Bernoulli **24**(3), 2018, 2278–2327
https://doi.org/10.3150/17-BEJ928

Schwarz type model comparison for LAQ models

SHOICHI EGUCHI[1] and HIROKI MASUDA[2]

[1]*Graduate School of Mathematics, Kyushu University, 744 Motooka Nishi-ku Fukuoka 819-0395, Japan. E-mail: s-eguchi@math.kyushu-u.ac.jp*
[2]*Faculty of Mathematics, Kyushu University, 744 Motooka Nishi-ku Fukuoka 819-0395, Japan. E-mail: hiroki@math.kyushu-u.ac.jp*

For model-comparison purpose, we study asymptotic behavior of the marginal quasi-log likelihood associated with a family of locally asymptotically quadratic (LAQ) statistical experiments. Our result entails a far-reaching extension of applicable scope of the classical approximate Bayesian model comparison due to Schwarz, with frequentist-view theoretical foundation. In particular, the proposed statistics can deal with both ergodic and non-ergodic stochastic process models, where the corresponding M-estimator may of multi-scaling type and the asymptotic quasi-information matrix may be random. We also deduce the consistency of the multistage optimal-model selection where we select an optimal sub-model structure step by step, so that computational cost can be much reduced. Focusing on some diffusion type models, we illustrate the proposed method by the Gaussian quasi-likelihood for diffusion-type models in details, together with several numerical experiments.

Keywords: approximate Bayesian model comparison; Gaussian quasi-likelihood; locally asymptotically quadratic family; quasi-likelihood; Schwarz's criterion

1. Introduction

The objective of this paper is Bayesian model comparison for a general class of statistical models, which includes various kinds of stochastic process models that cannot be handled by preceding results. There are two classical principles of model selection: the Kullback–Leibler divergence (KL divergence) principle and the Bayesian one, acted over Akaike information criterion (AIC, [1,2]) and Schwarz or Bayesian information criterion (BIC, [35]), respectively. A common knowledge is that there are no universal politic between AIC and BIC type statistics, and they are indeed used for different purposes. On the one hand, the AIC is a predictive model selection criterion minimizing the KL divergence between prediction and true models, not intended to pick up the true model consistently even if it does exist in the candidate-model set. On the other hand, the BIC is used to look for better model description, putting importance not only on underfitting but also on overfitting. The BIC usually takes the form

$$\mathrm{BIC}_n = -2\ell_n\big(\hat{\theta}_n^{\mathrm{MLE}}\big) + p\log n,$$

where ℓ_n, $\hat{\theta}_n^{\mathrm{MLE}}$, and p denote the log-likelihood function, the maximum-likelihood estimator (MLE), and the dimension of the parameter space of the statistical model to be assessed, respectively. The model selection consistency via BIC type statistics has been studied by many authors

1350-7265

in several different model setups, for example, [5,8], and [34], to mention just a few old ones. An extension of the BIC-derivation logic to subsume smoothly regularized likelihood estimation can be found in [26].

There also do exist many studies of the BIC methodology in the time series context. The underlying principles, such as maximization of posterior model selection probability, remain the same in this case. It should be mentioned that [9] demonstrated that derivation of the classical BIC could be generalized into general $\sqrt{n}$-consistent framework with constant asymptotic information. Their argument supposes the almost-sure behaviors of the likelihood characteristics, especially of the observed information matrix. Our stance is similar to theirs, but more general so as to subsume a much broader spectrum of models that cannot be handled by [9]. We note that much less has been known about theoretically guaranteed information criteria concerning sampled data from stochastic process models; to mention some of them, we refer to [37–40] and [44].

Our primary interest is to extend the range of application of Schwarz's BIC to a large degree in a unified way, so as to be able to target a wide class of dependent data models especially including the locally asymptotically mixed-normal family of statistical experiments. The Bayesian principle of model selection amounts choosing the model that is most likely in terms of the posterior model selection probability, which is typically measured by approximating the (expected) marginal quasi-log likelihood. Unfortunately, a mathematically rigorous derivation of BIC type statistics is sometimes missing in the literature, especially when the underlying model is non-ergodic. In this paper, we will focus on *locally asymptotically quadratic* (*LAQ*) statistical models. We will introduce the quasi-BIC (QBIC) through the stochastic expansion of the marginal quasi-likelihood. Here, we use the terminology "quasi" to mean that the model may be misspecified in the sense that none of candidate models may not include the true one; see [31] for information criteria for a class of generalized linear models for independent data. Our proof of the expansion essentially utilizes the polynomial type large deviation inequality of [47]; quite importantly, the asymptotic information matrix then may be random (i.e., suitably scaled observed information (random bilinear form) has a random limit in probability), enabling us to deal with non-ergodic models in a unified way. We note the two things, though we do not go into any detail in this paper: the popular cointegration models (see [4] and the references therein) would be in the scope of the QBIC as well; the QBIC may be closely related to the correct BIC in the context of non-stationary time series models [25], where the observed information matrix is involved in the bias-correction term. Further, it is worth mentioning that QBIC may be used even for semiparametric models, where possibly infinite-dimensional nuisance element, whenever a suitable quasi-likelihood is available.

There are many other works on the model selection, which includes the risk information criterion [17], the generalized information criterion [27], the "parametricness" index [30], and many extensions of AIC and BIC including [12,31]. We refer to [7,13], and [28] for comprehensive accounts of information criteria, and also to [14] for an illustration from practical point of view.

This paper is organized as follows. Section 2 describes the basic model setup and some related backgrounds. In Section 3, we will present the asymptotic expansions of the marginal quasi-log likelihood (equivalently, the Bayes factor or the Kullback–Leibler divergence); the presentation contains a revised and extended version of [15]. In Section 4, we illustrate the proposed model selection method by the Gaussian quasi-likelihoods, with focuses on estimation of an ergodic

diffusion process and volatility-parameter estimation for a class of continuous semimartingales, both based on high-frequency sampling; to the best of our knowledge, this is the first place that mathematically validates Schwarz's methodology of model comparison for high-frequency data from a stochastic process. Section 5 is devoted to the model selection consistency with respect to the optimal model, which is naturally defined to be a minimal model among those minimizing the quasi-entropy quantities. When in particular the quasi-maximum likelihood estimator is of multi-scaling type, we prove the consistency of the multistage optimal model selection procedure, where we partially select an optimal model structure step by step, resulting in a reduced computational cost. Section 6 give some numerical experiments supporting our asymptotic results. All the proofs are presented in Section 7.

2. Preliminaries

2.1. Basic model setup

We begin with describing our basic Bayesian-model setup used throughout this paper. Denote by $\mathbf{X}_n$ an observation random variable defined on an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and by $G_n(dx) = g_n(x)\mu_n(dx)$ the true distribution $\mathcal{L}(\mathbf{X}_n)$, where μ_n is a σ-finite dominating measure on a Borel state space of $\mathbf{X}_n$, that is, $G_n(dx) = \mathbb{P} \circ \mathbf{X}_n^{-1}(dx)$.

Suppose that we are given a set of M candidate Bayesian models $\mathcal{M}_1, \ldots, \mathcal{M}_M$:

$$\mathcal{M}_m = \big\{\big(\mathfrak{p}_m, \pi_{m,n}(\theta_m), \mathbb{H}_{m,n}(\cdot|\theta_m)\big)|\theta_m \in \Theta_m\big\}, \qquad m = 1, \ldots, M,$$

where the ingredients in each $\mathcal{M}_m$ are given as follows.

- $\mathfrak{p}_m > 0$ denotes the relative likeliness of the model-$\mathcal{M}_m$ occurrence among $\mathcal{M}_1, \ldots, \mathcal{M}_M$; we have $\sum_{m=1}^M \mathfrak{p}_m = 1$.
- $\pi_{m,n} : \Theta_m \to (0, \infty)$ is the prior distribution $\mathcal{L}(\theta_m)$ of mth-model parameter θ_m, here defined to be a probability density function possibly depending on the sample size n, with respect to the Lebesgue density on a bounded convex domain $\Theta_m \subset \mathbb{R}^{p_m}$.
- The measurable function $x \mapsto \mathbb{H}_{m,n}(x|\theta_m)$ for each $\theta_m \in \Theta_m$ defines a logarithmic regular conditional probability density of $\mathcal{L}(\mathbf{X}_n|\theta_m)$ with respect to $\mu_n(dx)$.

Each $\mathcal{M}_m$ may be misspecified in the sense that the true data generating model $g_n(x)$ does not belong to the family $\{\exp\{\mathbb{H}_{m,n}(\cdot|\theta_m)\}|\theta_m \in \Theta_m\}$; we will, however, assume suitable regularity conditions for the associated statistical random fields.

Concerning the model $\mathcal{M}_m$, the random function $\theta_m \mapsto \exp\{\mathbb{H}_{m,n}(\mathbf{X}_n|\theta_m)\}$, assumed to be a.s. well-defined, is referred to as the *quasi-likelihood* of $\mathcal{L}(\mathbf{X}_n|\theta_m)$. The *quasi-maximum likelihood estimator* (*QMLE*) $\hat{\theta}_{m,n}$ associated with $\mathbb{H}_{m,n}$ is defined to be any maximizer of $\mathbb{H}_{m,n}$:

$$\hat{\theta}_{m,n} \in \operatorname*{argmax}_{\theta \in \bar{\Theta}_m} \mathbb{H}_{m,n}(\mathbf{X}_n|\theta).$$

We will assume the a.s. continuity of $\mathbb{H}_{m,n}$ over the compact set $\bar{\Theta}_m$, so that $\hat{\theta}_{m,n}$ always exists.

Our objective includes estimators of multi-scaling type, meaning that the components of $\hat{\theta}_{m,n}$ converges at different rates, which can often occur when considering high-frequency asymptotics. A typical example is the Gaussian quasi-likelihood estimation of ergodic diffusion process: see [24], also Section 4.2. Let $K_m \in \mathbb{N}$ be a given number, which represents the number of the components having different convergence rates in $\mathcal{M}_m$, and assume that the mth-model parameter vector is divided into K_m parts:

$$\theta_m = (\theta_{m,1}, \dots, \theta_{m,K_m}) \in \prod_{k=1}^{K_m} \Theta_{m,k} = \Theta_m,$$

with each $\Theta_{m,k}$ being a bounded convex domain in $\mathbb{R}^{p_{m,k}}$, $k \in \{1, \dots, K_m\}$, where $p_m = \sum_{k=1}^{K_m} p_{m,k}$. Then the QMLE in the mth model takes the form $\hat{\theta}_{m,n} = (\hat{\theta}_{m,1,n}, \dots, \hat{\theta}_{m,K_m,n})$. The optimal value of θ_m associated with $\mathbb{H}_{m,n}$, to be precisely defined later on, is denoted by $\theta_{m,0} = (\theta_{m,1,0}, \dots, \theta_{m,K_m,0})$, $\theta_{m,k,0} \in \Theta_{m,k}$. The rate matrix in the model $\mathcal{M}_m$ is then given in the form

$$A_{m,n}(\theta_{m,0}) = \operatorname{diag}\bigl(a_{m,1,n}(\theta_{m,0}) I_{p_{m,1}}, \dots, a_{m,K_m,n}(\theta_{m,0}) I_{p_{m,K_m}}\bigr), \tag{2.1}$$

where I_p denotes the p-dimensional identity matrix and $a_{m,k,n}\ (\theta_{m,0})$ are deterministic positive sequences satisfying that

$$a_{m,k,n}(\theta_{m,0}) \to 0, \qquad a_{m,i,n}(\theta_{m,0}) / a_{m,j,n}(\theta_{m,0}) \to 0 \qquad (i < j), n \to \infty. \tag{2.2}$$

The diagonality of $A_{m,n}(\theta_0)$ is just for simplicity.

Since we are allowing not only data dependency but also the possibility of model misspecification, we may deal with a wide range of quasi-likelihoods $\mathbb{H}_{m,n}$, even including semiparametric situations such as the Gaussian quasi-likelihood; see Section 4 for related models.

2.2. Bayesian model selection principle

The quasi-marginal distribution of $\mathbf{X}_n$ in the mth model $\mathcal{M}_m$ is given by the density

$$x \mapsto f_{m,n}(x) := \int_{\Theta_m} \exp\bigl\{\mathbb{H}_{m,n}(x|\theta_m)\bigr\} \pi_{m,n}(\theta_m)\, d\theta_m,$$

which is sometimes referred to as the model evidence of $\mathcal{M}_i$. Typical reasoning in Bayesian principle of model selection in $\mathcal{M}_1, \dots, \mathcal{M}_M$ is to choose the model that is most likely to occur in terms of the posterior model selection probability, namely to choose the model maximizing

$$\log\left(\frac{f_{m,n}(x)\mathfrak{p}_{m,n}}{\sum_{i=1}^{M} f_{i,n}(x)\mathfrak{p}_{i,n}}\right) = \log f_{m,n}(x) + \log \mathfrak{p}_m - \log\left(\sum_{i=1}^{M} f_{i,n}(x)\mathfrak{p}_i\right)$$

over $m = 1, \dots, M$. This is equivalent to finding

$$\operatorname*{argmax}_{m \le M} \bigl\{\log f_{m,n}(x) + \log \mathfrak{p}_m\bigr\}.$$

Then one proceeds with suitable almost-sure ($\Omega \ni \omega$-wise) asymptotic expansion of the logarithm of the quasi-marginal likelihood $\log f_{m,n}(x)$ for $n \to \infty$ around a suitable estimator, a measurable function of $x = x_n$ for each n: when $\sqrt{n}(\hat{\theta}_{m,n} - \theta_{m,0}) = O_p(1)$, the resulting form may be quite often given by

$$\log f_{m,n}(x) + \log \mathfrak{p}_m \approx \mathbb{H}_{m,n}(x|\hat{\theta}_{m,n}) - \frac{p_m}{2}\log n + O(1) \qquad \text{a.s.} \tag{2.3}$$

This is the usual scenario of derivation of the classical-BIC type statistics; see [9] and [31] as well as [35].

We recall that the expansion (2.3) is also used to approximate the Bayes factor. The logarithmic Bayes factor of $\mathcal{M}_i$ against $\mathcal{M}_j$ is defined by the (random) ratio of posterior and prior odds of model selection probabilities: letting $\mathbb{P}(\mathcal{M}_i|\mathbf{X}_n)$ denote the posterior model selection probability of the ith model, we have

$$\log \mathrm{BF}_n(i, j) := \log \frac{\mathbb{P}(\mathcal{M}_i|\mathbf{X}_n)/\mathbb{P}(\mathcal{M}_j|\mathbf{X}_n)}{\mathfrak{p}_i/\mathfrak{p}_j} = \log \frac{f_{i,n}(\mathbf{X}_n)}{f_{j,n}(\mathbf{X}_n)}. \tag{2.4}$$

The Bayes factor measures change in model selection odds between $\mathcal{M}_i$ and $\mathcal{M}_j$ when observing $\mathbf{X}_n$. We relatively prefer $\mathcal{M}_i$ to $\mathcal{M}_j$ if $\log \mathrm{BF}_n(i, j) > 0$, and vice versa. A selected model via the Bayes factor minimizes the total error rates compounding false-positive and false-negative probabilities, while, different from the AIC, it has no theoretical implication for predictive performance of the selected model. For a more detailed account of the philosophy of the Bayes factor, we refer to [29].

As was explained in [31], we have yet another interpretation based on the Kullback–Leibler (KL) divergence between the true distribution g_n and the mth quasi-marginal distribution $f_{m,n}$:

$$\begin{aligned} \mathrm{KL}(f_{m,n}; g_n) &:= -\int \left(\log \frac{f_{m,n}(x)}{g_n(x)}\right) g_n(x)\mu_n(dx) \\ &= \int \big\{\log g_n(x)\big\} g_n(x)\mu_n(dx) - \int \big\{\log f_{m,n}(x)\big\} g_n(x)\mu_n(dx); \end{aligned} \tag{2.5}$$

recall that in the classical AIC methodology we instead look at $\mathrm{KL}\{f_{m,n}(\cdot; \hat{\theta}_{m,n}); g_n\}$ where $\hat{\theta}_{m,n} = \hat{\theta}_{m,n}(\tilde{\mathbf{X}}_n)$ denotes the MLE in the mth correctly specified model, constructed from an i.i.d. copy $\tilde{\mathbf{X}}_n$ of $\mathbf{X}_n$. Based on (2.5), we choose a relatively optimal one among $\mathcal{M}_1, \ldots, \mathcal{M}_M$, the model index of which equals

$$\operatorname*{argmin}_{m \le M} \mathrm{KL}(f_{m,n}; g_n) = \operatorname*{argmax}_{m \le M} \int \big\{\log f_{m,n}(x)\big\} g_n(x)\mu_n(dx).$$

Comparison of $f_{i,n}$ and $f_{j,n}$ is equivalent to looking at the sign of

$$\begin{aligned} \mathrm{KL}(f_{j,n}; g_n) - \mathrm{KL}(f_{i,n}; g_n) &= \int \log\left(\frac{f_{i,n}(x)}{f_{j,n}(x)}\right) g_n(x)\mu_n(dx) \\ &= \mathbb{E}\left\{\log\left(\frac{f_{i,n}(\mathbf{X}_n)}{f_{j,n}(\mathbf{X}_n)}\right)\right\}. \end{aligned} \tag{2.6}$$

As was noted in [31], it is important to notice that this reasoning remains valid even when any of candidate models does not coincide with the true model. We also refer to [21] for another Bayesian variable selection device based on the KL projection.

We will introduce a set of regularity conditions under which explicit statistics $\mathrm{QBIC}_n^{\sharp,i}$ for each model $\mathcal{M}_i$, $i = 1, \ldots, M$ (see (3.7) below) satisfy the stochastic expansion

$$\log \mathrm{BF}_n(i, j) = \frac{1}{2}\big(\mathrm{QBIC}_n^{\sharp,j} - \mathrm{QBIC}_n^{\sharp,i}\big) + o_p(1).$$

In the classical treatment originating [35], the almost-sure expansion was relevant; see also Remark 3.5.

2.3. Expected logarithmic Bayes factor

Comparing (2.6) with (2.4), we see that the expected Bayes factor is directly related to the KL divergence difference:

$$\mathbb{E}\big\{\log \mathrm{BF}_n(i, j)\big\} = \mathrm{KL}(f_{j,n}; g_n) - \mathrm{KL}(f_{i,n}; g_n).$$

For Bayesian model comparison in our model setting, we wish to estimate the quantity $\mathbb{E}\{\log \mathrm{BF}_n(i, j)\}$ for each pair (i, j). In Section 3.2, we will derive statistics $\mathrm{QBIC}_n^{\sharp,1}, \ldots, \mathrm{QBIC}_n^{\sharp,M}$ such that for each $i, j \in \{1, \ldots, M\}$ with $i \neq j$:

$$\mathbb{E}\left(\left|\log \mathrm{BF}_n(i, j) - \frac{1}{2}\big(\mathrm{QBIC}_n^{\sharp,j} - \mathrm{QBIC}_n^{\sharp,i}\big)\right|\right) = o(1). \tag{2.7}$$

In particular, it follows that the statistics $(\mathrm{QBIC}_n^{\sharp,j} - \mathrm{QBIC}_n^{\sharp,i})/2$ serves as an asymptotically unbiased estimator of the expected Bayes factor:

$$\mathbb{E}\left(\mathbb{E}\big\{\log \mathrm{BF}_n(i, j)\big\} - \frac{1}{2}\big(\mathrm{QBIC}_n^{\sharp,j} - \mathrm{QBIC}_n^{\sharp,i}\big)\right) = o(1),$$

or, of the raw (random) Bayes factor:

$$\mathbb{E}\left(\log \mathrm{BF}_n(i, j) - \frac{1}{2}\big(\mathrm{QBIC}_n^{\sharp,j} - \mathrm{QBIC}_n^{\sharp,i}\big)\right) = o(1).$$

Obviously it suffices for (2.7) to show that

$$\mathbb{E}\big\{\big|\mathrm{QBIC}_n^{\sharp,i} - \big(-2\log f_{i,n}(\mathbf{X}_n)\big)\big|\big\} = o(1)$$

for each i (see Theorem 3.15). We are thus led to the basic rule about an optimal model $\mathcal{M}_{m_0}$ in the sense of approximate Bayesian model description:

$$m_0 \in \operatorname*{argmin}_{1 \le i \le M} \mathrm{QBIC}_n^{\sharp,i}.$$

Remark 2.1. To perform a model comparison based on Bayesian prediction, we should replace the marginal likelihood $f_{m,n}$ in (2.5) by a Bayesian predictive model and also the "$g_n\,d\mu_n$"-integral by suitable one. We refer to [46] for an extensive review of Bayesian prediction, and also to [37] for a study in this direction for the LAMN models.

Remark 2.2. While we here focus on the finite model comparison among M candidates, it would be possible to consider a *continuum* of models, say $(\mathcal{M}_\lambda)_{\in\Lambda}$ for some (possibly uncountable) model-index set Λ. This is relevant when considering continuous fine-tuning in regularization methods, for example, [3]. Although we do not treat such a setting here, it is readily expected that our claims remain valid in an analogous form.

3. Quasi-Bayesian information criterion

We here focus on a single model $\mathcal{M}_m$ and consider the asymptotic expansion related to $\mathbb{H}_{m,n}$. From now on, we will omit the model index "m" from the notation, simply denoting the prior density and the quasi-log likelihood by $\pi_n(\theta)$ and $\mathbb{H}_n(\theta) = \mathbb{H}_n(\mathbf{X}_n|\theta)$, respectively. The parameter $\theta \in \Theta \subset \mathbb{R}^p$ is graded into K parts, say

$$\theta = (\theta_1, \ldots, \theta_K), \qquad \theta_k \in \mathbb{R}^{p_k}.$$

Here we wrote $p = \sum_{k=1}^K p_k$. Let $\theta_0 \in \Theta$ be a constant, which will serve as the optimal parameter defined in Section 5. We are thinking of situations where the contrast function $\mathbb{H}_n$ provides an M-estimator $\hat{\theta}_n$ such that the $A_n(\theta_0)(\hat{\theta}_n - \theta_0)$ tends in distribution to a non-trivial asymptotic distribution. The rate matrix $A_n(\theta_0)$ is of the form (2.1) satisfying (2.2):

$$A_n(\theta_0) = \operatorname{diag}\bigl(a_{1,n}(\theta_0) I_{p_1}, \ldots, a_{K,n}(\theta_0) I_{p_K}\bigr),$$

where positive decreasing sequences $a_{k,n}(\theta_0)$ such that $a_{k,n}^{-1}(\theta_0)/a_{l,n}^{-1}(\theta_0) \to 0$ for $k > l$; we will assume that $A_n(\hat{\theta}_n) - A_n(\theta_0) \xrightarrow{P} 0$ (see Theorem 3.7(ii)), so that $\log|A_n(\hat{\theta}_n)| = \sum_{k=1}^K p_k \log a_{k,n}(\hat{\theta}_n)$; here and in what follows, with a slight abuse of notation we often write $|A|$ instead of $\det(A)$ for a square matrix A. The statistical random field associated with $\mathbb{H}_n$ is given by

$$\mathbb{Z}_n(u) = \mathbb{Z}_n(u;\theta_0) := \exp\bigl\{\mathbb{H}_n\bigl(\theta_0 + A_n(\theta_0)u\bigr) - \mathbb{H}_n(\theta_0)\bigr\}, \tag{3.1}$$

which is defined on the admissible domain

$$\mathbb{U}_n(\theta_0) := \bigl\{u \in \mathbb{R}^p;\, \theta_0 + A_n(\theta_0)u \in \Theta\bigr\}.$$

The objective here is to deduce the asymptotic behavior of the marginal quasi-log likelihood function

$$\log\biggl(\int_\Theta \exp\bigl\{\mathbb{H}_n(\theta)\bigr\}\pi_n(\theta)\,d\theta\biggr),$$

and then derive an extension of the classical BIC.

3.1. Stochastic expansion

We begin with the stochastic expansion of the marginal quasi-log likelihood function. The polynomial type large deviation inequality (PLDI, [47]), mentioned in the introduction, is a powerful tool for ensuring the $L^q(\mathbb{P})$-boundedness of scaled M- and Bayes estimators stemming from the quasi-likelihood $\mathbb{H}_n$. As seen below, the PLDI argument can be effectively used also to verify the key Laplace-approximation type argument in a unified manner.

Write $\partial_\theta = \partial/\partial\theta$, and denote by θ^j the jth element of θ and by $A_{n,ii}(\theta_0)$ the (i,i)th element of $A_n(\theta_0)$ (i.e., $A_{n,ii}(\theta_0) = a_{j,n}(\theta_0)$ for some $j \in \{1,\dots,K\}$).

Assumption 3.1. $\mathbb{H}_n(\theta)$ is of class $\mathcal{C}^3(\Theta)$ and satisfies the following conditions:

(i) $\Delta_n = \Delta_n(\theta_0) := A_n(\theta_0)\partial_\theta \mathbb{H}_n(\theta_0) = O_p(1)$;
(ii) $\Gamma_n = \Gamma_n(\theta_0) := -A_n(\theta_0)\partial_\theta^2 \mathbb{H}_n(\theta_0) A_n(\theta_0) = \Gamma_0 + o_p(1)$ where $\mathbb{P}(\Gamma_0 > 0) = 1$;
(iii) $\max_{i,j,k\in\{1,\dots,p\}} \sup_\theta |A_{n,ii}(\theta_0)A_{n,jj}(\theta_0)A_{n,kk}(\theta_0)\partial_{\theta^i}\partial_{\theta^j}\partial_{\theta^k}\mathbb{H}(\theta)| = o_p(1)$.

Assumption 3.1 implicitly sets down the optimal value θ_0; of course, as in the usual M-estimation theory (e.g. [45], Chapter 5) it is possible to put more specific conditions in terms of the uniform-in-θ limits of suitable scaled quasi-log likelihoods function, but we omit them. The quadratic form Γ_0 is the asymptotic quasi-Fisher information matrix, which may be random. A truly random example is the volatility-parameter estimation of a continuous semimartingale (see Section 6.2). In particular, Assumption 3.1 leads to the LAQ approximation of $\log \mathbb{Z}_n$:

$$\sup_{u\in A}\left|\log \mathbb{Z}_n(u) - \left(\Delta_n[u] - \frac{1}{2}\Gamma_0[u,u]\right)\right| = o_p(1) \tag{3.2}$$

for each compact set $A \subset \mathbb{R}^p$.

Assumption 3.2. The prior density π_n satisfies the following:

(i) $\pi_n(\theta_0) > 0$ for all n, and $\sup_n \sup_\theta \pi_n(\theta) < \infty$;
(ii) $\sup_{|u|<M} |\pi_n(\theta_0 + A_n(\theta_0)u) - \pi_n(\theta_0)| \to 0$ as $n \to \infty$ for each $M > 0$.

Assumption 3.3. For any $\varepsilon > 0$ there exist $M > 0$ and $N \in \mathbb{N}$ such that

$$\sup_{n\geq N} \mathbb{P}\left(\int_{\mathbb{U}_n(\theta_0)\cap\{|u|\geq M\}} \mathbb{Z}_n(u)\,du > \varepsilon\right) < \varepsilon.$$

Thanks to Assumption 3.3, we can consider general LAQ models in a unified manner. Let us mention some sufficient conditions for the key assumption Assumption 3.3. To this end, we need to introduce further notation. Write

$$\Gamma_n(\theta) = -A_n(\theta_0)\partial_\theta^2 \mathbb{H}_n(\theta) A_n(\theta_0),$$

and denote by $\lambda_{\min}(A)$ the smallest eigenvalues of a given matrix A. We write $\underline{\theta}_k = (\theta_1,\dots,\theta_k)$ and $\overline{\theta}_k = (\theta_k,\dots,\theta_K)$, with $\underline{\theta}_{k,0}$ and $\overline{\theta}_{k,0}$ in a similar manner. Let $u := (u_1,\dots,u_K) \in \mathbb{R}^{p_1} \times$

$\cdots \times \mathbb{R}^{p_K}$. The kth random field is defined by

$$\mathbb{Z}_n^k(u_k; \underline{\theta}_{k-1}, \theta_{k,0}, \overline{\theta}_{k+1}) = \exp\{\mathbb{H}_n(\underline{\theta}_{k-1}, \theta_{k,0} + a_{k,n}(\theta_0)u_k, \overline{\theta}_{k+1}) - \mathbb{H}_n(\underline{\theta}_{k-1}, \theta_{k,0}, \overline{\theta}_{k+1})\}.$$

The random fields $\mathbb{Z}_n^k$ is designed to focus on the kth-graded parameters, when we have more than one rate of convergence, that is, when $K \geq 2$ (we neglect symbols with index $K+1$ like $\overline{\theta}_{K+1}$ and ones with index 0 like $\underline{\theta}_0$).

Theorem 3.4. *Let Assumption 3.1 hold. Then, Assumption 3.3 follows if at least one of the following conditions holds*:

(i) *There exist constants $L > 1$ and $C_L > 0$ such that*

$$\sup_n \mathbb{P}\Big(\sup_{(u_k, \overline{\theta}_{k+1}) \in \{|u_k| \geq r\} \times \prod_{j=k+1}^K \Theta_j} \mathbb{Z}_n^k(u_k; \underline{\theta}_{k-1,0}, \theta_{k,0}, \overline{\theta}_{k+1}) \geq e^{-r} \Big) \leq \frac{C_L}{r^L} \tag{3.3}$$

for $r > 0$ and $k = 1, \ldots, K$;

(ii) *We have*

$$\limsup_{\delta \to 0} \limsup_{n \to \infty} \mathbb{P}\Big(\inf_{\theta \in \Theta} \lambda_{\min}\big(\Gamma_n(\theta)\big) < \delta \Big) = 0. \tag{3.4}$$

The proof of Theorem 3.4(i) can be found in [47], Theorem 6; what is important in the proof is that, inside the probability, we are bounding the supremum of the random field from below by the quickly decreasing "e^{-r}" (see also Remark 3.6). The proof of Theorem 3.4(ii) is given in Section 7.1; the condition (3.4) is a sort of *global* non-degeneracy condition of the asymptotic information matrix. Since we are dealing with the integral-type functional, the non-degeneracy condition may not be *local* in u.

Remark 3.5. As already mentioned in (3.2), Assumption 3.1 ensures the LAQ structure of the random field $\mathbb{Z}_n$, so that, in the verification of Assumption 3.3 the uniform-in-θ asymptotic non-degeneracy of the quasi-observed-information matrix $\Gamma_n(\theta)$ plays a crucial role. In the literature, among others: the original [35] considered genuinely Bayesian situation, where data was regarded as *non-random* quantities; [9] proved the key Laplace approximation for the marginal log-likelihood under the assumption that the minimum eigenvalue of the observed information matrix is *almost surely* bounded away from zero and infinity; [31] considered the quasi-likelihood estimation in the generalized linear models where the observed information matrix is *non-random*. When attempting to directly follow such routes, in general we would need to impose almost-sure type condition instead of (3.4), such as the existence of $\delta > 0$ for which

$$\mathbb{P}\Big(\limsup_{n \to \infty} \inf_{\theta \in \Theta} \lambda_{\min}\big(\Gamma_n(\theta)\big) < \delta \Big) = 0.$$

Remark 3.6. For reference, let us mention the tail-probability estimate about the normalized estimator

$$\hat{u}_n := A_n^{-1}(\theta_0)(\hat{\theta}_n - \theta_0).$$

We can consider the stepwise probability estimates of $\log \mathbb{Z}_n(u)$ through successive applications of the PLDI result [47]. Namely, the following statement holds for a constant $L > 0$: *if there exists a universal constant $C_L > 0$ such that*

$$\sup_n \mathbb{P}\Big(\sup_{(u_k, \overline{\theta}_{k+1}) \in \{|u_k| \geq r\} \times \prod_{j=k+1}^K \Theta_j} \mathbb{Z}_n^k(u_k; \hat{\underline{\theta}}_{k-1}, \theta_{k,0}, \overline{\theta}_{k+1}) \geq 1 \Big) \leq \frac{C_L}{r^L} \tag{3.5}$$

for all $r > 0$ and $k = 1, \ldots, K$, then $\hat{u}_n$ satisfies the estimate

$$\sup_n \mathbb{P}\big(|\hat{u}_n| \geq r\big) \leq \frac{C_L}{r^L}, \qquad r > 0, \tag{3.6}$$

which implies the tightness of $(\hat{u}_n)$, hence in particular $\hat{\theta}_n \xrightarrow{P} \theta_0$. As in [47], Proposition 2, we can derive (3.6) as follows: we have

$$\mathbb{P}\big(|\hat{u}_n| \geq r\big) \leq \sum_{k=1}^K \mathbb{P}\bigg(\big|a_{k,n}^{-1}(\theta_0)(\hat{\theta}_{k,n} - \theta_{k,0})\big| \geq \frac{r}{K} \bigg),$$

and each $\mathbb{P}(|a_{k,n}^{-1}(\theta_0)(\hat{\theta}_{k,n} - \theta_{k,0})| \geq \frac{r}{K})$ can be bounded by

$$\begin{aligned} &\mathbb{P}\Big(\sup_{\frac{r}{K} \leq |u_k|} \big\{ \mathbb{H}_n\big(\hat{\underline{\theta}}_{k-1}, \theta_{k,0} + a_{k,n}(\theta_0)u_k, \overline{\hat{\theta}}_{k+1}\big) - \mathbb{H}_n(\hat{\underline{\theta}}_{k-1}, \theta_{k,0}, \overline{\hat{\theta}}_{k+1}) \big\} \geq 0 \Big) \\ &\quad \leq \mathbb{P}\Big(\sup_{(u_k, \overline{\theta}_{k+1}) \in \{\frac{r}{K} \leq |u_k|\} \times \prod_{j=k+1}^K \Theta_j} \mathbb{Z}_n^k(u_k; \hat{\underline{\theta}}_{k-1}, \theta_{k,0}, \overline{\theta}_{k+1}) \geq 1 \Big) \leq \frac{C_L}{r^L} K^L \end{aligned}$$

for all $n > 0$ and $r > 0$. Sufficient conditions for the PLDI (3.3) and (3.5) to hold can be found in [47], Theorem 2. The asymptotic mixed normality of $\hat{u}_n$ then follows from a functional weak convergence of $\mathbb{Z}_n$ on compact sets to a suitable exponential quadratic random field, which often follows through a stable convergence in law of the random linear form $\Delta_n = A_n(\theta_0)\partial_\theta \mathbb{H}_n(\theta_0)$.

Now we are in position to state the stochastic expansion result.

Theorem 3.7. *Suppose that Assumptions 3.1 to 3.3 are satisfied and that $\hat{\theta}_n \xrightarrow{P} \theta_0$.*

(i) *We have the asymptotic expansion*

$$\begin{aligned} \log\bigg(\int_\Theta \exp\big\{\mathbb{H}_n(\theta)\big\} \pi_n(\theta)\, d\theta \bigg) &= \mathbb{H}_n(\theta_0) + \sum_{k=1}^K p_k \log a_{k,n}(\theta_0) - \frac{1}{2}\log|\Gamma_0| \\ &\quad + \frac{p}{2}\log 2\pi + \frac{1}{2}\big\|\Gamma_0^{-\frac{1}{2}}\Delta_n\big\|^2 + \log \pi_n(\theta_0) + o_p(1). \end{aligned}$$

(ii) *If further* $\log a_{k,n}(\hat{\theta}_n) = \log a_{k,n}(\theta_0) + o_p(1)$ *and* $\log \pi_n(\hat{\theta}_n) = \log \pi_n(\theta_0) + o_p(1)$, *then*

$$\begin{aligned}
&\log\left(\int_\Theta \exp\{\mathbb{H}_n(\theta)\}\pi_n(\theta)\,d\theta\right) \\
&\quad = \mathbb{H}_n(\hat{\theta}_n) + \sum_{k=1}^K p_k \log a_{k,n}(\hat{\theta}_n) + \frac{p}{2}\log 2\pi + \log \pi_n(\hat{\theta}_n) \\
&\qquad - \frac{1}{2}\log\left|-A_n(\hat{\theta}_n)\partial_\theta^2 \mathbb{H}_n(\hat{\theta}_n) A_n(\hat{\theta}_n)\right| + o_p(1) \\
&\quad = \mathbb{H}_n(\hat{\theta}_n) + \frac{p}{2}\log 2\pi - \frac{1}{2}\log\left|-\partial_\theta^2 \mathbb{H}_n(\hat{\theta}_n)\right| + \log \pi_n(\hat{\theta}_n) + o_p(1).
\end{aligned}$$

It follows from Theorem 3.7(ii) that the statistics

$$\mathrm{QBIC}_n^\sharp := -2\mathbb{H}_n(\hat{\theta}_n) + \log\left|-\partial_\theta^2 \mathbb{H}_n(\hat{\theta}_n)\right| - 2\log \pi_n(\hat{\theta}_n) - p\log 2\pi \tag{3.7}$$

is a consistent estimator of the marginal quasi-log likelihood function multiplied by "−2". Then, ignoring the $O_p(1)$ parts, we define the *quasi-Bayesian information criterion* (*QBIC*) by

$$\mathrm{QBIC}_n = \mathrm{QBIC}_n(\mathbf{X}_n) := -2\mathbb{H}_n(\hat{\theta}_n) + \log\left|-\partial_\theta^2 \mathbb{H}_n(\hat{\theta}_n)\right|. \tag{3.8}$$

As long as π_n is not so dominant and n is moderately large, using QBIC_n instead of $\mathrm{QBIC}_n^\sharp$ would be enough in practice. We compute QBIC for each candidate model, say $\mathrm{QBIC}_n^{(1)}, \ldots, \mathrm{QBIC}_n^{(M)}$, and then define the best model $\mathcal{M}_{m_0}$ in the sense of approximate Bayesian model description:

$$\{m_0\} = \operatorname*{argmin}_{1\le m\le M} \mathrm{QBIC}_n^{(m)},$$

the uniqueness being implicitly assumed. In view of Assumption 3.1, we see that

$$\mathrm{QBIC}_n = -2\mathbb{H}_n(\hat{\theta}_n) + 2\sum_{k=1}^K p_k \log a_{k,n}^{-1}(\hat{\theta}_n) + O_p(1). \tag{3.9}$$

Since the second term in the right-hand side diverges in probability, we could more simply define QBIC to be the the sum of the first two terms in the right-hand side of (3.9). We may thus define Schwarz's BIC in our context by

$$\mathrm{BIC}_n = -2\mathbb{H}_n(\hat{\theta}_n) + 2\sum_{k=1}^K p_k \log a_{k,n}^{-1}(\hat{\theta}_n). \tag{3.10}$$

Note that in the classical case of single $\sqrt{n}$-scaling (3.10) reduces to the familiar form

$$\mathrm{BIC}_n = -2\mathbb{H}_n(\hat{\theta}_n) + p\log n. \tag{3.11}$$

The statistics QBIC_n thus provides us with a far-reaching extension of derivation machinery of the classical BIC.

Although the original definition (3.8) has higher computational load than (3.10), it enables us to incorporate a model-complexity bias correction taking the volume of observed information into account. In particular, to reflect data information for dependent-data models, (3.8) would be more suitable than (3.10) whose bias correction is only based on the rate of convergence.

Remark 3.8. Making use of the observed information matrix (3.8) for regularization has been already mentioned in the literature; for example, [5,23], and [36] contain such statistics for some variants of the AIC statistics. Further, it is worth mentioning that using the observed-information is a right way for some non-stationary models (see [25]).

Remark 3.9. At the beginning, the prior model selection probabilities $\mathfrak{p}_1, \ldots, \mathfrak{p}_M$ are to be set in a subjective manner. As usual, using the QBIC of the candidate models we may estimate the posterior model selection probabilities in the data-driven manner through the quantities

$$\frac{\mathfrak{p}_m \exp\{-\mathrm{QBIC}_n^{(m)}/2\}}{\sum_{l=1}^{M} \mathfrak{p}_l \exp\{-\mathrm{QBIC}_n^{(l)}/2\}}, \qquad m = 1, \ldots, M,$$

or those with QBIC replaced by BIC, where $\mathrm{QBIC}_n^{(m)}$ denotes the QBIC of the mth model.

Remark 3.10 (Variants of QBIC). In practice, we may conveniently consider several variants of the QBIC (3.8). When Γ_0 takes the form $\Gamma_0 = \mathrm{diag}(\Gamma_{10}, \ldots, \Gamma_{K0})$ with each $\Gamma_{k0} \in \mathbb{R}^{p_k} \otimes \mathbb{R}^{p_k}$ being a.s. positive definite, we may slightly simplify the form of the QBIC as follows. We can see that under Assumption 3.1,

$$-a_{k,n}(\theta_0) a_{l,n}(\theta_0) \partial_{\theta_k} \partial_{\theta_l} \mathbb{H}_n(\hat{\theta}_n) = o_p(1), \qquad k \neq l.$$

Taking logarithmic determinant of a positive definite matrix is continuous, the asymptotic expansion in Theorem 3.7(ii) becomes

$$\log\left(\int_{\Theta} \exp\{\mathbb{H}_n(\theta)\} \pi_n(\theta)\, d\theta \right) = \mathbb{H}_n(\hat{\theta}_n) - \frac{1}{2} \sum_{k=1}^{K} \log\left| -\partial_{\theta_k}^2 \mathbb{H}_n(\hat{\theta}_n) \right| + O_p(1),$$

resulting in the QBIC of the form

$$-2\mathbb{H}_n(\hat{\theta}_n) + \sum_{k=1}^{K} \log\left| -\partial_{\theta_k}^2 \mathbb{H}_n(\hat{\theta}_n) \right|. \tag{3.12}$$

In particular, this is the case if $A_n^{-1}(\theta_0)(\hat{\theta}_n - \theta_0)$ is asymptotically mixed normally distributed, with a block diagonal asymptotic (random) covariance matrix $\Sigma_0 = \mathrm{diag}(\Sigma_{10}, \ldots, \Sigma_{K0})$ where each $\Sigma_{k0} \in \mathbb{R}^{p_k} \otimes \mathbb{R}^{p_k}$ is a.s. positive definite. We will deal with such an example in Section 4.2.

We may also consider finite-sample manipulations of QBIC without breaking its asymptotic behavior. For example, the problem caused by $|-\partial_\theta^2 \mathbb{H}_n(\hat{\theta}_n)| \le 0$ can be avoided by using

$$-2\mathbb{H}_n(\hat{\theta}_n) + I\{|-\partial_\theta^2 \mathbb{H}_n(\hat{\theta}_n)| > 0\} \log|-\partial_\theta^2 \mathbb{H}_n(\hat{\theta}_n)| + I\{|-\partial_\theta^2 \mathbb{H}_n(\hat{\theta}_n)| \le 0\} \sum_{k=1}^{K} p_k \log\bigl(a_{k,n}^{-2}(\hat{\theta}_{k,n})\bigr)$$

instead of (3.8); obviously, the difference between this quantity and QBIC_n is of $o_p(1)$. Further, we may use *any* $\hat{\Gamma}_n$ such that $\hat{\Gamma}_n \xrightarrow{P} \Gamma_0$:

$$-2\mathbb{H}_n(\hat{\theta}_n) - 2\log|A_n(\hat{\theta}_n)| + \log|\hat{\Gamma}_n|,$$

which would be convenient if $\hat{\Gamma}_n$ is more likely to be stable than $-A_n(\hat{\theta}_n)\partial_\theta^2 \mathbb{H}_n(\hat{\theta}_n) A_n(\hat{\theta}_n)$; for example, if we beforehand know the specific form of $\Gamma_0 = \Gamma_0(\theta)$, then it would be numerically more stable to use $\Gamma_0(\hat{\theta}_n)$ instead of $-A_n(\hat{\theta}_n)\partial_\theta^2 \mathbb{H}_n(\hat{\theta}_n) A_n(\hat{\theta}_n)$.

3.2. Convergence of the expected values

From the frequentist point of view where $\mathbf{X}_n$ is regarded as a random element, it is desirable to verify the convergence of *expected* marginal quasi-log likelihood, which follows from the asymptotic uniform integrability of the sequence

$$\left\{\left|-2\log\left(\int_\Theta \exp\{\mathbb{H}_n(\theta)\}\pi_n(\theta)\,d\theta\right) - \mathrm{QBIC}_n^\sharp\right|\right\}_n.$$

In particular, $\mathrm{QBIC}_n^\sharp$ is then an asymptotically unbiased estimator of the expected logarithmic Bayes factor; see Section 2.3, in particular (2.7).

Let us recall the notation $\Delta_n = A_n(\theta_0)\partial_\theta \mathbb{H}_n(\theta_0)$ and $\Gamma_n(\theta) = -A_n(\theta_0)\partial_\theta^2 \mathbb{H}_n(\theta) A_n(\theta_0)$. We replace Assumptions 3.1 to 3.3 as follows.

Assumption 3.11. The random function $\mathbb{H}_n$ is of class $\mathcal{C}^3(\Theta)$ a.s. and for every $r > 0$

$$\sup_n \mathbb{E}\left(|\Delta_n|^r + \sup_\theta |\Gamma_n(\theta)|^r + \sum_{i=1}^{p} \sup_\theta |A_n(\theta_0)\partial_{\theta^i}\partial_\theta^2 \mathbb{H}_n(\theta) A_n(\theta_0)|^r\right) < \infty.$$

Assumption 3.12. In addition to Assumption 3.2, we have $0 < \inf_{n,\theta} \pi_n(\theta) \le \sup_{n,\theta} \pi_n(\theta) < \infty$.

Assumption 3.13. There exists an a.s. positive definite random matrix Γ_0 such that $\Gamma_n(\theta_0) \xrightarrow{P} \Gamma_0$, and for some $q > 3p$ we have

$$\limsup_n \mathbb{E}\Bigl(\sup_\theta \lambda_{\min}^{-q}\bigl(\Gamma_n(\theta)\bigr)\Bigr) < \infty.$$

The moment bounds in Assumption 3.13 was studied in [10] and [11] for some time series models, with a view toward prediction. The integrability in Assumption 3.13 is related to the key index χ_0 of [43] in case of volatility estimation of a continuous Itô process.

Under Assumptions 3.11 and 3.13, we have $\lambda_{\min}^{-q}(\Gamma_n(\theta_0)) \xrightarrow{P} \lambda_{\min}^{-q}(\Gamma_0)$ by the continuous mapping theorem, and also $\lambda_{\min}^{-1}(\Gamma_0) \in L^q(\mathbb{P})$ as well as $\Gamma_0 \in \bigcap_{r>0} L^r(\mathbb{P})$.

Finally, we impose the boundedness of moments of the normalized estimator; see Remark 3.6.

Assumption 3.14. $\sup_n \mathbb{E}(|\hat{u}_n|^r) < \infty$ for some $r > 3$.

We can now state the $L^1(\mathbb{P})$-converge result.

Theorem 3.15. *If Assumptions 3.11 to 3.14 hold, then we have*

$$\lim_{n\to\infty} \mathbb{E}\left\{\left|-2\log\left(\int_\Theta \exp\{\mathbb{H}_n(\theta)\}\pi_n(\theta)\,d\theta\right) - \mathrm{QBIC}_n^\sharp\right|\right\} = 0.$$

In particular, $\mathrm{QBIC}_n^\sharp$ *is an asymptotically unbiased estimator of the logarithmic quasi-marginal likelihood.*

4. Gaussian quasi-likelihood

This section is devoted to the Gaussian quasi-likelihood.

4.1. General framework

A general setting for the Gaussian quasi-likelihood estimation is described as follows. Let $\mathbf{X}_n = (X_{n,j})_{j=0}^n = (X_{n,0}, \ldots, X_{n,n})$ be an array of random variables, where $X_{n,j} \in \mathbb{R}$ for brevity. Let $\mathcal{F}_{n,j} := \sigma(X_{n,k}; k \le j)$ denote the σ-field representing the data information at stage j when the total number of data is n. The Gaussian quasi-likelihood (in the univariate case) is constructed as if the conditional distribution of $X_{n,j}$ given past information $\mathcal{F}_{n,j-1}$ is Gaussian, say

$$\mathcal{L}(X_{n,j}|X_{n,0}, \ldots, X_{n,j-1}) \approx N\big(\mu_{n,j-1}(\theta), \sigma_{n,j-1}(\theta)\big),$$

where $\mu_{n,j-1}$ and $\sigma_{n,j-1}$ are $\mathcal{F}_{n,j-1}$-measurable (predictable) random function on Θ; typically,

$$\mu_{n,j-1}(\theta) = \mathbb{E}(X_{n,j}|\mathcal{F}_{n,j-1}), \qquad \sigma_{n,j-1}(\theta) = \mathrm{var}(X_{n,j}|\mathcal{F}_{n,j-1}),$$

where the conditional expectation and variance are taken under the image measure of $\mathbf{X}_n$ associated with the parameter value θ. In what follows, we will suppress the subscript "n".

Because the quasi-likelihood is given by

$$\begin{aligned}\theta \mapsto &\sum_{j=1}^n \log \frac{1}{\sqrt{2\pi\sigma_{j-1}^2(\theta)}} \exp\left\{-\frac{1}{2\sigma_{j-1}^2(\theta)}\big(X_j - \mu_{j-1}(\theta)\big)^2\right\} \\ &= (\text{const.}) + \left[-\frac{1}{2}\sum_{j=1}^n \left\{\log\sigma_{j-1}^2(\theta) + \frac{(X_j - \mu_{j-1}(\theta))^2}{\sigma_{j-1}^2(\theta)}\right\}\right],\end{aligned}$$

we may define the Gaussian quasi-likelihood function by

$$\mathbb{H}_n(\theta) = -\frac{1}{2}\sum_{j=1}^{n}\left\{\log\sigma_{j-1}^2(\theta) + \frac{(X_j - \mu_{j-1}(\theta))^2}{\sigma_{j-1}^2(\theta)}\right\}.$$

Then, supposing that $\mathbb{H}_n$ and its partial derivatives can be continuously extended to the boundary $\partial\Theta$, we define the Gaussian QMLE (GQMLE) by any maximizer of $\mathbb{H}_n$ over $\bar{\Theta}$.

The Gaussian quasi-likelihood is designed to fit not full joint distribution but only conditional-mean and conditional-covariance structures. The simplest case is the location-parameter estimation by the sample mean in the i.i.d.-data setting, where $\sigma_{j-1}^2(\theta) \equiv 1$ (set for brevity) and $\mu_{j-1}(\theta) = \theta$, namely the least-squares estimation without "full" specification of the underlying population distribution. Although the GQMLE is not (possibly far from being) asymptotically efficient when the model is misspecified and/or the conditional distribution is deviating from being normal, the GQMLE quite often exhibits asymptotic (mixed-)normality under appropriate conditions.

4.2. Ergodic diffusion process

Let $\mathbf{X}_n = (X_{t_j})_{j=0}^n$ with $t_j = jh_n$, where h_n is the discretization step and $nh_n = T_n$ and X_t is a solution to the d-dimensional strictly stationary diffusion process defined by the stochastic differential equation

$$dX_t = a(X_t)\,dt + b(X_t)\,dw_t, \qquad t \in [0, T_n].$$

Here a is an $\mathbb{R}^d$-valued function defined on $\mathbb{R}^d$, b is an $\mathbb{R}^d \otimes \mathbb{R}^d$-valued function defined on $\mathbb{R}^d$, and w_t is an d-dimensional standard Wiener process. We assume that $T_n = nh_n \to \infty$ and $nh_n^2 \to 0$ as $n \to \infty$, and that for some positive constant ε_0, $nh_n \geq n^{\varepsilon_0}$ for every large n. Let us consider the following stochastic differential equation as a statistical model $\mathcal{M}_{m_1,m_2}$:

$$dX_t = a_{m_2}(X_t, \theta_{m_2})\,dt + b_{m_1}(X_t, \theta_{m_1})\,dw_t, \qquad t \in [0, T_n],\ X_0 = x_0, \tag{4.1}$$

where a_{m_2} is an $\mathbb{R}^d$-valued function defined on $\mathbb{R}^d \times \Theta_{m_2}$, b_{m_1} is an $\mathbb{R}^d \otimes \mathbb{R}^d$-valued function defined on $\mathbb{R}^d \times \Theta_{m_1}$ and $(m_1, m_2) \in \{1, \ldots, M_1\} \times \{1, \ldots, M_2\}$; namely, we consider $M_1 \times M_2$ models in total. In each model $\mathcal{M}_{m_1,m_2}$, the coefficients b_{m_1} and a_{m_2} are assumed to be known up to the finite-dimensional parameter $\theta_{m_1,m_2} := (\theta_{m_1}, \theta_{m_2}) \in \Theta_{m_1} \times \Theta_{m_2} \subset \mathbb{R}^{p_{m_1}} \times \mathbb{R}^{p_{m_2}}$. We focus on the case of correctly specified parametric coefficients: we assume that for each m there exists the true value $(\theta_{m_1,0}, \theta_{m_2,0})$ for which $b_{m_1}(\cdot, \theta_{m_1,0}) = b(\cdot)$ and $a_{m_2}(\cdot, \theta_{m_2,0}) = a(\cdot)$.

Below, we omit the model index "m_1" and "m_2" from the notation. That is, the stochastic differential equation (4.1) is expressed by

$$dX_t = a(X_t, \theta_2)\,dt + b(X_t, \theta_1)\,dw_t, \qquad t \in [0, T_n],\ X_0 = x_0.$$

Let $B(x,\theta_1) := b(x,\theta_1)b'(x,\theta_1)$ and $\Delta_j X := X_{t_j} - X_{t_{j-1}}$. We consider the quasi-likelihood function based on the small-time Gaussian approximation:

$$\prod_{j=1}^{n}(2\pi h_n)^{-\frac{d}{2}}\big|B(X_{t_{j-1}},\theta_1)\big|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2h_n}B(X_{t_{j-1}},\theta_1)^{-1}\big[\big(\Delta_j X - h_n a(X_{t_{j-1}},\theta_2)\big)^{\otimes 2}\big]\right\},$$

where $x^{\otimes 2} := xx'$. Then, up to an additive constant common to all the candidate models, the quasi-log likelihood function is given by

$$\mathbb{H}_n(\theta) = -\frac{1}{2}\sum_{j=1}^{n}\left\{\log\big|B(X_{t_{j-1}},\theta_1)\big| + \frac{1}{h_n}B(X_{t_{j-1}},\theta_1)^{-1}\big[\big(\Delta_j X - h_n a(X_{t_{j-1}},\theta_2)\big)^{\otimes 2}\big]\right\}.$$

We set $A_n = A_n(\theta_0) := \operatorname{diag}(\frac{1}{\sqrt{n}}I_{p_1}, \frac{1}{\sqrt{nh_n}}I_{p_2})$ for the rate matrix.

We assume the following conditions ([47], Section 6):

Assumption 4.1. (i) For some constant C,

$$\sup_{\theta_2\in\Theta_2}\big|\partial_{\theta_2}^{i}a(x,\theta_2)\big| \le C\big(1+|x|\big)^{C} \qquad (0\le i\le 4),$$

$$\sup_{\theta_1\in\Theta_1}\big|\partial_{x}^{j}\partial_{\theta_1}^{i}b(x,\theta_1)\big| \le C\big(1+|x|\big)^{C} \qquad (0\le i\le 4, 0\le j\le 2).$$

(ii) $\inf_{|u|=1}\inf_{(x,\theta_1)} B(x,\theta_1)[u,u] > 0$.

(iii) There exists a constant C such that for every $x_1, x_2 \in \mathbb{R}^p$,

$$\sup_{\theta_2\in\Theta_2}\big|a(x_1,\theta_2)-a(x_2,\theta_2)\big| + \sup_{\theta_1\in\Theta_1}\big|b(x_1,\theta_1)-b(x_2,\theta_1)\big| \le C|x_1-x_2|.$$

(iv) $X_0 \in \bigcap_{p>0} L^p(\mathbb{P})$.

Assumption 4.2. For some constant $a > 0$,

$$\sup_{t\in\mathbb{R}_+}\sup_{\substack{A\in\sigma[X_r;r\le t]\\ B\in\sigma[X_r;r\ge t+h]}}\big|\mathbb{P}(A\cap B)-\mathbb{P}(A)\mathbb{P}(B)\big| \le a^{-1}e^{-ah} \qquad (h>0).$$

Assumption 4.2 ensures the ergodicity: there exists a unique invariant probability measure $\nu = \nu_{\theta_0}$ of X_t such that

$$\frac{1}{T}\int_0^T g(X_t)\,dt \xrightarrow{P} \int_{\mathbb{R}^d} g(x)\nu(dx), \qquad T\to\infty,$$

for any measurable function g of at most polynomial growth.

Assumption 4.3. There exists a positive constant $\chi > 0$ such that $\mathbb{Y}_{1,0}(\theta_1) \le -\chi|\theta_1 - \theta_{1,0}|^2$ for all $\theta_1 \in \Theta_1$, where

$$\mathbb{Y}_{1,0}(\theta_1) = -\frac{1}{2}\int_{\mathbb{R}^d}\left\{\operatorname{tr}\bigl(B(x,\theta_1)^{-1}B(x,\theta_{1,0}) - I_p\bigr) + \log\frac{|B(x,\theta_1)|}{|B(x,\theta_{1,0})|}\right\}\nu(dx).$$

Assumption 4.4. There exists a positive constant $\chi' > 0$ such that $\mathbb{Y}_{2,0}(\theta_2) \le -\chi'|\theta_2 - \theta_{2,0}|^2$ for all $\theta_2 \in \Theta_2$, where

$$\mathbb{Y}_{2,0}(\theta_2) = -\frac{1}{2}\int_{\mathbb{R}^d} B(x,\theta_{1,0})^{-1}\bigl[\bigl(a(x,\theta_2) - a(x,\theta_{2,0})\bigr)^{\otimes 2}\bigr]\nu(dx).$$

The second-order partial derivatives of $\mathbb{H}_n$ are given as follows: for $u_1 \in \mathbb{R}^{m_1}$ and $u_2 \in \mathbb{R}^{m_2}$,

$$\begin{aligned}
\partial_{\theta_1}^2\mathbb{H}_n(\theta_1,\theta_2)\bigl[u_1^{\otimes 2}\bigr] &= -\frac{1}{2}\sum_{j=1}^n\biggl\{\partial_{\theta_1}^2\log\frac{|B(X_{t_{j-1}},\theta_1)|}{|B(X_{t_{j-1}},\theta_{1,0})|}\bigl[u_1^{\otimes 2}\bigr] \\
&\qquad + \frac{1}{h_n}\partial_{\theta_1}^2 B(X_{t_{j-1}},\theta_1)^{-1}\bigl[u_1^{\otimes 2},\bigl(\Delta_j X - h_n a(X_{t_{j-1}},\theta_2)\bigr)^{\otimes 2}\bigr]\biggr\}, \\
\partial_{\theta_2}^2\mathbb{H}_n(\theta_1,\theta_2)\bigl[u_2^{\otimes 2}\bigr] &= -\sum_{j=1}^n B(X_{t_{j-1}},\theta_1)^{-1} \\
&\qquad \times\bigl\{\bigl[\partial_{\theta_2}a(X_{t_{j-1}},\theta_2)[u_2],\partial_{\theta_2}h_n a(X_{t_{j-1}},\theta_2)[u_2]\bigr] \\
&\qquad -\bigl[\partial_{\theta_2}^2 a(X_{t_{j-1}},\theta_2)\bigl[u_2^{\otimes 2}\bigr],\Delta_j X - h_n a(X_{t_{j-1}},\theta_2)\bigr]\bigr\}, \\
\partial_{\theta_1}\partial_{\theta_2}\mathbb{H}_n(\theta_1,\theta_2)[u_1,u_2] &= \sum_{j=1}^n \partial_{\theta_1}B(X_{t_{j-1}},\theta_1) \\
&\qquad \times\bigl[u_1,\partial_{\theta_2}a(X_{t_{j-1}},\theta_2)[u_2],\Delta_j X - h_n a(X_{t_{j-1}},\theta_2)\bigr].
\end{aligned}$$

Then, we obtain the corresponding QBIC as in the following theorem.

Theorem 4.5. *Suppose that Assumptions 4.1 to 4.4 are satisfied. Then, the assumptions in Theorem 3.7 are satisfied and the corresponding QBIC is given by*

$$\begin{aligned}
\mathrm{QBIC}_n = \sum_{j=1}^n\biggl\{\log\bigl|B(X_{t_{j-1}},\hat\theta_{1,n})\bigr| + \frac{1}{h_n}B(X_{t_{j-1}},\hat\theta_{1,n})^{-1}\bigl[\bigl(\Delta_j X - h_n a(X_{t_{j-1}},\hat\theta_{2,n})\bigr)^{\otimes 2}\bigr]\biggr\} \\
+\log\left|-\begin{pmatrix}\partial_{\theta_1}^2\mathbb{H}_n(\hat\theta_n) & \partial_{\theta_1}\partial_{\theta_2}\mathbb{H}_n(\hat\theta_n)\\ \partial_{\theta_1}\partial_{\theta_2}\mathbb{H}_n(\hat\theta_n) & \partial_{\theta_2}^2\mathbb{H}_n(\hat\theta_n)\end{pmatrix}\right|.
\end{aligned}$$

In the present case of ergodic diffusion process, the convergence in probability

$$\frac{1}{\sqrt{n^2h_n}}\partial_{\theta_1}\partial_{\theta_2}\mathbb{H}_n(\hat{\theta}_n)\xrightarrow{P}0\qquad(n\to\infty)\tag{4.2}$$

is satisfied, so that

$$\begin{aligned}\log\bigl|-A_n\partial_\theta^2\mathbb{H}_n(\hat{\theta}_n)A_n\bigr| &= \log\begin{vmatrix}-\frac{1}{n}\partial_{\theta_1}^2\mathbb{H}_n(\hat{\theta}_n) & -\frac{1}{\sqrt{n^2h_n}}\partial_{\theta_1}\partial_{\theta_2}\mathbb{H}_n(\hat{\theta}_n)\\ -\frac{1}{\sqrt{n^2h_n}}\partial_{\theta_1}\partial_{\theta_2}\mathbb{H}_n(\hat{\theta}_n)' & -\frac{1}{nh_n}\partial_{\theta_2}^2\mathbb{H}_n(\hat{\theta}_n)\end{vmatrix}\\ &= \log\begin{vmatrix}-\frac{1}{n}\partial_{\theta_1}^2\mathbb{H}_n(\hat{\theta}_n) & 0\\ 0 & -\frac{1}{nh_n}\partial_{\theta_2}^2\mathbb{H}_n(\hat{\theta}_n)\end{vmatrix}+o_p(1)\\ &= \log\bigl|A_n\,\mathrm{diag}\bigl(-\partial_{\theta_1}^2\mathbb{H}_n(\hat{\theta}_n),-\partial_{\theta_2}^2\mathbb{H}_n(\hat{\theta}_n)\bigr)A_n\bigr|+o_p(1).\end{aligned}$$

In the asymptotic framework, statistics $\hat{S}_n$ such that $\hat{S}_n$ is easier to compute and that $\hat{S}_n = \mathrm{QBIC}_n + O_p(1)$ may be used as a variant of QBIC_n; recall (3.8) and (3.9), and also Remark 3.10.

Theorem 4.6. *Assume that Assumptions 4.1 to 4.4 hold, then the difference between the statistics*

$$\begin{aligned}&\sum_{j=1}^n\Bigl\{\log\bigl|B(X_{t_{j-1}},\hat{\theta}_{1,n})\bigr|+\frac{1}{h_n}B(X_{t_{j-1}},\hat{\theta}_{1,n})^{-1}\bigl[\bigl(\Delta_jX-h_na(X_{t_{j-1}},\hat{\theta}_{2,n})\bigr)^{\otimes2}\bigr]\Bigr\}\\ &\quad+\log\bigl|-\partial_{\theta_1}^2\mathbb{H}_n(\hat{\theta}_n)\bigr|+\log\bigl|-\partial_{\theta_2}^2\mathbb{H}_n(\hat{\theta}_n)\bigr|\end{aligned}$$

and the QBIC_n *given in Theorem 4.5 is* $o_p(1)$.

The BIC corresponding to (3.10) takes the form

$$\begin{aligned}&\sum_{j=1}^n\Bigl\{\log\bigl|B(X_{t_{j-1}},\hat{\theta}_{1,n})\bigr|+\frac{1}{h_n}B(X_{t_{j-1}},\hat{\theta}_{1,n})^{-1}\bigl[\bigl(\Delta_jX-h_na(X_{t_{j-1}},\hat{\theta}_{2,n})\bigr)^{\otimes2}\bigr]\Bigr\}\\ &\quad+p_1\log n+p_2\log(nh_n)\\ &=\sum_{j=1}^n\Bigl\{\log\bigl|B(X_{t_{j-1}},\hat{\theta}_{1,n})\bigr|+\frac{1}{h_n}B(X_{t_{j-1}},\hat{\theta}_{1,n})^{-1}\bigl[\bigl(\Delta_jX-h_na(X_{t_{j-1}},\hat{\theta}_{2,n})\bigr)^{\otimes2}\bigr]\Bigr\}\\ &\quad+p\log n+p_2\log h_n,\end{aligned}$$

clarifying that the high frequency of data indeed has a significant impact through the term "$p_2\log h_n$" (diverging to $-\infty$): this point is quite important since one might wrongly set BIC-correction term to be "$p\log n$".

Remark 4.7. It follows from [24] that we have

$$\big(\sqrt{n}(\hat{\theta}_{1,n}-\theta_{1,0}),\sqrt{nh_n}(\hat{\theta}_{2,n}-\theta_{2,0})\big)\xrightarrow{\mathcal{L}} N_p\big(0,\operatorname{diag}\big(\Gamma_{1,0}(\theta_{1,0})^{-1},\Gamma_{2,0}(\theta_{1,0},\theta_{2,0})^{-1}\big)\big),$$

where

$$\begin{aligned}\Gamma_{1,0}(\theta_{1,0})\big[u_1^{\otimes 2}\big]&=\frac{1}{2}\int \operatorname{tr}\big\{B(x,\theta_{1,0})^{-1}\big(\partial_{\theta_1}B(x,\theta_{1,0})\big)\\&\qquad\times B(x,\theta_{1,0})^{-1}\big(\partial_{\theta_1}B(x,\theta_{1,0})\big)\big[u_1^{\otimes 2}\big]\big\}\nu(dx),\\ \Gamma_{2,0}(\theta_{1,0},\theta_{2,0})\big[u_2^{\otimes 2}\big]&=\int B(x,\theta_{1,0})^{-1}\big[\partial_{\theta_2}a(x,\theta_{2,0})[u_2],\partial_{\theta_2}a(x,\theta_{2,0})[u_2]\big]\nu(dx)\end{aligned}$$

for $u_1\in\mathbb{R}^{m_1}$, $u_2\in\mathbb{R}^{m_2}$. We know from [20] that this GQMLE is asymptotically efficient in the sense of Hajék–Le Cam.

Remark 4.8 (Partially convex example). Consider the following class of univariate stochastic differential equations:

$$dX_t=\left(\sum_{k=1}^{p_2}\theta_{2,k}a_k(X_t)\right)dt+\exp\left(\frac{1}{2}\sum_{\ell=1}^{p_1}\theta_{1,\ell}b_\ell(X_t)\right)dw_t,$$

where a_k and b_l are known functions. Write $\theta_1=(\theta_{1,1},\dots,\theta_{1,p_1})'$, $\theta_2=(\theta_{2,1},\dots,\theta_{2,p_2})'$, $a(x)=(a_1(x),\dots,a_{p_2}(x))'$, and $b(x)=(b_1(x),\dots,b_{p_1}(x))'$. Then the quasi-likelihood function is given by

$$\mathbb{H}_n(\theta_1,\theta_2)=-\frac{1}{2}\sum_{j=1}^{n}\left\{\theta_1'b(X_{t_{j-1}})+\frac{1}{h_n}\big(\Delta_jX-h_na(X_{t_{j-1}})'\theta_2\big)^2\exp\big\{-b(X_{t_{j-1}})'\theta_1\big\}\right\}.$$

The corresponding QBIC of Theorem 4.6 is given by

$$\begin{aligned}\mathrm{QBIC}_n=&-2\mathbb{H}_n(\hat{\theta}_{1,n},\hat{\theta}_{2,n})+\log\left|h_n\sum_{j=1}^{n}\exp\big\{-\hat{\theta}_{1,n}'b(X_{t_{j-1}})\big\}a^{\otimes 2}(X_{t_{j-1}})\right|\\&+\log\left|\frac{1}{2}\sum_{j=1}^{n}\frac{1}{h_n}\exp\big\{-\hat{\theta}_{1,n}'b(X_{t_{j-1}})\big\}\big(\Delta_jX-h_n\hat{\theta}_{2,n}'a(X_{t_{j-1}})\big)^2b^{\otimes 2}(X_{t_{j-1}})\right|.\end{aligned}$$

Several adaptive-estimation methodologies for general parametric ergodic diffusions have been developed in the literature; see [22] and [42] as well as the references therein. We here remark that, under mild conditions on the functions a and b, the optimization may be made even simpler and more efficient by using an adaptive estimation strategy. This is because of the convexity of each of the random functions to be optimized: specifically, we first get an estimate $\hat{\theta}_{1,n}$ of θ_1 as

an minimizer of the convex random function

$$\theta_1 \mapsto \sum_{j=1}^{n}\left(\theta_1' b(X_{t_{j-1}}) + \frac{1}{h_n}(\Delta_j X)^2 \exp\{-\theta_1' b(X_{t_{j-1}})\}\right)$$

(regarding $a(x) \equiv 0$ in the original $\mathbb{H}_n(\theta_1, \theta_2)$). Second, we get an estimate $\hat{\theta}_{2,n}$ by the explicit minima of the convex random function

$$\theta_2 \mapsto -2\mathbb{H}_n(\hat{\theta}_{1,n}, \theta_2) = \sum_{j=1}^{n}\frac{1}{h_n}\big(\Delta_j X - h_n\theta_2' a(X_{t_{j-1}})\big)^2 \exp\{-\hat{\theta}_{1,n}' b(X_{t_{j-1}})\}.$$

This framework naturally provides us with an adaptive model selection procedure. See Section 5.2.1 for details.

4.3. Volatility-parameter estimation for continuous semimartingales

In this section, we deal with the stochastic volatility-regression model [19,43]:

$$dY_t = b_t\,dt + \sigma(X_t, \theta)\,dw_t, \qquad t \in [0, T],$$

where w is an r-dimensional standard Wiener process, b and X are progressively measurable processes with values in $\mathbb{R}^m$ and $\mathbb{R}^d$, respectively, σ is an $\mathbb{R}^m \otimes \mathbb{R}^r$-valued function defined on $\mathbb{R}^d \times \Theta$ with $\Theta \in \mathbb{R}^p$. A data set consists of discrete observations $\mathbf{X}_n = (X_{t_j}, Y_{t_j})_{j=0}^n$ with $t_j = jh_n$, where $h_n = T/n$ with T fixed. The process b is completely unobservable and unknown. All processes are defined on a filtered probability space $\mathcal{B} := (\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\le T}, P)$.

Let $S(x, \theta) := \sigma(x, \theta)\sigma(x, \theta)'$ and $\Delta_j Y := Y_j - Y_{j-1}$. Then the quasi-likelihood function is given by

$$\mathbb{H}_n(\theta) = -\frac{1}{2}\sum_{j=1}^{n}\left\{\log\big|S(X_{t_{j-1}}, \theta)\big| + \frac{1}{h_n}S(X_{t_{j-1}}, \theta)^{-1}\big[(\Delta_j Y)^{\otimes 2}\big]\right\}.$$

Under appropriate conditions, the asymptotic distribution of $A_n^{-1}(\theta_0)(\hat{\theta}_n - \theta_0) = \sqrt{n}(\hat{\theta}_n - \theta_0)$ is mixed normal, that is,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} \Sigma_\theta^{-1/2} Z,$$

where Σ_θ is a symmetric $p \times p$-matrix which is a.s. positive definite, and Z is a p-variate standard-normal random variable which is defined on an extension of $\mathcal{B}$ and is independent of $\mathcal{F}$: see [19] and [43] for details.

The QBIC is computed as

$$\begin{aligned}\mathrm{QBIC}_n = {} & \sum_{j=1}^{n}\left\{\log\big|S(X_{t_{j-1}}, \hat{\theta}_n)\big| + \frac{1}{h_n}S^{-1}(X_{t_{j-1}}, \hat{\theta}_n)\big[(\Delta_j Y)^{\otimes 2}\big]\right\} \\ & + \log\left|\frac{1}{2}\sum_{j=1}^{n}\left\{\partial_\theta^2 \log\big|S(X_{t_{j-1}}, \hat{\theta}_n)\big| + \frac{1}{h_n}\partial_\theta^2\big(S^{-1}\big)(X_{t_{j-1}}, \hat{\theta}_n)\big[(\Delta_j Y)^{\otimes 2}\big]\right\}\right|.\end{aligned}$$

Let us consider the conditions for the QBIC to be valid, when $d = p$ and $m = r = 1$ with $\sigma(x, \theta) = \exp(x'\theta/2)$. The quasi-likelihood function is then given by

$$\mathbb{H}_n(\theta) = -\frac{1}{2}\sum_{j=1}^{n}\left\{X'_{t_{j-1}}\theta + \frac{1}{h_n}(\Delta_j Y)^2 \exp\bigl(-X'_{t_{j-1}}\theta\bigr)\right\}, \tag{4.3}$$

with $-\partial_\theta^2 \mathbb{H}_n(\theta) = \frac{1}{2}\sum_{j=1}^{n} \frac{(\Delta_j Y)^2}{h_n} \exp(-X'_{t_{j-1}}\theta) X_{t_{j-1}} X'_{t_{j-1}} \geq 0$ a.s.

Assumption 4.9. (i) $\sup_{\omega\in\Omega} \sup_{t\leq T} |X_t| < \infty$.
(ii) $\forall q > 0, \exists C > 0, \forall s, t \in [0, T], \mathbb{E}(|X_t - X_s|^q) \leq C|t - s|^{q/2}$.
(iii) $\forall q > 0, \sup_{0\leq t\leq T} \mathbb{E}(|b_t|^q) < \infty$.

Assumption 4.10. $\forall L > 0, \exists C_L > 0, \forall r > 0, \mathbb{P}\{\lambda_{\min}(\int_0^T X_t X'_t\, dt) \leq \frac{1}{r}\} \leq \frac{C_L}{r^L}$.

It will be seen that Assumptions 4.9 and 4.10 ensure Assumption 3.1 and inequality (3.3):

Theorem 4.11. *Let Assumptions 4.9 and 4.10 hold. Then, the assumptions in Theorem 3.7 are satisfied and the corresponding QBIC is given by*

$$\begin{aligned}\mathrm{QBIC}_n = &\sum_{j=1}^{n}\left\{X'_{t_{j-1}}\hat{\theta}_n + \frac{1}{h_n}(\Delta_j Y)^2 \exp\bigl(-X'_{t_{j-1}}\hat{\theta}_n\bigr)\right\} \\ &+ \log\left|\frac{1}{2h_n}\sum_{j=1}^{n}(\Delta_j Y)^2 \exp\bigl(-X'_{t_{j-1}}\hat{\theta}_n\bigr) X_{t_{j-1}} X'_{t_{j-1}}\right|.\end{aligned}$$

5. Model selection consistency

As long as concerned with good prediction performance, model selection consistency itself does not matter in an essential way. Given a set of models, it does when attempting to find the one "closest" (in the sense of KL divergence) to the true data-generating model structure itself as much as possible. For example, estimation of daily integrated volatility in econometrics would be the case, for econometricians usually builds up daily-volatility prediction model through a time series model such as, among others, ARFIMA models; an underlying continuous-time dynamics and a daily-volatility time series are separately modeled. This section is devoted to studying the validity of model selection consistency in our general setting. In particular, we propose an adaptive (stepwise) model selection strategy when we have more than one scaling rate. We start with a single-norming case, and then, before moving on to the multi-scaling case, we look at the case of ergodic diffusions since it well illustrates the proposed method.

5.1. Single-scaling case

We first consider cases where

$$a_n = a_{m,k,n}(\theta_0) \to 0$$

for each $m \in \{1, \ldots, M\}$ and $k \in \{1, \ldots, K_m\}$. Suppose that there exists a random function $\mathbb{H}_{m,0}$ such that

$$a_n^2 \mathbb{H}_{m,n}(\theta_m) \xrightarrow{P} \mathbb{H}_{m,0}(\theta_m) \tag{5.1}$$

uniformly in $\theta_m \in \bar{\Theta}_m$ as $n \to \infty$ ($m = 1, \ldots, M$). Moreover, we assume that the optimal parameter $\theta_{m,0} \in \Theta_m$ in the model $\mathcal{M}_m$ is the unique maximizer of $\mathbb{H}_{m,0}$:

$$\{\theta_{m,0}\} = \operatorname*{argmax}_{\theta_m \in \Theta_m} \mathbb{H}_{m,0}(\theta_m) \qquad \text{a.s.}$$

If m_0 satisfies

$$\{m_0\} = \operatorname*{argmin}_{m \in \mathfrak{M}} \dim(\Theta_m),$$

where $\mathfrak{M} = \operatorname{argmax}_{1 \le m \le M} \mathbb{H}_{m,0}(\theta_{m,0})$, we say that $\mathcal{M}_{m_0}$ is the *optimal model*. That is, the optimal model is, if exists, an element of the optimal model set $\mathfrak{M}$ which has the smallest dimension.

Let $\Theta_i \subset \mathbb{R}^{p_i}$ and $\Theta_j \subset \mathbb{R}^{p_j}$ be the parameter space associated with $\mathcal{M}_i$ and $\mathcal{M}_j$, respectively. We say that Θ_i *is nested in* Θ_j when $p_i < p_j$ and there exist a matrix $F \in \mathbb{R}^{p_j \times p_i}$ with $F'F = I_{p_i \times p_i}$ and a constant $c \in \mathbb{R}^{p_j}$ such that $\mathbb{H}_{i,n}(\theta_i) = \mathbb{H}_{j,n}(F\theta_i + c)$ for all $\theta_i \in \Theta_i$. That is, when Θ_i is nested in Θ_j, any model given by a parameter in Θ_i can also be generated by a parameter in Θ_j, so that $\mathcal{M}_j$ includes $\mathcal{M}_i$. Denote by $\mathrm{QBIC}_n^{(m)}$ the QBIC in $\mathcal{M}_m$.

Theorem 5.1. *Assume that* (5.1) *is satisfied and that* $\mathcal{M}_{m_0}$ *is the optimal model. Let* $m \in \{1, \ldots, M\} \setminus \{m_0\}$, *and let Assumptions* 3.1 *to* 3.3 *hold, and suppose that either*

(i) Θ_{m_0} *is nested in* Θ_m, *or*
(ii) $\mathbb{H}_{m,0}(\theta_m) \neq \mathbb{H}_{m_0,0}(\theta_{m_0,0})$ *a.s. for any* $\theta_m \in \Theta_m$.

Then we have

$$\lim_{n \to \infty} \mathbb{P}\big(\mathrm{QBIC}_n^{(m_0)} - \mathrm{QBIC}_n^{(m)} < 0\big) = 1, \tag{5.2}$$

$$\lim_{n \to \infty} \mathbb{P}\big(\mathrm{BIC}_n^{(m_0)} - \mathrm{BIC}_n^{(m)} < 0\big) = 1. \tag{5.3}$$

This theorem indicates that the probability that QBIC and BIC choose the optimal model tends to 1 as $n \to \infty$.

5.2. Multi-scaling case: Adaptive model comparison

For simplicity of exposition, we consider the two-scaling case, that is, $K = 2$. We propose a multi-step model selection procedure, which seems natural and more effective especially when an adaptive estimation procedure is possible in such a way that we can estimate a first component θ_{m_1} without knowledge of a second one θ_{m_2}. That is to say, it should be possible to select an optimal "partial" model structure associated with θ_{m_1}, with regarding θ_{m_2} as a nuisance element.

We suppose that the full model is "decomposed" into two parts, each consisting of M_1 and M_2 candidates, resulting in $M_1 \times M_2$ models in total. Write $(\mathcal{M}_{m_1,m_2})_{m_1 \le M_1; m_2 \le M_2}$ for the set of all the candidate models. We are given the "full" quasi-log likelihood function $\mathbb{H}_{m_1,m_2,n}(\theta_{m_1}, \theta_{m_2})$. Roughly speaking, we proceed as follows.

- First, introducing an auxiliary quasi-log likelihood which is only associated with the first-component parameter θ_{m_1} and does *not* involve θ_{m_2}, we obtain an estimate $\hat{\theta}_{m_1,n}$ of θ_{m_1}. Then we compare the corresponding (Q)BICs to select a first-stage optimal index, say $m^*_{1,n} \in \{1, \dots, M_1\}$; note that this strategy reduces the model-candidate set from $\{\mathbb{H}_{m_1,m_2,n}(\theta_{m_1}, \theta_{m_2})\}_{m_1,m_2}$ to $\{\mathbb{H}_{m^*_{1,n},m_2,n}(\hat{\theta}_{m^*_{1,n},n}, \theta_{m_2})\}_{m_2}$.
- Second, based on the "partly optimized" full quasi-log likelihoods $\mathbb{H}_{m^*_{1,n},1,n}, \dots, \mathbb{H}_{m^*_{1,n},M_2,n}$, we find a second-stage optimal index $m^*_{2,n} \in \{1, \dots, M_2\}$ through (Q)BIC again.
- Finally, we pick the model $\mathcal{M}_{m^*_{1,n},m^*_{2,n}}$ as our optimal model.

This adaptive procedure apparently reduces the computational cost (the number of comparison) to much extent compared with the joint-(Q)BIC case, that is, from "$O(M_1 \times M_2)$" to "$O(M_1 + M_2)$"; needless to say, the amount of reduction becomes larger for $K \ge 3$.

Remark 5.2. It is not essential in the above argument that the final step is based on the original quasi-log likelihood $\mathbb{H}_{m_1,m_2,n}$. What is essential for the model selection consistency is that at each stage we have a suitable auxiliary quasi-likelihood function based on which we can estimate a suitably separated optimal model. We here do not go into this direction.

To be specific, we here focus on the ergodic diffusion discussed in Section 4.2, and then briefly mention the general case in Section 5.2.2.

5.2.1. *Example*: *Ergodic diffusion*

Here we consider the same setting as in Section 4.2, that is, the model $\mathcal{M}_{m_1,m_2}$ is given by (4.1):

$$dX_t = a_{m_2}(X_t, \theta_{m_2})\,dt + b_{m_1}(X_t, \theta_{m_1})\,dw_t, \qquad t \in [0, T_n],\ X_0 = x_0.$$

Let $B_{m_1}(x, \theta_{m_1}) := b_{m_1}(x, \theta_{m_1}) b_{m_1}(x, \theta_{m_1})'$. Up to an additive constant term, the quasi-likelihood function $\mathbb{H}_{m_1,m_2,n}$ is given by

$$\begin{aligned}\mathbb{H}_{m_1,m_2,n}(\theta_{m_1,m_2}) = -\frac{1}{2}\sum_{j=1}^{n}\Bigg\{&\log\left|B_{m_1}(X_{t_{j-1}}, \theta_{m_1})\right| \\ &+ \frac{1}{h_n} B_{m_1}(X_{t_{j-1}}, \theta_{m_1})^{-1}\left[\left(\Delta_j X - h_n a_{m_2}(X_{t_{j-1}}, \theta_{m_2})\right)^{\otimes 2}\right]\Bigg\}.\end{aligned} \tag{5.4}$$

Then,

$$\frac{1}{n}\mathbb{H}_{m_1,m_2,n}(\theta_{m_1,m_2}) \xrightarrow{P} -\frac{1}{2}\int_{\mathbb{R}^d}\left[\operatorname{tr}\left\{B(x)B_{m_1}(x,\theta_{m_1})^{-1}\right\} + \log\left|B_{m_1}(x,\theta_{m_1})\right|\right]\nu(dx)$$
$$=: \mathbb{H}^1_{m_1,0}(\theta_{m_1})$$

uniformly in θ_{m_1}, where $B(x) := b(x)b(x)'$. Now we define an optimal model in the present setting: the situation is somewhat different from the single-scaling case because of the different rates of convergence, and we here need to introduce "partial" optimality step by step. First, we assume that optimal parameter $\theta_{m_1,0}$ and optimal model index $m_{1,0}$ are a.s. uniquely determined, that is,

$$\{\theta_{m_1,0}\} = \operatorname*{argmax}_{\theta_{m_1}\in\Theta_{m_1}} \mathbb{H}^1_{m_1,0}(\theta_{m_1}),$$
$$\{m_{1,0}\} = \operatorname*{argmin}_{m_1\in\mathfrak{M}_1} \dim(\Theta_{m_1}),$$

respectively, where $\mathfrak{M}_1 := \operatorname{argmax}_{1\le m_1\le M_1} \mathbb{H}^1_{m_1,0}(\theta_{m_1,0})$. Furthermore, we have

$$\frac{1}{nh_n}\left\{\mathbb{H}_{m_1,m_2,n}(\theta_{m_1,m_2}) - \mathbb{H}_{m_1,m_2,n}(\theta_{m_1},\theta_{m_2,0})\right\}$$
$$\xrightarrow{P} -\frac{1}{2}\int_{\mathbb{R}^d} B_{m_1}(x,\theta_{m_1})^{-1}\left[\left(a(x)-a_{m_2}(x,\theta_{m_2})\right)^{\otimes 2}\right]\nu(dx) =: \mathbb{H}_{m_1,m_2,0}(\theta_{m_1,m_2})$$

uniformly in θ_{m_1,m_2}, and that the optimal parameter $\theta_{m_2,0}$ is the unique maximizer of $\mathbb{H}_{m_{1,0},m_2,0}$:

$$\{\theta_{m_2,0}\} = \operatorname*{argmax}_{\theta_{m_2}\in\Theta_{m_2}} \mathbb{H}_{m_{1,0},m_2,0}(\theta_{m_{1,0},0},\theta_{m_2}).$$

If $m_{2,0}$ satisfies

$$\{m_{2,0}\} = \operatorname*{argmin}_{m_2\in\mathfrak{M}_2} \dim(\Theta_{m_2}),$$

where $\mathfrak{M}_2 := \operatorname{argmax}_{1\le m_2\le M_2} \mathbb{H}_{m_{1,0},m_2,0}(\theta_{m_{1,0},0},\theta_{m_2,0})$, we say that $\mathcal{M}_{m_{1,0},m_{2,0}}$ is the optimal model. Note that $b_{m_{1,0}}(\cdot,\theta_{m_{1,0},0}) = b(\cdot)$ and $a_{m_{2,0}}(\cdot,\theta_{m_{2,0},0}) = a(\cdot)$ since we are only considering correctly specified models.

Let $\Theta_{i_1}\times\Theta_{i_2}\subset\mathbb{R}^{p_{i_1}}\times\mathbb{R}^{p_{i_2}}$, $\Theta_{j_1}\times\Theta_{j_2}\subset\mathbb{R}^{p_{j_1}}\times\mathbb{R}^{p_{j_2}}$ be the parameter spaces associated with $\mathcal{M}_{i_1,i_2}$ and $\mathcal{M}_{j_1,j_2}$, respectively. We say that Θ_{i_1} *is nested in* Θ_{j_1} if $p_{i_1} < p_{j_1}$ and there exists a matrix $F_1\in\mathbb{R}^{p_{j_1}\times p_{i_1}}$ with $F_1'F_1 = I_{p_{i_1}\times p_{i_1}}$ as well as a $c_1\in\mathbb{R}^{p_{j_1}}$ such that $\mathbb{H}_{i_1,m_2,n}(\theta_{i_1},\theta_{m_2}) = \mathbb{H}_{j_1,m_2,n}(F_1\theta_{i_1}+c_1,\theta_{m_2})$ for all $\theta_{i_1}\in\Theta_{i_1}$ and $m_2\in\{1,\dots,M_2\}$. It is defined in a similar manner that Θ_{i_2} *is nested in* Θ_{j_2}.

First, we consider the *joint QBIC* of (3.12):

$$\mathrm{QBIC}_n^{(m_1,m_2)} = -2\mathbb{H}_{m_1,m_2,n}(\hat\theta_{m_1,m_2,n})$$
$$+\log\left|-\partial^2_{\theta_{m_1}}\mathbb{H}_{m_1,m_2,n}(\hat\theta_{m_1,m_2,n})\right| + \log\left|-\partial^2_{\theta_{m_2}}\mathbb{H}_{m_1,m_2,n}(\hat\theta_{m_1,m_2,n})\right|.$$

If $(m_{1,n}^*, m_{2,n}^*) = \operatorname{argmin}_{(m_1,m_2)\in\{1,\dots,M_1\}\times\{1,\dots,M_2\}} \mathrm{QBIC}_n^{(m_1,m_2)}$, we choose the model $\mathcal{M}_{m_{1,n}^*,m_{2,n}^*}$, which we again call the *optimal model*, as the optimal model among the candidate models. The details of $\mathrm{QBIC}_n^{(m_1,m_2)}$ are given in Theorems 4.5 and 4.6. Likewise, we define

$$\mathrm{BIC}_n^{(m_1,m_2)} = -2\mathbb{H}_{m_1,m_2,n}(\hat{\theta}_{m_1,m_2,n}) + p_{m_1}\log n + p_{m_2}\log T_n.$$

Theorem 5.3. *Suppose that Assumptions 4.1 to 4.4 hold for the models $\mathcal{M}_{m_{1,0},m_{2,0}}$ and $\mathcal{M}_{m_1,m_2}$ where $\mathcal{M}_{m_{1,0},m_{2,0}}$ is the optimal model, and at least one of m_1 and m_2 differs from $m_{1,0}$ and $m_{2,0}$, respectively. Moreover, suppose that $\Theta_{m_{1,0}}$ and $\Theta_{m_{2,0}}$ are nested in Θ_{m_1} and Θ_{m_2}, respectively. Then we have*

$$\lim_{n\to\infty} \mathbb{P}\big(\mathrm{QBIC}_n^{(m_{1,0},m_{2,0})} - \mathrm{QBIC}_n^{(m_1,m_2)} < 0\big) = 1,$$

*and the same statement with "*QBIC*" replaced by "*BIC*".*

Remark 5.4. Here we have derived the validity and the selection consistency of the (Q)BIC for correctly specified (nested) ergodic diffusion process. Concerned with misspecified (non-nested) ergodic diffusion processes, [41] proved that the asymptotic behavior of the GQMLE can be essentially different; in that case, the tail-probability estimate for Assumption 3.3 has not yet been established. We do not consider the misspecified case in this paper. Nevertheless, it is easily expected that, as in [33], the tail-probability estimate can be deduced in a two-step manner by applying [32], Theorem 3.5, twice, first for the diffusion part and then for the drift one.

Next, we turn to the *two-step QBIC*. In the present case, we apply the previous single-scaling result twice for the single true data-generating model. First, we focus on the diffusion coefficient, which we can estimate more quickly than the drift one. Under suitable conditions, $\mathrm{QBIC}_n^{(m_1)}$ and $\mathrm{BIC}_n^{(m_1)}$ are given by

$$\begin{aligned}\mathrm{QBIC}_n^{(m_1)} &= -2\mathbb{H}^1_{m_1,n}(\hat{\theta}_{m_1,n}) + \log\big|-\partial^2_{\theta_{m_1}}\mathbb{H}^1_{m_1,n}(\hat{\theta}_{m_1,n})\big|,\\ \mathrm{BIC}_n^{(m_1)} &= -2\mathbb{H}^1_{m_1,n}(\hat{\theta}_{m_1,n}) + p_{m_1}\log n,\end{aligned}$$

where $\mathbb{H}^1_{m_1,n}$ is defined by the joint quasi-likelihood (5.4) with a_{m_2} being ignored:

$$\mathbb{H}^1_{m_1,n}(\theta_{m_1,m_2}) = -\frac{1}{2}\sum_{j=1}^n\left\{\log\big|B_{m_1}(X_{t_{j-1}},\theta_{m_1})\big| + \frac{1}{h_n}B_{m_1}(X_{t_{j-1}},\theta_{m_1})^{-1}\big[(\Delta_j X)^{\otimes 2}\big]\right\},$$

and where $\hat{\theta}_{m_1,n}$ is the QMLE associated with $\mathbb{H}^1_{m_1,n}$. Note that we can write

$$\frac{1}{n}\mathbb{H}_n(\theta) = \frac{1}{n}\mathbb{H}^1_n(\theta_1) + \delta^1_n(\theta)$$

with $\sup_\theta |\delta^1_n(\theta)| \xrightarrow{P} 0$. We proceed as follows.

- First, assuming that Θ_{i_1} is nested in Θ_{j_1}, i.e. $\mathbb{H}^1_{i_1,n}(\theta_{i_1}) = \mathbb{H}^1_{j_1,n}(F_1\theta_{i_1} + c_1)$, we set

$$\{m^*_{1,n}\} = \operatorname*{argmin}_{1\le m_1\le M_1} \mathrm{QBIC}^{(m_1)}_n.$$

- Next, we consider the stochastic differential equation

$$dX_t = a_{m_2}(X_t, \theta_{m_2})\,dt + b_{m^*_{1,n}}(X_t, \hat{\theta}_{m^*_{1,n},n})\,dw_t. \tag{5.5}$$

Assuming that $\{\hat{\theta}_{m_2,n}\} = \operatorname{argmax}_{\theta_{m_2}\in\Theta_{m_2}} \mathbb{H}_{m^*_{1,n},m_2,n}(\hat{\theta}_{m^*_{1,n},n}, \theta_{m_2})$ and that $\{m^*_{2,n}\} = \operatorname{argmin}_{1\le m_2\le M_2} \mathrm{QBIC}^{(m_2|m^*_{1,n})}_n$, where

$$\begin{aligned}\mathrm{QBIC}^{(m_2|m^*_{1,n})}_n := {} & -2\mathbb{H}_{m^*_{1,n},m_2,n}(\hat{\theta}_{m^*_{1,n},n}, \hat{\theta}_{m_2,n}) \\ & + \log\bigl|-\partial^2_{\theta_{m_2}} \mathbb{H}_{m^*_{1,n},m_2,n}(\hat{\theta}_{m^*_{1,n},n}, \hat{\theta}_{m_2,n})\bigr|,\end{aligned}$$

we select the model $\mathcal{M}_{m^*_{1,n},m^*_{2,n}}$ as the final optimal model.

When we use BIC, the best model is selected by a similar procedure with

$$\mathrm{BIC}^{(m_2|m^*_{1,n})}_n := -2\mathbb{H}_{m^*_{1,n},m_2,n}(\hat{\theta}_{m^*_{1,n}}, \hat{\theta}_{m_2,n}) + p_{m_2}\log T_n$$

used instead of $\mathrm{QBIC}^{(m_2|m^*_{1,n})}_n$.

Joint (Q)BIC and two-step (Q)BIC may select different models for a fixed sample size, however, the model selection consistency property is asymptotically shared.

Theorem 5.5. *Suppose that Assumptions 4.1 to 4.4 hold for the models $\mathcal{M}_{m_{1,0},m_{2,0}}$ and $\mathcal{M}_{m_1,m_2}$ where $\mathcal{M}_{m_{1,0},m_{2,0}}$ is the optimal model and $(m_1, m_2) \in (\{1,\dots,M_1\}\setminus\{m_{1,0}\}) \times (\{1,\dots,M_2\}\setminus\{m_{2,0}\})$, and that $\Theta_{m_{1,0}}$ and $\Theta_{m_{2,0}}$ are nested in Θ_{m_1} and Θ_{m_2}, respectively. Then we have*

$$\lim_{n\to\infty} \mathbb{P}\bigl(\mathrm{QBIC}^{(m_{1,0})}_n - \mathrm{QBIC}^{(m_1)}_n < 0\bigr) = 1,$$

$$\lim_{n\to\infty} \mathbb{P}\bigl(\mathrm{QBIC}^{(m_{2,0}|m^*_{1,n})}_n - \mathrm{QBIC}^{(m_2|m^*_{1,n})}_n < 0\bigr) = 1,$$

*and the same statements with "*QBIC*" replaced by "*BIC*".*

5.2.2. *Remark on general case*

Without essential change, we may follow the same scenario as in the previous section for general LAQ models under the setting described in the beginning of this section: instead of the original "full" quasi-likelihood $\mathbb{H}_n(\theta)$, we solely look at some "auxiliary" random fields $\theta_1 \mapsto \mathbb{H}^1_n(\theta_1)$ and $\theta_2 \mapsto \mathbb{H}_n(\hat{\theta}_{1,n}, \theta_2)$ in this order, based on which we successively define the two-step QMLE as $\hat{\theta}_{1,n} \in \operatorname{argmax} \mathbb{H}^1_n$ and $\hat{\theta}_{2,n} \in \operatorname{argmax}_{\theta_2} \mathbb{H}^2_n(\hat{\theta}_{1,n}, \theta_2)$. More specifically, we let Assumptions 3.1

to 3.3 hold and further assume that there exist positive sequences $a_{i,n}(\theta_0) \to 0$ $(i = 1, 2)$ such that $a_{1,n}(\theta_0)/a_{2,n}(\theta_0) \to 0$ and random functions $\mathbb{Y}_0^1(\theta_1)$ and $\mathbb{Y}_0^2(\theta)$, for which:

1. $a_{1,n}^2(\theta_0)\mathbb{H}_n(\theta) = a_{1,n}^2(\theta_0)\mathbb{H}_n^1(\theta_1) + \delta_n^1(\theta)$ where $\sup_\theta |\delta_n^1(\theta)| \xrightarrow{P} 0$;
2. $\sup_{\theta_1} |a_{1,n}(\theta_0)^2\{\mathbb{H}_n^1(\theta_1) - \mathbb{H}_n^1(\theta_{1,0})\} - \mathbb{Y}_0^1(\theta_1)| \xrightarrow{P} 0$, where $\{\theta_{1,0}\} = \operatorname{argmax} \mathbb{Y}_0^1$ a.s.;
3. $\sup_\theta |a_{2,n}(\theta_0)^2\{\mathbb{H}_n(\theta) - \mathbb{H}_n(\theta_1, \theta_{2,0})\} - \mathbb{Y}_0^2(\theta)| \xrightarrow{P} 0$, where $\{\theta_{2,0}\} = \operatorname{argmax}_{\theta_2} \mathbb{Y}_0^2(\theta_{1,0}, \theta_2)$ a.s.

Under these conditions, we may deduce the consistency $\hat{\theta}_n = (\hat{\theta}_{1,n}, \hat{\theta}_{2,n}) \xrightarrow{P} \theta_0$, which combined with Assumption 3.1 gives the tightness $A_n^{-1}(\theta_0)(\hat{\theta}_n - \theta_0) = O_p(1)$; in this case, the LAQ structure (3.2) typically takes the form

$$\sup_{u=(u_1,u_2)\in A} \left| \log \mathbb{Z}_n(u) - \left(\Delta_{1,n}[u_1] + \Delta_{2,n}[u_2] - \frac{1}{2}\big(\Gamma_{1,0}(\theta_{1,0})[u_1, u_1] + \Gamma_{2,0}(\theta_0)[u_2, u_2]\big) \right) \right| = o_p(1)$$

for each compact $A \in \mathbb{R}^p$, with $\Delta_1 =: (\Delta_{1,n}, \Delta_{2,n})$ and $\Gamma_0 =: \operatorname{diag}\{\Gamma_{1,0}(\theta_{1,0}), \Gamma_{2,0}(\theta_0)\}$. Then, we can prove the model selection consistency in analogy with the proofs of Theorems 5.1, 5.3 and 5.5.

6. Simulation results

We here conduct a number of simulations to observe finite-sample performance of the (Q)BIC proposed in this paper. While what to be looked at is quasi-Bayes factors for candidate models, for conciseness we focus on the selection frequency as well as the estimation performance of the quasi-maximum likelihood estimates. In Section 6.1, we use the R package `yuima` [6] for generating data and estimating the parameter. We set the initial value in numerical optimization to be random numbers drawn from uniform distributions $U(\theta_{i,0} - 0.5, \theta_{i,0} + 0.5)$, $1 \le i \le p$. All the SDE coefficients considered here are of the partially convex type mentioned in Remark 4.8. In the examples below, w denotes a one-dimensional standard Wiener process, and all the Monte Carlo trials are based on 1000 independent sample paths.

6.1. Ergodic diffusion process

Suppose that we have a sample $\mathbf{X}_n = (X_{t_j})_{j=0}^n$ with $t_j = jn^{-2/3}$ from the true model

$$dX_t = -X_t\,dt + \exp\left\{\frac{1}{2}(-2\cos X_t + 1)\right\} dw_t, \qquad t \in [0, T_n],\ X_0 = 1,$$

where $T_n = n^{1/3}$. We consider the following as the diffusion part:

Diff 1 : $\exp\left\{\frac{1}{2}(\theta_{11}\cos X_t + \theta_{12}\sin X_t + \theta_{13})\right\}$;

$$\textbf{Diff 2}: \quad \exp\left\{\frac{1}{2}(\theta_{11}\cos X_t + \theta_{12}\sin X_t)\right\}; \qquad \textbf{Diff 3}: \quad \exp\left\{\frac{1}{2}(\theta_{11}\cos X_t + \theta_{13})\right\};$$

$$\textbf{Diff 4}: \quad \exp\left\{\frac{1}{2}(\theta_{12}\sin X_t + \theta_{13})\right\}; \qquad \textbf{Diff 5}: \quad \exp\left\{\frac{1}{2}\theta_{11}\cos X_t\right\};$$

$$\textbf{Diff 6}: \quad \exp\left\{\frac{1}{2}\theta_{12}\sin X_t\right\}; \qquad \textbf{Diff7}: \quad \exp\left\{\frac{1}{2}\theta_{13}\right\},$$

and also the following for the drift one:

$$\textbf{Drif 1}: \quad \theta_{21}X_t + \theta_{22}; \qquad \textbf{Drif 2}: \quad \theta_{21}X_t; \qquad \textbf{Drif 3}: \quad \theta_{22}.$$

Each candidate model is given through a combination of the diffusion and drift parts; for example, in the case of Diff 1 and Drif 1, we consider the statistical model

$$dX_t = (\theta_{21}X_t + \theta_{22})\,dt + \exp\left\{\frac{1}{2}(\theta_{11}\cos X_t + \theta_{12}\sin X_t + \theta_{13})\right\}dw_t.$$

That is, the true model consists of Diff 3 and Drif 2.

We compare model selection frequency through QBIC, BIC, and the contrast-based information criterion (CIC), which is an AIC-type criterion introduced by [38] under the rapidly increasing experimental design $nh_n^2 \to 0$ (see also [18] for CIC under a weaker sampling-design condition $nh^q \to 0$ for some $q \geq 2$). We simulate the number of the model selected by using joint QBIC, joint BIC, two-step QBIC, two-step BIC, and CIC among the candidate models. The simulations are done for $n = 1000, 3000$, and 5000.

Tables 1 and 2 summarize the comparison results of the model selection frequencies; they show quite similar tendencies, in particular, the frequencies that the true model defined by Diff 3 and Drif 2 is selected by QBIC and BIC become larger as n increases. Also observed is that BIC often takes values between QBIC and CIC; in particular, QBIC chooses the full model consisting of Diff 1 and Drif 1 more frequently than BIC. Moreover, joint (Q)BIC gets close to two-step (Q)BIC as n increases.

It is worth mentioning that computation time of joint (Q)BIC was overall about twice of that of two-step (Q)BIC. This superiority of the two-step (Q)BIC should become more significant for higher-dimensional models.

6.2. Volatility-parameter estimation for continuous semimartingale

Let $(X_{t_j}, Y_{t_j})_{j=0}^n$ be a data set with $t_j = j/n$ and the number of data n. We consider a solution to the stochastic regression model

$$dY_t = \exp\left(\frac{1}{2}X_t'\theta_0\right)dw_t = \exp\left\{\frac{1}{2}(-2X_{2,t} + 3X_{3,t})\right\}dw_t, \qquad t \in [0,1],$$

Table 1. The number of models selected by joint QBIC, joint BIC and CIC in Section 6.1 over 1000 simulations for various n. The true model consists of Diff 3 and Drif 2

	Criteria	Diff 1	Diff 2	Diff 3*	Diff 4	Diff 5	Diff 6	Diff 7
		$n = 1000$						
Drif 1	QBIC	7	8	109	1	15	0	1
	BIC	0	20	105	1	49	0	2
	fAIC	25	23	136	3	19	0	2
Drif 2*	QBIC	19	17	**741**	0	76	0	1
	BIC	1	22	**523**	0	248	0	1
	fAIC	92	43	**559**	0	73	0	1
Drif 3	QBIC	0	0	5	0	0	0	0
	BIC	0	0	28	0	0	0	0
	fAIC	5	0	19	0	0	0	0
		$n = 3000$						
Drif 1	QBIC	1	2	102	0	0	0	0
	BIC	0	2	126	0	10	0	0
	fAIC	24	5	173	0	2	0	0
Drif 2*	QBIC	12	4	**867**	0	12	0	0
	BIC	1	4	**786**	0	63	0	0
	fAIC	110	6	**667**	0	7	0	0
Drif 3	QBIC	0	0	0	0	0	0	0
	BIC	0	0	8	0	0	0	0
	fAIC	0	0	6	0	0	0	0
		$n = 5000$						
Drif 1	QBIC	1	0	80	0	0	0	0
	BIC	0	0	113	0	3	0	0
	fAIC	30	1	166	0	2	0	0
Drif 2*	QBIC	16	0	**900**	0	3	0	0
	BIC	1	0	**863**	0	20	0	0
	fAIC	135	0	**666**	0	7	0	0
Drif 3	QBIC	0	0	0	0	0	0	0
	BIC	0	0	8	0	0	0	0
	fAIC	0	0	0	0	0	0	0

where $X_t = (X_{1,t}, X_{2,t}, X_{3,t})'$ and the true parameter $\theta_0 = (0, -2, 3)'$. We consider the following models:

$$\textbf{Model 1}: \quad dY_t = \exp\left\{\frac{1}{2}(\theta_1 X_{1,t} + \theta_2 X_{2,t} + \theta_3 X_{3,t})\right\} dw_t;$$

Table 2. The number of models selected by two-step QBIC and two-step BIC in Section 6.1 over 1000 simulations for various n. The true model consists of Diff 3 and Drif 2

	Criteria	Diff 1	Diff 2	Diff 3*	Diff 4	Diff 5	Diff 6	Diff 7
		$n = 1000$						
Drif 1	QBIC	4	3	108	0	0	0	0
	BIC	1	6	120	0	41	0	0
Drif 2*	QBIC	19	10	**798**	0	45	0	0
	BIC	3	16	**588**	0	199	0	0
Drif 3	QBIC	0	0	1	0	0	0	0
	BIC	0	0	26	0	0	0	0
		$n = 3000$						
Drif 1	QBIC	6	0	77	0	0	0	0
	BIC	0	3	111	0	3	0	0
Drif 2*	QBIC	19	1	**892**	0	4	0	0
	BIC	1	1	**836**	0	36	0	0
Drif 3	QBIC	0	0	1	0	0	0	0
	BIC	0	0	9	0	0	0	0
		$n = 5000$						
Drif 1	QBIC	1	0	80	0	0	0	0
	BIC	0	3	115	0	1	0	0
Drif 2*	QBIC	14	0	**904**	0	1	0	0
	BIC	2	0	**864**	0	18	0	0
Drif 3	QBIC	0	0	0	0	0	0	0
	BIC	0	0	0	0	0	0	0

$$\textbf{Model 2}: \quad dY_t = \exp\left\{\frac{1}{2}(\theta_1 X_{1,t} + \theta_2 X_{2,t})\right\} dw_t;$$

$$\textbf{Model 3}: \quad dY_t = \exp\left\{\frac{1}{2}(\theta_1 X_{1,t} + \theta_3 X_{3,t})\right\} dw_t;$$

$$\textbf{Model 4}: \quad dY_t = \exp\left\{\frac{1}{2}(\theta_2 X_{2,t} + \theta_3 X_{3,t})\right\} dw_t; \qquad \textbf{Model 5}: \quad dY_t = \exp\left\{\frac{\theta_1}{2} X_{1,t}\right\} dw_t;$$

$$\textbf{Model 6}: \quad dY_t = \exp\left\{\frac{\theta_2}{2} X_{2,t}\right\} dw_t; \qquad \textbf{Model 7}: \quad dY_t = \exp\left\{\frac{\theta_3}{2} X_{3,t}\right\} dw_t.$$

Then the true model is Model 4. Note that Models 2, 3, 5, 6, and 7 are misspecified models.

For each model, an estimator is obtained from the quasi-likelihood (4.3). In the Model 1 (full model), the statistics QBIC, BIC and formal AIC (fAIC) are given by

$$\begin{aligned}
\mathrm{QBIC}_n &= \sum_{j=1}^{n}\big\{(\hat\theta_{1,n}X_{1,t_{j-1}} + \hat\theta_{2,n}X_{2,t_{j-1}} + \hat\theta_{3,n}X_{3,t_{j-1}}) \\
&\quad + n(\Delta_j Y)^2 \exp(-\hat\theta_{1,n}X_{1,t_{j-1}} - \hat\theta_{2,n}X_{2,t_{j-1}} - \hat\theta_{3,n}X_{3,t_{j-1}})\big\} \\
&\quad + \log\left|\frac{n}{2}\sum_{j=1}^{n}(\Delta_j Y)^2\right. \\
&\quad \left.\times \exp(-\hat\theta_{1,n}X_{1,t_{j-1}} - \hat\theta_{2,n}X_{2,t_{j-1}} - \hat\theta_{3,n}X_{3,t_{j-1}})X_{t_{j-1}}X'_{t_{j-1}}\right|, \\
\mathrm{BIC}_n &= \sum_{j=1}^{n}\big\{(\hat\theta_{1,n}X_{1,t_{j-1}} + \hat\theta_{2,n}X_{2,t_{j-1}} + \hat\theta_{3,n}X_{3,t_{j-1}}) \\
&\quad + n(\Delta_j Y)^2 \exp(-\hat\theta_{1,n}X_{1,t_{j-1}} - \hat\theta_{2,n}X_{2,t_{j-1}} - \hat\theta_{3,n}X_{3,t_{j-1}})\big\} + 3\log n, \\
\mathrm{fAIC}_n &= \sum_{j=1}^{n}\big\{(\hat\theta_{1,n}X_{1,t_{j-1}} + \hat\theta_{2,n}X_{2,t_{j-1}} + \hat\theta_{3,n}X_{3,t_{j-1}}) \\
&\quad + n(\Delta_j Y)^2 \exp(-\hat\theta_{1,n}X_{1,t_{j-1}} - \hat\theta_{2,n}X_{2,t_{j-1}} - \hat\theta_{3,n}X_{3,t_{j-1}})\big\} + 3\times 2,
\end{aligned}$$

where $\hat\theta_n = (\hat\theta_{1,n}, \hat\theta_{2,n}, \hat\theta_{3,n})$ is the quasi-maximum likelihood estimator.

6.2.1. *Non-random covariate process*

First, we set

$$X_{t_j} = \left(1, \cos\left(\frac{2j\pi}{n}\right), \sin\left(\frac{2j\pi}{n}\right)\right)', \qquad j = 0, 1, \ldots, n.$$

Then the model is Gaussian. We readily get

$$\begin{aligned}
\int_0^T X_t X_t'\,dt &= \int_0^1 \begin{pmatrix} 1 & \cos(2t\pi) & \sin(2t\pi) \\ \cos(2t\pi) & \cos^2(2t\pi) & \cos(2t\pi)\sin(2t\pi) \\ \sin(2t\pi) & \cos(2t\pi)\sin(2t\pi) & \sin^2(2t\pi) \end{pmatrix} dt \\
&= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 1/2 \end{pmatrix},
\end{aligned}$$

so that $\det(\int_0^T X_t X_t'\,dt) = \frac{1}{4}$.

Table 3. The number of models selected by QBIC, BIC and fAIC in Section 6.2.1 over 1000 simulations for various n (1–7 express the model labels, and the true model is model 4)

	$n=50$							$n=100$							$n=200$						
Criterion	1	2	3	4*	5	6	7	1	2	3	4*	5	6	7	1	2	3	4*	5	6	7
QBIC	74	0	0	**925**	0	0	0	57	0	0	**943**	0	0	0	37	0	0	**963**	0	0	0
BIC	67	0	0	**933**	0	0	0	39	0	0	**961**	0	0	0	25	0	0	**975**	0	0	0
fAIC	183	0	0	**817**	0	0	0	178	0	0	**822**	0	0	0	179	0	0	**821**	0	0	0

In Table 3, Model 4 is selected with high frequency as the best model for all cases. Note that in this case $\hat{\theta}_n$ is approximately the MLE, so that fAIC is approximately the true AIC. fAIC shows the tendency to choose a model larger than the true one even for large sample size, while QBIC and BIC do the model selection consistency; these phenomena are common in the classical information criteria based on likelihood function and MLE. Recall that the model selection inconsistency is not a defect as the AIC is not intended to estimate the true model consistently.

Table 4 summarizes the mean and the standard deviation of estimators in each model. In the case of the correctly specified models, the estimators get closer to the true value.

Table 4. The mean and the standard deviation (s.d.) of the estimator $\hat{\theta}_{1,n}$, $\hat{\theta}_{2,n}$ and $\hat{\theta}_{3,n}$ in Section 6.2.1 for various n (1–7 express the models, and the true parameter $\theta_0 = (0, -2, 3)$)

		$n=50$			$n=100$			$n=200$		
		$\hat{\theta}_{1,n}$	$\hat{\theta}_{2,n}$	$\hat{\theta}_{3,n}$	$\hat{\theta}_{1,n}$	$\hat{\theta}_{2,n}$	$\hat{\theta}_{3,n}$	$\hat{\theta}_{1,n}$	$\hat{\theta}_{2,n}$	$\hat{\theta}_{3,n}$
1	mean	−0.0564	−1.8312	3.1129	−0.0218	−1.8972	3.0748	−0.0183	−1.9586	3.0299
	s.d.	0.2086	0.2879	0.2978	0.1509	0.2006	0.2052	0.1026	0.1461	0.1442
2	mean	1.5872	−1.8314	–	1.6145	−1.8839	–	1.5834	−1.9602	–
	s.d.	0.3579	0.4852	–	0.2497	0.3505	–	0.1750	0.2525	–
3	mean	0.6473	–	3.1168	0.7259	–	3.0660	0.7734	–	3.0328
	s.d.	0.2829	–	0.3981	0.2054	–	0.2705	0.1486	–	0.1966
4*	mean	–	−1.8312	3.1129	–	−1.8972	3.0748	–	−1.9586	3.0299
	s.d.	–	0.2879	0.2978	–	0.2006	0.2052	–	0.1461	0.1441
5	mean	0.4871	–	–	0.5045	–	–	0.4915	–	–
	s.d.	0.2858	–	–	0.2866	–	–	0.2948	–	–
6	mean	–	−1.892	–	–	−1.916	–	–	−1.9726	–
	s.d.	–	0.3427	–	–	0.2814	–	–	0.2157	–
7	mean	–	–	3.0867	–	–	3.0498	–	–	3.0262
	s.d.	–	–	0.3026	–	–	0.2267	–	–	0.1716

6.2.2. *Random covariate process*

Next, we consider the following two cases:

(i–1) $X_t = (X_{1,t}, X_{2,t}, X_{3,t})' = (1, \cos(B_t), \sin(B_t))'$;
(i–2) $X_t = (X_{1,t}, X_{2,t}, X_{3,t})' = (10, \cos(B_t), \sin(B_t))'$,

where B is a one-dimensional standard Wiener process independent of w. For data generation, we use the 3-dimensional stochastic differential equation for $(X_{2,t}, X_{3,t}, Y_t)$

$$d\begin{pmatrix} X_{2,t} \\ X_{3,t} \\ Y_t \end{pmatrix} = -\frac{1}{2}\begin{pmatrix} X_{2,t} \\ X_{3,t} \\ 0 \end{pmatrix} dt + \begin{pmatrix} -X_{3,t} & 0 \\ X_{2,t} & 0 \\ 0 & \exp\left\{\frac{1}{2}(-2X_{2,t} + 3X_{3,t})\right\} \end{pmatrix} d\begin{pmatrix} B_t \\ w_t \end{pmatrix}.$$

Tables 5 and 7 summarize the comparison results of model selection frequency. In the case of (i–1) (Table 5), the probability that a full model is chosen by QBIC seems to be too high when the sample size is small. This phenomenon in QBIC would be caused by the problem that $|-\partial_\theta^2 \mathbb{H}_n(\hat{\theta}_n)| \approx 0$; as a matter of fact, we did observe that the values of the determinant in simulations were so small. Nevertheless, judging from the whole Table 5, tendencies of QBIC, BIC and fAIC for $n \to \infty$ are overall the same as in Section 6.2.1.

In the case of (i–2) (Table 7), QBIC tends to perform better not only than fAIC but also than BIC for all n; indeed, in this case we observed that the values of $|-\partial_\theta^2 \mathbb{H}_n(\hat{\theta}_n)|$ were far from zero. Moreover, the true model was selected by using QBIC with high probability even for small sample size.

Tables 6 and 8 show that a tendency of the estimators for Model 1 and Model 4 is analogous to the previous non-random case. As is easily expected from the result [41] concerning parametric estimation of a diffusion with misspecified coefficients, we need to let $T_n \to \infty$ in order to consistently estimate optimal parameter values.

Table 5. The number of models selected by QBIC, BIC and fAIC in Section 6.2.2 (i–1) over 1000 simulations for various n (1–7 express the models, and the true model is model 4)

Criterion	1	2	3	4*	5	6	7	1	2	3	4*	5	6	7	1	2	3	4*	5	6	7
	$n = 200$							$n = 500$							$n = 1000$						
QBIC	831	0	5	**164**	0	0	0	657	1	8	**334**	0	0	0	500	0	7	**493**	0	0	0
BIC	8	29	234	**729**	0	0	0	8	5	141	**846**	0	0	0	5	0	117	**878**	0	0	0
fAIC	75	24	224	**677**	0	0	0	107	4	132	**757**	0	0	0	129	0	105	**766**	0	0	0
	$n = 3000$							$n = 5000$							$n = 10\,000$						
QBIC	250	0	7	**743**	0	0	0	217	0	8	**775**	0	0	0	123	0	3	**874**	0	0	0
BIC	0	0	43	**957**	0	0	0	4	0	40	**956**	0	0	0	4	0	8	**988**	0	0	0
fAIC	111	0	38	**851**	0	0	0	156	0	30	**814**	0	0	0	153	0	5	**842**	0	0	0

Table 6. The mean and the standard deviation (s.d.) of the estimator $\hat{\theta}_{1,n}$, $\hat{\theta}_{2,n}$ and $\hat{\theta}_{3,n}$ in Section 6.2.2 (i–1) for various n (1–7 express the models, and the true parameter $\theta_0 = (0, -2, 3)$)

		$\hat{\theta}_{1,n}$	$\hat{\theta}_{2,n}$	$\hat{\theta}_{3,n}$	$\hat{\theta}_{1,n}$	$\hat{\theta}_{2,n}$	$\hat{\theta}_{3,n}$	$\hat{\theta}_{1,n}$	$\hat{\theta}_{2,n}$	$\hat{\theta}_{3,n}$
		$n = 200$			$n = 500$			$n = 1000$		
1	mean	−0.2910	−1.7237	2.9497	−0.0683	−1.9422	2.9842	−0.0458	−1.9548	2.9974
	s.d.	2.2703	2.2300	0.8309	1.5636	1.5293	0.5462	1.0542	1.0335	0.3824
2	mean	0.9085	−2.7435	–	0.8747	−2.7212	–	0.8685	−2.7301	–
	s.d.	6.4524	6.2657	–	6.3784	6.1702	–	6.2699	6.0750	–
3	mean	−2.0247	–	2.9749	−2.0092	–	3.0041	−2.0528	–	2.9834
	s.d.	0.5204	–	1.2907	0.5638	–	1.2907	0.3399	–	1.2396
4*	mean	–	−2.0000	2.9724	–	−1.9989	3.0037	–	−1.9986	2.9890
	s.d.	–	0.1691	0.3137	–	0.1075	0.2044	–	0.0712	0.1384
5	mean	−0.2429	–	–	−0.2132	–	–	−0.2409	–	–
	s.d.	0.4806	–	–	0.4905	–	–	0.4746	–	–
6	mean	–	−1.8835	–	–	−1.8974	–	–	−1.9033	–
	s.d.	–	0.4777	–	–	0.4701	–	–	0.4763	–
7	mean	–	–	3.0079	–	–	2.9923	–	–	2.9776
	s.d.	–	–	0.5366	–	–	0.5401	–	–	0.5385
		$n = 3000$			$n = 5000$			$n = 10\,000$		
1	mean	−0.0098	−1.9915	3.0000	−0.0432	−1.9569	2.9971	−0.0040	−1.9957	2.9964
	s.d.	0.5700	0.5601	0.2043	0.4887	0.4802	0.1722	0.3244	0.3193	0.1145
2	mean	0.9311	−2.7877	–	0.7949	−2.6941	–	0.7585	−2.5841	–
	s.d.	6.2734	6.0827	–	6.4740	6.2692	–	6.4844	6.3014	–
3	mean	−2.0158	–	3.0173	−2.0390	–	2.9579	−2.0177	–	3.0653
	s.d.	0.4537	–	1.2141	0.4679	–	1.2394	0.4491	–	1.2283
4*	mean	–	−2.0002	2.9953	–	−1.9986	2.9987	–	−1.9999	2.9994
	s.d.	–	0.0433	0.0797	–	0.0335	0.0617	–	0.0220	0.0450
5	mean	−0.2391	–	–	−0.2603	–	–	−0.2159	–	–
	s.d.	0.4794	–	–	0.4765	–	–	0.4806	–	–
6	mean	–	−1.8916	–	–	−1.9019	–	–	−1.8775	–
	s.d.	–	0.4627	–	–	0.4842	–	–	0.4521	–
7	mean	–	–	2.9736	–	–	3.0103	–	–	2.9793
	s.d.	–	–	0.5533	–	–	0.5483	–	–	0.5326

We also conducted similar simulations for the case where X is instead given by

$$X_t = \left(1, \frac{1}{1 + B_t^2}, \frac{B_t}{1 + B_t^2}\right)',$$

and quite similar tendencies were observed.

Table 7. The number of models selected by QBIC, BIC and fAIC in Section 6.2.2 (i–2) over 1000 simulations for various n (1–7 express the models, and the true model is model 4)

	$n=200$							$n=500$							$n=1000$						
Criterion	1	2	3	4*	5	6	7	1	2	3	4*	5	6	7	1	2	3	4*	5	6	7
QBIC	78	1	1	**920**	0	0	0	38	0	7	**954**	1	0	0	27	1	3	**969**	0	0	0
BIC	6	42	245	**703**	4	0	0	7	5	161	**826**	1	0	0	4	1	122	**873**	0	0	0
fAIC	74	40	236	**648**	2	0	0	94	2	155	**748**	1	0	0	119	1	115	**765**	0	0	0

Remark 6.1. In each case of Section 6.2, we have not paid attention to Assumption 4.10, which may not be so easy-to-verify; we refer to [43] for several general criterion for the non-degeneracy of the statistical random fields in the present context. Let us mention almost sure lower bounds of $\det(\int_0^1 X_t X_t' \, dt)$ for the models considered in Sections 6.2.2. Let $X_{1,0} = a > 0$ (either 1 or 10; case of $a = 0$ is not relevant here). Then, because of the Schwarz inequality

$$\det\left(\int_0^1 X_t X_t' \, dt\right) = \det \begin{pmatrix} a^2 & a\int_0^1 X_{2,t}\, dt & a\int_0^1 X_{3,t}\, dt \\ a\int_0^1 X_{2,t}\, dt & \int_0^1 X_{2,t}^2\, dt & \int_0^1 X_{2,t}X_{3,t}\, dt \\ a\int_0^1 X_{3,t}\, dt & \int_0^1 X_{2,t}X_{3,t}\, dt & \int_0^1 X_{3,t}^2\, dt \end{pmatrix}$$

Table 8. The mean and the standard deviation (s.d.) of the estimator $\hat{\theta}_{1,n}$, $\hat{\theta}_{2,n}$ and $\hat{\theta}_{3,n}$ in Section 6.2.2 (i–2) for various n (1–7 express the models, and the true parameter $\theta_0 = (0, -2, 3)$)

		$n=200$			$n=500$			$n=1000$		
		$\hat{\theta}_{1,n}$	$\hat{\theta}_{2,n}$	$\hat{\theta}_{3,n}$	$\hat{\theta}_{1,n}$	$\hat{\theta}_{2,n}$	$\hat{\theta}_{3,n}$	$\hat{\theta}_{1,n}$	$\hat{\theta}_{2,n}$	$\hat{\theta}_{3,n}$
1	mean	−0.0242	−1.7665	2.9373	−0.0091	−1.9145	2.9985	−0.0080	−1.9306	2.9845
	s.d.	0.2290	2.2493	0.8445	0.1366	1.3373	0.5130	0.1028	1.0048	0.3855
2	mean	0.0620	−2.5040	–	0.0870	−2.7296	–	0.1050	−2.8925	–
	s.d.	0.6529	6.3498	–	0.6370	6.1847	–	0.6329	6.1223	–
3	mean	−0.2047	–	2.9181	−0.2029	–	2.9897	−0.2048	–	3.0437
	s.d.	0.0471	–	1.2826	0.0473	–	1.2787	0.0503	–	1.2950
4*	mean	–	−1.9948	2.9628	–	−1.9976	2.9963	–	−2.0038	2.9883
	s.d.	–	0.1712	0.3275	–	0.1053	0.1918	–	0.0751	0.1424
5	mean	−0.1029	–	–	−0.0946	–	–	−0.0871	–	–
	s.d.	0.1484	–	–	0.1527	–	–	0.1518	–	–
6	mean	–	−1.9063	–	–	−1.8867	–	–	−1.8722	–
	s.d.	–	0.4801	–	–	0.4622	–	–	0.4716	–
7	mean	–	–	3.0147	–	–	2.9964	–	–	2.9648
	s.d.	–	–	0.5319	–	–	0.5530	–	–	0.5440

$$
\begin{aligned}
&= a^2\Bigg[\left\{\int_0^1\left(X_{2,t}-\int_0^1 X_{2,t}\,dt\right)^2 dt\right\}\left\{\int_0^1\left(X_{3,t}-\int_0^1 X_{3,t}\,dt\right)^2 dt\right\}\\
&\quad-\left(\int_0^1 X_{2,t}X_{3,t}\,dt-\int_0^1 X_{2,t}\,dt\int_0^1 X_{3,t}\,dt\right)^2\Bigg]\\
&\geq a^2\Bigg[\left\{\int_0^1\left(X_{2,t}-\int_0^1 X_{2,t}\,dt\right)\left(X_{3,t}-\int_0^1 X_{3,t}\,dt\right)dt\right\}^2\\
&\quad-\left(\int_0^1 X_{2,t}X_{3,t}\,dt-\int_0^1 X_{2,t}\,dt\int_0^1 X_{3,t}\,dt\right)^2\Bigg]=0.
\end{aligned}
$$

Hence, $\det(\int_0^1 X_t X_t'\,dt)=0$ holds if and only if at least one of the following conditions is satisfied:

(i) $X_{2,t}-\int_0^1 X_{2,t}\,dt=0$ for all $t\in[0,1]$;
(ii) $X_{3,t}-\int_0^1 X_{3,t}\,dt=0$ for all $t\in[0,1]$;
(iii) There exists a constant $c\neq 0$, $X_{2,t}-\int_0^1 X_{2,t}\,dt=c(X_{3,t}-\int_0^1 X_{3,t}\,dt)$ for all $t\in[0,1]$.

6.3. Non-ergodic diffusion process

Let $\mathbf{X}_n=(X_{t_j})_{j=0}^n$ be a data set with $t_j=j/n$ and the number of data n, sampled from a solution to

$$dX_t=\exp\left\{\frac{5+2X_t}{2(1+X_t^2)}\right\}dw_t,\qquad t\in[0,1],\ X_0=0.$$

We consider the following models:

$$\textbf{Model 1}:\quad dX_t=\exp\left\{\frac{\theta_1+\theta_2X_t+\theta_3X_t^2}{2(1+X_t^2)}\right\}dw_t;$$

$$\textbf{Model 2}:\quad dX_t=\exp\left\{\frac{\theta_1+\theta_2X_t}{2(1+X_t^2)}\right\}dw_t;\qquad \textbf{Model 3}:\quad dX_t=\exp\left\{\frac{\theta_1+\theta_3X_t^2}{2(1+X_t^2)}\right\}dw_t;$$

$$\textbf{Model 4}:\quad dX_t=\exp\left\{\frac{\theta_2X_t+\theta_3X_t^2}{2(1+X_t^2)}\right\}dw_t;\qquad \textbf{Model 5}:\quad dX_t=\exp\left\{\frac{\theta_1}{2(1+X_t^2)}\right\}dw_t;$$

$$\textbf{Model 6}:\quad dX_t=\exp\left\{\frac{\theta_2X_t}{2(1+X_t^2)}\right\}dw_t;\qquad \textbf{Model 7}:\quad dX_t=\exp\left\{\frac{\theta_2X_t^2}{2(1+X_t^2)}\right\}dw_t.$$

Then the optimal model is Model 2, the true parameter being $\theta_0=(5,2,0)$. Table 9 shows that Model 2 is chosen with high probability as the best model for all criteria, where QBIC tends to take values between BIC and fAIC. As before, the larger the sample size becomes, the higher the frequency that the true model is selected by QBIC and BIC become. In Table 10, the estimators exhibit similar tendencies to Tables 6 and 8.

Table 9. The number of models selected by QBIC, BIC and AIC in Section 6.3 over 1000 simulations for various n (1–7 express the models, and the true model is Model 2)

	$n=200$							$n=500$							$n=1000$						
Criterion	1	2*	3	4	5	6	7	1	2*	3	4	5	6	7	1	2*	3	4	5	6	7
QBIC	291	**690**	19	0	0	0	0	151	**832**	17	0	0	0	0	115	**874**	11	0	0	0	0
BIC	30	**733**	237	0	0	0	0	15	**842**	143	0	0	0	0	20	**892**	88	0	0	0	0
fAIC	130	**642**	228	0	0	0	0	135	**728**	137	0	0	0	0	151	**767**	82	0	0	0	0

Remark 6.2. We write $Z_t = (\frac{1}{1+X_t^2}, \frac{X_t}{1+X_t^2}, \frac{X_t^2}{1+X_t^2})'$. In a similar way to Remark 6.1, we can see that

$$\det\left(\int_0^T Z_t Z_t'\,dt\right) = \det\begin{pmatrix} \int_0^1 \frac{1}{(1+X_t^2)^2}\,dt & \int_0^1 \frac{X_t}{(1+X_t^2)^2}\,dt & \int_0^1 \frac{X_t^2}{(1+X_t^2)^2}\,dt \\ \int_0^1 \frac{X_t}{(1+X_t^2)^2}\,dt & \int_0^1 \frac{X_t^2}{(1+X_t^2)^2}\,dt & \int_0^1 \frac{X_t^3}{(1+X_t^2)^2}\,dt \\ \int_0^1 \frac{X_t^2}{(1+X_t^2)^2}\,dt & \int_0^1 \frac{X_t^3}{(1+X_t^2)^2}\,dt & \int_0^1 \frac{X_t^4}{(1+X_t^2)^2}\,dt \end{pmatrix}$$

Table 10. The mean and the standard deviation (s.d.) of the estimator $\hat{\theta}_{1,n}$, $\hat{\theta}_{2,n}$ and $\hat{\theta}_{3,n}$ in Section 6.3 for various n (1–7 express the models, and the true parameter $\theta_0 = (5, 2, 0)$)

		$n=200$			$n=500$			$n=1000$		
		$\hat{\theta}_{1,n}$	$\hat{\theta}_{2,n}$	$\hat{\theta}_{3,n}$	$\hat{\theta}_{1,n}$	$\hat{\theta}_{2,n}$	$\hat{\theta}_{3,n}$	$\hat{\theta}_{1,n}$	$\hat{\theta}_{2,n}$	$\hat{\theta}_{3,n}$
1	mean	4.7972	1.6082	0.0057	4.9333	1.7426	−0.0238	4.9679	1.8711	−0.0051
	s.d.	0.8198	1.2508	0.5725	0.5754	0.8357	0.3637	0.4783	0.6970	0.2835
2*	mean	4.9551	1.8910	–	5.0031	1.9385	–	5.0135	1.9713	–
	s.d.	0.7156	0.4938	–	0.4686	0.3138	–	0.3416	0.2260	–
3	mean	4.7796	–	−0.0972	4.8136	–	−0.1076	4.8542	–	−0.1369
	s.d.	1.0889	–	0.7943	0.9700	–	0.7772	0.9593	–	0.7626
4	mean	–	−0.3651	1.1604	–	−0.2268	1.4376	–	0.0324	1.2284
	s.d.	–	4.1514	2.8986	–	3.6764	2.7531	–	3.8005	2.6084
5	mean	4.9133	–	–	4.9294	–	–	4.9231	–	–
	s.d.	0.5360	–	–	0.5259	–	–	0.5477	–	–
6	mean	–	1.6946	–	–	1.7188	–	–	1.7393	–
	s.d.	–	0.4614	–	–	0.4670	–	–	0.4742	–
7	mean	–	–	0.4908	–	–	0.4826	–	–	0.4926
	s.d.	–	–	0.2881	–	–	0.2782	–	–	0.2872

$$
\begin{aligned}
&= \left(\int_0^1 \frac{1}{(1+X_t^2)^2}\,dt\right) \\
&\quad \times \left[\left\{\int_0^1 \frac{1}{(1+X_t^2)^2}\left(X_t - \left(\int_0^1 \frac{1}{(1+X_t^2)^2}\,dt\right)^{-1}\int_0^1 \frac{X_t}{(1+X_t^2)^2}\,dt\right)^2 dt\right\}\right. \\
&\quad \times \left\{\int_0^1 \frac{1}{(1+X_t^2)^2}\left(X_t^2 - \left(\int_0^1 \frac{1}{(1+X_t^2)^2}\,dt\right)^{-1}\int_0^1 \frac{X_t^2}{(1+X_t^2)^2}\,dt\right)^2 dt\right\} \\
&\quad \left. - \left(\int_0^1 \frac{X_t^3}{(1+X_t^2)^2}\,dt - \left(\int_0^1 \frac{1}{(1+X_t^2)^2}\,dt\right)^{-1}\int_0^1 \frac{X_t}{(1+X_t^2)^2}\,dt\int_0^1 \frac{X_t^2}{(1+X_t^2)^2}\,dt\right)^2\right] \\
&\geq 0,
\end{aligned}
$$

with the last equality holding if and only if at least one of the following holds true for all $t \in [0, 1]$:

(i) $X_t - (\int_0^1 \frac{1}{(1+X_t^2)^2}\,dt)^{-1} \int_0^1 \frac{X_t}{(1+X_t^2)^2}\,dt = 0$;

(ii) $X_t^2 - (\int_0^1 \frac{1}{(1+X_t^2)^2}\,dt)^{-1} \int_0^1 \frac{X_t^2}{(1+X_t^2)^2}\,dt = 0$;

(iii) There exists a constant $c \neq 0$ such that

$$
\begin{aligned}
&X_t - \left(\int_0^1 \frac{1}{(1+X_t^2)^2}\,dt\right)^{-1}\int_0^1 \frac{X_t}{(1+X_t^2)^2}\,dt \\
&\quad = c\left\{X_t^2 - \left(\int_0^1 \frac{1}{(1+X_t^2)^2}\,dt\right)^{-1}\int_0^1 \frac{X_t^2}{(1+X_t^2)^2}\,dt\right\}.
\end{aligned}
$$

7. Proofs

Recall that $\mathbb{U}_n(\theta_0) = \{u \in \mathbb{R}^p; \theta_0 + A_n(\theta_0)u \in \Theta\}$. In what follows, we deal with the zero-extended version of $\mathbb{Z}_n$ and use the same notation: $\mathbb{Z}_n$ vanishes outside $\mathbb{U}_n(\theta_0)$, so that

$$
\int_{\mathbb{R}^p \setminus \mathbb{U}_n(\theta_0)} \mathbb{Z}_n(u)\,du = 0.
$$

7.1. Proof of Theorem 3.4(ii)

By using the Taylor expansion, we obtain

$$
\mathbb{Z}_n(u) = \exp\left(\Delta_n[u] - \frac{1}{2}\Gamma_n(\tilde{\theta}_n)[u, u]\right)
$$

for a random point $\tilde{\theta}_n$ on the segment connecting θ_0 and $\theta_0 + A_n(\theta_0)u$. Then, for any positive ε, δ and M, we have

$$\begin{aligned}
&\mathbb{P}\biggl(\int_{\mathbb{U}_n(\theta_0)\cap\{|u|\geq M\}} \mathbb{Z}_n(u)\,du > \varepsilon\biggr)\\
&\quad\leq \mathbb{P}\biggl(\int_{|u|\geq M} \mathbb{Z}_n(u)\,du > \varepsilon;\ \inf_{\theta\in\Theta} \lambda_{\min}\bigl(\Gamma_n(\theta)\bigr) < \delta\biggr)\\
&\qquad + \mathbb{P}\biggl(\int_{|u|\geq M} \mathbb{Z}_n(u)\,du > \varepsilon;\ \inf_{\theta\in\Theta} \lambda_{\min}\bigl(\Gamma_n(\theta)\bigr) \geq \delta\biggr)\\
&\quad\leq \mathbb{P}\Bigl(\inf_{\theta\in\Theta} \lambda_{\min}\bigl(\Gamma_n(\theta)\bigr) < \delta\Bigr) + \mathbb{P}\biggl\{\int_{|u|\geq M} \exp\biggl(\Delta_n[u] - \frac{\delta}{2}[u,u]\biggr)\,du > \varepsilon\biggr\}.
\end{aligned}$$

Under (3.4), we can find δ and N' for which the first term in the rightmost side can be bounded by $\varepsilon/2$ for $n \geq N'$. Given such a δ, making use of the tightness of (Δ_n) we can take N'' and $M > 0$ large enough to ensure

$$\sup_{n\geq N''} \mathbb{P}\biggl\{\int_{|u|\geq M} \exp\biggl(\Delta_n[u] - \frac{\delta}{2}[u,u]\biggr)\,du > \varepsilon\biggr\} < \frac{\varepsilon}{2}.$$

Hence, we have $\sup_{n\geq N} \mathbb{P}(\int_{\mathbb{U}_n(\theta_0)\cap\{|u|\geq M\}} \mathbb{Z}_n(u)\,du > \varepsilon) < \varepsilon$ for $N := N' \vee N''$, completing the proof.

7.2. Proof of Theorem 3.7

(i) By the change of variable $\theta = \theta_0 + A_n(\theta_0)u$, the marginal quasi-log likelihood function equals

$$\mathbb{H}_n(\theta_0) + \sum_{k=1}^{K} p_k \log a_{k,n}(\theta_{k,0}) + \log\biggl(\int_{\mathbb{U}_n(\theta_0)} \mathbb{Z}_n(u)\pi_n\bigl(\theta_0 + A_n(\theta_0)u\bigr)\,du\biggr).$$

Consequently,

$$\begin{aligned}
&\log\biggl(\int_\Theta \exp\bigl\{\mathbb{H}_n(\theta)\bigr\}\pi_n(\theta)\,d\theta\biggr) - \biggl(\mathbb{H}_n(\theta_0) + \sum_{k=1}^{K} p_k \log a_{k,n}(\theta_{k,0}) + \log \bar{Q}_n\biggr)\\
&\quad= \log(\bar{Q}_n + \bar{\varepsilon}_n) - \log \bar{Q}_n,
\end{aligned}$$

where

$$\begin{aligned}
\bar{Q}_n &= \pi_n(\theta_0)\int_{\mathbb{R}^p} \exp\biggl(\Delta_n[u] - \frac{1}{2}\Gamma_0[u,u]\biggr)\,du,\\
\bar{\varepsilon}_n &= \int_{\mathbb{U}_n(\theta_0)} \mathbb{Z}_n(u)\bigl(\pi_n\bigl(\theta_0 + A_n(\theta_0)u\bigr) - \pi_n(\theta_0)\bigr)\,du\\
&\qquad + \pi_n(\theta_0)\int_{\mathbb{R}^p} \biggl\{\mathbb{Z}_n(u) - \exp\biggl(\Delta_n[u] - \frac{1}{2}\Gamma_0[u,u]\biggr)\biggr\}\,du.
\end{aligned}$$

First, we note that

$$\begin{aligned}\bar{Q}_n &= \pi_n(\theta_0)\exp\left(\frac{1}{2}\left\|\Gamma_0^{-\frac{1}{2}}\Delta_n\right\|^2\right)\int_{\mathbb{R}^p}\exp\left(-\frac{1}{2}\Gamma_0\big[u-\Gamma_0^{-1}\Delta_n, u-\Gamma_0^{-1}\Delta_n\big]\right)du\\ &= \pi_n(\theta_0)\exp\left(\frac{1}{2}\left\|\Gamma_0^{-\frac{1}{2}}\Delta_n\right\|^2\right)(2\pi)^{\frac{p}{2}}|\Gamma_0|^{-\frac{1}{2}},\end{aligned}$$

so that

$$\log\bar{Q}_n = \log\pi_n(\theta_0) + \frac{1}{2}\left\|\Gamma_0^{-\frac{1}{2}}\Delta_n\right\|^2 + \frac{p}{2}\log(2\pi) - \frac{1}{2}\log|\Gamma_0|.$$

Hence, it remains to show that $|\log(\bar{Q}_n+\bar{\varepsilon}_n) - \log\bar{Q}_n| \xrightarrow{P} 0$.

Observe that

$$\begin{aligned}|\bar{\varepsilon}_n| &\le \int_{\mathbb{U}_n(\theta_0)}\mathbb{Z}_n(u)\big|\pi_n\big(\theta_0+A_n(\theta_0)u\big)-\pi_n(\theta_0)\big|\,du\\ &\quad+\pi_n(\theta_0)\int_{\mathbb{R}^p}\left|\mathbb{Z}_n(u)-\exp\left(\Delta_n[u]-\frac{1}{2}\Gamma_0[u,u]\right)\right|du.\end{aligned}$$

Fix any $\varepsilon>0$. Then, for each $M>0$ we have

$$\begin{aligned}&\mathbb{P}\left(\int_{\mathbb{U}_n(\theta_0)}\mathbb{Z}_n(u)\big|\pi_n\big(\theta_0+A_n(\theta_0)u\big)-\pi_n(\theta_0)\big|\,du>\varepsilon\right)\\ &\le \mathbb{P}\left((2M)^p\sup_{|u|<M}\big|\pi_n\big(\theta_0+A_n(\theta_0)u\big)-\pi_n(\theta_0)\big|\sup_{|u|<M}\mathbb{Z}_n(u)>\frac{\varepsilon}{2}\right)\\ &\quad+\mathbb{P}\left(2\sup_\theta\pi_n(\theta)\int_{|u|\ge M}\mathbb{Z}_n(u)\,du>\frac{\varepsilon}{2}\right).\end{aligned}\tag{7.1}$$

Let $r_n(u) := \frac{1}{2}(\Gamma_0-\Gamma_n)[u,u] + \frac{1}{6}\sum_{i,j,k=1}^p(\partial_{\theta^i}\partial_{\theta^j}\partial_{\theta^k}\mathbb{H}_n(\tilde{\theta}_n))A_{n,ii}(\theta_0)A_{n,jj}(\theta_0)A_{n,kk}(\theta_0)\times u_iu_ju_k$ for a point $\tilde{\theta}_n$ between θ_0 and $\theta_0+A_n(\theta_0)u$. Then, under the assumptions we have $\sup_{|u|<K_0}|r_n(u)|\xrightarrow{P}0$ for every $K_0>0$. We may write

$$\sup_{|u|<M}\mathbb{Z}_n(u) = \sup_{|u|<M}\exp\left(\Delta_n[u]-\frac{1}{2}\Gamma_0[u,u]+r_n(u)\right)$$

and this quantity equals $O_p(1)$, so that $\sup_{|u|<M}|\pi_n(\theta_0+A_n(\theta_0)u)-\pi_n(\theta_0)|\sup_{|u|<M}\mathbb{Z}_n(u) = o_p(1)$ for each $M>0$. Under Assumption 3.3, we can take a sufficiently large M to conclude that

$$\mathbb{P}\left(\int_{\mathbb{U}_n(\theta_0)}\mathbb{Z}_n(u)\big|\pi_n\big(\theta_0+A_n(\theta_0)u\big)-\pi_n(\theta_0)\big|\,du>\varepsilon\right) < \frac{\varepsilon}{2(2M)^p}+\frac{\varepsilon}{4\sup_\theta\pi_n(\theta)} \lesssim \varepsilon,$$

where "$a_n \lesssim b_n$" for positive sequences a_n and b_n means that "$\sup_n (a_n/b_n) \le C$" for a universal positive constant C. Then it follows that $\int_{\mathbb{U}_n(\theta_0)} \mathbb{Z}_n(u)|\pi_n(\theta_0 + A_n(\theta_0)u) - \pi_n(\theta_0)|\,du \xrightarrow{P} 0$. Next, for any $\delta > 0$ and $K > 0$,

$$\begin{aligned}
&\mathbb{P}\left\{\int_{\mathbb{R}^p} \left|\mathbb{Z}_n(u) - \exp\left(\Delta_n[u] - \frac{1}{2}\Gamma_0[u,u]\right)\right| du > \delta\right\} \\
&\quad \le \mathbb{P}\left\{\int_{|u|<K} \left|\mathbb{Z}_n(u) - \exp\left(\Delta_n[u] - \frac{1}{2}\Gamma_0[u,u]\right)\right| du > \frac{\delta}{2}\right\} + \mathbb{P}\left(\int_{|u|\ge K} \mathbb{Z}_n(u)\,du > \frac{\delta}{4}\right) \\
&\qquad + \mathbb{P}\left\{\int_{|u|\ge K} \exp\left(\Delta_n[u] - \frac{1}{2}\Gamma_0[u,u]\right) du > \frac{\delta}{4}\right\}.
\end{aligned}$$

We can pick $K > 0$ and N'' large enough to ensure

$$\sup_{n \ge N''} \mathbb{P}\left(\int_{|u|\ge K} \mathbb{Z}_n(u)\,du > \frac{\delta}{4}\right) < \frac{\delta}{4}, \tag{7.2}$$

$$\sup_{n \ge N''} \mathbb{P}\left\{\int_{|u|\ge K} \exp\left(\Delta_n[u] - \frac{1}{2}\Gamma_0[u,u]\right) du > \frac{\delta}{4}\right\} < \frac{\delta}{4}. \tag{7.3}$$

Since $\Delta_n[u] - \frac{1}{2}\Gamma_0[u,u] \le \frac{1}{2}\Delta_n'\Gamma_0^{-1}\Delta_n$ with equality holding if and only if $u = \Gamma_0^{-1}\Delta_n$, for the same $K > 0$ as above we get

$$\begin{aligned}
&\int_{|u|<K} \left|\mathbb{Z}_n(u) - \exp\left(\Delta_n[u] - \frac{1}{2}\Gamma_0[u,u]\right)\right| du \\
&\quad \lesssim \sup_{|u|<K} \left|\exp\left(\Delta_n[u] - \frac{1}{2}\Gamma_0[u,u]\right)\left\{\exp\big(r_n(u)\big) - 1\right\}\right| \\
&\quad \le \sup_{|u|<K} \left|\exp\big(r_n(u)\big) - 1\right| \exp\left(\frac{1}{2}\Delta_n'\Gamma_0^{-1}\Delta_n\right) \\
&\quad \xrightarrow{P} 0.
\end{aligned} \tag{7.4}$$

Because of (7.2) to (7.4) we have $\int_{\mathbb{R}^p} |\mathbb{Z}_n(u) - \exp(\Delta_n[u] - \frac{1}{2}\Gamma_0[u,u])|\,du \xrightarrow{P} 0$, hence $\bar{\varepsilon}_n \xrightarrow{P} 0$ and it follows that

$$\log(\bar{Q}_n + \bar{\varepsilon}_n) - \log \bar{Q}_n = \big(\log \bar{Q}_n + o_p(1)\big) - \log \bar{Q}_n = o_p(1),$$

establishing the claim (i).

(ii) By the consistency of $\hat{\theta}_n$ we may focus on the event $\{\hat{\theta}_n \in \Theta\}$ ($\subset \{\partial_\theta \mathbb{H}_n(\hat{\theta}_n) = 0\}$). Then

$$\Delta_n = -A_n(\theta_0) \int_0^1 \partial_\theta^2 \mathbb{H}_n\big(\hat{\theta}_n + s(\theta_0 - \hat{\theta}_n)\big)\,ds A_n(\theta_0)[\hat{u}_n] = \big\{\Gamma_0 + o_p(1)\big\}[\hat{u}_n],$$

so that $\hat{u}_n = \Gamma_0^{-1}\Delta_n + o_p(1) = O_p(1)$. Therefore,

$$\begin{aligned}\mathbb{H}_n(\theta_0) &= \mathbb{H}_n(\hat{\theta}_n) - \frac{1}{2}\hat{u}_n'\Gamma_0\hat{u}_n + o_p(1)\\ &= \mathbb{H}_n(\hat{\theta}_n) - \frac{1}{2}\big(\Gamma_0^{-1}\Delta_n\big)'\Gamma_0\big(\Gamma_0^{-1}\Delta_n\big) + o_p(1)\\ &= \mathbb{H}_n(\hat{\theta}_n) - \frac{1}{2}\big\|\Gamma_0^{-\frac{1}{2}}\Delta_n\big\|^2 + o_p(1),\end{aligned}$$

which combined with the preceding result (i) and the fact $-A_n(\hat{\theta}_n)\partial_\theta^2\mathbb{H}_n(\hat{\theta}_n)A_n(\hat{\theta}_n) = \Gamma_0 + o_p(1)$ establishes (ii).

Remark 7.1. Recall that the Bayes point estimator associated with a loss function $\mathfrak{L} : \Theta \times \Theta \to \mathbb{R}$ is defined to be any statistics $\tilde{\theta}_n(\mathfrak{L})$ minimizing the random function

$$t \mapsto \int_\Theta \mathfrak{L}(t,\theta)\pi_n(\theta|\mathbf{X}_n)\,d\theta,$$

where

$$\pi_n(\theta|\mathbf{X}_n) := \frac{\exp\{\mathbb{H}_n(\theta)\}\pi_n(\theta)}{\int_\Theta \exp\{\mathbb{H}_n(\theta)\}\pi_n(\theta)\,d\theta}$$

denotes the quasi-posterior density of θ; in particular, the quadratic loss $\mathfrak{L}_2(t,\theta) := |t-\theta|^2$ gives rise to the posterior-mean:

$$\tilde{\theta}_n(\mathfrak{L}_2) := \frac{\int_\Theta \theta\exp\{\mathbb{H}_n(\theta)\}\pi_n(\theta)\,d\theta}{\int_\Theta \exp\{\mathbb{H}_n(\theta)\}\pi_n(\theta)\,d\theta}.$$

In the theoretical derivation of the QBIC, we made use of the fact that (at least with sufficiently high probability) $\partial_\theta\mathbb{H}_n(\hat{\theta}_n) = 0$, which does not hold true if we use an integral-type Bayes point estimator.

7.3. Proof of Theorem 3.15

Let

$$F_n := -2\log\bigg(\int_\Theta \exp\big\{\mathbb{H}_n(\theta)\big\}\pi_n(\theta)\,d\theta\bigg),$$

$$F_n' := -2\mathbb{H}_n(\theta_0) - 2\sum_{k=1}^K p_k\log a_{k,n}(\theta_{k,0}) + \log|\Gamma_0| - p\log 2\pi - \big\|\Gamma_0^{-\frac{1}{2}}\Delta_n\big\|^2 - 2\log\pi_n(\theta_0).$$

We complete the proof by showing that $\mathbb{E}(|F_n - F_n'|) \to 0$ and $\mathbb{E}(|F_n' - \mathrm{QBIC}_n^\sharp|) \to 0$ separately.

Proof of $\mathbb{E}(|F_n - F_n'|) \to 0$. Obviously Assumptions 3.11 and 3.13 imply Assumption 3.1, hence Theorem 3.7(i) yields that $F_n = F_n' + o_p(1)$.

Now pick any $\kappa \in (1, q/p)$ with q being the constant given in Assumption 3.13. To deduce the claim, it suffices to show that $\limsup_n \mathbb{E}(|F_n - F_n'|^\kappa) < \infty$. Then for some $\delta \in (0, 1/\kappa)$,

$$
\begin{aligned}
&\mathbb{E}\big(\big|F_n - F_n'\big|^\kappa\big) \\
&\lesssim 1 + \mathbb{E}\bigg(\bigg|\log\bigg(\int_{\mathbb{U}_n(\theta_0)} \mathbb{Z}_n(u)\,du\bigg)\bigg|^\kappa\bigg) + \mathbb{E}\big(\big|\log|\Gamma_0|\big|^\kappa\big) + \mathbb{E}\big(\big|\Gamma_0^{-1}\big[\Delta_n^{\otimes 2}\big]\big|^\kappa\big) \\
&\lesssim 1 + \mathbb{E}\bigg[\bigg\{-\log\bigg(\int_{\mathbb{U}_n(\theta_0)} \mathbb{Z}_n(u)\,du\bigg)\bigg\}^\kappa; \int_{\mathbb{U}_n(\theta_0)} \mathbb{Z}_n(u)\,du \le 1\bigg] \\
&\quad + \mathbb{E}\bigg[\bigg\{\log\bigg(\int_{\mathbb{U}_n(\theta_0)} \mathbb{Z}_n(u)\,du\bigg)\bigg\}^\kappa; \int_{\mathbb{U}_n(\theta_0)} \mathbb{Z}_n(u)\,du > 1\bigg] \\
&\quad + \mathbb{E}\big(\lambda_{\min}^{-p\kappa}(\Gamma_0) + |\Gamma_0|^\kappa\big) + \mathbb{E}\big(|\Delta_n|^{2\kappa}\lambda_{\min}^{-p\kappa}(\Gamma_0)\big) \\
&\lesssim 1 + \mathbb{E}\bigg\{\bigg(\int_{\mathbb{U}_n(\theta_0)} \mathbb{Z}_n(u)\,du\bigg)^{-\delta\kappa}; \int_{\mathbb{U}_n(\theta_0)} \mathbb{Z}_n(u)\,du \le 1\bigg\} \\
&\quad + \mathbb{E}\bigg[\bigg\{\log\bigg(\int_{\mathbb{U}_n(\theta_0)} \exp\bigg(\Delta_n[u] - \frac{1}{2}\inf_{\theta\in\Theta}\lambda_{\min}\big(\Gamma_n(\theta)\big)[u,u]\bigg)\,du\bigg)\bigg\}^\kappa; \\
&\quad \int_{\mathbb{U}_n(\theta_0)} \mathbb{Z}_n(u)\,du > 1\bigg] \qquad (7.5) \\
&\lesssim 1 + \mathbb{E}\bigg\{\bigg(\int_{\mathbb{U}_n(\theta_0)} \mathbb{Z}_n(u)\,du\bigg)^{-1}\bigg\} \\
&\quad + \mathbb{E}\bigg[\bigg\{\Big(\inf_{\theta\in\Theta}\lambda_{\min}\big(\Gamma_n(\theta)\big)\Big)^{-1}\big[\Delta_n^{\otimes 2}\big] + \log\bigg(\int_{\mathbb{R}^p} \exp\bigg(-\frac{1}{2}\inf_{\theta\in\Theta}\lambda_{\min}\big(\Gamma_n(\theta)\big) \\
&\quad \times \Big[\Big(u - \Big(\inf_{\theta\in\Theta}\lambda_{\min}\big(\Gamma_n(\theta)\big)\Big)^{-1}\Delta_n\Big)^{\otimes 2}\Big]\bigg)\,du\bigg)\bigg\}^\kappa\bigg] \\
&\lesssim 1 + \mathbb{E}\Big[\Big\{\Big(\inf_{\theta\in\Theta}\lambda_{\min}\big(\Gamma_n(\theta)\big)\Big)^{-1}\big[\Delta_n^{\otimes 2}\big]\Big\}^\kappa + \Big|\log\Big(\inf_{\theta\in\Theta}\lambda_{\min}\big(\Gamma_n(\theta)\big)\Big)\Big|^\kappa\Big] \\
&\le 1 + \mathbb{E}\Big[\Big(\inf_{\theta\in\Theta}\lambda_{\min}\big(\Gamma_n(\theta)\big)\Big)^{-\kappa}|\Delta_n|^{2\kappa} \\
&\quad + \Big(\inf_{\theta\in\Theta}\lambda_{\min}\big(\Gamma_n(\theta)\big)\Big)^{\kappa} + \Big(\inf_{\theta\in\Theta}\lambda_{\min}\big(\Gamma_n(\theta)\big)\Big)^{-\kappa}\Big] \\
&\le 1 + \mathbb{E}\Big[\Big(\sup_{\theta\in\Theta}\lambda_{\min}^{-\kappa}\big(\Gamma_n(\theta)\big)\Big)\big(|\Delta_n|^{2\kappa} + 1\big) + \sup_{\theta\in\Theta}\big|\Gamma_n(\theta)\big|^\kappa\Big] \\
&\lesssim 1,
\end{aligned}
$$

where: in the fifth step, we applied [47], Lemma 2, for the second term; in the second step and the sixth step, we made use of the inequality: for any $s > 0$, $|\log x| \lesssim x^{-s} + x^{s}$ for $x > 0$.

Proof of $\mathbb{E}(|F_n' - \mathrm{QBIC}_n^{\sharp}|) \to 0$. We have $|F_n' - \mathrm{QBIC}_n^{\sharp}| \lesssim \overline{R}_{1,n} + \overline{R}_{2,n} + \overline{R}_{3,n}$, where

$$\begin{aligned}
\overline{R}_{1,n} &:= \left|\log \frac{\pi_n(\hat{\theta}_n)}{\pi_n(\theta_0)}\right| + \sup_{\theta} \left|\partial_\theta^3 \mathbb{H}_n(\theta)\left[(\hat{\theta}_n - \theta_0)^{\otimes 3}\right]\right| + \left|\partial_\theta \mathbb{H}_n(\hat{\theta}_n)[\hat{\theta}_n - \theta_0]\right|, \\
\overline{R}_{2,n} &:= \left|\log\left|\Gamma_n(\hat{\theta}_n)\right| - \log|\Gamma_0|\right|, \\
\overline{R}_{3,n} &:= \left|\Gamma_n(\hat{\theta}_n)\left[\hat{u}_n^{\otimes 2}\right] - \Gamma_0^{-1}\left[\Delta_n^{\otimes 2}\right]\right|.
\end{aligned}$$

We will show that $\mathbb{E}(\overline{R}_{i,n}) \to 0$ for $i = 1, 2, 3$ separately.

First, we look at $\overline{R}_{1,n}$. The convergence $\pi_n(\hat{\theta}_n)/\pi_n(\theta_0) \xrightarrow{P} 1$ holds under Assumption 3.12: indeed, for any $\varepsilon > 0$ and $M > 0$ we have $\mathbb{P}(|\pi_n(\hat{\theta}_n)/\pi_n(\theta_0) - 1| > \varepsilon) \le \sup_n \mathbb{P}(|\hat{u}_n| > M) + \mathbb{P}\{(|\hat{u}_n| \le M) \cap (\sup_{|u| \le M} |\pi_n(\theta_0 + A_n(\theta_0)u) - \pi_n(\theta_0)| \ge C\varepsilon)\}$ for some constant $C > 0$, from which the claim follows on letting M large enough and then $n \to \infty$. Then,

$$\lim_n \mathbb{E}\left(\left|\log \frac{\pi_n(\hat{\theta}_n)}{\pi_n(\theta_0)}\right|\right) \to 0$$

by the bounded convergence theorem. We are assuming that $(\hat{u}_n)_n$ is $L^r(\mathbb{P})$-bounded for some $r > 3$ (Assumption 3.14), hence under the assumptions we can apply Hölder's inequality to deduce

$$\begin{aligned}
&\mathbb{E}\Big(\sup_{\theta} \left|\partial_\theta^3 \mathbb{H}_n(\theta)\left[(\hat{\theta}_n - \theta_0)^{\otimes 3}\right]\right|\Big) \\
&\lesssim \Big(\max_{i \le p} A_{n,ii}(\theta_0)\Big)\mathbb{E}\Big(\sup_{\theta} \left|A_n(\theta_0)\partial_\theta^3 \mathbb{H}_n(\theta) A_n(\theta_0)\right| |\hat{u}_n|^3\Big) \\
&\lesssim o(1)\left\{\mathbb{E}\left(|\hat{u}_n|^r\right)\right\}^{\frac{3}{r}} = o(1)\|\hat{u}_n\|_r^3 \to 0.
\end{aligned}$$

Also, for any $s > 1$ small enough we have

$$\begin{aligned}
\mathbb{E}\left(\left|\partial_\theta \mathbb{H}_n(\hat{\theta}_n)[\hat{\theta}_n - \theta_0]\right|^s\right) &\lesssim \mathbb{E}\left(|\Delta_n|^s |\hat{u}_n|^s\right) + \mathbb{E}\Big(\sup_{\theta} \left|\Gamma_n(\theta)\right|^s |\hat{u}_n|^{2s}\Big) \\
&\lesssim \|\hat{u}_n\|_r^s + \|\hat{u}_n\|_r^{2s} \lesssim 1 + \|\hat{u}_n\|_r^{2s},
\end{aligned}$$

thereby $\mathbb{E}(|\partial_\theta \mathbb{H}_n(\hat{\theta}_n)[\hat{\theta}_n - \theta_0]|) \to 0$, concluding that $\mathbb{E}(\overline{R}_{1,n}) \to 0$.

For handling $\mathbb{E}(\overline{R}_{2,n})$, it suffices to observe that $\overline{R}_{2,n} = |\log(|\Gamma_n(\hat{\theta}_n)|/|\Gamma_0|)| \xrightarrow{P} 0$ and that for any $s' \in (1, q/p)$,

$$\begin{aligned}
\mathbb{E}\left(\overline{R}_{2,n}^{s'}\right) &\le \mathbb{E}\left(\left|\log\left|\Gamma_n(\hat{\theta}_n)\right|\right|^{s'} + \left|\log|\Gamma_0|\right|^{s'}\right) \\
&\lesssim \mathbb{E}\Big(\sup_{\theta \in \Theta} \lambda_{\min}^{-ps'}\left(\Gamma_n(\theta)\right) + \sup_{\theta \in \Theta} \left|\Gamma_n(\theta)\right|^{s'} + \lambda_{\min}^{-ps'}(\Gamma_0) + |\Gamma_0|^{s'}\Big) \lesssim 1.
\end{aligned}$$

Here, in the second step, we used the same inequality as in the second step of (7.5).

To deduce $\mathbb{E}(\overline{R}_{3,n}) \to 0$, we note that $\overline{R}_{3,n} = |\{\Gamma_0 + o_p(1)\}[\hat{u}_n^{\otimes 2}] - \Gamma_0^{-1}[\{\Gamma_0[\hat{u}_n]\}^{\otimes 2}]| = o_p(1)$, since $\hat{u}_n = \Gamma_0^{-1}\Delta_n + o_p(1)$ as was mentioned in the proof of Theorem 3.7(ii). The uniform integrability of $(\overline{R}_{3,n})_n$ can be verified in a similar manner to the previous case. The proof is complete.

7.4. Proof of Theorem 4.5

Under Assumptions 4.1 to 4.4, the argument in [47], Section 6, ensures the PLDI: for every $L > 0$ we can find a constant $C_L > 0$ such that

$$\mathbb{P}\Big(\sup_{(u_1,\theta_2)\in\{r\le|u_1|\}\times\Theta_2} \mathbb{Z}_n^1(u_1;\theta_{1,0},\theta_2) \ge e^{-r}\Big) + \mathbb{P}\Big(\sup_{u_2\in\{r\le|u_2|\}} \mathbb{Z}_n^2(u_2;\theta_{1,0},\theta_{2,0}) \ge e^{-r}\Big) \le \frac{C_L}{r^L}$$

for any $n > 0$ and $r > 0$. This implies that the inequality (3.3) holds (see Remark 3.6). Assumption 3.1 readily follows by making use of the lemmas in [47], Section 6, hence we omit them (see [15], Section 5.3, for some details).

7.5. Proof of Theorem 4.11

It is enough to check the conditions [H1] and [H2] of [43].

The condition [H1] is a regularity conditions concerning the processes X and b, and the non-degeneracy of the diffusion-coefficient function $S(x,\theta)$. As a consequence of Assumption 4.9(i) and the compactness of Θ, we get

$$\inf_{\omega\in\Omega, t\le T, \theta\in\Theta} \exp\big(X_t'\theta\big) > 0.$$

Based on this inequality, it is straightforward to verify [H1].

The condition [H2] is the non-degeneracy of the random field in the limit: for every $L > 0$, there exists $C_L > 0$ such that

$$\mathbb{P}\big(\chi_0 \le r^{-1}\big) \le \frac{C_L}{r^L}, \qquad r > 0,$$

where

$$\chi_0 := \inf_{\theta\ne\theta_0} \frac{1}{2T|\theta-\theta_0|^2} \int_0^T \big\{X_t'(\theta-\theta_0) + \big(\exp\big(X_t'(\theta_0-\theta)\big) - 1\big)\big\}\,dt.$$

Since $\exp(x) = 1 + x + \frac{1}{2}\exp(\xi x)x^2$ for some ξ satisfying $0 < \xi < 1$, letting $x = X_t'(\theta_0-\theta)$ we obtain

$$\begin{aligned} &X_t'(\theta-\theta_0) + \big\{\exp\big(X_t'(\theta_0-\theta)\big) - 1\big\} \\ &\quad = \frac{1}{2}\exp\big(\xi X_t'(\theta_0-\theta)\big)\big(X_t'(\theta_0-\theta)\big)^2 \end{aligned}$$

$$= \frac{1}{2}\exp\bigl(\xi X_t'(\theta_0-\theta)\bigr)(\theta_0-\theta)'X_tX_t'(\theta_0-\theta)$$
$$\geq \frac{1}{2}\exp(-C_0)(\theta_0-\theta)'X_tX_t'(\theta_0-\theta)$$

for some $C_0 > 0$. Hence,

$$\chi_0 \geq \frac{\exp(-C_0)}{4T}\inf_{\theta\neq\theta_0}\frac{\int_0^T(\theta_0-\theta)'X_tX_t'(\theta_0-\theta)\,dt}{|\theta-\theta_0|^2} \gtrsim \lambda_{\min}\biggl(\int_0^T X_tX_t'\,dt\biggr),$$

so that $\mathbb{P}(\chi_0 \leq r^{-1}) \leq \mathbb{P}\{\lambda_{\min}(\int_0^T X_tX_t'\,dt) \leq r^{-1}\} \lesssim C_L r^{-L}$. The proof is complete.

7.6. Proof of Theorem 5.1

We only prove (5.2) because (5.3) can be handled analogously. We basically follow Fasen and Kimmig [16].

(i) Let Θ_{m_0} be nested in Θ_m ($p_{m_0} < p_m$). Define the map $f : \Theta_{m_0} \to \Theta_m$ by $f(\theta_{m_0}) = F\theta_{m_0} + c$, where F and c satisfy that $\mathbb{H}_{m_0,n}(\theta_{m_0}) = \mathbb{H}_{m,n}(f(\theta_{m_0}))$ for any $\theta_{m_0} \in \Theta_{m_0}$. If $f(\theta_{m_0,0}) \neq \theta_{m,0}$, then $\mathbb{H}_{m_0,0}(\theta_{m_0,0}) = \mathbb{H}_{m,0}(f(\theta_{m_0,0})) < \mathbb{H}_{m,0}(\theta_{m,0})$ and the assumption of the optimal model is not satisfied. Hence, we have $f(\theta_{m_0,0}) = \theta_{m,0}$.

By the Taylor expansion of $\mathbb{H}_{m,n}$ around $\hat{\theta}_{m,n}$, we may write

$$\begin{aligned}\mathbb{H}_{m_0,n}(\hat{\theta}_{m_0,n}) &= \mathbb{H}_{m,n}\bigl(f(\hat{\theta}_{m_0,n})\bigr)\\ &= \mathbb{H}_{m,n}(\hat{\theta}_{m,n})\\ &\quad - \frac{1}{2}\bigl(\hat{\theta}_{m,n} - f(\hat{\theta}_{m_0,n})\bigr)'\bigl(-\partial^2_{\theta_m}\mathbb{H}_{m,n}(\tilde{\theta}_{m,n})\bigr)\bigl(\hat{\theta}_{m,n} - f(\hat{\theta}_{m_0,n})\bigr),\end{aligned}$$

where $\tilde{\theta}_{m,n} \xrightarrow{P} \theta_{m,0}$ as $n \to \infty$. Also, $f(\hat{\theta}_{m_0,n}) - \theta_{m,0} = f(\hat{\theta}_{m_0,n}) - f(\theta_{m_0,0}) = F(\hat{\theta}_{m_0,n} - \theta_{m_0,0})$. Since $a_n^{-1}(\hat{\theta}_{m,n} - f(\hat{\theta}_{m_0,n})) = a_n^{-1}(\hat{\theta}_{m,n} - \theta_{m,0}) - Fa_n^{-1}(\hat{\theta}_{m_0,n} - \theta_{m_0,0}) = O_p(1)$, $\Gamma_{m_0,n}(\hat{\theta}_{m_0,n}) = O_p(1)$ and $\Gamma_{m,n}(\hat{\theta}_{m,n}) = O_p(1)$, we have

$$\begin{aligned}&\mathbb{P}\bigl(\mathrm{QBIC}_n^{(m_0)} - \mathrm{QBIC}_n^{(m)} < 0\bigr)\\ &= \mathbb{P}\bigl\{\bigl(\hat{\theta}_{m,n} - f(\hat{\theta}_{m_0,n})\bigr)'\bigl(-\partial^2_{\theta_m}\mathbb{H}_{m,n}(\tilde{\theta}_{m,n})\bigr)\bigl(\hat{\theta}_{m,n} - f(\hat{\theta}_{m_0,n})\bigr)\\ &\quad + \log\det\bigl(-\partial^2_{\theta_{m_0}}\mathbb{H}_{m_0,n}(\hat{\theta}_{m_0,n})\bigr) - \log\det\bigl(-\partial^2_{\theta_m}\mathbb{H}_{m,n}(\hat{\theta}_{m,n})\bigr) < 0\bigr\}\\ &= \mathbb{P}\bigl[\bigl\{a_n^{-1}\bigl(\hat{\theta}_{m,n} - f(\hat{\theta}_{m_0,n})\bigr)\bigr\}'\bigl(-a_n^2\partial^2_{\theta_m}\mathbb{H}_{m,n}(\tilde{\theta}_{m,n})\bigr)\bigl\{a_n^{-1}\bigl(\hat{\theta}_{m,n} - f(\hat{\theta}_{m_0,n})\bigr)\bigr\}\\ &\quad + \log\det\bigl(-a_n^2\partial^2_{\theta_{m_0}}\mathbb{H}_{m_0,n}(\hat{\theta}_{m_0,n})\bigr)\\ &\quad - \log\det\bigl(-a_n^2\partial^2_{\theta_m}\mathbb{H}_{m,n}(\hat{\theta}_{m,n})\bigr) < p_m\log a_n^{-2} - p_{m_0}\log a_n^{-2}\bigr]\\ &= \mathbb{P}\bigl[\bigl\{a_n^{-1}\bigl(\hat{\theta}_{m,n} - f(\hat{\theta}_{m_0,n})\bigr)\bigr\}'\Gamma_{m,n}(\tilde{\theta}_{m,n})\bigl\{a_n^{-1}\bigl(\hat{\theta}_{m,n} - f(\hat{\theta}_{m_0,n})\bigr)\bigr\}\end{aligned}$$

$$+\log\det\bigl(\Gamma_{m_0,n}(\hat{\theta}_{m_0,n})\bigr)-\log\det\bigl(\Gamma_{m,n}(\hat{\theta}_{m,n})\bigr)<(p_m-p_{m_0})\log a_n^{-2}\bigr]$$
$$\to 1$$

as $n\to\infty$.

(ii) Let $\mathbb{H}_{m,0}(\theta_m)\neq\mathbb{H}_{m_0,0}(\theta_{m_0,0})$ a.s. for every $\theta_m\in\Theta_m$. Because of (5.1) and the convergences $\hat{\theta}_{m_0,n}\xrightarrow{P}\theta_{m_0,0}$ and $\hat{\theta}_{m,n}\xrightarrow{P}\theta_{m,0}$, we have

$$a_n^2\mathbb{H}_{m_0,n}(\hat{\theta}_{m_0,n})=a_n^2\mathbb{H}_{m_0,n}(\theta_{m_0,0})+o_p(1)=\mathbb{H}_{m_0,0}(\theta_{m_0,0})+o_p(1),$$
$$a_n^2\mathbb{H}_{m,n}(\hat{\theta}_{m,n})=a_n^2\mathbb{H}_{m,n}(\theta_{m,0})+o_p(1)=\mathbb{H}_{m,0}(\theta_{m,0})+o_p(1).$$

Since $\mathbb{H}_{m_0,0}(\theta_{m_0,0})>\mathbb{H}_{m,0}(\theta_{m,0})$ a.s. and $a_n^2\log a_n^{-2}\to 0$,

$$\begin{aligned}
&\mathbb{P}\bigl(\mathrm{QBIC}_n^{(m_0)}-\mathrm{QBIC}_n^{(m)}<0\bigr)\\
&=\mathbb{P}\bigl\{-2\mathbb{H}_{m_0,n}(\hat{\theta}_{m_0,n})+2\mathbb{H}_{m,n}(\hat{\theta}_{m,n})+\log\det\bigl(-a_n^2\partial^2_{\theta_{m_0}}\mathbb{H}_{m_0,n}(\hat{\theta}_{m_0,n})\bigr)\\
&\quad-\log\det\bigl(-a_n^2\partial^2_{\theta_m}\mathbb{H}_{m,n}(\hat{\theta}_{m,n})\bigr)<(p_m-p_{m_0})\log a_n^{-2}\bigr\}\\
&=\mathbb{P}\bigl\{a_n^2\bigl(\mathbb{H}_{m_0,n}(\hat{\theta}_{m_0,n})-\mathbb{H}_{m,n}(\hat{\theta}_{m,n})\bigr)>o_p(1)\bigr\}\\
&=\mathbb{P}\bigl\{\mathbb{H}_{m_0,0}(\theta_{m_0,0})-\mathbb{H}_{m,0}(\theta_{m,0})>o_p(1)\bigr\}\\
&=\mathbb{P}\bigl\{\mathbb{H}_{m_0,0}(\theta_{m_0,0})-\mathbb{H}_{m,0}(\theta_{m,0})>0\bigr\}+o(1)\\
&\to 1
\end{aligned}$$

as $n\to\infty$.

7.7. Proof of Theorem 5.3

If both $m_1\neq m_{1,0}$ and $m_2\neq m_{2,0}$ hold, then we have

$$\begin{aligned}
&\mathbb{P}\bigl(\mathrm{QBIC}_n^{(m_{1,0},m_{2,0})}-\mathrm{QBIC}_n^{(m_1,m_2)}\geq 0\bigr)\\
&\leq\mathbb{P}\bigl(\mathrm{QBIC}_n^{(m_{1,0},m_{2,0})}-\mathrm{QBIC}_n^{(m_{1,0},m_2)}\geq 0\bigr)\\
&\quad+\mathbb{P}\bigl(\mathrm{QBIC}_n^{(m_{1,0},m_2)}-\mathrm{QBIC}_n^{(m_1,m_2)}\geq 0\bigr).
\end{aligned}\tag{7.6}$$

Applying the proof of Theorem 5.1(i) we see that the both terms in the right-hand side tends to zero, hence the claim. The other cases are similar and simpler.

7.8. Proof of Theorem 5.5

As with Theorem 5.1(i), under assumptions of Theorem 5.5 we can deduce that

$$\mathbb{P}\bigl(\mathrm{QBIC}_n^{(m_{1,0})}-\mathrm{QBIC}_n^{(m_1)}<0\bigr)\to 1,\tag{7.7}$$

which means that $\mathbb{P}(m^*_{1,n} = m_{1,0}) \to 1$. This together with Theorem 5.1(i) then gives

$$\begin{aligned}
&\mathbb{P}\big(\mathrm{QBIC}_n^{(m_{2,0}|m^*_{1,n})} - \mathrm{QBIC}_n^{(m_2|m^*_{1,n})} \geq 0\big) \\
&\quad = \mathbb{P}\big(\mathrm{QBIC}_n^{(m_{2,0}|m^*_{1,n})} - \mathrm{QBIC}_n^{(m_2|m^*_{1,n})} \geq 0, m^*_{1,n} = m_{1,0}\big) \\
&\qquad + \mathbb{P}\big(\mathrm{QBIC}_n^{(m_{2,0}|m^*_{1,n})} - \mathrm{QBIC}_n^{(m_2|m^*_{1,n})} \geq 0, m^*_{1,n} \neq m_{1,0}\big) \\
&\quad \leq \mathbb{P}\big(\mathrm{QBIC}_n^{(m_{2,0}|m_{1,0})} - \mathrm{QBIC}_n^{(m_2|m_{1,0})} \geq 0\big) + \mathbb{P}\big(m^*_{1,n} \neq m_{1,0}\big) \to 0,
\end{aligned}$$

completing the proof.

Acknowledgements

The authors thank the anonymous referee for careful reading, which led to substantial improvements. They are also grateful to Prof. Yoshinori Kawasaki for drawing authors' attention to some relevant literature in econometrics, and to Prof. Masayuki Uchida and Prof. Nakahiro Yoshida for their valuable comments. This work was partly supported by CREST, JST.

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (*Tsahkadsor*, 1971) 267–281. Budapest: Akadémiai Kiadó. MR0483125

[2] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* **AC-19** 716–723. MR0423716

[3] Bickel, P.J., Li, B., Tsybakov, A.B., van de Geer, S.A., Yu, B., Valdés, T., Rivelo, C., Fan, J. and van der Vaart, A. (2006). Regularization in statistics. *TEST* **15** 271–344.

[4] Boswijk, H.P. (2010). Mixed normal inference on multicointegration. *Econometric Theory* **26** 1565–1576. MR2684795

[5] Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika* **52** 345–370. MR0914460

[6] Brouste, A., Fukasawa, M., Hino, H., Iacus, S.M., Kamatani, K., Koike, Y., Masuda, H., Nomura, R., Ogihara, T., Shimizu, Y., Uchida, M. and Yoshida, N. (2014). The yuima project: A computational framework for simulation and inference of stochastic differential equations. *J. Stat. Softw.* **57** 1–51.

[7] Burnham, K.P. and Anderson, D.R. (2002). *Model Selection and Multimodel Inference*: *A Practical Information-Theoretic Approach*, 2nd ed. New York: Springer. MR1919620

[8] Casella, G., Girón, F.J., Martínez, M.L. and Moreno, E. (2009). Consistency of Bayesian procedures for variable selection. *Ann. Statist.* **37** 1207–1228.

[9] Cavanaugh, J.E. and Neath, A.A. (1999). Generalizing the derivation of the Schwarz information criterion. *Comm. Statist. Theory Methods* **28** 49–66. MR1669504

[10] Chan, N.H., Huang, S.-F. and Ing, C.-K. (2013). Moment bounds and mean squared prediction errors of long-memory time series. *Ann. Statist.* **41** 1268–1298. MR3113811

[11] Chan, N.H. and Ing, C.-K. (2011). Uniform moment bounds of Fisher's information with applications to time series. *Ann. Statist.* **39** 1526–1550. MR2850211

[12] Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95** 759–771. MR2443189
[13] Claeskens, G. and Hjort, N.L. (2008). *Model Selection and Model Averaging*. Cambridge: Cambridge Univ. Press.
[14] Dziak, J.J., Coffman, D.L., Lanze, S.T. and Li, R. (2012). Sensitivity and specificity of information criteria. *PeerJ PrePrints* **3**.
[15] Eguchi, S. and Masuda, H. (2015). Quasi-Bayesian model comparison for LAQ models. Technical report, MI Preprint Series 2015-7, Kyushu University.
[16] Fasen, K. and Kimmig, S. (2015). Information criteria for multivariate CARMA processes. *Bernoulli*. To appear. Available at arXiv:1505.00901.
[17] Foster, D.P. and George, E.I. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22** 1947–1975. MR1329177
[18] Fujii, T. and Uchida, M. (2014). AIC type statistics for discretely observed ergodic diffusion processes. *Stat. Inference Stoch. Process.* **17** 267–282. MR3256840
[19] Genon-Catalot, V. and Jacod, J. (1993). On the estimation of the diffusion coefficient for multi-dimensional diffusion processes. *Ann. Inst. Henri Poincaré Probab. Stat.* **29** 119–151.
[20] Gobet, E. (2002). LAN property for ergodic diffusions with discrete observations. *Ann. Inst. Henri Poincaré Probab. Stat.* **38** 711–737. MR1931584
[21] Goutis, C. and Robert, C.P. (1998). Model choice in generalised linear models: A Bayesian approach via Kullback–Leibler projections. *Biometrika* **85** 29–37.
[22] Kamatani, K. and Uchida, M. (2015). Hybrid multi-step estimators for stochastic differential equations based on sampled data. *Stat. Inference Stoch. Process.* **18** 177–204.
[23] Kashyap, R.L. (1982). Optimal choice of AR and MA parts in autoregressive moving average models. *IEEE Trans. Pattern Anal. Mach. Intell.* **4** 99–104.
[24] Kessler, M. (1997). Estimation of an ergodic diffusion from discrete observations. *Scand. J. Stat.* **24** 211–229. MR1455868
[25] Kim, J.Y. (1998). Large sample properties of posterior densities, Bayesian information criterion and the likelihood principle in nonstationary time series models. *Econometrica* **66** 359–380.
[26] Konishi, S., Ando, T. and Imoto, S. (2004). Bayesian information criteria and smoothing parameter selection in radial basis function networks. *Biometrika* **91** 27–43. MR2050458
[27] Konishi, S. and Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika* **83** 875–890. MR1440051
[28] Konishi, S. and Kitagawa, G. (2008). *Information Criteria and Statistical Modeling. Springer Series in Statistics*. New York: Springer. MR2367855
[29] Lavine, M. and Schervish, M.J. (1999). Bayes factors: What they are and what they are not. *Amer. Statist.* **53** 119–122. MR1707756
[30] Liu, W. and Yang, Y. (2011). Parametric or nonparametric? A parametricness index for model selection. *Ann. Statist.* **39** 2074–2102.
[31] Lv, J. and Liu, J.S. (2014). Model selection principles in misspecified models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 141–167.
[32] Masuda, H. (2013). Convergence of Gaussian quasi-likelihood random fields for ergodic Lévy driven SDE observed at high frequency. *Ann. Statist.* **41** 1593–1641.
[33] Masuda, H. and Uehara, Y. (2016). On stepwise estimation of Lévy driven stochastic differential equation (Japanese). *Proc. Inst. Statist. Math.*
[34] Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.* **12** 758–765. MR0740928
[35] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.

[36] Sclove, S.L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika* **52** 333–343.
[37] Sei, T. and Komaki, F. (2007). Bayesian prediction and model selection for locally asymptotically mixed normal models. *J. Statist. Plann. Inference* **137** 2523–2534.
[38] Uchida, M. (2010). Contrast-based information criterion for ergodic diffusion processes from discrete observations. *Ann. Inst. Statist. Math.* **62** 161–187. MR2577445
[39] Uchida, M. and Yoshida, N. (2001). Information criteria in model selection for mixing processes. *Stat. Inference Stoch. Process.* **4** 73–98. MR1850590
[40] Uchida, M. and Yoshida, N. (2006). Asymptotic expansion and information criteria. *SUT J. Math.* **42** 31–58. MR2255535
[41] Uchida, M. and Yoshida, N. (2011). Estimation for misspecified ergodic diffusion processes from discrete observations. *ESAIM Probab. Stat.* **15** 270–290.
[42] Uchida, M. and Yoshida, N. (2012). Adaptive estimation of an ergodic diffusion process based on sampled data. *Stochastic Process. Appl.* **122** 2885–2924. MR2931346
[43] Uchida, M. and Yoshida, N. (2013). Quasi likelihood analysis of volatility and nondegeneracy of statistical random field. *Stochastic Process. Appl.* **123** 2851–2876.
[44] Uchida, M. and Yoshida, N. (2016). Model selection for volatility prediction. In *The Fascination of Probability, Statistics and Their Applications* 343–360. Cham: Springer. MR3495692
[45] van der Vaart, A.W. (1998). *Asymptotic Statistics*. Cambridge: Cambridge Univ. Press.
[46] Vehtari, A. and Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Stat. Surv.* **6** 142–228. MR3011074
[47] Yoshida, N. (2011). Polynomial type large deviation inequalities and quasi-likelihood analysis for stochastic differential equations. *Ann. Inst. Statist. Math.* **63** 431–479.

Received June 2016 and revised November 2016