Ann Inst Stat Math (2010) 62:161–187
DOI 10.1007/s10463-009-0245-1

Contrast-based information criterion for ergodic diffusion processes from discrete observations

Masayuki Uchida

Received: 24 June 2008 / Revised: 22 April 2009 / Published online: 28 July 2009

Abstract In this paper, we consider the model selection problem for discretely observed ergodic multi-dimensional diffusion processes. In order to evaluate the statistical models, Akaike's information criterion (AIC) is a useful tool. Since AIC is constructed by the maximum log likelihood and the dimension of the parameter space, it may look easy to get AIC even for discretely observed diffusion processes. However, there is a serious problem that a transition density of a diffusion process does not generally have an explicit form. Instead of the exact log-likelihood, we use a contrast function based on a locally Gaussian approximation of the transition density and we propose the contrast-based information criterion.

Keywords Akaike's information criteria · Model selection · Malliavin calculus · Maximum contrast estimator · Large deviation inequality · Discrete time observation

1 Introduction

We consider a d-dimensional diffusion process defined by the stochastic differential equation

$$\mathrm{d}X_t = B(X_t)\mathrm{d}t + S(X_t)\mathrm{d}w_t, \quad t \in [0, T], \quad X_0 = x_0, \tag{1}$$

where B is an $\mathbf{R}^d$-valued function defined on $\mathbf{R}^d$, S is an $\mathbf{R}^d \otimes \mathbf{R}^d$-valued function defined on $\mathbf{R}^d$, w is a d-dimensional standard Wiener process and x_0 is a deterministic initial condition. The data we treat are discrete observations $\mathbf{X}_n = (X_{t_k^n})_{0 \le k \le n}$ with $t_k^n = kh_n$, where h_n is the discretization step and $nh_n = T$. Based on the

M. Uchida (✉)
Graduate School of Engineering Science, Osaka University, Toyonaka, Osaka 560-8531, Japan
e-mail: uchida@sigmath.es.osaka-u.ac.jp

discrete observations, we consider a model selection problem among the following M parametric models. For $m = 1, \dots, M$,

$$\mathrm{d}X_t = b_m(X_t, \alpha_m)\mathrm{d}t + \sigma_m(X_t, \beta_m)\mathrm{d}w_t, \quad t \in [0, T], \quad X_0 = x_0, \tag{2}$$

where $\theta_m = (\alpha_m, \beta_m) \in \Theta_{\alpha_m} \times \Theta_{\beta_m} = \Theta_m$, Θ_{α_m} and Θ_{β_m} are, respectively, bounded domains in $\mathbf{R}^{p_m}$ and $\mathbf{R}^{q_m}$ with a locally Lipschitz boundary, which means that Θ_{α_m} and Θ_{β_m} satisfy the strong local Lipschitz condition, see Adams and Fournier (2003). Furthermore, b_m is an $\mathbf{R}^d$-valued function defined on $\mathbf{R}^d \times \Theta_{\alpha_m}$ and σ_m is an $\mathbf{R}^d \otimes \mathbf{R}^d$-valued function defined on $\mathbf{R}^d \times \Theta_{\beta_m}$, where the drift b_m and the diffusion coefficient σ_m are known apart from the parameters α_m and β_m, respectively. We assume that the process X is ergodic for every θ_m with invariant probability measure μ_{θ_m}. For details of ergodicity and the invariant probability measures for diffusion processes, see Kutoyants (2004). Moreover, we assume that all the parametric models (2) include the true model (1), that is, for $m = 1, \dots, M$, there exists $\theta_{m,0} = (\alpha_{m,0}, \beta_{m,0}) \in \Theta_m$ such that $b_m(x, \alpha_{m,0}) = B(x)$ and $\Xi_m(x, \beta_{m,0}) = [SS^\star](x)$ for all x, where $\Xi_m(x, \beta_m) = [\sigma_m \sigma_m^\star](x, \beta_m)$, $\star$ means the transpose, $\alpha_{m,0}$ and $\beta_{m,0}$ are true values of α_m and β_m, respectively. The type of asymptotics we treat is when $h_n \to 0$, $nh_n \to \infty$ and $nh_n^2 \to 0$ as $n \to \infty$. Moreover, we assume that for some $\epsilon_0 \in (0, 1/2)$, $n^{\epsilon_0} \leq nh_n$ for large n, see Yoshida (2005).

As is well known, Akaike (1973, 1974) proposed Akaike's information criterion (AIC) for the problem of choosing a statistical model among the correctly specified parametric models which include the true model. AIC for the mth model is

$$\mathrm{AIC}(\mathbf{X}_n, m) = -2l_{m,n}\left(\mathbf{X}_n, \hat{\theta}_m^{(\mathrm{ML})}(\mathbf{X}_n)\right) + 2\mathrm{dim}(\Theta_m), \tag{3}$$

where $l_{m,n}(\mathbf{X}_n, \theta_m)$ is the log likelihood function for the mth model and $\hat{\theta}_m^{\mathrm{ML}}(\mathbf{X}_n)$ is the maximum likelihood estimator (MLE) for an unknown parameter of the mth model. Under some regularity conditions, $\mathrm{AIC}(\mathbf{X}_n, m)$ is an asymptotically unbiased estimator for $-2 \times \mathrm{EL}(\mathbf{X}_n, m)$, where $\mathrm{EL}(\mathbf{X}_n, m)$ is the expected log likelihood of the m-th model defined as $\mathrm{EL}(\mathbf{X}_n, m) = E_{\mathbf{Z}_n}[l_{m,n}(\mathbf{Z}_n, \hat{\theta}_m^{(ML)}(\mathbf{X}_n))]$, $\mathbf{Z}_n$ is an independent copy of $\mathbf{X}_n$ and $E_{\mathbf{Z}_n}$ stands for the expectation under the law of $\mathbf{Z}_n$. Consequently, we choose a statistical model which minimizes the value of AIC among a set of competing models. Needless to say, there are many applications of model selection by means of information criteria, see, for example, Shibata (1976), Hall (1990), Burman and Nolan (1995), Hurvich et. al. (1998), Burnham and Anderson (2002), Sei and Komaki (2007), Konishi and Kitagawa (2008) and references therein. In order to obtain AIC defined by (3), it is enough to get both the log likelihood function and the MLE. For this reason, it may seem that there is no difficulty in constructing AIC even for multi-dimensional diffusion processes. For information criteria of continuously observed diffusion processes, see Uchida and Yoshida (2001, 2004, 2006). However, since the transition density $p_m(t, x, y; \theta_m)$ of the diffusion process (2) does not generally have an explicit form, there is a serious problem that we cannot explicitly get the log likelihood function $l_{m,n}(\mathbf{X}_n, \theta_m) = \sum_{k=1}^n \log p_m(h_n, X_{t_{k-1}^n}, X_{t_k^n}; \theta_m)$ for discretely observed diffusion model. The MLE cannot be also obtained because of it. Therefore, it is not a trivial problem to construct AIC for discretely observed diffusion models.

In this paper, in order to construct an information criterion for discretely observed ergodic diffusion models, we consider the contrast function based on a locally Gaussian approximation instead of the exact log likelihood function and we show that an information criterion based on the contrast function is an asymptotically unbiased estimator for the expected log likelihood. Asymptotically parametric estimation for discretely observed ergodic diffusions has been developed, see Prakasa Rao (1983, 1988), Florens-Zmirou (1989), Yoshida (1992b, 2005) and Kessler (1997). In order to show the asymptotic unbiasedness of the information criterion, however, we need the large deviation inequality for the normalized maximum contrast estimator and the estimates of moments for the derivatives of the log likelihood function $l_{m,n}(\theta_m)$ with respect to parameter θ_m. By using the similar argument as in Yoshida (2005), the large deviation inequality is obtained, see Lemma 1 below. In the analogous way as in Gobet (2001, 2002), the estimates for the derivatives of the log likelihood function are derived by means of Malliavin calculus, see Lemma 2 below. Moreover, we propose the contrast-based information criterion (CIC), and the asymptotic result of the difference between the contrast-based information criteria (CIC difference) is obtained.

The rest of this paper is organized as follows. In Sect. 2, the notation and assumptions are stated, and the asymptotic result of an information criterion constructed by the contrast function based on a locally Gaussian approximation is presented. Both CIC and CIC difference for discretely observed multi-dimensional ergodic diffusion processes are proposed. We also give examples of model selection problem based on CIC and simulation results. The result presented in Sect. 2 is proved in Sect. 3.

2 Contrast-based information criterion

2.1 Notation and assumptions

We introduce the notation used in this paper and the assumptions. Let $m \in \{1, \ldots, M\}$ for the suffix m in the notation and assumptions.

1. $\partial_x^{\mathbf{n}} = \partial_{x_1}^{n_1} \cdots \partial_{x_d}^{n_d}$ and $\partial_{\theta_m}^{\nu} = \partial_{\theta_{m,1}}^{\nu_1} \cdots \partial_{\theta_{m,p_m+q_m}}^{\nu_{p_m+q_m}}$, where $\mathbf{n} = (n_1, \ldots, n_d)$ and $\nu = (\nu_1, \ldots, \nu_{p_m+q_m})$ are multi-indices, $|\mathbf{n}| = n_1 + \cdots + n_d$, $|\nu| = \nu_1 + \cdots + \nu_{p_m+q_m}$, $\partial_{x_i} = \partial/\partial x_i$ and $\partial_{\theta_{m,j}} = \partial/\partial\theta_{m,j}$.
2. Let $C_b^{k,l}(\mathbf{R}^d \times \Theta_m; \mathbf{R}^d \otimes \mathbf{R}^d)$ be the space of functions f satisfying the following conditions: (i) $f(x, \theta_m)$ is an $\mathbf{R}^d \otimes \mathbf{R}^d$-valued function on $\mathbf{R}^d \times \Theta_m$ which is continuously differentiable with respect to (w.r.t.) x up to order k for all θ_m, and for $|\mathbf{n}| = 0, 1, \ldots, k$, $\partial_x^{\mathbf{n}} f(x, \theta_m)$ is continuously differentiable w.r.t. θ_m up to order l. (ii) For $|\mathbf{n}| = 0, 1, \ldots, k$ and $|\nu| = 0, 1, \ldots, l$, $\sup_{x,\theta_m} |\partial_{\theta_m}^{\nu} \partial_x^{\mathbf{n}} f(x, \theta_m)| < \infty$.
3. Let $\bar{C}_b^{k,l}(\mathbf{R}^d \times \Theta_m; \mathbf{R}^d)$ denote the space of functions f satisfying the following conditions: (i) $f(x, \theta_m)$ is an $\mathbf{R}^d$-valued function on $\mathbf{R}^d \times \Theta_m$ which is continuously differentiable w.r.t. x up to order k for all θ_m, and for $|\mathbf{n}| = 0, 1, \ldots, k$, $\partial_x^{\mathbf{n}} f(x, \theta_m)$ is continuously differentiable w.r.t. θ_m up to order l. (ii) $\sup_{x,\theta_m} |\partial_x^{\mathbf{n}} f(x, \theta_m)| < \infty$ for $|\mathbf{n}| = 1, \ldots, k$. (iii) For $|\nu| = 1$, there exists

a constant $C > 0$ such that $\sup_{\theta_m} |\partial^{\nu}_{\theta_m} f(x, \theta_m)| \leq C(1 + |x|)$ for all x, and for $|\nu| = 2, 3, \ldots, l$, $\sup_{x,\theta_m} |\partial^{\nu}_{\theta_m} f(x, \theta_m)| < \infty$. (iv) For $|\mathbf{n}| = 1, \ldots, k$ and $|\nu| = 1, \ldots, l$, $\sup_{x,\theta_m} |\partial^{\nu}_{\theta_m} \partial^{\mathbf{n}}_x f(x, \theta_m)| < \infty$.

4. P_{θ^*} and P_{θ_m} denote the laws of the processes defined by the Equations (1) and (2), respectively. Let E_{θ^*} and E_{θ_m} be the expectations under P_{θ^*} and P_{θ_m}, respectively. Note that $P_{\theta^*} = P_{\theta_{m,0}}$ and $E_{\theta^*} = E_{\theta_{m,0}}$. We make two sets of assumptions as follows.

A1 (i) $b_m(x, \alpha_m) \in \bar{C}^{5,5}_b(\mathbf{R}^d \times \Theta_{\alpha_m}; \mathbf{R}^d)$. (ii) $\sigma_m(x, \beta_m) \in C^{5,5}_b(\mathbf{R}^d \times \Theta_{\beta_m}; \mathbf{R}^d \otimes \mathbf{R}^d)$.
A2 (i) There exist constants $c_0 > 0$ and $K > 0$ such that for any $(x, \alpha_m) \in \mathbf{R}^d \times \Theta_{\alpha_m}$,

$$x^{\star} b_m(x, \alpha_m) \leq -c_0 |x|^2 + K.$$

(ii) $\sigma_m(x, \beta_m)$ is symmetric and there exists a constant $c_1 \geq 1$ such that for any $(x, \beta_m, \lambda) \in \mathbf{R}^d \times \Theta_{\beta_m} \times \mathbf{R}$,

$$\frac{1}{c_1}|\lambda|^2 \leq \lambda^{\star} \sigma_m(x, \beta_m) \lambda \leq c_1 |\lambda|^2.$$

Remark 1 It follows from $A1$ and $A2$ that we can show the following results. (i) For every $\theta_m \in \Theta_m$, the process X admits a unique invariant probability measure. We denote it by μ_{θ_m}. Moreover, for every $\theta_m \in \Theta_m$, $\int_{\mathbf{R}^d} |x|^p \mu_{\theta_m}(\mathrm{d}x) < \infty$ for all $p \geq 0$. (ii) $\sup_t E_{\theta_m}[|X_t|^p] < \infty$ for all $p \geq 0$. (iii) For every $\theta_m \in \Theta_m$, the process X is ergodic, in special, for every measurable function satisfying $|f| \leq c(1 + |x|)^c$ for some $c > 0$, one has $\frac{1}{T}\int_0^T f(X_t)\mathrm{d}t \xrightarrow{p} \int_{\mathbf{R}^d} f(x)\mu_{\theta_{m,0}}(\mathrm{d}x)$ as $T \to \infty$.

Let $I_m(\theta_{m,0})$ denote the Fisher information matrix as follows:

$$I_m(\theta_{m,0}) = \begin{pmatrix} ((I_{b,m}(\theta_{m,0}))_{ij})_{i,j=1,\ldots,p_m} & 0 \\ 0 & ((I_{\sigma,m}(\theta_{m,0}))_{ij})_{i,j=1,\ldots,q_m} \end{pmatrix},$$

where

$$(I_{b,m}(\theta_{m,0}))_{ij} = \int_{\mathbf{R}^d} (\partial_{\alpha_{m,i}} b_m)^{\star}(x, \alpha_{m,0})\, \Xi_m^{-1}(x, \beta_{m,0}) \partial_{\alpha_{m,j}} b_m(x, \alpha_{m,0}) \mu_{\theta_{m,0}}(\mathrm{d}x),$$
$$(I_{\sigma,m}(\theta_{m,0}))_{ij} = \frac{1}{2}\int_{\mathbf{R}^d} \mathrm{tr}\left[(\partial_{\beta_{m,i}} \Xi_m)\Xi_m^{-1}(\partial_{\beta_{m,j}} \Xi_m)\Xi_m^{-1}(x, \beta_{m,0})\right] \mu_{\theta_{m,0}}(\mathrm{d}x).$$

Moreover, we make the following two assumptions, which are needed to estimate unknown parameter θ_m.
A3 $I_m(\theta_{m,0})$ is positive definite.
A4 If $b_m(x, \alpha_m) = b_m(x, \alpha_{m,0})$ for $\mu_{\theta_{m,0}}$ a.s. all x, then $\alpha_m = \alpha_{m,0}$. If $\Xi_m(x, \beta_m) = \Xi_m(x, \beta_{m,0})$ for $\mu_{\theta_{m,0}}$ a.s. all x, then $\beta_m = \beta_{m,0}$.

2.2 Main result

In this section, we consider an information criterion for correctly specified parametric models under the situation when $h_n \to 0$, $nh_n \to \infty$ and $nh_n^2 \to 0$ as $n \to \infty$, and for some $\epsilon_0 \in (0, 1/2)$, $n^{\epsilon_0} \le nh_n$ for large n. In order to get the information criterion, we use the following contrast function based on a locally Gaussian approximation:

$$\begin{aligned} g_{m,n}(\theta_m) = & -\frac{nd}{2}\log(2\pi h_n) - \frac{1}{2}\sum_{k=1}^{n}\log\det\left(\Xi_m\left(X_{t_{k-1}^n}, \beta_m\right)\right) \\ & -\frac{1}{2h_n}\sum_{k=1}^{n}\Xi_m^{-1}\left(X_{t_{k-1}^n}, \beta_m\right)\left[\left(X_{t_k^n} - X_{t_{k-1}^n} - h_n b_m\left(X_{t_{k-1}^n}, \alpha_m\right)\right)^{\otimes 2}\right] \end{aligned} \tag{4}$$

for $m \in \{1, 2, \ldots, M\}$. The maximum contrast estimator $\hat{\theta}_{m,n}$ is defined as $g_{m,n}(\hat{\theta}_{m,n}) = \sup_{\theta_m} g_{m,n}(\theta_m)$. Moreover, we set that an information criterion based on the contrast function $\mathrm{IC}(\mathbf{X}_n, m)$ and the expected log likelihood $\mathrm{EL}(\mathbf{X}_n, m)$ are

$$\begin{aligned} \mathrm{IC}(\mathbf{X}_n, m) &= g_{m,n}(\hat{\theta}_{m,n}(\mathbf{X}_n)) - g_{m,n}(\theta_{m,0}) + E_{\mathbf{Z}_n}\left[l_{m,n}(\mathbf{Z}_n, \theta_{m,0})\right] - \dim(\Theta_m), \\ \mathrm{EL}(\mathbf{X}_n, m) &= E_{\mathbf{Z}_n}\left[l_{m,n}(\mathbf{Z}_n, \hat{\theta}_{m,n}(\mathbf{X}_n))\right], \end{aligned}$$

where $\mathbf{Z}_n$ is an independent copy of $\mathbf{X}_n$ and $E_{\mathbf{Z}_n}$ means the expectation under the law of $\mathbf{Z}_n$.

The asymptotic result of $\mathrm{IC}(\mathbf{X}_n, m)$ is as follows.

Theorem 1 *Let $m \in \{1, 2, \ldots, M\}$. Suppose that* A1–A4 *hold true. Then, as $n \to \infty$,*

$$E_{\theta^*}\left[\mathrm{IC}(\mathbf{X}_n, m) - \mathrm{EL}(\mathbf{X}_n, m)\right] = o(1).$$

It follows from Theorem 1 that $\mathrm{IC}(\mathbf{X}_n, m)$ is an asymptotically unbiased estimator for $\mathrm{EL}(\mathbf{X}_n, m)$, and by the similar argument as AIC, the optimal model m^* among competing models is selected by $\mathrm{IC}(\mathbf{X}_n, m^*) = \max_{m=1,\ldots,M}\mathrm{IC}(\mathbf{X}_n, m)$. However, $\mathrm{IC}(\mathbf{X}_n, m)$ is an impracticable criterion since $\theta_{m,0}$ is unknown and both $g_{m,n}(\theta_{m,0})$ and $E_{\mathbf{Z}_n}[l_{m,n}(\mathbf{Z}_n, \theta_{m,0})]$ cannot be calculated. Fortunately, one has that for $i, j \in \{1, \ldots, M\}$, $g_{i,n}(\theta_{i,0}) = g_{j,n}(\theta_{j,0})$ and $l_{i,n}(\theta_{i,0}) = l_{j,n}(\theta_{j,0})$ because we consider the correctly specified parametric models. Thus, we define the CIC for the mth model as

$$\mathrm{CIC}(\mathbf{X}_n, m) = -2g_{m,n}(\hat{\theta}_{m,n}(\mathbf{X}_n)) + 2\dim(\Theta_m),$$

and the optimal model m^* among competing models is selected by $\mathrm{CIC}(\mathbf{X}_n, m^*) = \min_{m=1,\ldots,M}\mathrm{CIC}(\mathbf{X}_n, m)$ since one has that $\arg\min_{m=1,\ldots,M}\mathrm{CIC}(\mathbf{X}_n, m) = \arg\max_{m=1,\ldots,M}\mathrm{IC}(\mathbf{X}_n, m)$.

Let $\mathrm{KL}(\mathbf{X}_n, m)$ denote the estimated Kullback–Leibler information for the true model $l_n(\cdot)$ and the mth statistical model $l_{m,n}(\cdot, \hat{\theta}_m(\mathbf{X}_n))$ defined by $\mathrm{KL}(\mathbf{X}_n, m) =$

$E_{\mathbf{Z}_n}[l_n(\mathbf{Z}_n)] - E_{\mathbf{Z}_n}[l_{m,n}(\mathbf{Z}_n, \hat{\theta}_m(\mathbf{X}_n))]$. We set that for $i, j \in \{1, 2, \ldots, M\}$,

$$\mathrm{DCIC}(\mathbf{X}_n, i, j) = \mathrm{CIC}(\mathbf{X}_n, i) - \mathrm{CIC}(\mathbf{X}_n, j) = -2(\mathrm{IC}(\mathbf{X}_n, i) - \mathrm{IC}(\mathbf{X}_n, j)),$$
$$\mathrm{DKL}(\mathbf{X}_n, i, j) = 2(\mathrm{KL}(\mathbf{X}_n, i) - \mathrm{KL}(\mathbf{X}_n, j)) = -2(\mathrm{EL}(\mathbf{X}_n, i) - \mathrm{EL}(\mathbf{X}_n, j)).$$

As a corollary of Theorem 1, we obtain the following result of the CIC difference.

Corollary 1 *Let* $i, j \in \{1, 2, \ldots, M\}$. *Suppose that* A1–A4 *hold true. Then, as* $n \to \infty$,

$$E_{\theta^*}[\mathrm{DCIC}(\mathbf{X}_n, i, j) - \mathrm{DKL}(\mathbf{X}_n, i, j)] = o(1).$$

Remark 2 (i) By the analogous argument as CIC together with Theorem 1, we choose the statistical model i^* as the optimal model among the competing models if DCIC $(\mathbf{X}_n, i^*, j) < 0$ for all $j \neq i^*$. (ii) In the same way as in Inagaki and Ogata (1975), if some of the competing models do not include a true model, the probability that the misspecified statistical model is selected by CIC converges to zero as $n \to \infty$, see Sect. 2.3.

Note that a bounded drift coefficient does not meet A2. Instead of A1–A2, we assume A1′ and A2′ as follows. A1′: A1–(ii) and $b_m(x, \alpha_m) \in C_b^{5,5}(\mathbf{R}^d \times \Theta_{\alpha_m}; \mathbf{R}^d)$. A2′: A2–(ii) and there exist constants $c_0 > 0$ and $K_0 > 0$ such that $x^\star b_m(x, \alpha_m) \leq -c_0|x|$ for every $(x, \alpha_m) \in \mathbf{R}^d \times \Theta_{\alpha_m}$ satisfying $|x| \geq K_0$. Then, using a version of Lemma 2 below together with the results of Section 5 in Gobet (2002), we can show that under A1′–A2′ and A3–A4, the assertions of Theorem 1 and Corollary 1 hold true.

The proposed criterion may be appropriate to be called an AIC-type criterion because the criterion is constructed by the maximum contrast and the dimension of the parameter space. However, both TIC (Takeuchi 1976) and GIC (Konishi and Kitagawa 1996), which are general information criteria for misspecified models, are also based on the exact log likelihood. On the other hand, the proposed information criterion is based on the contrast function instead of the exact log likelihood, and we call it the CIC for correctly specified parametric models in order to distinguish the existing information criteria and the proposed information criterion.

The predictive distribution of the mth diffusion model from discrete observations is $p_m(t, x, y; \hat{\theta}_{m,n})$, where $p_m(t, x, y; \theta_m)$ is the transition density of the mth diffusion model and $\hat{\theta}_{m,n}$ is the maximum contrast estimator obtained from the contrast function $g_{m,n}$. However, $p_m(t, x, y; \theta_m)$ does not generally have an explicit form. Fortunately, it is possible to choose the optimal diffusion model m^* in the sense of CIC, which is based on the minimization of the estimated Kullback–Leibler information. Using a suitable approximation method, we need to obtain an approximate transition density of the m^*th diffusion model. For example, Aït-Sahalia (2008) derived closed-form expansions for the log-transition density and Beskos et al. (2006) provided numerical approximations of the transition density. The approximation of the transition density of a diffusion process is a very challenging problem and the validity of the approximate predictive distribution is a future work.

Under the assumption that all the parametric models contain the true model, we proposed CIC whose validity was based on Theorem 1. As seen from the derivation of CIC, however, it is impossible to extend CIC to an information criterion for misspecified diffusion models such as TIC and GIC. In order to obtain a general information criterion for misspecified diffusion models, we will need a different approach from the derivation of CIC. This generalization will be also a future project.

2.3 Examples and simulation results

As an example of a model selection based on CIC, we treat the following setting. The data $\mathbf{X}_n = (X_{t_k^n})_{0\le k\le n}$ are obtained from the true model defined by

$$\mathrm{d}X_t = -(X_t - 5)\mathrm{d}t + 30\mathrm{d}w_t, \quad t \in [0, T], \quad X_0 = 5.$$

We consider the model selection problem for the following three statistical models:

$$\text{(Model 1)} \quad \mathrm{d}X_t = -\alpha_1(X_t - \alpha_2)\mathrm{d}t + \beta_1 \mathrm{d}w_t, \tag{5}$$

$$\text{(Model 2)} \quad \mathrm{d}X_t = -\alpha_1(X_t - \alpha_2)\mathrm{d}t + \left(\frac{\beta_1 + \beta_2 X_t^2}{1 + X_t^2}\right)\mathrm{d}w_t, \tag{6}$$

$$\text{(Model 3)} \quad \mathrm{d}X_t = -\alpha_1(X_t - \alpha_2)\mathrm{d}t + \left(\frac{\beta_1 + \beta_3 X_t + \beta_2 X_t^2}{1 + X_t^2}\right)\mathrm{d}w_t, \tag{7}$$

where $\alpha_1 > 0$, $\alpha_2 > 0$, $\beta_1 > 0$, $\beta_2 > 0$ and $\beta_3^2 < 4\beta_1\beta_2$.

It follows from (4) that for the models (5), (6) and (7), we have the contrast functions $g_{1,n}(\theta_1)$, $g_{2,n}(\theta_2)$ and $g_{3,n}(\theta_3)$, respectively. Thus, CIC of the models (5), (6), (7) are $\mathrm{CIC}(\mathbf{X}_n, 1) = -2g_{1,n}(\mathbf{X}_n, \hat{\theta}_{1,n}) + 2 \times 3$, $\mathrm{CIC}(\mathbf{X}_n, 2) = -2g_{2,n}(\mathbf{X}_n, \hat{\theta}_{2,n}) + 2 \times 4$ and $\mathrm{CIC}(\mathbf{X}_n, 3) = -2g_{3,n}(\mathbf{X}_n, \hat{\theta}_{3,n}) + 2 \times 5$, respectively, where $\hat{\theta}_{i,n}$ is the maximum contrast estimator obtained from the contrast function $g_{i,n}$ for $i = 1, 2, 3$.

We examine the number of models selected by CIC among the competing models (5), (6), (7) for 10,000 independent sample paths generated by the Milstein scheme through simulations. For details of the Milstein scheme, see Kloeden and Platen (1992). The simulations are done for each $T = 10, 30$ and $h_n = 1/50, 1/200$.

In Table 1, Model 1 is selected with high frequency as the optimal model for all cases. However, either Model 2 or Model 3 is selected in a significant probability. This result implies that CIC does not have consistency for estimating the minimal model i^* such that $i^* = \arg\min_i \dim(\Theta_i)$ among the i competing models including the true model. Note that in this example, the minimal model is Model 1. However, this inconsistency is not a weak point of CIC. We must note that CIC is a tool to choose the optimal model among competing models from the aspect of both model-fitting and prediction.

Next, we consider the situation where the true model is defined by

$$\mathrm{d}X_t = -(X_t - 5)\mathrm{d}t + \left(\frac{30 + 5X_t^2}{1 + X_t^2}\right)\mathrm{d}w_t, \quad t \in [0, T], \quad X_0 = 5.$$

Table 1 The number of models selected by CIC for 10,000 independent simulated sample paths in the case that the true model is defined by $\mathrm{d}X_t = -(X_t - 5)\mathrm{d}t + 30\mathrm{d}w_t$, $t \in [0, T]$, $X_0 = 5$

T	h_n	Model 1	Model 2	Model 3
10	1/50	7,605	1,435	960
	1/200	7,848	1,361	791
30	1/50	7,676	1,366	958
	1/200	7,822	1,360	818

Table 2 The number of models selected by CIC for 10,000 independent simulated sample paths in the case that the true model is defined by $\mathrm{d}X_t = -(X_t - 5)\mathrm{d}t + \left(\frac{30+5X_t^2}{1+X_t^2}\right)\mathrm{d}w_t$, $t \in [0, T]$, $X_0 = 5$.

T	h_n	Model 1	Model 2	Model 3
10	1/50	108	6,305	3,587
	1/200	8	7,540	2,452
30	1/50	0	7,086	2,914
	1/200	0	7,497	2,503

The competing models, T and h_n are the same as the previous example. Note that Model 1 does not include the true model, which is the misspecified model.

In Table 2, the probability that Model 1 is selected by CIC is very low, while Model 2 is selected with high probability for all cases. This result indicates that if a model does not include the true model, the probability that the misspecified model is selected by CIC tends to zero as $n \to \infty$.

3 Proofs

Let $\theta_m = (\theta_{m,1}, \ldots, \theta_{m,p_m+q_m})^\star = (\alpha_{m,1}, \ldots, \alpha_{m,p_m}, \beta_{m,1}, \ldots, \beta_{m,q_m})^\star$, $\partial_{m,i} = \partial/\partial_{\theta_{m,i}}$ and

$$d_{m,l,n} = \begin{cases} \dfrac{1}{\sqrt{nh_n}} & \text{if } l = 1, \ldots, p_m, \\ \dfrac{1}{\sqrt{n}} & \text{if } l = p_m + 1, \ldots, p_m + q_m. \end{cases}$$

Set $\hat{u}_{m,l,n} = d_{m,l,n}^{-1}(\hat{\theta}_{m,l,n} - \theta_{m,l,0})$, $\partial \hat{g}_{m,l,n} = d_{m,l,n}\partial_{m,l}g_{m,n}(\theta_{m,0})$,

$$\hat{u}_{m,n} = \begin{pmatrix} \sqrt{nh_n}(\hat{\alpha}_{m,n} - \alpha_{m,0}) \\ \sqrt{n}(\hat{\beta}_{m,n} - \beta_{m,0}) \end{pmatrix} = (\hat{u}_{m,1,n}, \ldots, \hat{u}_{m,p_m+q_m,n})^\star,$$

$$\partial_{\theta_m}\hat{g}_{m,n}(\theta_{m,0}) = \begin{pmatrix} \frac{1}{\sqrt{nh_n}}\partial_{\alpha_m}g_{m,n}(\theta_{m,0}) \\ \frac{1}{\sqrt{n}}\partial_{\beta_m}g_{m,n}(\theta_{m,0}) \end{pmatrix} = (\partial\hat{g}_{m,1,n}, \ldots, \partial\hat{g}_{m,p_m+q_m,n})^\star.$$

Furthermore, we define that $\partial^2_{m,l_1,l_2} = \partial_{m,l_1}\partial_{m,l_2}$, $\partial^3_{m,l_1,l_2,l_3} = \partial_{m,l_1}\partial_{m,l_2}\partial_{m,l_3}$, $d^2_{m,l_1,l_2,n} = d_{m,l_1,n}d_{m,l_2,n}$ and $d^3_{m,l_1,l_2,l_3,n} = d_{m,l_1,n}d_{m,l_2,n}d_{m,l_3,n}$. Let $C^{k,l}_{\uparrow}(\mathbf{R}^d \times \Theta)$ denote the space of functions f satisfying the following conditions: (i) $f(x,\theta)$ is an $\mathbf{R}$-valued function on $\mathbf{R}^d \times \Theta$ which is continuously differentiable with respect to x up to order k for all θ, and for $|\mathbf{n}| = 0, 1, \ldots, k$, $\partial_x^{\mathbf{n}} f(x,\theta)$ is continuously differentiable w.r.t. θ up to order l. (ii) For $|\mathbf{n}| = 0, \ldots, k$ and for $|\nu| = 0, \ldots, l$, $\sup_\theta |\partial_x^{\mathbf{n}}\partial_\theta^\nu f(x,\theta)| \le C(1+|x|)^C$. Let $\mathcal{F}_{\uparrow}(\mathbf{R}^d \times \Theta)$ be the space of $\mathbf{R}$-valued measurable functions f on $\mathbf{R}^d \times \Theta$ such that $\sup_\theta |f(x,\theta)|$ is at most polynomial growth w.r.t. x.

In order to prove Theorem 1, we use the following lemma about large deviation inequalities.

Lemma 1 *Let* $m \in \{1, 2, \ldots, M\}$. *Suppose that* A1–A4 *hold true. Then,*

(i) *for any* $L > 0$, *there exists a constant* $C_L > 0$ *such that*

$$P_{\theta^*}[|\hat{u}_{m,n}| > r] \le \frac{C_L}{r^L}$$

for all $n \in \mathbf{N}$ *and* $r > 0$. *Moreover,* $\sup_n E[|\hat{u}_{m,n}|^\mu] < \infty$ *for all* $\mu > 0$.

(ii) *for any continuous function* f *of at most polynomial growth, as* $n \to \infty$,

$$E_{\theta^*}[f(\hat{u}_{m,n})] \to E[f(G_m)],$$

where G_m *is the* $(p_m + q_m)$*-dimensional Gaussian random variable with mean* 0 *and covariance matrix* $I_m^{-1}(\theta_{m,0})$.

(iii) $\hat{u}_{m,n} = I_m^{-1}(\theta_{m,0})\partial_{\theta_m}\hat{g}_{m,n}(\theta_{m,0}) + R_{m,n}$, *where for any* $E_0 \in (0, \epsilon_0^2/2)$ *and* $E_1 \ge 1$, *as* $n \to \infty$,

$$P_{\theta^*}\left[|R_{m,n}| > n^{-E_0}\right] = O(1/n^{E_1}).$$

Proof For simplicity, we fix m and consider the following stochastic differential equation

$$\mathrm{d}X_t = b(X_t, \alpha)\mathrm{d}t + \sigma(X_t, \beta)\mathrm{d}w_t, \quad t \in [0, T], \quad X_0 = x_0, \tag{8}$$

where b is an $\mathbf{R}^d$-valued function defined on $\mathbf{R}^d \times \Theta_\alpha$, σ is an $\mathbf{R}^d \otimes \mathbf{R}^d$-valued function defined on $\mathbf{R}^d \times \Theta_\beta$, w is a d-dimensional standard Wiener process, x_0 is a deterministic initial condition, Θ_α and Θ_β are, respectively, bounded domains in $\mathbf{R}^p$ and $\mathbf{R}^q$ with a locally Lipschitz boundary, and $\theta = (\alpha, \beta) \in \Theta_\alpha \times \Theta_\beta = \Theta$. Let $\theta^* = (\alpha^*, \beta^*)$ denote the true value of θ and we assume that $\theta^* \in \Theta$. Here, we note that for the model (8), A1–A4 are satisfied. Moreover, since m is fixed, for the notation with m stated above, the suffix m is omitted, for example, $g_n = g_{m,n}$, $I(\theta^*) = I_m(\theta^*)$, $d_{i,n} = d_{m,i,n}$ and $\partial_i = \partial_{m,i}$.

Let $\epsilon_1 = \epsilon_0/2$. For stationary ergodic diffusion processes, both (i) and (ii) have been proved in Section 6 of Yoshida (2005). Even if we do not assume stationarity, we can show (i) and (ii) by the analogous argument as the proof of Yoshida (2005). For the proofs of (i) and (ii), we need to show that for $f \in C_{\uparrow}^{1,1}(\mathbf{R}^d \times \Theta)$ and for any $\mu > 1$,

$$\sup_n E_{\theta^*}\left[\left(\sup_\theta n^{\epsilon_1}\left|\frac{1}{n}\sum_{k=1}^n f(X_{t_{k-1}^n},\theta) - \int_{\mathbf{R}^d} f(x,\theta)\mu_{\theta^*}(\mathrm{d}x)\right|\right)^\mu\right] < \infty. \tag{9}$$

For the proof of (9), we set that $g(x,\theta) = f(x,\theta) - \int_{\mathbf{R}^d} f(x,\theta)\mu_{\theta^*}(\mathrm{d}x)$ and $\mathcal{G}_n(\theta) = \frac{1}{n}\sum_{k=1}^n g(X_{t_{k-1}^n},\theta)$. By Theorem 1 in Pardoux and Veretennikov (2001), under A1–A2, there exist $G_f(x,\theta)$, $\partial_i G_f(x,\theta) \in \mathcal{F}_\uparrow(\mathbf{R}^d \times \Theta)$ such that $L_\theta G_f(x,\theta) = g(x,\theta)$, where $L_\theta = \sum_{i,j=1}^d \Xi(x,\beta)_{ij}\partial_{x_i}\partial_{x_j} + \sum_{i=1}^d b(x,\alpha)_i\partial_{x_i}$. Since it follows from Itô's formula that $G_f(X_t,\theta) - G_f(X_0,\theta) = \int_0^t g(X_s,\theta)\mathrm{d}s + \int_0^t (\partial_x G_f)^\star(X_s,\theta)\,\sigma(X_s,\beta)\mathrm{d}w_t$, noting that $n^{\epsilon_1} = n^{\epsilon_0/2} \le (nh_n)^{1/2}$, one has that for any $\mu > 1$,

$$\begin{aligned}
E_{\theta^*}\left[n^{\epsilon_1\mu}\,|\mathcal{G}_n(\theta)|^\mu\right] &\le C\frac{n^{\epsilon_1\mu}}{nh_n}\sum_{k=1}^n\int_{t_{k-1}^n}^{t_k^n} E_{\theta^*}\left[\left|f(X_s,\theta) - f\left(X_{t_{k-1}^n},\theta\right)\right|^\mu\right]\mathrm{d}s\\
&\quad + C\frac{n^{\epsilon_1\mu}}{(nh_n)^\mu}E_{\theta^*}\left[|G_f(X_T,\theta)|^\mu + |G_f(X_0,\theta)|^\mu\right]\\
&\quad + C\frac{n^{\epsilon_1\mu}}{(nh_n)^\mu}T^{\mu/2-1}\int_0^T E_{\theta^*}\left[|(\partial_x G_f)^\star(X_s,\theta)\sigma(X_s,\beta)|^\mu\right]\mathrm{d}s\\
&\le C\left(n^{\epsilon_1\mu}h_n^{\mu/2} + \frac{n^{\epsilon_1\mu}}{(nh_n)^\mu} + \frac{n^{\epsilon_1\mu}}{(nh_n)^{\mu/2}}\right) < \infty.
\end{aligned}$$

Thus, $\sup_n\sup_\theta E_{\theta^*}[n^{\epsilon_1\mu}|\mathcal{G}_n(\theta)|^\mu] < \infty$ for any $\mu > 1$. In the same way, $\sup_n\sup_\theta E_{\theta^*}[n^{\epsilon_1\mu}|\partial_\theta\mathcal{G}_n(\theta)|^\mu] < \infty$ for any $\mu > 1$. By the Sobolev inequality,

$$E_{\theta^*}\left[\sup_\theta\left|n^{\epsilon_1}\mathcal{G}_n(\theta)\right|^\mu\right] \le C_\Theta\left\{\sup_\theta E_{\theta^*}\left[\left|n^{\epsilon_1}\mathcal{G}_n(\theta)\right|^\mu\right] + \sup_\theta E_{\theta^*}\left[\left|n^{\epsilon_1}\partial_\theta\mathcal{G}_n(\theta)\right|^\mu\right]\right\}$$

for $\mu > p+q$, and consequently, one has $\sup_n E_{\theta^*}\left[\sup_\theta|n^{\epsilon_1}\mathcal{G}_n(\theta)|^\mu\right] < \infty$ for $\mu > p+q$, which completes the proof of (9). Following the proof of Section 6 in Yoshida (2005) with the estimate (9), we can show (i) and (ii).

(iii) Let $B_n = \{|\hat{\theta}_n - \theta^*| \le n^{-\beta_0}\}$ for $\beta_0 \in (0,\epsilon_0/2)$. By noting that $nh_n \ge n^{\epsilon_0}$, Lemma 1–(i) yields that

$$\begin{aligned}
P_{\theta^*}\left[\left|\hat{\theta}_n - \theta^*\right| > n^{-\beta_0}\right] &\le P_{\theta^*}\left[\left|\sqrt{nh_n}(\hat{\theta}_n - \theta^*)\right| > \sqrt{nh_n}n^{-\beta_0}\right]\\
&\le P_{\theta^*}\left[\left|\hat{u}_n\right| > n^{\epsilon_0/2-\beta_0}\right] = O\left(\frac{1}{n^{(\epsilon_0/2-\beta_0)L}}\right).
\end{aligned} \tag{10}$$

By an easy computation, $\sum_{j=1}^{p+q} I(\theta^*)_{ij}\hat{u}_{j,n}1_{B_n} = \partial\hat{g}_{i,n}1_{B_n} + R_{i,n}^{(1)}1_{B_n} + R_{i,n}^{(2)}1_{B_n}$, where $Q_{i,j,n}^{(1)} = (d_{i,j,n}^2\partial_{i,j}^2 g_n(\theta^*) + I(\theta^*)_{ij})\hat{u}_{j,n}$, $Q_{i,j,l,n}^{(2)} = d_{i,j,l,n}^3\int_0^1\partial_{i,j,l}^3 g_n(\theta^* +$

$t(\hat{\theta}_n-\theta^*))\mathrm{d}t\hat{u}_{j,n}\hat{u}_{l,n}$, $R^{(1)}_{i,n}=\sum_{j=1}^{p+q}Q^{(1)}_{i,j,n}$ and $R^{(2)}_{i,n}=\sum_{j,l=1}^{p+q}Q^{(2)}_{i,j,l,n}$. First, we estimate $R^{(1)}_{i,n}$. Noting that $n^{\epsilon_0}\le nh_n$, one has that for $i=1,\dots,p$ and $E_0\in(0,\epsilon_0^2/2)$,

$$\begin{aligned}
&P_{\theta^*}\left[\left|R^{(1)}_{i,n}1_{B_n}\right|>n^{-E_0}\right]\\
&\le P_{\theta^*}\left[\left|\sum_{j=1}^{p}(nh_n)^{\epsilon_1}\left(\frac{1}{nh_n}\partial^2_{i,j}g_n(\theta^*)+I(\theta^*)_{ij}\right)\hat{u}_{j,n}\right|>\frac{n^{\epsilon_0\epsilon_1-E_0}}{2}\right]\\
&\quad+P_{\theta^*}\left[\left|\sum_{j=p+1}^{p+q}n^{\epsilon_1}\left(\frac{1}{n\sqrt{h_n}}\partial^2_{i,j}g_n(\theta^*)\right)\hat{u}_{j,n}\right|>\frac{n^{\epsilon_1-E_0}}{2}\right].
\end{aligned}$$

Lemma 1–(i) together with a version of (9) yields that for any $\mu>1$ and for $i=1,\dots,p$,

$$\begin{aligned}
\sup_n E_{\theta^*}\left[\left|\sum_{j=1}^{p}(nh_n)^{\epsilon_1}\left(\frac{1}{nh_n}\partial^2_{i,j}g_n(\theta^*)+I(\theta^*)_{ij}\right)\hat{u}_{j,n}\right|^{\mu}\right]<\infty,\\
\sup_n E_{\theta^*}\left[\left|\sum_{j=p+1}^{p+q}n^{\epsilon_1}\left(\frac{1}{n\sqrt{h_n}}\partial^2_{i,j}g_n(\theta^*)\right)\hat{u}_{j,n}\right|^{\mu}\right]<\infty.
\end{aligned}$$

Therefore, setting that $E_1=(\epsilon_0\epsilon_1-E_0)\mu$, one has that for $\mu\ge 1/(\epsilon_0\epsilon_1-E_0)$,

$$P_{\theta^*}\left[\left|R^{(1)}_{i,n}1_{B_n}\right|>n^{-E_0}\right]=O\left(\frac{1}{n^{(\epsilon_0\epsilon_1-E_0)\mu}}\right)=O\left(\frac{1}{n^{E_1}}\right)$$

for $i=1,\dots,p$. In the same argument as above, we obtain that for $i=p+1,\dots,p+q$, $P_{\theta^*}\left[\left|R^{(1)}_{i,n}1_{B_n}\right|>n^{-E_0}\right]=O\left(\frac{1}{n^{E_1}}\right)$.

Next, in order to estimate $R^{(2)}_{i,n}$, note that

$$\begin{aligned}
&P_{\theta^*}\left[\left|R^{(2)}_{i,n}1_{B_n}\right|>n^{-E_0}\right]\\
&\le P_{\theta^*}\left[\sum_{j,l=1}^{p}\left|Q^{(2)}_{i,j,l,n}1_{B_n}\right|>\frac{n^{-E_0}}{4}\right]+P_{\theta^*}\left[\sum_{j=1}^{p}\sum_{l=p+1}^{p+q}\left|Q^{(2)}_{i,j,l,n}1_{B_n}\right|>\frac{n^{-E_0}}{4}\right]\\
&\quad+P_{\theta^*}\left[\sum_{j=p+1}^{p+q}\sum_{l=1}^{p}\left|Q^{(2)}_{i,j,l,n}1_{B_n}\right|>\frac{n^{-E_0}}{4}\right]+P_{\theta^*}\left[\sum_{j,l=p+1}^{p+q}\left|Q^{(2)}_{i,j,l,n}1_{B_n}\right|>\frac{n^{-E_0}}{4}\right].
\end{aligned}$$

Using the standard estimates and the Sobolev inequality, it is easy to show that for $\mu > p+q$,

$$\sup_n E_{\theta^*}\left[\left(\sum_{i,j,l=1}^{p}\frac{1}{nh_n}\sup_\theta\left|\partial^3_{i,j,l}g_n(\theta)\right|\right)^\mu\right] < \infty,$$

$$\sup_n E_{\theta^*}\left[\left(\sum_{i,j=1}^{p}\sum_{l=p+1}^{p+q}\frac{1}{n\sqrt{h_n}}\sup_\theta\left|\partial^3_{i,j,l}g_n(\theta)\right|\right)^\mu\right] < \infty,$$

$$\sup_n E_{\theta^*}\left[\left(\sum_{i=1}^{p+q}\sum_{j,l=p+1}^{p+q}\frac{1}{n}\sup_\theta\left|\partial^3_{i,j,l}g_n(\theta)\right|\right)^\mu\right] < \infty.$$

Thus, by the analogous argument as $R^{(1)}_{i,n}$, we have that $P_{\theta^*}\left[\left|R^{(2)}_{i,n}1_{B_n}\right| > n^{-E_0}\right] = O\left(\frac{1}{n^{E_1}}\right)$. By Lemma 1–(i), we can take L in (10) satisfying that $(\epsilon_0/2-\beta_0)L/2-2E_0 > 1$. By setting $E_1 = (\epsilon_0/2-\beta_0)L/2-2E_0$,

$$P_{\theta^*}\left[\left|\sum_{j=1}^{p+q}I(\theta^*)_{ij}\hat{u}_{j,n}1_{B_n^c}\right| > n^{-E_0}\right] = O\left(\frac{1}{n^{(\epsilon_0/2-\beta_0)L/2-2E_0}}\right) = O\left(\frac{1}{n^{E_1}}\right),$$

$$P_{\theta^*}\left[\left|\partial\hat{g}_{i,n}1_{B_n^c}\right| > n^{-E_0}\right] = O\left(\frac{1}{n^{E_1}}\right).$$

Since $\sum_{j=1}^{p+q}I(\theta^*)_{ij}\hat{u}_{j,n} = \partial\hat{g}_{i,n}+\tilde{R}_{i,n}$ and $\tilde{R}_{i,n} = \sum_{j=1}^{p+q}I(\theta^*)_{ij}\hat{u}_{j,n}1_{B_n^c} - \partial\hat{g}_{i,n}1_{B_n^c} + R^{(1)}_{i,n}1_{B_n} + R^{(2)}_{i,n}1_{B_n}$ for $i = 1,\dots,p+q$, one has

$$\hat{u}_{i,n} = \sum_{j=1}^{p+q}(I^{-1}(\theta^*))_{ij}\left(\partial\hat{g}_{j,n}+\tilde{R}_{j,n}\right),$$

which completes the proof. □

In order to estimate moments for derivatives of log likelihood function w.r.t. θ, we will use the Malliavin calculus. For this purpose, we briefly review the Malliavin calculus and present several basic properties used in the proof of Lemma 2. For details of the Malliavin calculus, see Malliavin (1997) and Nualart (2006). For the use of the Malliavin calculus techniques in statistics, we can refer to Yoshida (1992a, 1997, 2004), Gobet (2001, 2002) and Sakamoto and Yoshida (2004).

Fix a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), P)$ and let $(W_t)_{t\geq 0}$ be a d-dimensional Wiener process. For $h(\cdot) \in H = L^2([0,T];\mathbf{R}^d)$, denote by $W(h)$ the Wiener stochastic integral $\int_0^T h(t)\cdot dW_t$. Let $\mathcal{S}$ be the space of random variables of the form $F = f(W(h_1),\dots,W(h_N))$, where f is a C^∞-function with derivatives of at most polynomial growth, $(h_1,\dots,h_N) \in H^N$ and $n \geq 1$. For $F \in \mathcal{S}$, we define its derivative $\mathcal{D}F = (\mathcal{D}_t F)_{t\in[0,T]}$ as the H-valued random variable given by $\mathcal{D}_t F = \sum_{i=1}^N \partial_{x_i} f(W(h_1), \ldots, W(h_N))h_i(t)$. The operator $\mathcal{D}$ is closable as an operator from $L^p(\Omega)$ to $L^p(\Omega; H)$, for any $p \geq 1$. We denote its domain by $\mathbf{D}^{1,p}$, which means the closure of $\mathcal{S}$ with respect to the norm $\| F \|_{1,p} = [E(|F|^p) + E(\| \mathcal{D}F \|_H^p)]^{1/p}$. We can define the iteration of the operator $\mathcal{D}$ in such a way that for $F \in \mathcal{S}$, the iterated derivative $\mathcal{D}^k F$ is a random variable with values in $H^{\otimes k}$. Then, the seminorm on $\mathcal{S}$ is defined by $\| F \|_{k,p} = [E|F|^p + \sum_{j=1}^k E(\| \mathcal{D}^j F \|_{H^{\otimes j}}^p)]^{1/p}$ for all $p \geq 1$ and any natural number $k \geq 1$. As in the case that $k = 1$, the operator $\mathcal{D}^k$ is closable from $S \subset L^p(\Omega)$ into $L^p(\Omega; H^{\otimes k})$ for $p \geq 1$. We define $\mathbf{D}^{k,p}$ as the completion of $\mathcal{S}$ with respect to the norm $\| \cdot \|_{k,p}$.

δ is the Skorohod integral defined by the adjoint operator of $\mathcal{D}$ as follows.

Definition 1 δ is a linear operator on $L^2([0,T] \times \Omega; \mathbf{R}^d)$ with values in $L^2(\Omega)$ such that

(i) The domain of δ, denoted by Dom(δ), is the set of processes $u \in L^2([0,T] \times \Omega; \mathbf{R}^d)$ such that for all $F \in \mathbf{D}^{1,2}$, $E\big[\int_0^T \mathcal{D}_t F \cdot u_t \mathrm{d}t\big] \leq c \| F \|_{L^2}$, where c is some constant depending on u.

(ii) If u belongs to Dom(δ), then $\delta(u)$ is the element of $L^2(\Omega)$ characterized by $E[F\delta(u)] = E\big[\int_0^T \mathcal{D}_t F \cdot u_t \mathrm{d}t\big]$ for all $F \in \mathbf{D}^{1,2}$.

Basic properties of δ used in this paper are as follows. For the proofs, see Nualart (2006).

Proposition 1 (i) *The space of weakly differentiable H-valued variables $\mathbf{D}^{1,2}(H)$ is included in Dom(δ).*

(ii) *For all $k \geq 1$ and $p > 1$, the operator δ is continuous from $\mathbf{D}^{k,p}(H)$ into $\mathbf{D}^{k-1,p}$. In particular, in the case that $k = 1$, for $p > 1$,*

$$\| \delta(u) \|_p \leq c_p(\| u \|_{L^p(\Omega;H)} + \| \mathcal{D}u \|_{L^p(\Omega;H\otimes H)}).$$

(iii) *Let $F \in \mathbf{D}^{1,2}$. Then, for any $u \in$ Dom(δ) such that $E\left[F^2 \int_0^T |u_t|^2 \mathrm{d}t\right] < \infty$, one has*

$$\delta(Fu) = F\delta(u) - \int_0^T \mathcal{D}_t F \cdot u_t \mathrm{d}t, \tag{11}$$

provided the right-hand side of (11) is square integrable.

(iv) *Let u be an adapted process in $L^2([0,T] \times \Omega; \mathbf{R}^d)$. Then, the Skorohod integral coincides with the Itô integral: $\delta(u) = \int_0^T u_t \mathrm{d}W_t$.*

(v) *For $u \in \mathbf{D}^{2,2}(H)$, the following commutation relation holds:*

$$\mathcal{D}_t(\delta(u)) = u_t + \delta(\mathcal{D}_t u).$$

Next, we state the results on estimates of moments for derivatives of log likelihood function w.r.t. θ.

Lemma 2 *Let* $m \in \{1, 2, \dots, M\}$. *Suppose that* A1–A2 *hold true. Then,*

(i) *for* $i = 1 \dots, p_m + q_m$,

$$E_{\theta^*}\left[\partial_{m,i} l_{m,n}(\theta_{m,0})\right] = 0.$$

(ii) *for* $i, j = 1 \dots, p_m + q_m$, *as* $h_n \to 0$, $nh_n \to \infty$ *and* $nh_n^2 \to 0$,

$$E_{\theta^*}\left[d^2_{m,i,j,n}\partial^2_{m,i,j} l_{m,n}(\theta_{m,0})\right] \to -(I_m(\theta_{m,0}))_{ij}.$$

(iii) *for* $i, j, l = 1 \dots, p_m + q_m$ *and for any* $\mu > 1$, *as* $h_n \to 0$, $nh_n \to \infty$ *and* $nh_n^2 \to 0$,

$$E_{\theta^*}\left[\sup_{\theta_m}\left|d^3_{m,i,j,l,n}\partial^3_{m,i,j,l} l_{m,n}(\theta_m)\right|^{\mu}\right] = o(1).$$

Proof As in the proof of Lemma 1, we fix m and consider the following stochastic differential equation

$$\mathrm{d}X_t^\theta = b\left(X_t^\theta, \alpha\right)\mathrm{d}t + \sigma\left(X_t^\theta, \beta\right)\mathrm{d}W_t, \quad t \in [0, T], \quad X_0^\theta = x_0, \tag{12}$$

where W is a d-dimensional standard Wiener process independent of w and the others are the same as (8). In what follows, for the process with the true value θ^*, the suffix θ^* of the process is omitted, for example, $X_t = X_t^{\theta^*}$. We denote by $\mathcal{D}_t^j F$, $t \in [0, T]$, $j = 1, \dots, d$, the derivative of a random variable F as an element of $L^2([0, T] \times \Omega; \mathbf{R}^d) \simeq L^2(\Omega; H)$. Similarly, we denote by $\mathcal{D}_{t_1,\dots,t_N}^{j_1,\dots,j_N} F$ the Nth derivative of F. Under $A1$, X_t^θ is differentiable with respect to x_0, α and β, see Kunita (1984) or Jacod (2006). Let Y_t^θ be the Jacobian matrix $\frac{\partial X_t^\theta}{\partial x_0}$. Under $A1$, for any $t \geq 0$, X_t^θ, Y_t^θ and $(Y_t^\theta)^{-1}$ belong to $\bigcap_{p \geq 1} \mathbf{D}^{4,p}$, see Section 2.2 of Nualart (2006). Furthermore, for any $\mu > 1$ and for $N = 1, 2, 3, 4$,

$$E_\theta\left[\sup_{0 \leq t \leq 1} \| Z_t^\theta \|^{\mu} \middle| X_0 = x\right] + \sup_{r_1,\dots,r_N \in [0,1]} E_\theta\left[\sup_{r_1 \vee \dots \vee r_N \leq t \leq 1} \| \mathcal{D}_{r_1,\dots,r_N} Z_t^\theta \|^{\mu} \middle| X_0 = x\right] = R(\theta, 1, x) \tag{13}$$

for $Z_t^\theta = X_t^\theta$, Y_t^θ or $(Y_t^\theta)^{-1}$, where $R(\theta, a, x)$ is a real-valued function on $\Theta \times (0, 1] \times \mathbf{R}^d$ for which there exists a constant C such that $|R(\theta, a, x)| \leq aC(1 + |x|)^C$ for all θ, a, x. Let $i, j, l, m \in \{1, \dots, p\}$ and $i', j', l', m' \in \{1, \dots, q\}$. By using the standard estimates in Section 3.1 of Jacod (2006), we obtain that for any $\mu > 1$,

$$E_\theta\left[\sup_{0 \leq t \leq h_n} \| Z_t^{(\alpha)} \|^{\mu} \middle| X_0 = x\right] = R\left(\theta, h_n^{\mu}, x\right) \tag{14}$$

for $Z_t^{(\alpha)} = \partial_{\alpha_i} X_t^\theta,\ \partial_{\alpha_i}\partial_{\alpha_j} X_t^\theta,\ \partial_{\alpha_i}\partial_{\alpha_j}\partial_{\alpha_l} X_t^\theta$ or $\partial_{\alpha_i}\partial_{\alpha_j}\partial_{\alpha_l}\partial_{\alpha_m} X_t^\theta$. Similarly, one has that for any $\mu > 1$,

$$E_\theta\left[\sup_{0\le t\le h_n}\left\|Z_t^{(\beta)}\right\|^\mu \middle| X_0 = x\right] = R\left(\theta, h_n^{\mu/2}, x\right) \tag{15}$$

for $Z_t^{(\beta)} = \partial_{\beta_{i'}} X_t^\theta,\ \partial_{\beta_{i'}}\partial_{\beta_{j'}} X_t^\theta,\ \partial_{\beta_{i'}}\partial_{\beta_{j'}}\partial_{\beta_{l'}} X_t^\theta$ or $\partial_{\beta_{i'}}\partial_{\beta_{j'}}\partial_{\beta_{l'}}\partial_{\beta_{m'}} X_t^\theta$, and that for any $\mu > 1$,

$$E_\theta\left[\sup_{0\le t\le h_n}\left\|Z_t^{(\alpha\beta)}\right\|^\mu \middle| X_0 = x\right] = R\left(\theta, h_n^{3\mu/2}, x\right) \tag{16}$$

for $Z_t^{(\alpha\beta)} = \partial_{\alpha_i}\partial_{\beta_{i'}} X_t^\theta,\ \partial_{\alpha_i}\partial_{\alpha_j}\partial_{\beta_{i'}} X_t^\theta,\ \partial_{\alpha_i}\partial_{\beta_{i'}}\partial_{\beta_{j'}} X_t^\theta,\ \partial_{\alpha_i}\partial_{\alpha_j}\partial_{\alpha_l}\partial_{\beta_{i'}} X_t^\theta, \partial_{\alpha_i}\partial_{\alpha_j}\partial_{\beta_{i'}}\partial_{\beta_{j'}} X_t^\theta$ or $\partial_{\alpha_i}\partial_{\beta_{i'}}\partial_{\beta_{j'}}\partial_{\beta_{l'}} X_t^\theta$.

(i) By Proposition 2.2 in Gobet (2002),

$$\frac{\partial_{\theta_i} p}{p}(h_n, x, y; \theta) = E_\theta\left[\frac{1}{h_n}\sum_{l_1=1}^{d} \delta\left(\partial_{\theta_i} X_{h_n,l_1}^\theta \cdot U_{l_1}^\theta\right) \middle| X_0^\theta = x, X_{h_n}^\theta = y\right], \tag{17}$$

where $U_{l_1}^\theta = (U_{t,l_1}^\theta)_{t\in[0,h_n]}$ is the $\mathbf{R}^d$-valued process whose l_2-th component is $U_{t,l_1,l_2}^\theta = (\sigma^{-1}(X_t^\theta, \beta) Y_t^\theta (Y_{h_n}^\theta)^{-1})_{l_2 l_1}$. Let $\mathcal{G}_k$ denote the *history* up to the time t_k^n. Since

$$\begin{aligned} & E_{\theta^*}\left[\frac{\partial_{\theta_i} p}{p}\left(h_n, X_{t_{k-1}^n}, X_{t_k^n}; \theta\right) \middle| \mathcal{G}_{k-1}\right] \\ & = E_\theta\left[\frac{p\left(h_n, X_0^\theta, X_{h_n}^\theta; \theta^*\right)}{p\left(h_n, X_0^\theta, X_{h_n}^\theta; \theta\right)} \frac{1}{h_n}\sum_{l_1=1}^{d} \delta\left(\partial_{\theta_i} X_{h_n,l_1}^\theta \cdot U_{l_1}^\theta\right) \middle| X_0^\theta = X_{t_{k-1}^n}\right] \end{aligned} \tag{18}$$

and $E_{\theta^*}\left[\delta(\partial_{\theta_i} X_{h_n,l_1} \cdot U_{l_1}) \middle| X_0 = x\right] = 0$ by the definition of δ, one has that E_{θ^*} $[\frac{\partial_{\theta_i} p}{p}(h_n, X_{t_{k-1}^n}, X_{t_k^n}; \theta^*)] = 0$, which completes the proof of (i).
(ii) In the same way as the proof of Proposition 4.1 in Gobet (2001),

$$\frac{\partial_{\theta_i}\partial_{\theta_j} p}{p}(h_n, x, y; \theta) = E_\theta\left[\frac{1}{h_n}\Psi_{1,i,j}^\theta + \frac{1}{h_n^2}\Psi_{2,i,j}^\theta \middle| X_0^\theta = x, X_{h_n}^\theta = y\right], \tag{19}$$

where $\Psi_{1,i,j}^\theta = \sum_{l_1=1}^d \delta(\partial_{\theta_i}\partial_{\theta_j} X_{h_n,l_1}^\theta \cdot U_{l_1}^\theta)$ and $\Psi_{2,i,j}^\theta = \sum_{l_1,l_2=1}^d \delta(\delta(\partial_{\theta_i} X_{h_n,l_1}^\theta \cdot \partial_{\theta_j} X_{h_n,l_2}^\theta \cdot U_{l_1}^\theta) U_{l_2}^\theta)$. As in (i), one has $E_{\theta^*}\big[\frac{\partial_{\theta_i}\partial_{\theta_j} p}{p}(h_n, X_{t_{k-1}^n}, X_{t_k^n}; \theta^*)\big] = 0$. Therefore, it is enough to show that for any $\mu > 1$,

$$d_{i,j,n}^2 \sum_{k=1}^{n} \frac{\partial_{\theta_i} p}{p} \frac{\partial_{\theta_j} p}{p} \left(h_n, X_{t_{k-1}^n}, X_{t_k^n}; \theta^*\right) \to^p (I(\theta^*))_{ij}, \tag{20}$$

$$\limsup_{n\to\infty} E_{\theta^*}\left[\left|d_{i,j,n}^2 \sum_{k=1}^{n} \frac{\partial_{\theta_i} p}{p} \frac{\partial_{\theta_j} p}{p} \left(h_n, X_{t_{k-1}^n}, X_{t_k^n}; \theta^*\right)\right|^{\mu}\right] < \infty. \tag{21}$$

For the proof of (20), we note that from (3.23) and (3.29) in Gobet (2002),

$$\begin{aligned}
&\delta\left(\partial_{\alpha_i} X_{h_n,l_1}^{\theta} \cdot U_{l_1}^{\theta}\right) \\
&\quad = h_n[\partial_{\alpha_i} b(x,\alpha)]_{l_1}\left[\sigma^{-2}(x,\beta)\left(X_{h_n}^{\theta} - m^{\theta}(x)\right)\right]_{l_1} + H_{1,i,l_1}^{\theta}, &(22)\\
&\delta\left(\partial_{\beta_j} X_{h_n,l_1}^{\theta} \cdot U_{l_1}^{\theta}\right) \\
&\quad = \left[(\partial_{\beta_j}\sigma)\sigma^{-1}(x,\beta)\left(X_{h_n}^{\theta} - m^{\theta}(x)\right)\right]_{l_1}\left[\sigma^{-2}(x,\beta)(X_{h_n}^{\theta} - m^{\theta}(x))\right]_{l_1} \\
&\qquad - \left[(\partial_{\beta_j}\sigma)\sigma^{-1}(x,\beta)V^{\theta}(x)\sigma^{-2}(x,\beta)\right]_{l_1 l_1} + H_{2,j,l_1}^{\theta}, &(23)
\end{aligned}$$

where $m^{\theta}(x) = E_{\theta}[X_{h_n}^{\theta}|X_0^{\theta} = x]$, $V^{\theta}(x) = E_{\theta}[(X_{h_n}^{\theta} - m^{\theta}(x))(X_{h_n}^{\theta} - m^{\theta}(x))^{\star}|X_0^{\theta} = x]$, and $E_{\theta}[|H_{1,i,l_1}^{\theta}|^{\mu}|X_0 = x]^{1/\mu} = R(\theta, h_n^2, x)$ and $E_{\theta}[|H_{2,j,l_1}^{\theta}|^{\mu}|X_0 = x]^{1/\mu} = R(\theta, h_n^{3/2}, x)$ for any $\mu > 1$. By (17) and (22),

$$\begin{aligned}
&\frac{1}{nh_n}\frac{\partial_{\alpha_i} p}{p}\frac{\partial_{\alpha_j} p}{p}(h_n, x, y; \theta) \\
&\quad = \frac{1}{nh_n^3}\sum_{l_1,l_2=1}^{d}\left\{h_n^2 \mathcal{A}_{i,j,l_1,l_2}^{(1)}(x,y;\theta) + h_n \mathcal{A}_{i,j,l_1,l_2}^{(2)}(x,y;\theta)\right. \\
&\qquad \left. + h_n \mathcal{A}_{j,i,l_2,l_1}^{(2)}(x,y;\theta) + \mathcal{A}_{i,j,l_1,l_2}^{(3)}(x,y;\theta)\right\},
\end{aligned}$$

where $\mathcal{A}_{i,j,l_1,l_2}^{(1)}(x,y;\theta) = \mathcal{A}_{i,l_1}(x,y;\theta)\mathcal{A}_{j,l_2}(x,y;\theta)$,

$$\begin{aligned}
\mathcal{A}_{i,j,l_1,l_2}^{(2)}(x,y;\theta) &= \mathcal{A}_{i,l_1}(x,y;\theta)E_{\theta}\left[H_{1,j,l_2}^{\theta}\,\middle|\, X_0^{\theta} = x, X_{h_n}^{\theta} = y\right], \\
\mathcal{A}_{i,j,l_1,l_2}^{(3)}(x,y;\theta) &= E_{\theta}\left[H_{1,i,l_1}^{\theta}\,\middle|\, X_0^{\theta} = x, X_{h_n}^{\theta} = y\right]E_{\theta}\left[H_{1,j,l_2}^{\theta}\,\middle|\, X_0^{\theta} = x, X_{h_n}^{\theta} = y\right], \\
\mathcal{A}_{i,l_1}(x,y;\theta) &= [\partial_{\alpha_i} b(x,\alpha)]_{l_1}\left[\sigma^{-2}(x,\beta)(y - m^{\theta}(x))\right]_{l_1}.
\end{aligned}$$

For any $\mu > 1$,

$$E_{\theta^*}\left[\left|\frac{1}{nh_n^2}\sum_{k=1}^{n}\mathcal{A}_{i,j,l_1,l_2}^{(2)}\left(X_{t_{k-1}^n}, X_{t_k^n}, \theta^*\right)\right|^{\mu}\right] \le C\frac{1}{nh_n^{2\mu}}\sum_{k=1}^{n} h_n^{\mu/2}h_n^{2\mu} \to 0, \tag{24}$$

$$E_{\theta^*}\left[\left|\frac{1}{nh_n^3}\sum_{k=1}^n \mathcal{A}^{(3)}_{i,j,l_1,l_2}\left(X_{t_{k-1}^n}, X_{t_k^n}, \theta^*\right)\right|^\mu\right] \le C\frac{1}{nh_n^{3\mu}}\sum_{k=1}^n h_n^{4\mu} \to 0. \tag{25}$$

Hence, one has that $\frac{1}{nh_n^2}\sum_{k=1}^n \sum_{l_1,l_2=1}^d \mathcal{A}^{(2)}_{i,j,l_1,l_2}\big(X_{t_{k-1}^n}, X_{t_k^n}, \theta^*\big) = o_p(1)$ and $\frac{1}{nh_n^3}\sum_{k=1}^n \sum_{l_1,l_2=1}^d \mathcal{A}^{(3)}_{i,j,l_1,l_2}(X_{t_{k-1}^n}, X_{t_k^n}, \theta^*) = o_p(1)$. Moreover,

$$\frac{1}{nh_n}\sum_{k=1}^n E_{\theta^*}\left[\sum_{l_1,l_2=1}^d \mathcal{A}^{(1)}_{i,j,l_1,l_2}\left(X_{t_{k-1}^n}, X_{t_k^n}; \theta^*\right)\middle|\mathcal{G}_{k-1}\right] \to^p (I_b(\theta^*))_{ij},$$

$$\frac{1}{n^2h_n^2}\sum_{k=1}^n E_{\theta^*}\left[\left(\sum_{l_1,l_2=1}^d \mathcal{A}^{(1)}_{i,j,l_1,l_2}(X_{t_{k-1}^n}, X_{t_k^n}; \theta^*)\right)^2\middle|\mathcal{G}_{k-1}\right] \to^p 0.$$

Lemma 9 of Genon-Catalot and Jacod (1993) yields that

$$\frac{1}{nh_n}\sum_{k=1}^n \frac{\partial_{\alpha_i} p}{p}\frac{\partial_{\alpha_j} p}{p}\left(h_n, X_{t_{k-1}^n}, X_{t_k^n}; \theta^*\right) \to^p (I_b(\theta^*))_{ij}.$$

In the same way, we obtain that

$$\frac{1}{n}\sum_{k=1}^n \frac{\partial_{\beta_i} p}{p}\frac{\partial_{\beta_j} p}{p}\left(h_n, X_{t_{k-1}^n}, X_{t_k^n}; \theta^*\right) \to^p (I_\sigma(\theta^*))_{ij},$$

$$\frac{1}{n\sqrt{h_n}}\sum_{k=1}^n \frac{\partial_{\alpha_i} p}{p}\frac{\partial_{\beta_j} p}{p}\left(h_n, X_{t_{k-1}^n}, X_{t_k^n}; \theta^*\right) \to^p 0,$$

and the proof of (20) is complete.

Next, we will show (21). By Proposition 1.2 in Gobet (2002), for any $\mu > 1$, there exist constants $c > 1$, $K > 1$ and $r > 1$ such that

$$\sup_\theta E_{\theta^*}\left[\left|\frac{\partial_{\alpha_i} p}{p}(h_n, x, X_{h_n}; \theta)\right|^\mu \middle| X_0 = x\right] \le Kh_n^{\mu/2}\exp\left(ch_n|x|^2\right)(1+|x|)^r, \tag{26}$$

$$\sup_\theta E_{\theta^*}\left[\left|\frac{\partial_{\beta_{i'}} p}{p}(h_n, x, X_{h_n}; \theta)\right|^\mu \middle| X_0 = x\right] \le K\exp\left(ch_n|x|^2\right)(1+|x|)^r, \tag{27}$$

for $x \in \mathbf{R}^d$, $1 \le i \le p$ and $1 \le i' \le q$. It follows from (26) that for any $\mu > 1$,

$$E_{\theta^*}\left[\left|\frac{1}{nh_n}\sum_{k=1}^n \frac{\partial_{\alpha_i} p}{p}\frac{\partial_{\alpha_j} p}{p}\left(h_n, X_{t_{k-1}^n}, X_{t_k^n}; \theta^*\right)\right|^\mu\right]$$
$$\le C'\sup_t E_{\theta^*}\left[\exp\left(C'h_n|X_t|^2\right)\right].$$

It follows from Proposition 1.1 in Gobet (2002) that for sufficiently large n satisfying that $C'h_n < c_0/c_1^2$, $\sup_t E_{\theta^*}[\exp(C'h_n|X_t|^2)] < \infty$, where c_0 and c_1 are defined in $A2$. Therefore,

$$\limsup_{n\to\infty} E_{\theta^*}\left[\left|\frac{1}{nh_n}\sum_{k=1}^{n}\frac{\partial_{\alpha_i}p}{p}\frac{\partial_{\alpha_j}p}{p}\left(h_n, X_{t_{k-1}^n}, X_{t_k^n};\theta^*\right)\right|^{\mu}\right] < \infty$$

for any $\mu > 1$. In the same way, it follows from (26) to (27) that for any $\mu > 1$,

$$\limsup_{n\to\infty} E_{\theta^*}\left[\left|\frac{1}{n\sqrt{h_n}}\sum_{k=1}^{n}\frac{\partial_{\alpha_i}p}{p}\frac{\partial_{\beta_j}p}{p}\left(h_n, X_{t_{k-1}^n}, X_{t_k^n};\theta^*\right)\right|^{\mu}\right] < \infty,$$
$$\limsup_{n\to\infty} E_{\theta^*}\left[\left|\frac{1}{n}\sum_{k=1}^{n}\frac{\partial_{\beta_i}p}{p}\frac{\partial_{\beta_j}p}{p}\left(h_n, X_{t_{k-1}^n}, X_{t_k^n};\theta^*\right)\right|^{\mu}\right] < \infty.$$

Thus, one has (21), which completes the proof of (ii).
(iii) Let $\Phi_{1,i,j,l}^{\theta} = \sum_{l_1=1}^{d}\delta(\partial_{\theta_i}\partial_{\theta_j}\partial_{\theta_l}X_{h_n,l_1}^{\theta}\cdot U_{l_1}^{\theta})$, $\tilde{\Phi}_{2,i,j,l}^{\theta} = \Phi_{2,i,j,l}^{\theta} + \Phi_{2,i,l,j}^{\theta} + \Phi_{2,j,l,i}^{\theta}$, $\Phi_{2,i,j,l}^{\theta} = \sum_{l_1,l_2=1}^{d}\delta(\delta(\partial_{\theta_i}\partial_{\theta_j}X_{h_n,l_1}^{\theta}\cdot\partial_{\theta_l}X_{h_n,l_2}^{\theta}\cdot U_{l_1}^{\theta})U_{l_2}^{\theta})$ and $\Phi_{3,i,j,l}^{\theta} = \sum_{l_1,l_2,l_3=1}^{d}\delta(\delta(\delta(\partial_{\theta_i}X_{h_n,l_1}^{\theta}\cdot\partial_{\theta_j}X_{h_n,l_2}^{\theta}\cdot\partial_{\theta_l}X_{h_n,l_3}^{\theta}\cdot U_{l_1}^{\theta})U_{l_2}^{\theta})U_{l_3}^{\theta})$. As in (19),

$$\begin{aligned}&\frac{\partial_{\theta_i}\partial_{\theta_j}\partial_{\theta_l}p}{p}(h_n, x, y;\theta)\\&\quad = E_{\theta}\left[\frac{1}{h_n}\Phi_{1,i,j,l}^{\theta} + \frac{1}{h_n^2}\tilde{\Phi}_{2,i,j,l}^{\theta} + \frac{1}{h_n^3}\Phi_{3,i,j,l}^{\theta}\,\middle|\, X_0^{\theta} = x, X_{h_n}^{\theta} = y\right].\end{aligned} \tag{28}$$

First, we estimate $E_{\theta}[|\Psi_{1,i,j}^{\theta}|^{\mu}|X_0^{\theta} = x]$ for $i, j = 1, \ldots, p$ and for any $\mu > 1$, where $\Psi_{1,i,j}^{\theta}$ is defined in (19). It follows from Proposition 1 that

$$\delta\left(\partial_{\theta_i}\partial_{\theta_j}X_{h_n,l_1}^{\theta}\cdot U_{l_1}^{\theta}\right) = \partial_{\theta_i}\partial_{\theta_j}X_{h_n,l_1}^{\theta}\,\delta\left(U_{l_1}^{\theta}\right) - \int_0^{h_n}\mathcal{D}_{r_1}\partial_{\theta_i}\partial_{\theta_j}X_{h_n,l_1}^{\theta}\cdot U_{l_1,r_1}^{\theta}\mathrm{d}r_1,$$

and using (13) and the estimate that $E_{\theta}\left[\left|\delta(U_{l_1}^{\theta})\right|^{\mu}|X_0^{\theta} = x\right] = R(\theta, h_n^{\mu/2}, x)$, one has that for any $\mu, \mu_1 > 1$,

$$\begin{aligned}&E_{\theta}\left[\left|\delta\left(\partial_{\theta_i}\partial_{\theta_j}X_{h_n,l_1}^{\theta}\cdot U_{l_1}^{\theta}\right)\right|^{\mu}|X_0^{\theta} = x\right]\\&\quad\le C\left\{E_{\theta}\left[\left|\partial_{\theta_i}\partial_{\theta_j}X_{h_n,l_1}^{\theta}\right|^{\mu\mu_1}|X_0^{\theta} = x\right]^{1/\mu_1}R\left(\theta, h_n^{\mu/2}, x\right)\right.\end{aligned}$$

$$
\begin{aligned}
&+ h_n^{\mu-1} \int_0^{h_n} E_\theta \big[\big| \mathcal{D}_{r_1} \partial_{\theta_i} \partial_{\theta_j} X_{h_n, l_1}^\theta \big|^{\mu\mu_1} | X_0^\theta = x \big]^{1/\mu_1} \mathrm{d}r_1 R(\theta, 1, x) \Big\} \\
&= R\left(\theta, h_n^{3\mu/2}, x\right),
\end{aligned}
$$

where in the last estimate, we used (14) and the following estimate that $\sup_{0 \leq r_1 \leq h_n} E_\theta [\| \mathcal{D}_{r_1} \partial_{\theta_i} \partial_{\theta_j} X_{h_n}^\theta \|^\mu \, | X_0^\theta = x] = R(\theta, h_n^\mu, x)$. Thus, for any $\mu > 1$ and for $i, j = 1, \ldots, p$, $E_\theta[|\Psi_{1,i,j}^\theta|^\mu | X_0^\theta = x] = R(\theta, h_n^{3\mu/2}, x)$.

Next, we estimate $E_\theta[|\Psi_{2,i,j}^\theta|^\mu | X_0^\theta = x]$ for $i, j = 1, \ldots, p$ and for any $\mu > 1$. By Proposition 1–(iii) and (v),

$$
\begin{aligned}
&E_\theta \left[\left| \delta\left(\delta \left(\partial_{\theta_i} X_{h_n, l_1}^\theta \cdot \partial_{\theta_j} X_{h_n, l_2}^\theta \cdot U_{l_1}^\theta \right) U_{l_2}^\theta \right) \right|^\mu \middle| X_0^\theta = x \right] \\
&\leq C \left\{ E_\theta \left[\left| \delta \left(\partial_{\theta_i} X_{h_n, l_1}^\theta \cdot \partial_{\theta_j} X_{h_n, l_2}^\theta \cdot U_{l_1}^\theta \right) \delta(U_{l_2}^\theta) \right|^\mu \middle| X_0^\theta = x \right] \right. \\
&\quad + E_\theta \left[\left| \int_0^{h_n} \partial_{\theta_i} X_{h_n, l_1}^\theta \cdot \partial_{\theta_j} X_{h_n, l_2}^\theta \cdot U_{l_1, r_1}^\theta \cdot U_{l_2, r_1}^\theta \mathrm{d}r_1 \right|^\mu \middle| X_0^\theta = x \right] \\
&\quad \left. + E_\theta \left[\left| \int_0^{h_n} \delta\left(\mathcal{D}_{r_1} \left(\partial_{\theta_i} X_{h_n, l_1}^\theta \cdot \partial_{\theta_j} X_{h_n, l_2}^\theta \cdot U_{l_1}^\theta \right) \right) \cdot U_{l_2, r_1}^\theta \mathrm{d}r_1 \right|^\mu \middle| X_0^\theta = x \right] \right\}.
\end{aligned}
$$

It follows from Proposition 1–(ii) that

$$
E_\theta \left[\left| \delta\left(\mathcal{D}_{r_1} \left(\partial_{\theta_i} X_{h_n, l_1}^\theta \cdot \partial_{\theta_j} X_{h_n, l_2}^\theta \cdot U_{l_1}^\theta \right) \right) \right|^\mu | X_0^\theta = x \right] = R\left(\theta, h_n^{5\mu/2}, x\right),
$$

where we used (13), (14) and the estimates that $\sup_{r_1 \in [0, h_n]} E_\theta [\| \mathcal{D}_{r_1} \partial_{\theta_i} X_{h_n}^\theta \|^\mu \, | X_0 = x] = R(\theta, h_n^\mu, x)$ and that $\sup_{r_1, r_2 \in [0, h_n]} E_\theta \big[\| \mathcal{D}_{r_1, r_2} \partial_{\theta_i} X_{h_n}^\theta \|^\mu \big| X_0 = x \big] = R(\theta, h_n^\mu, x)$. Furthermore, by using the previous arguments,

$$
E_\theta \left[\left| \delta(\delta(\partial_{\theta_i} X_{h_n, l_1}^\theta \cdot \partial_{\theta_j} X_{h_n, l_2}^\theta \cdot U_{l_1}^\theta) U_{l_2}^\theta) \right|^\mu | X_0^\theta = x \right] = R(\theta, h_n^{3\mu}, x)
$$

and $E_\theta[|\Psi_{2,i,j}^\theta|^\mu | X_0^\theta = x] = R(\theta, h_n^{3\mu}, x)$ for any $\mu > 1$ and for $i, j = 1, \ldots, p$.

In the similar way as above, we obtain that for $i = 1, \ldots, p$ and $j, l, = 1, \ldots, p+q$, and for any $\mu > 1$,

$$
\begin{aligned}
E_\theta \left[\left| \frac{1}{h_n} \Psi_{1,i,j}^\theta \right|^\mu + \left| \frac{1}{h_n^2} \Psi_{2,i,j}^\theta \right|^\mu \middle| X_0^\theta = x \right] &= R\left(\theta, h_n^{\mu/2}, x\right), \\
E_\theta \left[\left| \frac{1}{h_n} \Phi_{1,i,j,l}^\theta \right|^\mu + \left| \frac{1}{h_n^2} \tilde{\Phi}_{2,i,j,l}^\theta \right|^\mu + \left| \frac{1}{h_n^3} \Phi_{3,i,j,l}^\theta \right|^\mu \middle| X_0^\theta = x \right] &= R\left(\theta, h_n^{\mu/2}, x\right).
\end{aligned}
$$

Moreover, for $i, j, l, = p+1, \dots, p+q$, and for any $\mu > 1$,

$$E_\theta \left[\left| \frac{1}{h_n} \Psi^\theta_{1,i,j} \right|^\mu + \left| \frac{1}{h_n^2} \Psi^\theta_{2,i,j} \right|^\mu \middle| X_0^\theta = x \right] = R(\theta, 1, x),$$

$$E_\theta \left[\left| \frac{1}{h_n} \Phi^\theta_{1,i,j,l} \right|^\mu + \left| \frac{1}{h_n^2} \tilde{\Phi}^\theta_{2,i,j,l} \right|^\mu + \left| \frac{1}{h_n^3} \Phi^\theta_{3,i,j,l} \right|^\mu \middle| X_0^\theta = x \right] = R(\theta, 1, x).$$

By using the same argument as the proof of Proposition 1.2 in Gobet (2002), for any $\mu > 1$, there exist constants $c > 1$, $K > 1$ and $r > 1$ such that

$$\sup_\theta E_{\theta^*} \left[\left| \frac{\partial_{\alpha_i} \partial_{\theta_j} p}{p} (h_n, x, X_{h_n}; \theta) \right|^\mu \middle| X_0 = x \right] \le K h_n^{\frac{\mu}{2}} \exp\left(c h_n |x|^2 \right) (1 + |x|)^r, \quad (29)$$

$$\sup_\theta E_{\theta^*} \left[\left| \frac{\partial_{\beta_{i'}} \partial_{\beta_{j'}} p}{p} (h_n, x, X_{h_n}; \theta) \right|^\mu \middle| X_0 = x \right] \le K \exp\left(c h_n |x|^2 \right) (1 + |x|)^r, \quad (30)$$

$$\sup_\theta E_{\theta^*} \left[\left| \frac{\partial_{\alpha_i} \partial_{\theta_j} \partial_{\theta_l} p}{p} (h_n, x, X_{h_n}; \theta) \right|^\mu \middle| X_0 = x \right] \le K h_n^{\frac{\mu}{2}} \exp\left(c h_n |x|^2 \right) (1 + |x|)^r, \quad (31)$$

$$\sup_\theta E_{\theta^*} \left[\left| \frac{\partial_{\beta_{i'}} \partial_{\beta_{j'}} \partial_{\beta_{l'}} p}{p} (h_n, x, X_{h_n}; \theta) \right|^\mu \middle| X_0 = x \right] \le K \exp\left(c h_n |x|^2 \right) (1 + |x|)^r \quad (32)$$

for $x \in \mathbf{R}^d$, $1 \le i \le p$, $1 \le j, l \le p+q$ and $1 \le i', j', l' \le q$. It follows from (26) and (29) that for any $\mu > 1$,

$$\lim_{n\to\infty} \sup \sup_\theta E_{\theta^*} \left[\left| \frac{1}{(nh_n)^{3/2}} \sum_{k=1}^n \frac{\partial_{\alpha_i} \partial_{\alpha_j} p}{p} \frac{\partial_{\alpha_l} p}{p} \left(h_n, X_{t^n_{k-1}}, X_{t^n_k}; \theta \right) \right|^\mu \right] = 0.$$

By the analogous way, for $i, j, l = 1, \dots, p+q$, and for any $\mu > 1$,

$$\lim_{n\to\infty} \sup \sup_\theta E_{\theta^*} \left[\left| d^3_{i,j,l,n} \sum_{k=1}^n \frac{\partial_{\theta_i} \partial_{\theta_j} p}{p} \frac{\partial_{\theta_l} p}{p} \left(h_n, X_{t^n_{k-1}}, X_{t^n_k}; \theta \right) \right|^\mu \right] = 0, \quad (33)$$

$$\lim_{n\to\infty} \sup \sup_\theta E_{\theta^*} \left[\left| d^3_{i,j,l,n} \sum_{k=1}^n \frac{\partial_{\theta_i} p}{p} \frac{\partial_{\theta_j} p}{p} \frac{\partial_{\theta_l} p}{p} \left(h_n, X_{t^n_{k-1}}, X_{t^n_k}; \theta \right) \right|^\mu \right] = 0. \quad (34)$$

Furthermore, it follows from (31) and (32) that for $i, j = 1, \dots, p$ and for $i', j', l' = 1, \dots, q$,

$$\limsup_{n\to\infty}\sup_{\theta} E_{\theta^*}\left[\left|\frac{1}{n^{3/2}h_n}\sum_{k=1}^{n}\frac{\partial_{\alpha_i}\partial_{\alpha_j}\partial_{\beta_{i'}}p}{p}\left(h_n, X_{t^n_{k-1}}, X_{t^n_k};\theta\right)\right|^{\mu}\right]=0, \quad (35)$$

$$\limsup_{n\to\infty}\sup_{\theta} E_{\theta^*}\left[\left|\frac{1}{n^{3/2}h_n^{1/2}}\sum_{k=1}^{n}\frac{\partial_{\alpha_i}\partial_{\beta_{i'}}\partial_{\beta_{j'}}p}{p}\left(h_n, X_{t^n_{k-1}}, X_{t^n_k};\theta\right)\right|^{\mu}\right]=0, \quad (36)$$

$$\limsup_{n\to\infty}\sup_{\theta} E_{\theta^*}\left[\left|\frac{1}{n^{3/2}}\sum_{k=1}^{n}\frac{\partial_{\beta_{i'}}\partial_{\beta_{j'}}\partial_{\beta_{l'}}p}{p}\left(h_n, X_{t^n_{k-1}}, X_{t^n_k};\theta\right)\right|^{\mu}\right]=0 \quad (37)$$

for any $\mu > 1$. Next, note that

$$E_{\theta}\left[\left|\partial_{\alpha_i}\partial_{\alpha_j}\partial_{\alpha_l}X^{\theta}_{h_n,l_1} - h_n[\partial_{\alpha_i}\partial_{\alpha_j}\partial_{\alpha_l}b(x,\alpha)]_{l_1}\right|^{\mu}|X^{\theta}_0 = x\right] = R\left(\theta, h_n^{3\mu/2}, x\right)$$

and that by Proposition 1, $\delta(\partial_{\alpha_i}\partial_{\alpha_j}\partial_{\alpha_l}X^{\theta}_{h_n,l_1}\cdot U^{\theta}_{l_1}) = \partial_{\alpha_i}\partial_{\alpha_j}\partial_{\alpha_l}X^{\theta}_{h_n,l_1}\delta(U^{\theta}_{l_1}) - \int_0^{h_n}\mathcal{D}_t\, \partial_{\alpha_i}\partial_{\alpha_j}\partial_{\alpha_l}X^{\theta}_{h_n,l_1}\cdot U^{\theta}_{t,l_1}\mathrm{d}t$. Moreover, it follows from the proof of (3.23) in Gobet (2002) that $E_{\theta}[|\delta(U_{l_1} - \hat{U}_{l_1})|^{\mu}|X^{\theta}_0 = x] = R(\theta, h_n^{\mu}, x)$ and

$$E_{\theta}\left[\left|\delta\left(\hat{U}^{\theta}_{l_1}\right) - \left[\sigma^{-2}(x,\beta)\left(X^{\theta}_{h_n} - m^{\theta}(x)\right)\right]_{l_1}\right|^{\mu}|X^{\theta}_0 = x\right] = R\left(\theta, h_n^{\mu}, x\right),$$

where $\hat{U}^{\theta}_{l_1} = (\hat{U}^{\theta}_{t,l_1})_{t\in[0,h_n]}$ is the $\mathbf{R}^d$-valued process whose l_2-th component is $\hat{U}^{\theta}_{t,l_1,l_2} = (\sigma^{-1}(X^{\theta}_t,\beta))_{l_2 l_1}$. Therefore, using (13) and the fact that

$$\sup_{0\le t\le h_n} E_{\theta}\left[\|\,\mathcal{D}_t\partial_{\alpha_i}\partial_{\alpha_j}\partial_{\alpha_l}X^{\theta}_{h_n}\,\|^{\mu}\ |X^{\theta}_0 = x\right] = R\left(\theta, h_n^{\mu}, x\right),$$

we obtain that $\Phi^{\theta}_{1,i,j,l} = h_n M_{1,i,j,l}(x, X^{\theta}_{h_n};\theta) + \mathcal{H}^{\theta}_{1,i,j,l}$, where for any $\mu > 1$, $E_{\theta}[|\mathcal{H}^{\theta}_{1,i,j,l}|^{\mu}|X^{\theta}_0 = x]^{1/\mu} = R(\theta, h_n^2, x)$ and

$$M_{1,i,j,l}(x,y;\theta) = \sum_{l_1=1}^{d}[\partial_{\alpha_i}\partial_{\alpha_j}\partial_{\alpha_l}b(x,\alpha)]_{l_1}[\sigma^{-2}(x,\beta)(y - m^{\theta}(x))]_{l_1}.$$

As in (18), for $i, j, l = 1, \ldots, p$, and for any $\mu > 1$,

$$E_{\theta^*}\left[\left|\frac{1}{nh_n}\sum_{k=1}^{n}\left\{\frac{\partial_{\alpha_i}\partial_{\alpha_j}\partial_{\alpha_l}p}{p}\left(h_n, X_{t^n_{k-1}}, X_{t^n_k};\theta\right) - M_{1,i,j,l}\left(X_{t^n_{k-1}}, X_{t^n_k};\theta\right)\right\}\right|^{\mu}\right]$$
$$\le \frac{C}{nh_n^{4\mu}}\sum_{k=1}^{n}E_{\theta^*}\left[E_{\theta}\left[\frac{p\left(h_n, X^{\theta}_0, X^{\theta}_{h_n};\theta^*\right)}{p\left(h_n, X_0, X_{h_n};\theta\right)}\left|\Phi^{\theta}_{3,i,j,l}\right|^{\mu}\middle| X^{\theta}_0 = X_{t^n_{k-1}}\right]\right]$$

$$+\frac{C}{nh_n^{3\mu}}\sum_{k=1}^{n}E_{\theta^*}\left[E_\theta\left[\frac{p\left(h_n,X_0^\theta,X_{h_n}^\theta;\theta^*\right)}{p\left(h_n,X_0,X_{h_n};\theta\right)}\left|\tilde{\Phi}_{2,i,j,l}^\theta\right|^\mu\middle|X_0^\theta=X_{t_{k-1}^n}\right]\right]$$
$$+\frac{C}{nh_n^{2\mu}}\sum_{k=1}^{n}E_{\theta^*}\left[E_\theta\left[\frac{p\left(h_n,X_0^\theta,X_{h_n}^\theta;\theta^*\right)}{p\left(h_n,X_0,X_{h_n};\theta\right)}\left|\mathcal{H}_{1,i,j,l}^\theta\right|^\mu\middle|X_0^\theta=X_{t_{k-1}^n}\right]\right].$$

From the result in the proofs of (1.8)–(1.9) in Gobet (2002),

$$\sup_\theta E_\theta\left[\left|\frac{p\left(h_n,x,X_{h_n}^\theta;\theta^*\right)}{p\left(h_n,x,X_{h_n}^\theta;\theta\right)}\right|^{\mu_1}\middle|X_0^\theta=x\right]\leq C\exp\left(Ch_n|x|^2\right)$$

for μ_1 closed to 1 and for some $C>1$. Moreover, we note that for any $\mu>1$, $E_\theta[|\Phi_{3,i,j,l}^\theta|^\mu|X_0^\theta=x]=R(\theta,h_n^{9\mu/2},x)$ and $E_\theta[|\tilde{\Phi}_{2,i,j,l}^\theta|^\mu|X_0^\theta=x]=R(\theta,h_n^{3\mu},x)$ for $1\leq i,j,l\leq p$. It follows from the Burkholder inequality that $\limsup_{n\to\infty}\sup_\theta E_{\theta^*}[|\frac{1}{nh_n}\sum_{k=1}^n\sum_{l_1=1}^d M_{1,i,j,l,l_1}(X_{t_{k-1}^n},X_{t_k^n};\theta)|^\mu]<\infty$ for any $\mu>1$. Therefore,

$$\limsup_{n\to\infty}\sup_\theta E_{\theta^*}\left[\left|\frac{1}{nh_n}\sum_{k=1}^{n}\frac{\partial_{\alpha_i}\partial_{\alpha_j}\partial_{\alpha_l}p}{p}\left(h_n,X_{t_{k-1}^n},X_{t_k^n};\theta\right)\right|^\mu\right]<\infty \qquad (38)$$

for any $\mu>1$. By (35)–(38), one has that for $i,j,l=1,\ldots,p+q$ and for any $\mu>1$,

$$\limsup_{n\to\infty}\sup_\theta E_{\theta^*}\left[\left|d_{i,j,l,n}^3\sum_{k=1}^{n}\frac{\partial_{\theta_i}\partial_{\theta_j}\partial_{\theta_l}p}{p}\left(h_n,X_{t_{k-1}^n},X_{t_k^n};\theta\right)\right|^\mu\right]=0. \qquad (39)$$

From (33), (34) and (39), $\limsup_{n\to\infty}\sup_\theta E_{\theta^*}[|d_{i,j,l,n}^3\partial_{i,j,l}^3 l_n(\theta)|^\mu]=0$ for $i,j,l=1,\ldots,p+q$ and for any $\mu>1$. By the analogous techniques with the proofs of (33)–(38), it follows that for $i,j,l,m=1,\ldots,p+q$, and for any $\mu>1$,

$$\limsup_{n\to\infty}\sup_\theta E_{\theta^*}\left[\left|d_{i,j,l,n}^3\sum_{k=1}^{n}\partial_{\theta_m}\left(\frac{\partial_{\theta_i}p}{p}\frac{\partial_{\theta_j}p}{p}\frac{\partial_{\theta_l}p}{p}\right)\left(h_n,X_{t_{k-1}^n},X_{t_k^n};\theta\right)\right|^\mu\right]=0.$$
$$\limsup_{n\to\infty}\sup_\theta E_{\theta^*}\left[\left|d_{i,j,l,n}^3\sum_{k=1}^{n}\partial_{\theta_m}\left(\frac{\partial_{\theta_i}\partial_{\theta_j}p}{p}\frac{\partial_{\theta_l}p}{p}\right)\left(h_n,X_{t_{k-1}^n},X_{t_k^n};\theta\right)\right|^\mu\right]=0,$$
$$\limsup_{n\to\infty}\sup_\theta E_{\theta^*}\left[\left|d_{i,j,l,n}^3\sum_{k=1}^{n}\partial_{\theta_m}\left(\frac{\partial_{\theta_i}\partial_{\theta_j}\partial_{\theta_l}p}{p}\right)\left(h_n,X_{t_{k-1}^n},X_{t_k^n};\theta\right)\right|^\mu\right]=0.$$

Thus, we obtain that $\limsup_{n\to\infty}\sup_\theta E_{\theta^*}[|d_{i,j,l,n}^3\partial_{i,j,l,m}^4 l_n(\theta)|^\mu]=0$ for any $\mu>1$ and for $i,j,l,m=1,\ldots,p+q$. The Sobolev inequality implies that for $\mu>p+q$, $\limsup_{n\to\infty}E_{\theta^*}[\sup_\theta|d_{i,j,l,n}^3\partial_{i,j,l}^3 l_n(\theta)|^\mu]=0$. This completes the proof.

Proof of Theorem 1. It follows from (10) that $P_{\theta^*}[B_{m,n}^c] = O(\frac{1}{n^{(\epsilon_0/2-\beta_0)L}})$, where $B_{m,n} = \{|\hat{\theta}_{m,n} - \theta_{m,0}| \leq n^{-\beta_0}\}$ for $\beta_0 \in (0, \epsilon_0/2)$. Setting $\mathrm{BIAS}(m) = g_{m,n}(\hat{\theta}_{m,n}(\mathbf{X}_n)) - g_{m,n}(\theta_{m,0}) + E_{\mathbf{Z}_n}[l_{m,n}(\mathbf{Z}_n, \theta_{m,0})] - E_{\mathbf{Z}_n}[l_{m,n}(\mathbf{Z}_n, \hat{\theta}_{m,n}(\mathbf{X}_n))]$, one has that $\mathrm{IC}(\mathbf{X}_n, m) - \mathrm{EL}(\mathbf{X}_n, m) = \mathrm{BIAS}(m)1_{B_{m,n}} + \mathrm{BIAS}(m)1_{B_{m,n}^c} - \dim(\Theta_m)$. The Hölder inequality implies that for $\mu > 1$,

$$\begin{aligned}
&E_{\theta^*}[|\mathrm{BIAS}(m)1_{B_{m,n}^c}|] \\
&\quad \leq 2E_{\theta^*}\left[\sup_{\theta_m}\left|g_{m,n}(\theta_m) + \frac{nd}{2}\log(2\pi h_n)\right|^{\mu}\right]^{1/\mu} P_{\theta^*}\left[B_{m,n}^c\right]^{1-1/\mu} \\
&\qquad +2E_{\mathbf{Z}_n}\left[\sup_{\theta_m}\left|l_{m,n}(\mathbf{Z}_n, \theta_m) + \frac{nd}{2}\log(h_n)\right|^{\mu}\right]^{1/\mu} P_{\theta^*}\left[B_{m,n}^c\right]^{1-1/\mu}.
\end{aligned}$$

From Proposition 1.2 in Gobet (2002), we obtain that for some positive constants K and c,

$$\sup_{\theta_m}\left|\log p_m(h_n, x, y; \theta_m) + \frac{d}{2}\log(h_n)\right| \leq K + \frac{c|x-y|^2}{h_n} + ch_n|x|^2,$$

and consequently, for sufficiently large $L > 0$,

$$\begin{aligned}
&E_{\mathbf{Z}_n}\left[\sup_{\theta_m}\left|l_{m,n}(\mathbf{Z}_n, \theta_m) + \frac{nd}{2}\log(h_n)\right|^{\mu}\right]^{1/\mu} P_{\theta^*}\left[B_{m,n}^c\right]^{1-1/\mu} \\
&\quad = O\left(\frac{1}{n^{(\epsilon_0/2-\beta_0)L(1-1/\mu)-1}}\right) = o(1)
\end{aligned}$$

for $\mu > 1$. In the similar way,

$$E_{\theta^*}\left[\sup_{\theta_m}\left|g_{m,n}(\theta_m) + \frac{nd}{2}\log(2\pi h_n)\right|^{\mu}\right]^{1/\mu} P_{\theta^*}\left[B_{m,n}^c\right]^{1-1/\mu} = o(1)$$

and we obtain that $E_{\theta^*}[|\mathrm{BIAS}(m)1_{B_{m,n}^c}|] = o(1)$. Hence,

$$E_{\theta^*}\left[\mathrm{IC}(\mathbf{X}_n, m) - \mathrm{EL}(\mathbf{X}_n, m)\right] = E_{\theta^*}[\xi_{m,n} - \eta_{m,n}] - \dim(\Theta_m) + o(1),$$

where $\xi_{m,n} = \left(g_{m,n}(\hat{\theta}_{m,n}(\mathbf{X}_n)) - g_{m,n}(\theta_{m,0})\right)1_{B_{m,n}}$ and

$$\eta_{m,n} = \left(E_{\mathbf{Z}_n}[l_{m,n}(\mathbf{Z}_n, \hat{\theta}_{m,n}(\mathbf{X}_n))] - E_{\mathbf{Z}_n}[l_{m,n}(\mathbf{Z}_n, \theta_{m,0})]\right)1_{B_{m,n}}.$$

By the Taylor expansion, one has that $\xi_{m,n} = \xi_{m,n}^{(1)} + \xi_{m,n}^{(2)} + \xi_{m,n}^{(3)}$ and $\eta_{m,n} = \eta_{m,n}^{(1)} + \eta_{m,n}^{(2)} + \eta_{m,n}^{(3)}$, where $\tilde{\theta}_{m,t} = \theta_{m,0} + t(\hat{\theta}_{m,n} - \theta_{m,0})$,

$$\begin{aligned}
\xi_{m,n}^{(1)} &= \sum_{l_1=1}^{p_m+q_m} d_{m,l_1,n}(\partial_{m,l_1} g_{m,n})(\theta_{m,0})\hat{u}_{m,l_1,n} 1_{B_{m,n}},\\
\xi_{m,n}^{(2)} &= \frac{1}{2}\sum_{l_1,l_2=1}^{p_m+q_m} d_{m,l_1,l_2,n}^2 \left(\partial_{m,l_1,l_2}^2 g_{m,n}\right)(\theta_{m,0}) \prod_{m=1}^{2} \hat{u}_{m,l_m,n} 1_{B_{m,n}},\\
\xi_{m,n}^{(3)} &= \frac{1}{2}\sum_{l_1,l_2,l_3=1}^{p_m+q_m} \int_0^1 (1-t)^2 d_{m,l_1,l_2,l_3,n}^3 \left(\partial_{m,l_1,l_2,l_3}^3 g_{m,n}\right)(\tilde{\theta}_{m,t})\mathrm{d}t \prod_{i=1}^{3} \hat{u}_{m,l_i,n} 1_{B_{m,n}},\\
\eta_{m,n}^{(1)} &= \sum_{l_1=1}^{p_m+q_m} E_{\mathbf{Z}_n}[d_{m,l_1,n}(\partial_{m,l_1} l_n)(\mathbf{Z}_n, \theta_{m,0})]\hat{u}_{m,l_1,n} 1_{B_{m,n}},\\
\eta_{m,n}^{(2)} &= \frac{1}{2}\sum_{l_1,l_2=1}^{p_m+q_m} E_{\mathbf{Z}_n}\left[d_{m,l_1,l_2,n}^2 \left(\partial_{m,l_1,l_2}^2 l_{m,n}\right)(\mathbf{Z}_n, \theta_{m,0})\right] \prod_{i=1}^{2} \hat{u}_{m,l_i,n} 1_{B_{m,n}},\\
\eta_{m,n}^{(3)} &= \frac{1}{2}\sum_{l_1,l_2,l_3=1}^{p_m+q_m} E_{\mathbf{Z}_n}\left[\int_0^1 (1-t)^2 d_{m,l_1,l_2,l_3,n}^3 \left(\partial_{m,l_1,l_2,l_3}^3 l_{m,n}\right)(\mathbf{Z}_n, \tilde{\theta}_{m,t})\mathrm{d}t\right]\\
&\quad \times \prod_{i=1}^{3} \hat{u}_{m,l_i,n} 1_{B_{m,n}}.
\end{aligned}$$

We first estimate the moment of $\xi_{m,n}^{(j)}$ for $j = 1, 2, 3$. Let $A_{m,n} = \{|R_{m,n}| \leq n^{-E_0}\}$ for $E_0 \in (0, \epsilon_0^2/2)$, where $R_{m,n}$ is defined in Lemma 1–(iii). It then follows that

$$\begin{aligned}
\xi_{m,n}^{(1)} 1_{A_{m,n}} = &\sum_{l_1,l_2=1}^{p_m+q_m} \partial\hat{g}_{m,l_1,n} \left(I_m^{-1}(\theta_{m,0})\right)_{l_1 l_2} \partial\hat{g}_{m,l_2,n} 1_{B_{m,n}} 1_{A_{m,n}}\\
&+ \sum_{l_1=1}^{p_m+q_m} \partial\hat{g}_{m,l_1,n} R_{m,l_1,n} 1_{B_{m,n}} 1_{A_{m,n}},
\end{aligned}$$

where $\partial\hat{g}_{m,l_1,n} = d_{m,l_1,n}(\partial_{m,l_1} g_{m,n})(\theta_{m,0})$. Note that $1_{A_{m,n}} \to^p 1$, $1_{B_{m,n}} \to^p 1$, and we obtain that $\sum_{l_1,l_2=1}^{p_m+q_m} \partial\hat{g}_{m,l_1,n}(I_m^{-1}(\theta_{m,0}))_{l_1 l_2}\partial\hat{g}_{m,l_2,n} \to^p \dim(\Theta_m)$, $\limsup_{n\to\infty} E_{\theta^*}[|\partial\hat{g}_{m,l_1,n}|^{\mu}] < \infty$, and $E_{\theta^*}[|R_{m,l_1,n}|^{\mu} 1_{A_{m,n}}] = O(n^{-\mu E_0})$ for any $\mu > 1$. Therefore, one has that $E_{\theta^*}[\xi_{m,n}^{(1)} 1_{A_{m,n}}] \to \dim(\Theta_m)$. Lemma 1 yields that $E_{\theta^*}[|\xi_{m,n}^{(1)} 1_{A_{m,n}^c}|] = o(1)$. Thus, $E_{\theta^*}[\xi_{m,n}^{(1)}] = \dim(\Theta_m) + o(1)$. Next, it is easy to show that $d_{m,l_1,l_2,n}^2(\partial_{m,l_1,l_2}^2 g_{m,n})(\theta_{m,0}) + (I_m(\theta_{m,0}))_{l_1 l_2} = o_p(1)$, for $\mu > 1$, $\limsup\limits_{n\to\infty} E_{\theta^*}[|d_{m,l_1,l_2,n}^2(\partial_{m,l_1,l_2}^2 g_{m,n})$

$(\theta_{m,0})|^{\mu}] < \infty$ and

$$E_{\theta^*}\left[\sum_{l_1,l_2=1}^{p_m+q_m}(I_m(\theta_{m,0}))_{l_1l_2}\prod_{i=1}^{2}\hat{u}_{m,l_i,n}1_{B_{m,n}}\right]=\dim(\Theta_m)+o(1).$$

Consequently, $E_{\theta^*}[\xi^{(2)}_{m,n}] = -\frac{1}{2}\dim(\Theta_m)+o(1)$. It follows from the standard arguments that $\limsup_{n\to\infty}\sup_{\theta_m}E_{\theta^*}[|d^3_{m,l_1,l_2,l_3,n}(\partial^3_{m,l_1,l_2,l_3}g_{m,n})(\theta_m)|^{\mu}] = o(1)$ and $\limsup_{n\to\infty}\sup_{\theta_m}E_{\theta^*}[|d^3_{m,l_1,l_2,l_3,n}(\partial^4_{m,l_1,l_2,l_3,l_4}g_{m,n})(\theta_m)|^{\mu}] = o(1)$. Lemma 1 together with Sobolev's inequality yields that $E_{\theta^*}[|\xi^{(3)}_{m,n}|]=o(1)$.

Finally, we consider the estimates of $E_{\theta^*}[\eta^{(j)}_{m,n}]$ for $j=1,2,3$. Lemmas 1 and 2 yield that $E_{\theta^*}[\hat{u}_{m,l_1,n}1_{B_{m,n}}]=o(1)$ and $E_{\mathbf{Z}_n}[d_{m,l_1,n}(\partial_{m,l_1}l_n)(\mathbf{Z}_n,\theta_{m,0})]=0$. Thus, we obtain that $E_{\theta^*}[\eta^{(1)}_{m,n}]=o(1)$. For the estimate of $E_{\theta^*}[\eta^{(2)}_{m,n}]$, Lemma 2 implies that $E_{\mathbf{Z}_n}[d^2_{m,l_1,l_2,n}(\partial^2_{m,l_1,l_2}l_{m,n})(\mathbf{Z}_n,\theta_{m,0})] \to -(I_m(\theta_{m,0}))_{l_1l_2}$. Moreover, Lemma 1 yields that $E_{\theta^*}\big[\prod_{m=1}^{2}\hat{u}_{m,l_m,n}1_{B_{m,n}}\big] \to (I_m^{-1}(\theta_{m,0}))_{l_1l_2}$. Thus, $E_{\theta^*}\big[\eta^{(2)}_{m,n}\big]=-\frac{1}{2}\dim(\Theta_m)+o(1)$. Since it follows from Lemmas 1 and 2 that $\limsup_{n\to\infty}E_{\theta^*}\big[|\prod_{i=1}^{3}\hat{u}_{m,l_i,n}|^{\mu}\big]<\infty$ and that

$$E_{\mathbf{Z}_n}\left[\left(d^3_{m,l_1,l_2,l_3,n}\sup_{\theta_m}\left|\left(\partial^3_{m,l_1,l_2,l_3}l_{m,n}\right)(\mathbf{Z}_n,\theta_m)\right|\right)^{\mu}\right]=o(1),$$

one has that $E_{\theta^*}[|\eta^{(3)}_{m,n}|]=o(1)$. This completes the proof.

Acknowledgments The author would like to express his thanks to Professor Nakahiro Yoshida for many helpful comments and stimulating discussions during the preparation of this work. He is also grateful to the referee for valuable comments which led to improvements of an earlier version of this paper. This work was partially supported by Grants-in-Aid for Scientific Research from the JSPS, and by Cooperative Research Program of the Institute of Statistical Mathematics.

References

Adams, R. A., Fournier, J. J. F. (2003). *Sobolev spaces* (2nd ed.). In *Pure and Applied Mathematics* (Vol. 140). Amsterdam: Elsevier/Academic Press.

Aït-Sahalia, Y. (2008). Closed-form likelihood expansions for multivariate diffusions. *The Annals of Statistics, 36*, 906–937.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov, F. Csaki (Eds.), *2nd international symposium in information theory*, pp. 267–281. Budapest: Akademiai Kaido.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control, AC-19*, pp. 716–723.

Beskos, A., Papaspiliopoulos, O., Roberts, G. O., Fearnhead, P. (2006). Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussions). *Journal of the Royal Statistical Society, Series B, Statistical Methodology, 68*, 333–382.

Burman, P., Nolan, D. (1995). A general Akaike-type criterion for model selection in robust regression. *Biometrika, 82*, 877–886.

Burnham, K. P., Anderson, D. R. (2002). *Model selection and multimodel Inference* (2nd ed.). New York: Springer.

Florens-Zmirou, D. (1989). Approximate discrete time schemes for statistics of diffusion processes. *Statistics, 20*, 547–557.

Genon-Catalot, V., Jacod, J. (1993). On the estimation of the diffusion coefficient for multidimensional diffusion processes. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques, 29*, 119–151.

Gobet, E. (2001). Local asymptotic mixed normality property for elliptic diffusion: a Malliavin calculus approach. *Bernoulli, 7*, 899–912.

Gobet, E. (2002). LAN property for ergodic diffusions with discrete observations. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques, 38*, 711–737.

Hall, P. (1990). Akaike's information criterion and Kullback–Leibler loss for histogram density estimation. *Probability Theory and Related Fields, 85*, 449–467.

Hurvich, C. M., Simonoff, J. S., Tsai, C. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society, Series B, Statistical Methodology, 60*, 271–293.

Inagaki, N., Ogata, Y. (1975). The weak convergence of likelihood ratio random fields and its applications. *Annals of the Institute of Statistical Mathematics, 27*, 391–419.

Jacod, J. (2006). Parametric inference for discretely observed non-ergodic diffusions. *Bernoulli, 12*, 383–401.

Kessler, M. (1997). Estimation of an ergodic diffusion from discrete observations. *Scandinavian Journal of Statistics, 24*, 211–229.

Kloeden, P. E., Platen, E. (1992) *Numerical solution of stochastic differential equations*. New York: Springer

Konishi, S., Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika, 83*, 875–890.

Konishi, S., Kitagawa, G. (2008). *Information criteria and statistical modeling*. New York: Springer.

Kunita, H. (1984). Stochastic differential equations and stochastic flows of diffeomorphisms. *Ecole d'eté de probabilités de Saint-Flour, XII—1982*, Lecture Notes in Mathematics (Vol. 1097) (pp. 143–303). Berlin: Springer.

Kutoyants, Y. A. (2004). *Statistical inference for ergodic diffusion processes*. London: Springer.

Malliavin, P. (1997). *Stochastic analysis*. Berlin: Springer.

Nualart, D. (2006). *The Malliavin calculus and related topics* (2nd ed.). Berlin: Springer.

Pardoux, E., Veretennikov, A. Y. (2001). On the Poisson equation and diffusion approximation 1. *The Annals of Probability, 29*, 1061–1085.

Prakasa Rao, B. L. S. (1983). Asymptotic theory for nonlinear least squares estimator for diffusion processes. *Mathematische Operationsforschung und Statistik Series Statistics, 14*, 195–209.

Prakasa Rao, B. L. S. (1988). Statistical inference from sampled data for stochastic processes. *Contemporary Mathematics* (Vol. 80, pp. 249–284). Providence, RI: American Mathematical Society.

Sakamoto, Y., Yoshida, N. (2004). Asymptotic expansion formulas for functionals of ϵ-Markov processes with a mixing property. *Annals of the Institute of Statistical Mathematics, 56*, 545–597.

Sei, T., Komaki, F. (2007). Bayesian prediction and model selection for locally asymptotically mixed normal models. *Journal of Statistical Planning and Inference, 137*, 2523–2534.

Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika, 63*, 117–126.

Takeuchi, K. (1976). Distribution of information statistics and criteria for adequacy of models. *Mathematical Sciences, 153*, 12–18 (in Japanese).

Uchida, M., Yoshida, N. (2001). Information criteria in model selection for mixing processes. *Statistical Inference for Stochastic Processes, 4*, 73–98.

Uchida, M., Yoshida, N. (2004). Information criteria for small diffusions via the theory of Malliavin–Watanabe. *Statistical Inference for Stochastic Processes, 7*, 35–67.

Uchida, M., Yoshida, N. (2006). Asymptotic expansion and information criteria. *SUT Journal of Mathematics, 42*, 31–58.

Yoshida, N. (1992a). Asymptotic expansion of maximum likelihood estimators for small diffusions via the theory of Malliavin-Watanabe. *Probability Theory and Related Fields, 92*, 275–311.

Yoshida, N. (1992b). Estimation for diffusion processes from discrete observation. *Journal of Multivariate Analysis, 41*, 220–242.

Yoshida, N. (1997) Malliavin calculus and asymptotic expansion for martingales. *Probability Theory and Related Fields, 109*, 301–342.

Yoshida, N. (2004). Partial mixing and Edgeworth expansion. *Probability Theory and Related Fields, 129*, 559–624.

Yoshida, N. (2005). Polynomial type large deviation inequalities and quasi-likelihood analysis for stochastic differential equations (to appear in *Annals of the Institute of Statistical Mathematics*).