_{y/y} - \tilde{l}_{y/yy} \right) \]

\[ + 2\mu_0 (\tilde{l}_{0/yt} - \tilde{l}_{yt}) + (\tilde{l}_{0/yt} - \tilde{l}_{yt}) + \sigma^2_0 (\tilde{l}_{0/yyt} - \tilde{l}_{yyt}) \],

where \( A \) is the infinitesimal generator defined in (2.4), the function \( f \) is the scaled log-likelihood ratio defined in (2.5) with the exact transition densities of \( X \) and \( M(\theta) \), and the derivatives of \( \tilde{l}_0 \) are similarly defined with \( \mu_0(x) \) and \( \sigma^2_0(x) \).

When an approximate transition is used, the exact form of \( C^S(\theta) \) would depend on the approximation method generally because the derivatives in (2.8)-(2.10) depend on the local properties of the transition of \( M(\theta) \) around \( y = x \) and \( t = 0 \). However, it is shown in Jeong and Park (2010) that these derivatives have identical forms for the exact transition and Aït-Sahalia’s closed-form approximate transition. Therefore, Aït-Sahalia’s closed-form approximation is sufficiently accurate to give the same \( C^S(\theta) \) as the exact transition would. Moreover, in the following corollary, we show that if diffusion function specification is correct, \( C^S(\theta) \) is invariant for a larger class of approximate transitions.

Corollary 2.1. If the diffusion function \( \sigma(\cdot; \beta^*) \) of \( M(\theta) \) is identical to the true diffusion function, i.e. \( \sigma(\cdot; \beta^*) = \sigma_0(\cdot) \) on \( \mathcal{D} \), then \( C^S(\theta) \) conditional on \( \beta = \beta^* \) is simplified to

\[ C^S(\theta) = \frac{1}{2} \mathbb{E} \left[ \frac{(\mu_0 - \mu)^2}{\sigma_0^2} \right] \]
for the exact transition as well as the Euler, Milstein, and Aït-Sahalia’s closed-form approximate transitions. Further, we have $C^S(\theta) \geq 0$ and the equality holds if and only if $\theta = (\alpha^*, \beta^*) \in \Theta$ such that $\mu(\cdot; \alpha^*) = \mu_0(\cdot)$ on $\mathcal{D}$.

As an example, we give the explicit forms of $C^P$ and $C^S$ for the CIR model

$$M(\theta) : dX_t = \alpha_2(\alpha_1 - X_t) dt + \beta \sqrt{X_t} dW_t.$$ 

Let the process with $\theta_0 = (\bar{\alpha}_1, \bar{\alpha}_2, \bar{\beta})$ be the true diffusion. In this example, the function $(A f)_0$ used to calculate the primary criterion becomes a constant that depends on $\beta$ only rather than random, and is simply given by

$$C^P(\beta) = -\frac{1}{2} \log \left(\frac{\bar{\beta}^2}{\beta^2}\right) + \frac{1}{2} \left(\frac{\bar{\beta}^2}{\beta^2} - 1\right).$$

The secondary criterion $C^S(\theta) = \frac{1}{2} \mathbb{E}(A^2 f)_0$ can be calculated from

$$(A^2 f)_0(x) = \left[\left(\bar{\alpha}_2^2 (\bar{\alpha}_1 - x)^2 + \bar{\alpha}_2 (\bar{\alpha}_1 - x) \bar{\beta}^2 + (\bar{\beta}^2) (-\bar{\alpha}_2)\right) \left(-\frac{1}{\beta^2 x} + \frac{1}{\beta^2 x}\right)
+ \frac{3}{2x} \left(\bar{\alpha}_2 (\bar{\alpha}_1 - x) + \bar{\beta}^2 \right)(1 - \frac{\bar{\beta}^2}{\beta^2}) + \frac{15 \bar{\beta}^2}{16 x} \left(1 - \frac{\bar{\beta}^2}{\beta^2}\right)
+ 2 \bar{\alpha}_2 (\bar{\alpha}_1 - x) \left(\frac{\bar{\alpha}_2 (\bar{\alpha}_1 - x)}{\beta^2 x} - \frac{\alpha_2 (\alpha_1 - x)}{\beta^2 x}\right)
+ \left(-\frac{\bar{\alpha}_2 (\bar{\alpha}_1 - x)^2}{\beta^2 x} + \frac{\alpha_2^2 (\alpha_1 - x)^2}{\beta^2 x}\right) + (\bar{\alpha}_2 - \alpha_2)
+ \frac{\bar{\alpha}_2 (\bar{\alpha}_1 - x) - \alpha_2 (\alpha_1 - x)}{x} - \frac{3}{16} \frac{\left(\bar{\beta}^2 - \beta^2\right)}{x}
+ \left(-\alpha_2 - \frac{\bar{\alpha}_2 (\bar{\alpha}_1 - x)}{x}\right) + \frac{\bar{\beta}^2}{\beta^2} \left(-\alpha_2 - \frac{\alpha_2 (\alpha_1 - x)}{x}\right)\right].$$

Let us set the true process as $(\bar{\alpha}_1, \bar{\alpha}_2, \bar{\beta}) = (0.06, 0.2, 0.1)$, and calculate $C^P(\beta)$ and $C^S(\theta)$. Figure 3 shows the values of $C^P(\beta)$ for different values of $\beta$. The closest diffusion function specification does not depend on drift specifications, and $C^P$ becomes zero if and only if the diffusion specification is correct ($\beta = 0.1$).

In Figure 4, we provide the contour plots of $C^S(\theta)$ from the true process to processes with $(\alpha_1, \alpha_2)$ given $\beta = 0.08, 0.10, 0.12, 0.14$. The true value of $(\alpha_1, \alpha_2)$ is shown as the
Figure 3: The primary criterion $C^P(\beta)$ from the true process $X$ to $M(\theta): dX_t = \alpha_2(\alpha_1 - X_t)dt + \beta\sqrt{X_t}dW_t$. The true process is defined by $(\alpha_1, \alpha_2, \beta) = (0.06, 0.2, 0.1)$. $C^P(\beta)$ does not depend on $(\alpha_1, \alpha_2)$, and it is minimized at the true parameter value $\beta = 0.1$.

We can see in the figure that $C^S(\theta)$ is minimized at the true value $(\alpha_1, \alpha_2) = (0.06, 0.2)$ when $\beta = 0.1$. However, if $\beta \neq 0.1$, $C^S(\theta)$ is minimized at a different $(\alpha_1, \alpha_2)$ point. As shown in this example, slight misspecification of diffusion functions leads to a large change of the closest drift specification in $C^S$. This implies that when diffusion specification is incorrect, the best drift specification in the sense of minimizing $C^S$ may not be the correct drift specification. Consequently, the diffusion process with correct drift specification may not be the closest diffusion to the true diffusion in this case.

For further development of our theory, we assume that the primary and secondary criteria satisfy the following.

**Assumption 5.** The functions $C^P(\beta)$ and $C^S(\theta)$ satisfy the following conditions.

(a) There is a unique $\beta^*$ that solves the minimization problem

$$C^P(\beta^*) = \min_{\beta} C^P(\beta).$$

(b) Let $\Theta_{\beta^*} = \{ (\alpha, \beta) | \beta = \beta^* \}$, then we have a unique $\theta^* = (\alpha^*, \beta^*)$ in the interior of $\Theta$.
Figure 4: Contour plot of the secondary criterion $C^S(\theta)$ from the true process $X$ to $M(\theta) : dX_t = \alpha_2(\alpha_1 - X_t)dt + \beta\sqrt{X_t}dW_t$. The true process is defined by $(\alpha_1, \alpha_2, \beta) = (0.06, 0.2, 0.1)$ and shown as the dot at the center. Each level represents the diffusion processes with $(\alpha_1, \alpha_2)$ in $M(\theta)$ with equal $C^S(\theta)$ from $X$ under four different values of $\beta = 0.08, 0.10, 0.12, 0.14$. Note that $\beta = 0.1$ is the true parameter value, and $C^S(\theta)$ is minimized at the true parameter $(\alpha_1, \alpha_2) = (0.06, 0.2)$ in this case.
that solves the minimization problem
\[ C^S(\theta^*) = \min_{\theta \in \Theta^*} C^S(\theta). \]

(c) \( C^P \) and \( C^S \) are twice differentiable at \( \theta^* \).

Let \( \theta^* = (\alpha^*, \beta^*) \) solve the system of equations
\[
\beta^* = \text{argmin}_\beta C^P(\beta), \tag{2.11}
\]
\[
\theta^* = \text{argmin}_\theta C^S(\theta). \tag{2.12}
\]

The sub-model \( M_{\beta^*}(\alpha) \) defined on \( \Theta_{\beta^*} \) represents the closest members in \( M(\theta) \) to the true diffusion process under the primary criterion \( C^P(\beta) \). The value \( \alpha^* \) gives the closest diffusion \( M(\theta^*) \) in \( M_{\beta^*}(\alpha) \) that minimizes \( C^S(\theta) \) given \( \beta^* \). The first order conditions for (2.11) and (2.12) are respectively given by
\[
E_A f_{/\beta} = E \left[ \frac{1}{2} \left( -\sigma^2 + \sigma_0^2 \right) (\sigma^{-2})_{/\beta} \right] = 0, \tag{2.13}
\]
\[
E_A^2 f_{/\alpha} = E \left[ \frac{2 \mu_{/\alpha}}{\sigma^2} (\mu - \mu_0) + \left( 1 - \frac{\sigma_0^2}{\sigma^2} \right) \left( \mu_{/\alpha} - \frac{\mu_{/\alpha} \sigma_{/x}^2}{\sigma^2} \right) \right] = 0. \tag{2.14}
\]

For example, if the diffusion function of \( M(\theta) \) is constant, \( \sigma(\cdot; \beta) = \beta \), then we have \( \beta^* = \left\{ E \sigma_0^2(X_0) \right\}^{1/2} \) implying that \( \beta^* \) is the mean of the true diffusion function. When an approximate transition is used, the first order condition for the drift parameter \( \alpha \) depends on the approximate transition. Aït-Sahalia’s closed-form approximate transition results in the same condition in (2.14) as the exact transition, but for the Euler and Milstein approximate transitions, the first order conditions for \( \alpha \) in (2.14) are given by
\[
E \left[ \frac{2 \mu_{/\alpha}}{\sigma_0^2} (\mu - \mu_0) \right] = 0, \quad \text{for the Euler}, \tag{2.15}
\]
\[
E \left[ \frac{2 \mu_{/\alpha}}{\sigma_0^2} (\mu - \mu_0) - \left( 1 - \frac{\sigma_0^2}{\sigma^2} \right) \frac{3 \mu_{/\alpha} \sigma_{/x}^2}{2 \sigma^2} \right] = 0, \quad \text{for the Milstein}.
\]

Compared to the first order condition with the Euler approximation, the Milstein approximation adds an additional term \( \left( 1 - \frac{\sigma_0^2}{\sigma^2} \right) \frac{3 \mu_{/\alpha} \sigma_{/x}^2}{2 \sigma^2} \) that incorporates the dependence of the diffusion function \( \sigma_{/x}^2 \) on \( x \), and the first order condition with the exact transition adds the
term \((1 - \frac{\sigma^2}{\sigma_0^2})(-\mu/\alpha + \frac{\mu/\alpha}{\sigma^2/\sigma_0^2})\) that incorporates the dependence on both diffusion and drift functions \(\sigma^2/\sigma_0^2\) and \(\mu/\alpha\) on \(x\). When diffusion function is misspecified but drift specification is correct, the solution \(\alpha^*\) with the exact or Milstein approximation is not the true drift parameter \(\alpha_0\) generally. However, the solution of (2.15) with the Euler approximation would be the same as the true parameter. In this sense, the Euler approximation gives a robust drift estimator under incorrect diffusion specification, which is useful when we are interested in finding correct drift specification rather than the closest diffusion process to the true process.

When diffusion specification is correct \((\sigma^2 = \sigma_0^2)\), from Corollary 2.1, the first order condition for \(\alpha\) is simplified to

\[E(\mathcal{A}^2 f)_{/\alpha} = E \left[ \frac{2\mu/\alpha}{\sigma_0^2} (\mu - \mu_0) \right] = 0\]

for the exact, Euler, Milstein approximations. Considering this result, we can expect that the MLE of drift parameters would be less sensitive to the choice of an approximate density in this case.

We study the asymptotic behavior of the ML estimators and the relationship of their asymptotic properties with the primary and secondary criteria in the following section.

### 3 Information theoretic analysis of maximum likelihood estimation

Suppose we have \(N+1\) observations \(\{X_i\Delta\}_{i=0}^{N}\) of the diffusion \(X\) from time zero to \(T = N\Delta\) at a non-random time interval \(\Delta\). Denote the log-likelihood function

\[L_0 \equiv \sum_{i=1}^{N} \log p_0(\Delta, X_{(i-1)\Delta}, X_i\Delta)\]

of the true diffusion process, and

\[L(\theta) \equiv \sum_{i=1}^{N} \log p(\Delta, X_{(i-1)\Delta}, X_i\Delta; \theta)\]
for $M(\theta)$ conditional on $X_0$. We suppress the dependency of the log-likelihoods on $N$ and $\Delta$ (or equivalently $T$ and $\Delta$) for notational simplicity. The MLE $\hat{\theta}$ is defined by

$$\hat{\theta} \equiv \arg\max_{\theta \in \Theta} L(\theta).$$

When $M(\theta)$ is misspecified, $L(\theta)$ and $\hat{\theta}$ are also called the quasi-log-likelihood function and the quasi-MLE respectively.

The exact transition density $p(t, x, y; \theta)$ of $M(\theta)$ is known in closed-form for only a few diffusion processes such as the Ornstein-Uhlenbeck or Feller’s square root processes. When $p(t, x, y; \theta)$ is not available in closed-form, we must use an approximate transition density such as the Euler or Milstein approximate transitions, the simulated likelihoods of Brandt and Santa-Clara (2002), or the closed-form approximation of Aït-Sahalia (2002) based on the Hermite expansion.

The log-likelihood ratio $L_0 - L(\theta)$ has an interesting relationship with the primary criterion. If we fix $\Delta$ and consider the limit of average log-likelihood ratio

$$\lim_{T \to \infty} \frac{1}{N}(L_0 - L(\theta)) = \lim_{T \to \infty} \frac{\Delta}{T}(L_0 - L(\theta)) = E\left( \log \frac{p(\Delta, X_0, X_\Delta)}{p(\Delta, X_0, X_\Delta; \theta)} \right) = D_\Delta(\theta)$$

as $T \to \infty$, then the sequential limit

$$\lim_{\Delta \to 0} \left( \lim_{T \to \infty} \frac{1}{N}(L_0 - L(\theta)) \right) = C^P(\beta)$$

becomes the primary criterion.

The following theorem shows that the MLE $\hat{\theta}$ is consistent to the minimizer $\theta^* = (\alpha^*, \beta^*)$ in (2.11)-(2.12), which is defined as the pseudo-true value.

**Theorem 3.1.** Suppose that Assumptions 1-5 are satisfied. Under $\Delta^3 T \to 0$ as $\Delta \to 0$ and $T \to \infty$, the MLE $\hat{\theta}$ converges in probability to the solution $\theta^* = (\alpha^*, \beta^*)$ of the system of equations in (2.11)-(2.12).

The above theorem shows that the probability limit of $\hat{\beta}$ does not depend on the specification of drift functions, and is robust to the choice of the approximation method for the transition density. But the limit of $\hat{\alpha}$ depends on the transition density approximation methods as well as diffusion specification. This explains the high sensitivity of drift estimators to the choice of an approximate transition in practice. We also can see that when
the diffusion function is misspecified but the drift function is correctly specified, $\hat{\alpha}$ may not converge the true parameter value because the true parameter value may not minimize the secondary criterion.

Based on this theorem, we illustrate the geometry of the ML estimation of diffusion models. Geometrically, our model $M(\theta)$ has the structure of a fiber bundle. A fiber bundle $E = B \times F$ consists of a base space $B$, a fiber space $F$, and a projection of $E$ onto $B$ that defines how to attach fibers on $B$. It is also denoted simply by $F \to E \xrightarrow{\pi} B$. For us, the total space $E$ of the fiber bundle represents all possible combinations of drift and diffusion specifications defined in $M(\theta)$. Thus, we can denote the total space as $E(\theta)$ indexed by $\theta \in \Theta$. The base space $B$ is the space of all diffusion functions $\sigma(\cdot; \beta)$ of $M(\theta)$. We denote the base space as $B(\beta)$ indexed by $\beta$. Over $B$, we define the fiber space $F$ of drift functions $\mu(\cdot; \alpha)$ of $M(\theta)$ attached to each element $\beta$ in $B$. We denote the fiber space as $F(\alpha)$ indexed by $\alpha$. Finally, the mapping $\pi : E \to B$ of the fiber bundle $E(\theta)$ onto $B(\beta)$ defines how each fiber space is attached to $B$. In our case, $\pi$ is the projection of $\theta = (\alpha, \beta)$ in $E$ onto the second factor $\beta$ in $B$, in which sense the fiber bundle is a trivial bundle. From these definitions, the fiber attached to $\beta$ on $B$ can be written as $\pi^{-1}(\beta)$, which represents the space of all the diffusion processes in $M(\theta)$ with diffusion function specification $\sigma(\cdot; \beta)$.

On the right side of Figure 5, we show the fiber bundle $E$ with the fiber space $F$ over the base space $B$. The primary criterion $C_P$ shown in Figure 5 measures the information divergence from the true process $(\mu_0, \sigma_0)$ to $\sigma(\cdot; \beta)$ in the base space $B(\beta)$. The $\sigma$-orbit represents the space of all diffusion functions with equal divergence in $C_P$ from the true diffusion. The divergence $C_P$ is defined on the horizontal plane that $\sigma$-orbit belongs to, which represents that $C_P$ does not depend on $F$ and is determined by diffusion functions only. Under the primary criterion $C_P$, the closest diffusion specification in $B(\beta)$ is shown as $\beta^*$ in Figure 5. Then given $\beta^*$, consider the fiber $\pi^{-1}(\beta^*)$. The secondary criterion $C_S$ measures the divergence from $(\mu_0, \sigma_0)$ to an element of $\pi^{-1}(\beta^*)$, and the minimizer of $C_S$ is shown as $\alpha^*$. Therefore, we can think the minimization with respect to $C_P$ and $C_S$ as two distinct projections, and the pseudo-true value $\theta^* = (\alpha^*, \beta^*)$ is obtained by sequential application of the two projections.

We now study the asymptotic distribution of the MLE. For correctly specified models, the primary and secondary criteria are closely related to the asymptotic variance of the MLE. Suppose we have $\sigma_0(\cdot) = \sigma(\cdot; \beta_0)$ and $\alpha_0(\cdot) = \mu(\cdot; \alpha_0)$ for the true parameter $\theta_0 = (\alpha_0, \beta_0)$. When we have $\Delta \to 0$ as $T \to \infty$, Jeong and Park (2013) show in their Theorem
Figure 5: The geometry of the ML estimation. On the right, the model $M(\theta)$ is shown as the fiber bundle $E = B \times F$ with the fiber space $F$ of the drift functions $\mu(\cdot; \alpha)$ over the base space $B$ of the diffusion functions $\sigma(\cdot; \beta)$ and the projection $\pi : E \to B$ that defines how the fibers are attached on $B$. The primary criterion $C^P$ measures the information divergence from the true process $(\mu_0, \sigma_0)$ to $\sigma(\cdot; \beta)$ in $B(\beta)$. The $\sigma$-orbit represents the space of all diffusion functions with equal divergence in $C^P$ from the true diffusion function $\sigma_0$. Under $C^P$, the closest member of the diffusion functions in $B(\beta)$ is shown as $\beta^*$. The secondary criterion $C^S$ measures the divergence from $(\mu_0, \sigma_0)$ to an element of the fiber $\pi^{-1}(\beta^*)$ given $\beta^*$, and the minimizer of $C^S$ is shown as $\alpha^*$. The pseudo-true value $\theta^* = (\alpha^*, \beta^*)$ is determined by the sequential minimization of $C^P$ and $C^S$. 
4.3 that

\[ \sqrt{\frac{T}{\Delta}} (\hat{\beta} - \beta_0) \xrightarrow{d} N \left( 0, \left[ 2\mathbb{E} \left( \frac{\sigma_{/\beta}^T \sigma_{/\beta}}{\sigma_0^2} \right) \right]^{-1} \right), \]  

(3.1)

\[ \sqrt{T}(\hat{\alpha} - \alpha_0) \xrightarrow{d} N \left( 0, \left[ \mathbb{E} \left( \frac{\mu_{/\alpha}^T \mu_{/\alpha}}{\sigma_0^2} \right) \right]^{-1} \right), \]  

(3.2)

where \( \sigma_{/\beta} \equiv \partial \sigma(X_0; \beta_0)/\partial \beta \) and \( \mu_{/\alpha} \equiv \partial \mu(X_0; \alpha_0)/\partial \alpha \). In (3.1)-(3.2), the asymptotic variances of the ML estimators are given by the inverse of the matrices

\[ \left[ \mathbb{E} \left( \frac{\sigma_{/\beta}^T \sigma_{/\beta}}{\sigma_0^2} \right) \right]^{-1} \]  

and

\[ \left[ \mathbb{E} \left( \frac{\mu_{/\alpha}^T \mu_{/\alpha}}{\sigma_0^2} \right) \right]^{-1} \]  

for \( \hat{\beta} \) and \( \hat{\alpha} \) respectively. We show that the matrices in (3.3) are derived from the limits as \( \Delta \to 0 \) of the Fisher information matrices calculated under discrete sampling with fixed \( \Delta \).

For a fixed sampling interval \( \Delta > 0 \), the Fisher information matrix \( G_\Delta \) is induced by the KLIC \( D_t(\theta) \) by the second derivative

\[ G_t(\theta) = \left\{ \frac{\partial^2 D_t(\theta)}{\partial \theta_i \partial \theta_j} \right\}, \]

where \( \{\cdot\} \) is the \((i, j)\) element of \( G_t \). Therefore, the asymptotic variance of the ML estimator \( \hat{\theta} \) under fixed \( \Delta \) would be given by

\[ \sqrt{T}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, G_\Delta(\theta_0)^{-1}), \]

as \( T \to \infty \). Based on this result, we consider the limit of the \( \Delta \)-sequence of the Fisher information matrix \( G_\Delta(\theta_0) \) for \( \beta \) as \( \Delta \to 0 \). Let \( G^P(\beta_0) \equiv \lim_{\Delta \to 0} \left\{ \frac{\partial^2 D_t(\theta)}{\partial \beta_i \partial \beta_j} \right\} \) be the limit, then we have

\[ G^P(\beta_0) = \left\{ \frac{\partial^2 C^P(\beta_0)}{\partial \beta_i \partial \beta_j} \right\} = \mathbb{E} \left\{ \frac{\partial^2 (A_f)/\theta}{\partial \beta_i \partial \beta_j} \right\}. \]  

(3.4)

Similarly, for the Fisher information for \( \alpha \), we consider the limit \( G^S(\theta) \equiv \lim_{\Delta \to 0} \left\{ \frac{\partial^2 D_t(\theta)}{\partial \alpha_i \partial \alpha_j} \right\}, \)
which is given by

\[ G^S(\theta_0) = \left\{ \frac{\partial^2 C^S(\theta_0)}{\partial \alpha_i \partial \alpha_j} \right\} = \frac{1}{2} E \left\{ \frac{\partial^2 (A^2 f)_{i0}}{\partial \alpha_i \partial \alpha_j} \right\}. \]  

(3.5)

This shows that the asymptotic Fisher information matrices are given by the curvatures of \( C^P \) and \( C^S \) at \( \theta_0 \). In the following theorem, we show that the curvatures given in (3.4) and (3.5) coincide with the inverse asymptotic variance matrices in (3.3) under \( \Delta \to 0 \) and \( T \to \infty \).

**Theorem 3.2.** Suppose that Assumptions (7) are satisfied, and \( M(\theta) \) is correctly specified with the true parameter \( \theta_0 = (\alpha_0, \beta_0) \). Let \( p, q \) be the constants defined in Assumption (1).

Under \( \Delta T^{4(pq + 1)} \to 0 \) as \( \Delta \to 0 \) and \( T \to \infty \), we have

\[ \sqrt{T/\Delta} (\hat{\beta} - \beta_0) \to N(0, G^P(\beta_0)^{-1}), \quad \sqrt{T}(\hat{\alpha} - \alpha_0) \to N(0, G^S(\theta_0)^{-1}), \]

where \( G^P \) and \( G^S \) are the asymptotic Fisher information matrices defined in (3.4) and (3.5) respectively.

For misspecified models, the second Bartlett identity does not hold because the Fisher information matrix is different from the negative expected Hessian matrix. Therefore, we derive the Fisher information and expected Hessian matrices separately. We first give the asymptotic expansion of the score functions.

**Theorem 3.3.** Suppose that Assumptions (7) are satisfied. For misspecified models, under \( \Delta^3 T \to 0 \) as \( \Delta \to 0 \) and \( T \to \infty \), we have the explicit form of the score functions

\[ s_{\beta}(\theta) \equiv \left\{ \frac{\partial L(\theta)}{\partial \beta_i} \right\} = \frac{1}{\Delta} \int_0^T \frac{\sigma_0^2 - \sigma^2}{\sigma_i^2} (\mu_{/a} - \mu) dt + O_p(T) + O_p \left( \sqrt{T/\Delta} \right), \]

(3.6)

\[ s_{\alpha}(\theta) \equiv \left\{ \frac{\partial L(\theta)}{\partial \alpha_i} \right\} = \frac{1}{2} \int_0^T \frac{2\mu_{/a}}{\sigma^2} (\mu_0 - \mu) + \left( 1 - \frac{\sigma_0^2}{\sigma^2} \right) \left( -\mu_{/x} + \frac{\mu_{/a} \sigma_x^2}{\sigma_0^2} \right) dt \]

\[ + \int_0^T \frac{\mu_{/a}}{\sigma_0^2} dW_t + O_p(\sqrt{\Delta T}) + O_p(\Delta T). \]

(3.7)

The asymptotic leading term of \( s_{\beta}(\theta) \) is \( O_p(T/\Delta) \), and it is generally larger than the second term which is \( O_p(T) \). But at \( \beta = \beta^* \), the first term becomes \( O_p(\sqrt{T/\Delta}) \) because
Therefore, the first term in (3.6) is larger than the second term only if \( \Delta^2 T \to 0 \). Similarly, the first term in the asymptotic expansion of \( s_\alpha(\theta) \) given (3.7) would be \( O_p(\sqrt{T}) \) at \( \theta = \theta^* \). This term becomes the largest term under \( \Delta^2 T \to 0 \). We will assume \( \Delta^2 T \to 0 \) for the rest of the paper to derive the asymptotic distribution of the MLE.

As shown in Jeong and Park (2013), the asymptotic order of \( s_\beta(\theta^*_0) \) for correctly specified models is \( O_p(\sqrt{T/\Delta}) \), or equivalently \( O_p(\sqrt{N}) \). From Theorem 3.3, we find that the score function \( s_\beta(\theta^*_0) \) for misspecified models becomes larger than \( O_p(\sqrt{N}) \). Intuitively, for correctly specified models, we identify the true parameter by finding \( \beta_0 \) that vanishes the first term in (3.6), but for misspecified models, we minimize the primary criterion to find the pseudo-true parameter \( \beta^* \) that vanishes the first term in (3.6) on the long run average. As shown later in this section, this fact reduces the convergence rate of \( \hat{\beta} \) to \( \sqrt{T} \) for misspecified models compared to the rate \( \sqrt{N} \) of \( \hat{\beta} \) for correctly specified models.

In the following theorem, we calculate the asymptotic Fisher information matrix from the asymptotic variances of the score functions

\[
    I_\beta(\beta) \equiv \lim_{T \to \infty, \Delta \to 0} \left( \frac{\Delta^2}{T} s_\beta(\theta) s_\beta(\theta)^T \right),
    \quad I_\alpha(\theta) \equiv \lim_{T \to \infty, \Delta \to 0} \left( \frac{1}{T} s_\alpha(\theta) s_\alpha(\theta)^T \right),
\]

and the asymptotic expected Hessian matrices

\[
    H_\beta(\beta) \equiv \lim_{T \to \infty, \Delta \to 0} \left( \frac{\Delta}{T} h_\beta(\theta) \right),
    \quad H_\alpha(\theta) \equiv \lim_{T \to \infty, \Delta \to 0} \left( \frac{1}{T} h_\alpha(\theta) \right),
\]

where \( h_\beta(\theta) = \left\{ \frac{\partial^2 L(\theta)}{\partial \beta_i \partial \beta_j} \right\} \) and \( h_\alpha(\theta) = \left\{ \frac{\partial^2 L(\theta)}{\partial \alpha_i \partial \alpha_j} \right\} \). We first define the derivative of the scale function

\[
    s'_0(x) \equiv \exp\left( -\int_{x_0}^x \frac{2\mu_0(v)}{\sigma_0^2(v)} dv \right)
\]

and the speed density

\[
    m_0(x) \equiv \frac{1}{\sigma_0^2(x)s'_0(x)},
\]

for any \( x_0 \in \mathcal{D} \) fixed. With the scale function \( s_0(x) = \int_{x_0}^x s'_0(w) dw \), the transformed diffusion \( s_0(X) \) of the true diffusion \( X \) becomes a driftless diffusion, and the stationary density of the diffusion \( X \) is given by \( m_0(x)/\int_{\mathcal{D}} m_0(x)dx \).
Theorem 3.4. Suppose that Assumptions 1-5 are satisfied. Let $\Delta^2T \to 0$ as $\Delta \to 0$ and $T \to \infty$. For a misspecified diffusion model with the pseudo-true parameter $\theta^* = (\alpha^*, \beta^*)$, we have the asymptotic Fisher information matrices given by

$$I_\beta(\beta^*) = \mathbf{E}V_\beta V_\beta^T, \quad I_\alpha(\theta^*) = \mathbf{E}V_\alpha V_\alpha^T,$$

where $V_\beta = V_\beta(X_0)$ and $V_\alpha = V_\alpha(X_0)$ are defined from the functions

$$V_\beta(x) = \sigma_0(x)s_0'(x) \int_{x_0}^x A\tilde{l}_\beta(v) m_0(v)dv,$$

$$V_\alpha(x) = \sigma_0(x)s_0'(x) \int_{x_0}^x \left( \frac{1}{2} A^2\tilde{l}_\alpha(v) \right) m_0(v)dv + \frac{\sigma_0(x)}{\sigma^2(x)} \mu_\alpha(x),$$

where

$$A\tilde{l}_\beta(v) = \frac{1}{2}(\sigma_0^2(v) - \sigma^2(v))\sigma_{0\beta}^{-2}(v),$$

$$A^2\tilde{l}_\alpha(v) = \frac{2\mu_\alpha(v)}{\sigma^2(v)}(\mu_0(v) - \mu(v)) + \left(1 - \frac{\sigma_0^2(v)}{\sigma^2(v)}\right)\left(-\mu_{0\alpha}(v) + \frac{\mu_\alpha(v)\sigma_0^2(v)}{\sigma^2(v)}\right).$$

The asymptotic expected Hessian matrices are the same as the negative second derivatives of the primary and secondary criteria, and their explicit forms are given by

$$\mathbf{H}_\beta(\beta^*) = -\left\{ \frac{\partial^2 C_\beta(\beta^*)}{\partial \beta_i \partial \beta_j} \right\} = \mathbf{E}A\tilde{l}_{\beta \beta} = \frac{1}{2} \mathbf{E} \left[ \sigma^4\sigma_{\beta \beta}^{-2}\sigma_{\beta \beta}^{-2} + \left(-\sigma^2 + \sigma_0^2\right)\sigma_{\beta \beta}^{-2} \right],$$

$$\mathbf{H}_\alpha(\theta^*) = -\left\{ \frac{\partial^2 C_\alpha(\theta^*)}{\partial \alpha_i \partial \alpha_j} \right\} = \frac{1}{2} \mathbf{E}A^2\tilde{l}_{\alpha \alpha} = \frac{1}{2} \mathbf{E} \left[ \frac{2\mu_{\alpha \alpha}^T(\mu_0 - \mu) - 2\mu_{\alpha \alpha}^T\mu_{0\alpha}^T + \left(1 - \frac{\sigma_0^2}{\sigma^2}\right)\left(-\mu_{0\alpha\alpha} + \frac{\mu_{0\alpha\alpha}^T\sigma_0^2}{\sigma^2}\right)}{\sigma^2} \right].$$

From the above theorem, we can get the asymptotic distribution of the MLE.

Corollary 3.1. Suppose that Assumptions 1-5 are satisfied. Let $\Delta^2T \to 0$ as $\Delta \to 0$ and $T \to \infty$. For a misspecified diffusion model with the pseudo-true parameter $\theta^* = (\alpha^*, \beta^*)$,
we have

\[ \sqrt{T}(\hat{\beta} - \beta^*) \xrightarrow{d} N \left( 0, H_\beta(\beta^*)^{-1} I_\beta(\beta^*) H_\beta(\beta^*)^{-1} \right), \]
\[ \sqrt{T}(\hat{\alpha} - \alpha^*) \xrightarrow{d} N \left( 0, H_\alpha(\theta^*)^{-1} I_\alpha(\theta^*) H_\alpha(\theta^*)^{-1} \right). \]

From the corollary, we can get the explicit form of the asymptotic distributions of the MLE in terms of the model’s drift and diffusion functions. It also clarifies that the differences of the asymptotic behaviors of the MLE under correct specification and mis-specification. This result also shows that the likelihood ratio test statistics would not have the Chi-squared distribution asymptotically. Consequently, the Lagrange Multiplier, and Wald test statistics would not be distributed as the Chi-squared distribution asymptotically either.

4 Conclusion

The information criteria developed in this paper provide the information theoretic foundation for studying asymptotic behaviors of the MLE and likelihood based test statistics for high frequency data. The new framework also allows us to analyze the estimation and inference problems geometrically. We surmise that we can derive the higher order properties of the MLE and other general estimators on this foundation using the differential geometric methods developed in Efron [1978] and Amari [1982]. Our new framework would make the development of these theories easier and the interpretation of them more intuitive.

Another important issue is asymptotic analysis of inference with decreasing sampling interval. Since the conventional form of the likelihood based tests would not be valid under model misspecification, a robust test statistic would be desirable. Based on the results of this paper, we can develop a robust version of the likelihood based test statistics and derive its limiting distribution. Also, asymptotic analysis of specification testing for diffusion models in our new framework is another important topic. These topics will be investigated in the future research. We did not consider in this paper more general processes such as jump diffusions, or Lévy processes. Although these processes are important, the development of the information theory for these processes would warrant a separate paper, and we leave it as a future research topic.
Appendices

A Notational convention

We summarize the notional convention used in this paper for reference. For a function $f(t, x, y)$, we define the functions

$$f_0(x) \equiv f(0, x, x), \quad f_t(x) \equiv \frac{\partial f(0, x, x)}{\partial t},$$

with the single argument $x$ as the functions evaluated at $y = x$ and $t = 0$. Other higher order derivatives such as $f_y(x), f_{yy}(x), \text{ or } f_{yt}(x)$ are defined similarly.

For a function $f(x)$ with the single argument $x$, we define $f \equiv f(X_0)$ omitting the argument $X_0$. For example, we write

$$\tilde{l}_0 = \tilde{l}_0(X_0), \quad \mu = \mu(X_0), \quad \sigma = \sigma(X_0).$$

Also, we suppress $\theta$ for the functions such as the log-likelihoods and the drift and diffusion functions of $M(\theta)$. Using the notations defined above, we define $\mu_x(x) \equiv \partial \mu(x)/\partial x$, $\sigma_{x}^2(x) \equiv \partial \sigma^2(x)/\partial x$, and $\sigma_{xx}^2(x) \equiv \partial^2 \sigma^2(x)/\partial x^2$, and denote

$$\mu_x \equiv \mu_x(X_0), \quad \sigma_{x}^2 \equiv \sigma_{x}^2(X_0), \quad \sigma_{xx}^2 \equiv \sigma_{xx}^2(X_0).$$

Other functions and their derivatives without arguments are defined in the same manner. For the infinitesimal generator $\mathcal{A}$, we define

$$\mathcal{A}f_0 \equiv (\mathcal{A}f)_0(X_0), \quad \mathcal{A}f_y \equiv (\mathcal{A}f)_y(X_0) = \frac{\partial \mathcal{A}f(0, X_0, X_0)}{\partial y}, \quad \mathcal{A}f_{y\beta} \equiv \frac{\partial (\mathcal{A}f)_{y}(X_0)}{\partial \beta}.$$

When these functions are used as integrands, the argument omitted is $X_t$ rather than $X_0$. 

25
Proof. We omit $\theta$ in $D_t(\theta)$ for notational simplicity. The following proof applies to all $\theta \in \Theta$. Define the conditional expectation

$$g_t(x) \equiv E^x \{ l_0(t, x, X_t) - l(t, x, X_t) \},$$

where $E^x$ is the conditional expectation with respect to $X_t$ given $X_0 = x$. Then we can rewrite $D_t = E \{ l_0(t, X_0, X_t) - l(t, X_0, X_t) \}$ as

$$D_t = E g_t(X_0).$$

Under Assumption 1, we get $\lim_{t \to 0} D_t = E \lim_{t \to 0} g_t(X_0)$ by the dominated convergence theorem. Therefore, we need show that $\lim_{t \to 0} g_t(x) = A f_{j0}(x)$. From (2.3) and (2.5), we can rewrite

$$g_t(x) = E^x \{ \tilde{l}_0(t, x, X_t) - \tilde{l}(t, x, X_t) \} / t$$

$$= E^x f(t, x, X_t) / t$$

$$= \frac{E^x f(t, x, X_t) - f_{j0}(x)}{t}.$$  

The last equality above is from $f_{j0}(x) = 0$ given by Assumption 3. Then we get the limit

$$\lim_{t \to 0} \frac{E^x f(t, x, X_t) - f_{j0}(x)}{t} = A f_{j0}(x)$$

by the definition of the infinitesimal generator $A$.

For the explicit expression in (2.6), we first apply

$$A = \frac{\partial}{\partial t} + \mu_0(y) \frac{\partial}{\partial y} + \frac{\sigma_0^2(y)}{2} \frac{\partial^2}{\partial y^2}$$

to $f(t, x, y) = \tilde{l}_0(t, x, y) - \tilde{l}(t, x, y)$ to evaluate $Af(t, x, y)$ at $t = 0$ and $y = x$. As the result,
we would get

\[ Af_0 = f_t + \mu_0 f_y + \frac{\sigma_0^2}{2} f_{yy} \]

\[ = (\tilde{l}_{0/t} - \tilde{l}_t) + \mu_0(\tilde{l}_{0/y} - \tilde{l}_y) + \frac{\sigma_0^2}{2}(\tilde{l}_{0/yy} - \tilde{l}_{yy}). \]

From Assumption 3, we have

\[ f_t(x) = -\frac{1}{2} \log \frac{\sigma_0^2(x)}{\sigma^2(x)}, \quad f_y(x) = 0, \]

\[ f_{yy}(x) = -\frac{1}{\sigma_0^2(x)} + \frac{1}{\sigma^2(x)}. \]

Therefore,

\[ Af_0(x) = -\frac{1}{2} \log \frac{\sigma_0^2(x)}{\sigma^2(x)} + \frac{\sigma_0^2(x)}{2} \left( -\frac{1}{\sigma_0^2(x)} + \frac{1}{\sigma^2(x)} \right), \]

which proves the second equality in the theorem.

**B.2 Proof of Theorem 2.2**

**Proof.** For both \( D_t(\theta) \) and \( D'_t(\theta) \), we omit \( \theta \) for notational simplicity. We define the conditional expectation

\[ g_t(x) \equiv \frac{E_x f(t, x, X_t) - f_0(x)}{t} \quad (B.1) \]

given \( X_0 = x \) as in the proof of Theorem 2.1. Since \( D_t = E g_t(X_0) \) by definition and \( D_0 = E Af_0 \) by Theorem 2.1 we can write

\[ D'_0 \equiv \lim_{t \to 0} \frac{D_t - D_0}{t} = \lim_{t \to 0} E \frac{g_t(X_0) - Af_0(X_0)}{t}. \]

Since \( \lim_{t \to 0} E \frac{g_t(X_0) - Af_0(X_0)}{t} = E \lim_{t \to 0} \frac{g_t(X_0) - Af_0(X_0)}{t} \) from the dominated convergence theorem, we only need to show

\[ \lim_{t \to 0} \frac{g_t(x) - Af_0(x)}{t} = \frac{1}{2} A^2 f_0(x). \quad (B.2) \]

27
To show the above equality, we use Dynkin’s formula (Dynkin (1965) p.133, Øksendal (1998) Theorem 7.4.1.),

\[ E^x \varphi(X_\tau) = \varphi(x) + E^x \int_0^\tau A\varphi(X_s) \, ds, \]

for a twice continuously differentiable function \( \varphi \) on \( \mathbb{R} \) and \( \tau \) is a stopping time such that \( E^x \tau < \infty \). By applying Dynkin’s formula to the numerator of \( g_t(x) \) in (B.1) after setting the stopping time \( \tau = t \), we get

\[ E^x f(t, x, X_t) - f/0(x) = E^x \int_0^t A f(s, x, X_s) \, ds \]

\[ = \int_0^t E^x A f(s, x, X_s) \, ds. \]  \hspace{1cm} \text{(B.3)}

We apply Dynkin’s formula twice sequentially to \( E^x A f(s, x, X_s) \) to have

\[ E^x A f(s, x, X_s) = A f/0(x) + E^x \int_0^s A^2 f(u, x, X_u) \, du \]

\[ = A f/0(x) + \int_0^s \{ A^2 f/0(x) + E^x \int_0^v A^3 f(v, x, X_v) \, dv \} \, dv \]

\[ = A f/0(x) + s A^2 f/0(x) + E^x \int_0^s \int_0^v A^3 f(v, x, X_v) \, dv \, du. \]

Therefore, (B.3) becomes

\[ E^x f(t, x, X_t) - f/0(x) = t A f/0(x) + \frac{1}{2} t^2 A^2 f/0(x) + E^x \int_0^t \int_0^s \int_0^v A^3 f(v, x, X_v) \, dv \, du \, ds. \]

From this, we can write

\[ \frac{g_t(x) - A f/0(x)}{t} = \frac{1}{2} A^2 f/0(x) + \frac{1}{t^2} E^x \int_0^t \int_0^s \int_0^v A^3 f(v, x, X_v) \, dv \, du \, ds. \] \hspace{1cm} \text{(B.4)}

Since \( E^x \int_0^t \int_0^s \int_0^v A^3 f(v, x, X_v) \, dv \, du \, ds = O(t^3) \) for any \( x \in \mathcal{D} \) from Assumption 1, we get (B.2) from (B.4) by taking \( t \to 0 \) on the both sides.

For the second equality in (2.7), we apply
\[ A^2 = \left( \mu_{0/x}(y)\mu_0(y) + \mu_{0/xx}(y)\frac{\sigma_0^2(y)}{2} \right) \frac{\partial}{\partial y} \]
\[ + \left\{ \mu_0(y) \left( \mu_0(y) + \frac{\sigma_0^2(y)}{2} \right) + \sigma_0^2(y) \left( \mu_0(y) + \frac{\sigma_0^2(y)}{4} \right) \right\} \frac{\partial^2}{\partial y^2} \]
\[ + \sigma_0^2(y) \left( \mu_0(y) + \frac{\sigma_0^2(y)}{2} \right) \frac{\partial^3}{\partial y^3} + \frac{\sigma_0^4(y)}{4} \frac{\partial^4}{\partial y^4} \]
\[ + \frac{\partial^2}{\partial t^2} + 2\mu_0(y) \frac{\partial^2}{\partial y \partial t} + \sigma_0^2(y) \frac{\partial^3}{\partial y^2 \partial t}, \]

to \( f(t, x, y) = \tilde{l}_0(t, x, y) - \tilde{l}_0(t, x, y) \). Then we evaluate \( A^2 f(t, x, y) \) at \( t = 0 \) and \( y = x \) to get

\[ A^2 f_{t=0} = \left( \mu_0 \left( \mu_0 + \frac{\sigma_0^2}{2} \right) + \frac{\sigma_0^2}{2} \left( 2\mu_0 + \frac{\sigma_0^2}{2} \right) \right) \left( -\frac{1}{\sigma_0^2} + \frac{1}{\sigma_0^2} \right) \]
\[ + \sigma_0^2 \left( \mu_0 + \frac{\sigma_0^2}{2} \right) (B_0 - B) + \frac{\sigma_0^4}{4} (C_0 - C) + (D_0 - D) + 2\mu_0(A_0 - A) + \sigma_0^2(E_0 - E), \]

where

\[ A = \tilde{l}_{ytt} = \frac{\mu}{\sigma^2} - \frac{3\sigma_{1x}^2}{4\sigma^2}, \quad B = \tilde{l}_{yyy} = \frac{3\sigma_{1x}^2}{2\sigma^4}, \]
\[ C = \tilde{l}_{yyyy} = \frac{2\sigma_{1x}^2}{\sigma^4} - \frac{15}{4} \left( \frac{\sigma_{1x}^2}{\sigma^2} \right)^2, \]
\[ D = \tilde{l}_{ytt} = \frac{\mu^2}{\sigma^2} - \frac{\mu}{\sigma^2} - \frac{\mu \sigma_{1x}^2}{\sigma^2} - \frac{3(\sigma_{1x}^2)^2}{16\sigma^2} + \frac{\sigma_{2x}^2}{4}, \]
\[ E = \tilde{l}_{yytt} = \frac{1}{\sigma^2} \left( \frac{\mu}{\sigma^2} - \frac{\mu \sigma_{1x}^2}{\sigma^2} + \frac{3(\sigma_{1x}^2)^2}{4\sigma^2} - \frac{3\sigma_{2x}^2}{4} \right), \]

and the terms \( A_0, B_0, \ldots, E_0 \) are defined similarly with \( \mu_0 \) and \( \sigma_0^2 \) from the true diffusion. The explicit forms of these terms are given from the corresponding terms \( A_j, \ldots, E_j \) in Section 3 of \cite{Choi2013} derived for the exact transition density.
B.3 Proof of Corollary 2.1

Proof. For \( C^S(\theta) \), we omit \( \theta \) for notational simplicity. Since \( \tilde{l}_{0/yyy} = \tilde{l}_{/yggg} \) and \( \tilde{l}_{0/yggg} = \tilde{l}_{/yggg} \) from \( \sigma = \sigma_0 \), we have

\[
C^S = \mathbb{E} \left[ (\tilde{l}_{0/y} - \tilde{l}_{/y}) \mu_0 + \frac{1}{2} (\tilde{l}_{0/tt} - \tilde{l}_{/tt}) + \frac{\sigma_0^2}{2} (\tilde{l}_{0/ygg} - \tilde{l}_{/ygg}) \right].
\]

From \( \sigma = \sigma_0 \), we also have

\[
\tilde{l}_{0/y} - \tilde{l}_{/y} = \frac{\mu_0 - \mu}{\sigma_0^2},
\]
\[
\tilde{l}_{0/tt} - \tilde{l}_{/tt} = - (\mu_0/\mu) - \frac{\mu_0^2 - \mu^2}{\sigma_0^2} + 2 \left( \frac{\mu_0 \sigma_0/\mu}{\sigma_0} - \frac{\mu \sigma_0/\mu}{\sigma_0} \right),
\]
\[
\tilde{l}_{0/ygg} - \tilde{l}_{/ygg} = \frac{\mu_0/\mu - \mu/\mu}{\sigma_0^2} - 2 \left( \frac{\mu_0 \sigma_0/\mu}{\sigma_0^2} - \frac{\mu \sigma_0/\mu}{\sigma_0^2} \right).
\]

Therefore,

\[
C^S = \mathbb{E} \left[ \frac{(\mu_0 - \mu) \mu_0}{\sigma_0^2} - \frac{1}{2} \frac{\mu_0^2 - \mu^2}{\sigma_0^2} \right] = \frac{1}{2} \mathbb{E} \left[ \frac{(\mu_0 - \mu)^2}{\sigma_0^2} \right].
\]

For Euler approximation, we use the explicit forms of \( \tilde{l}_{/y}, \tilde{l}_{/tt}, \) and \( \tilde{l}_{/ygg} \) which correspond to \( A_j, D_j, E_j \) respectively derived for the Euler approximate transition densities in Section 3 of Choi, Jeong, and Park (2013). From

\[
\tilde{l}_{0/y} - \tilde{l}_{/y} = \frac{\mu_0 - \mu}{\sigma_0^2}, \quad \tilde{l}_{0/tt} - \tilde{l}_{/tt} = - \frac{\mu_0^2 - \mu^2}{\sigma_0^2}, \quad \tilde{l}_{0/ygg} - \tilde{l}_{/ygg} = 0.
\]

we get

\[
C^S = \mathbb{E} \left[ \frac{(\mu_0 - \mu) \mu_0}{\sigma_0^2} - \frac{1}{2} \frac{\mu_0^2 - \mu^2}{\sigma_0^2} \right] = \frac{1}{2} \mathbb{E} \left[ \frac{(\mu_0 - \mu)^2}{\sigma_0^2} \right].
\]

For Milstein approximation, we use the explicit forms of \( \tilde{l}_{/y}, \tilde{l}_{/tt}, \) and \( \tilde{l}_{/ygg} \) which correspond to \( A_j, D_j, E_j \) respectively derived for the Milstein approximate transition densities
in Section 3 of Choi, Jeong, and Park (2013). From
\[
\begin{align*}
\bar{I}_{0/yt} - \bar{I}_{yt} &= \frac{\mu_0 - \mu}{\sigma_0^2}, \\
\bar{I}_{0/yt} - \bar{I}_{yt} &= -\frac{\mu_0^2 - \mu^2}{\sigma_0^2} + \frac{3(\mu_0 - \mu)\sigma_{0/x}^2}{2\sigma_0^2}, \\
\bar{I}_{0/yt} - \bar{I}_{yt} &= -\frac{3(\mu_0 - \mu)\sigma_{0/x}^2}{2\sigma_0^4},
\end{align*}
\]
we get
\[
C^S = \mathbb{E}\left[\frac{(\mu_0 - \mu)\mu_0}{\sigma_0^2} - \frac{1}{2}\frac{\mu_0^2 - \mu^2}{\sigma_0^2}\right] = \frac{1}{2} \mathbb{E}\left[\frac{(\mu_0 - \mu)^2}{\sigma_0^2}\right].
\]

**B.4 Proof of Theorem 3.2**

**Proof.** The limiting distribution is derived in Theorem 4.3 in Jeong and Park (2013). We only need to prove that the inverse asymptotic variances of diffusion and drift parameter estimators are the same as the second derivatives of the primary and secondary criteria respectively. For the Fisher information matrix for \(\hat{\beta}\), we first get the first derivative with respect to \(\beta_i\)
\[
\frac{\partial C^P(\beta)}{\partial \beta_i} = \frac{1}{2} \mathbb{E}\left[\frac{1}{\sigma^2} \left(1 - \frac{\sigma_0^2}{\sigma^2}\right) \frac{\partial \sigma^2}{\partial \beta_i}\right]
\]
by the dominated convergence theorem. Similarly, we get the second derivative with respect to \(\beta_j\)
\[
\frac{\partial^2 C^P(\beta)}{\partial \beta_i \partial \beta_j} = \frac{1}{2} \mathbb{E}\left[\frac{-\sigma^2}{\sigma^4} \left(1 - \frac{\sigma_0^2}{\sigma^2}\right) \frac{\partial \sigma^2}{\partial \beta_i} + \frac{1}{\sigma^2} \left(\frac{\sigma_0^2}{\sigma^4} \frac{\partial \sigma^2}{\partial \beta_i}\right) \frac{\partial \sigma^2}{\partial \beta_j} + \frac{1}{\sigma^2} \left(1 - \frac{\sigma_0^2}{\sigma^2}\right) \frac{\partial^2 \sigma^2}{\partial \beta_i \partial \beta_j}\right].
\]
Since \(\sigma^2 = \sigma_0^2\) at \(\beta = \beta_0\), we get
\[
\begin{align*}
g^P_{ij}(\beta) &= \frac{\partial^2 C^P(\beta)}{\partial \beta_i \partial \beta_j} = \frac{1}{2} \mathbb{E}\left[\frac{\sigma_0^2}{\sigma^6} \frac{\partial \sigma^2}{\partial \beta_i} \frac{\partial \sigma^2}{\partial \beta_j}\right] = \frac{1}{2} \mathbb{E}\left[\frac{1}{\sigma_0^2} \left(\frac{\partial \sigma^2}{\partial \beta_i} \frac{\partial \sigma^2}{\partial \beta_j}\right)\right],
\end{align*}
\]
which is the inverse of the asymptotic variance of the diffusion parameter estimator \(\hat{\beta}\) in Theorem 4.3 in Jeong and Park (2013).

For the Fisher information matrix for \(\hat{\alpha}\), we first collect the part of \(C^S(\theta)\) that depends
on \( \alpha \) with \( \sigma = \sigma_0 \). That part is given by

\[
\frac{1}{2} E \left[ \frac{1}{\sigma^2} (\mu^2 - 2\mu_0 \mu) \right].
\]

From the dominated convergence theorem, the second derivative is given by

\[
g_{i,j}(\theta) = \frac{\partial^2 C_S(\theta)}{\partial \alpha_i \partial \alpha_j} = E \left[ \frac{1}{\sigma^2} \left( \frac{\partial \mu}{\partial \alpha_i} \frac{\partial \mu}{\partial \alpha_j} \right) \right],
\]

which is the inverse of the asymptotic variance of the drift parameter estimator \( \hat{\alpha} \) in Theorem 4.3 in Jeong and Park (2013).

### B.5 Proof of Theorem 3.1

**Proof.** Lemma B.7 of Choi, Jeong, and Park (2013) gives the asymptotic first order condition that the probability limit \( \theta^* \) of the MLE satisfies under \( \Delta^3 T \to 0 \) as \( \Delta \to 0 \) and \( T \to \infty \). The asymptotic first order order condition in their Lemma B.7 can be rewritten as

\[
E \bar{A}_{i/\beta} = E \bar{A}_{i/\alpha} = 0,
\]

which is the first order condition of minimizing the primary and secondary criteria. Therefore, \( \theta^* \) also minimizes the criteria.

### B.6 Proof of Theorem 3.3

**Proof.** Define the operator

\[ B = \sigma_0(y) \frac{\partial}{\partial y}. \]

Let \( \bar{A}_{i/\beta} = \partial(\bar{A})_{i/\beta} / \partial \beta \) and \( \sigma^{-2}_{i/\beta} = \partial \sigma^{-2} / \partial \beta \) and define other derivatives such as \( \bar{A}_{i/\beta} \), \( \bar{A}_{i/\alpha} \), \( \bar{A}_{i/\alpha} \), \( \bar{A}_{i/\alpha} \) similarly. When these functions are used in integrands, the argument is \( X_t \) which is omitted for notational simplicity. From the proof of Theorem 3.1 of Choi, Jeong, and Park (2013), under \( \Delta^3 T \to 0 \), the score functions of \( \beta \) and \( \alpha \) are given by

\[
s_\beta(\theta) = \frac{1}{\Delta} \int_0^T A_{i/\beta} dt + \frac{1}{2} \int_0^T A_{i/\beta}^2 dt + \frac{1}{\sqrt{2\Delta}} \int_0^T B_{i/\beta} dW_t + O_p(\sqrt{T}),
\]

\[
s_\alpha(\theta) = \frac{1}{2} \int_0^T A_{i/\alpha} dt + \int_0^T A_{i/\alpha} dW_t + O_p(\sqrt{\Delta T}) + O_p(\Delta T),
\]

32
where $V$ is a standard Brownian motion independent of $W$. The asymptotic orders of the terms in the above theorem are

$$\frac{1}{\Delta} \int_0^T \mathcal{A}_t dt = O_p\left(\frac{\sqrt{T}}{\Delta}\right), \quad \frac{1}{2} \int_0^T \mathcal{A}_t^2 dt = O_p(T), \quad \frac{1}{\sqrt{2\Delta}} \int_0^T B_t^2 dt = O_p\left(\frac{\sqrt{T}}{\Delta}\right),$$

$$\frac{1}{\Delta} \int_0^T \mathcal{A}_t^2 dt = O_p(\sqrt{T}), \quad \int_0^T \mathcal{A}_t B_t dt = O_p(\sqrt{T}).$$

The explicit forms of the leading terms of $s_\beta(\theta)$ and $s_\alpha(\theta)$ are calculated from

$$\mathcal{A}_t = \frac{1}{2}(\sigma_0^2 - \sigma^2)\sigma_0^{-2},$$

$$\mathcal{A}_t^2 = \frac{2\mu_\alpha}{\sigma_0^2}(\mu_0 - \mu) + \left(1 - \frac{\sigma_0^2}{\sigma^2}\right)\left(-\mu_0 + \frac{\mu_\alpha \sigma_0^2}{\sigma^2}\right),$$

$$\mathcal{AB}_t = \sigma_0 \frac{\mu_\alpha}{\sigma_0^2}.$$

We get the main result of the theorem by plugging in these terms in (B.6)-(B.7).

**B.7 Proof of Theorem 3.4**

From (B.6)-(B.7), under $\Delta^2 T \to 0$ as $\Delta \to 0$ and $T \to \infty$, we have

$$\frac{\Delta}{\sqrt{T}} s_\beta(\theta^*) - \left(\frac{1}{\sqrt{T}} \int_0^T \mathcal{A}_t dt\right) \xrightarrow{p} 0$$

and

$$\frac{1}{\sqrt{T}} s_\alpha(\theta^*) - \left(\frac{1}{\sqrt{T}} \int_0^T \frac{1}{2} \mathcal{A}_t^2 dt + \frac{1}{\sqrt{T}} \int_0^T \mathcal{A}_t B_t dt\right) \xrightarrow{p} 0.$$

Therefore, we only need to get the asymptotic distributions of

$$\frac{1}{\sqrt{T}} \int_0^T \mathcal{A}_t dt$$

for $s_\beta(\theta^*)$ and

$$\frac{1}{\sqrt{T}} \int_0^T \frac{1}{2} \mathcal{A}_t^2 dt + \frac{1}{\sqrt{T}} \int_0^T \mathcal{A}_t B_t dt$$
for $s_{\alpha}(\theta^*)$. We can derive the limiting distributions of $\frac{1}{\sqrt{T}} \int_0^T A_{i/\beta} \, dt$ and $\frac{1}{\sqrt{T}} \int_0^T \frac{1}{2} A^2 l_{i/\alpha} \, dt$ from Lemma B.1 of Choi, Jeong, and Park (2013). Their lemma proves that for $f(X_t)$ such that $\mathbf{E} f(X_t) = 0$, we have

$$\frac{1}{\sqrt{T}} \int_0^T f(X_t) \, dt \overset{d}{\to} \sqrt{\mathbf{E} V(X_t)^2} W_1,$$

where

$$V(x) = \sigma_0(x)s_0(x) \int_{x_0}^x f(v)m_0(v)dv$$

for $x_0 \in \mathcal{D}$. Using the fact $\mathbf{E} A_{i/\beta} = \mathbf{E} A^2 l_{i/\alpha} = 0$ from (B.5), we apply their lemma to $\frac{1}{\sqrt{T}} \int_0^T A_{i/\beta} \, dt$ and $\frac{1}{\sqrt{T}} \int_0^T \frac{1}{2} A^2 l_{i/\alpha} \, dt$, which gives us

$$\frac{\Delta}{\sqrt{T}} s_{\beta}(\theta^*) \overset{d}{\to} N(0, \mathbf{E} V_{\beta} V_{\beta}^T), \quad (B.8)$$

$$\frac{1}{\sqrt{T}} s_{\alpha}(\theta^*) \overset{d}{\to} N(0, \mathbf{E} V_{\alpha} V_{\alpha}^T), \quad (B.9)$$

where $V_{\beta} = V_{\beta}(X_0)$ and $V_{\alpha} = V_{\alpha}(X_0)$ are defined from the functions

$$V_{\beta}(x) = \sigma_0(x)s_0(x) \int_{x_0}^x A_{i/\beta}(v) m_0(v)dv,$$

$$V_{\beta}(x) = \sigma_0(x)s_0(x) \int_{x_0}^x \frac{1}{2} A^2 l_{i/\alpha}(v) m_0(v)dv + \sigma_0(x) \frac{\mu_{i/\alpha}(x)}{\sigma^2(x)}.$$

For the expected Hessian matrices, we use the relationships

$$h_{\beta}(\theta^*) = \frac{1}{\Delta} \int_0^T A_{i/\beta} \, dt + O_p \left( \frac{\sqrt{T}}{\Delta} \right) + O_p(T),$$

$$h_{\alpha}(\theta^*) = \frac{1}{2} \int_0^T A^2 l_{i/\alpha} \, dt + O_p(\sqrt{T}) + O_p(\Delta T),$$
which are shown in the proof of Lemma B.6 of [Choi, Jeong, and Park (2013)]. From

\[
\frac{\Delta}{T} h_{\beta}(\theta^*) - \left( \frac{1}{T} \int_0^T A\tilde{l}_{/\beta\beta} \, dt \right) \xrightarrow{p} 0,
\]

\[
\frac{1}{T} h_{\alpha}(\theta^*) - \left( \frac{1}{T} \int_0^T \frac{1}{2} A^2\tilde{l}_{/\alpha\alpha} \, dt \right) \xrightarrow{p} 0,
\]
as \(\Delta \to 0\) and \(T \to \infty\), and

\[
\frac{1}{T} \int_0^T A\tilde{l}_{/\beta\beta} \, dt \xrightarrow{p} E\tilde{l}_{/\beta\beta},
\]

\[
\frac{1}{T} \int_0^T \frac{1}{2} A^2\tilde{l}_{/\alpha\alpha} \, dt \xrightarrow{p} \frac{1}{2} E A^2\tilde{l}_{/\alpha\alpha},
\]
as \(T \to \infty\), the asymptotic expected Hessian matrices of \(\beta\) and \(\alpha\) are the same as the negative second derivatives of the primary and secondary criteria respectively. The explicit forms are calculated by taking the second derivatives of the criteria.

**B.8 Proof of Corollary 3.1**

**Proof.** Under \(\Delta^2 T \to 0\) as \(\Delta \to 0\) and \(T \to \infty\), we get

\[
\sqrt{T}(\hat{\beta} - \beta^*) - \left( -\frac{\Delta}{T} h_{\beta}(\theta^*) \right)^{-1} \left( \frac{\Delta}{\sqrt{T}} s_{\beta}(\theta^*) \right) \xrightarrow{p} 0,
\]

\[
\sqrt{T}(\hat{\alpha} - \alpha^*) - \left( -\frac{1}{T} h_{\alpha}(\theta^*) \right)^{-1} \left( \frac{1}{\sqrt{T}} s_{\alpha}(\theta^*) \right) \xrightarrow{p} 0,
\]

from the proof of Theorem 3.1 of [Choi, Jeong, and Park (2013)]. We only need to show the asymptotic distributions of \(\left( -\frac{\Delta}{T} h_{\beta}(\theta^*) \right)^{-1} \left( \frac{\Delta}{\sqrt{T}} s_{\beta}(\theta^*) \right)\) and \(\left( -\frac{1}{T} h_{\alpha}(\theta^*) \right)^{-1} \left( \frac{1}{\sqrt{T}} s_{\alpha}(\theta^*) \right)\). Their asymptotic distributions are from the results of Theorem 3.4 and (B.8)-(B.9). \(\blacksquare\)
References


