Annals of the Institute of Statistical Mathematics (2024) 76:111–157
https://doi.org/10.1007/s10463-023-00878-2

Gaussian quasi-information criteria for ergodic Lévy driven SDE

Shoichi Eguchi[1] · **Hiroki Masuda**[2,3]

Received: 21 October 2022 / Revised: 22 May 2023 / Accepted: 22 June 2023 /
Published online: 11 August 2023

Abstract

We consider relative model comparison for the parametric coefficients of an ergodic Lévy driven model observed at high-frequency. Our asymptotics is based on the fully explicit two-stage Gaussian quasi-likelihood function (GQLF) of the Euler-approximation type. For selections of the scale and drift coefficients, we propose explicit Gaussian quasi-AIC and Gaussian quasi-BIC statistics through the stepwise inference procedure, and prove their asymptotic properties. In particular, we show that the mixed-rates structure of the joint GQLF, which does not emerge in the case of diffusions, gives rise to the non-standard forms of the regularization terms in the selection of the scale coefficient, quantitatively clarifying the relation between estimation precision and sampling frequency. Also shown is that the stepwise strategies are essential for both the tractable forms of the regularization terms and the derivation of the asymptotic properties of the Gaussian quasi-information criteria. Numerical experiments are given to illustrate our theoretical findings.

Keywords AIC · BIC · Ergodic Lévy driven SDE · Stepwise Gaussian quasi-likelihood estimation

✉ Shoichi Eguchi
shoichi.eguchi@oit.ac.jp

Hiroki Masuda
hmasuda@ms.u-tokyo.ac.jp

[1] Faculty of Information Science and Technology, Osaka Institute of Technology, 1-79-1 Kitayama, Hirakata, Osaka 573-0196, Japan

[2] Faculty of Mathematics, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan

[3] Graduate School of Mathematical Sciences, The University of Tokyo, 3-8-1 Komaba Meguro-ku, Tokyo 153-8914, Japan

1 Introduction

Suppose that we observe an equally spaced high-frequency sample $\boldsymbol{X}_n = (X_{t_j^n})_{j=0}^n$ for $t_j^n = t_j = jh$, where $X = (X_t)_{t\in\mathbb{R}_+}$ is a solution to the stochastic differential equation (SDE)

$$dX_t = A(X_t)dt + C(X_{t-})dZ_t, \tag{1}$$

where $A : \mathbb{R}^d \to \mathbb{R}^d$ and $C : \mathbb{R}^d \to \mathbb{R}^d \otimes \mathbb{R}^r$, and $Z = (Z_t)_{t\in\mathbb{R}_+}$ is an r-dimensional Lévy process independent of the initial value X_0. We will suppose that Z is standardized in the sense that Z_1 is zero-mean and has the identity covariance matrix. The sampling stepsize $h = h_n > 0$ is a known real such that

$$T_n := nh \to \infty, \quad nh^2 \to 0$$

as $n \to \infty$. We want to infer the coefficients A and C based on a sample $\boldsymbol{X}_n$, without specifying $\mathcal{L}(Z)$, the distribution of the process Z. Also, suppose that we are given the following candidates

$$c_1(x, \gamma_1), \ldots, c_{M_1}(x, \gamma_{M_1}),$$
$$a_1(x, \alpha_1), \ldots, a_{M_2}(x, \alpha_{M_2}),$$

for the scale and drift coefficients, respectively. Then, each candidate SDE model $\mathcal{M}_{m_1,m_2}$ is described by

$$dX_t = c_{m_1}(X_{t-}, \gamma_{m_1})dZ_t + a_{m_2}(X_t, \alpha_{m_2})dt. \tag{2}$$

The distribution of $\boldsymbol{X}_n$ is seldom explicitly given, hence we need to resort to some approximation. In this paper, we will consider the Gaussian approximation

$$\mathcal{L}(X_{t_j}|X_{t_{j-1}} = x) \sim N\big(x + ha_{m_2}(x, \alpha_{m_2}), hc_{m_1}(x, \gamma_{m_1})c_{m_1}(x, \gamma_{m_1})^\top\big) \tag{3}$$

as our statistical model corresponding to $\mathcal{M}_{m_1,m_2}$, where $\gamma_{m_1} \in \Theta_{\gamma_{m_1}} \subset \mathbb{R}^{p_{\gamma_{m_1}}}$ $(m_1 = 1, \ldots, M_1)$ and $\alpha_{m_2} \in \Theta_{\alpha_{m_2}} \subset \mathbb{R}^{p_{\alpha_{m_2}}}$ $(m_2 = 1, \ldots, M_2)$ are finite-dimensional unknown parameters. The parameter spaces $\Theta_{\gamma_{m_1}}$ and $\Theta_{\alpha_{m_2}}$ are assumed to be bounded convex domains. The main objective of this paper is to develop a model selection procedure for selecting the best model $\mathcal{M}_{\hat{m}_{1,n},\hat{m}_{2,n}}$ among the candidate models. For selecting an appropriate model, we will develop the Akaike information criterion (AIC, Akaike (1973)) and Bayesian information criterion (BIC, Schwarz (1978)) type model comparison for semiparametric Lévy driven SDE (1). Although we are interested in the SDE model (2), we consider the Gaussian (logarithmic) quasi-likelihood function (GQLF) based on (3) instead of the true likelihood for the inference. In this sense, our statistical models are all misspecified.

The information criteria are one of the most convenient and powerful tools for model selection, and the AIC and BIC are derived from two different classical principles: the AIC and the GIC (generalized information criterion (Konishi and Kitagawa 1996)), which is an extension of AIC, are predictive model selection criteria

minimizing the Kullback-Leibler divergence which measures the deviation from the true model to the prediction model; the BIC is given by the Bayesian principle and used for finding better model descriptions. The AIC is not intended to select the true model consistently even if the true model is included in the set of candidate models, while the BIC puts importance on both underfitting and overfitting. Based on the same classical principles as AIC and BIC, several studies have been conducted on model selection for SDEs: the contrast-based information criterion for ergodic diffusion processes (Uchida 2010), the BIC type information criterion for locally asymptotically quadratic models (Eguchi and Masuda 2018), and the BIC type information criterion for possibly misspecified ergodic SDEs (Eguchi and Uehara 2021).

Asymptotic inference based on the GQLF for Lévy driven SDE has been developed by several previous works, of course at the expense of asymptotic estimation efficiency: see (Masuda 2013) and Masuda and Uehara (2017), as well as the references therein. The *Gaussian quasi-AIC(BIC)*, which we will introduce and term *GQAIC (GQBIC)* for short, is based on the fully explicit two-stage GQLF of the Euler-approximation type (3). Our study develops the GQAIC (GQBIC) under the condition that the scale and drift coefficients are correctly specified and clarifies that taking the two steps will be inevitable for the simple form of the GQAIC (GQBIC) to be in force in the sense that they provide us with the specific asymptotic selection probabilities for GQAIC and the selection consistency for GQBIC: the details will be given in Sects. 3, 4, and 5. The GQAIC and GQBIC which are derived through the GQLF of the first stage will be called GQAIC_1 and GQBIC_1, respectively. Also, the GQAIC and GQBIC based on the GQLF of the second stage will be called GQAIC_2 and GQBIC_2, respectively.

The two-stage procedure proposed in this paper is summarized as follows: for the AIC type, first, we select a scale-coefficient model as a minimizer of GQAIC_1 over the candidates $c_1, \dots, c_{M_1}$, and then select a drift-coefficient model as a minimizer of GQAIC_2 over the candidates $a_1 \dots, a_{M_2}$. Our model comparisons will be presented in Sect. 5. There, we will consider the cases where the candidate coefficients $c_1, \dots, c_{M_1}$ and $a_1, \dots, a_{M_2}$ contain both correctly specified coefficients and misspecified coefficients. Also, we formally use the $\mathrm{GQAIC}_{1,n}$ and $\mathrm{GQAIC}_{2,n}$ even for the possibly misspecified coefficients, although the assumption that the candidate scale and drift coefficients are correctly specified is necessary for the derivation of GQAIC. As for the BIC-type, we follow the same way, replacing GQAIC_1 and GQAIC_2 by GQBIC_1 and GQBIC_2, respectively. In particular, concerned with both AIC- and BIC-type model comparisons of the scale coefficient, it turned out that we should employ some non-standard forms of the regularization term. Especially for the BIC type methodology, it turned out that the conventional stochastic expansion of the marginal (quasi-)likelihood is not appropriate for *consistent* model selection: we needed to "heated up" free energy (Sect. 4.1).

The very different features compared with the ergodic diffusions will be presented in this paper. They are essentially due to the mixed-rates structure (Radchenko 2008) of the joint GQLF given by (8) below, which does not emerge for the case of diffusions where Z is a standard Wiener process; see Remark 6(1). Informally speaking, the use of the GQLF against non-Gaussian Lévy processes causes the non-standard phenomena in inference for the scale coefficient, quantitatively clarifying

the relation between estimation precision and sampling frequency. Remarkably, we could still obtain explicit results of practical value.

In the rest of this paper, we give some prerequisites in Sect. 2. Then, in Sects. 3 and 4 we present how the classical AIC- and BIC-type arguments can work in our model setup, respectively. In Sect. 5, we introduce the stepwise model comparison procedure and discuss the asymptotic probability of relative model selection. Section 6 presents illustrative numerical results supporting our findings. The proofs are gathered in Sect. 7.

2 Preliminaries

2.1 Basic notation

The following basic notation will be used throughout this paper. We denote by $|A|$ the determinant of a square matrix A, and by $\|A\|$ the Frobenius norm of a matrix A. Write $A^{\otimes 2} = AA^\top$ for any matrix A, with $\top$ denoting transposition. For a Kth-order multilinear form $M = \{M^{(i_1 \dots i_K)} : i_k = 1, \dots, d_k; k = 1, \dots, K\} \in \mathbb{R}^{d_1} \otimes \cdots \otimes \mathbb{R}^{d_K}$ and d_k-dimensional vectors $u_k = \{u_k^{(j)}\}$, we let $M[u_1, \dots, u_K] := \sum_{i_1=1}^{d_1} \cdots \sum_{i_K=1}^{d_K} M^{(i_1,\dots,i_K)} u_1^{(i_1)} \dots u_K^{(i_K)}$; in particular, $A[B] := \mathrm{trace}(AB^\top)$ in case of $K = 2$ for matrices A and B of the same sizes. The symbol ∂_a^k stands for k-times partial differentiation with respect to variable a, and I_r denotes the $r \times r$-identity matrix. We write $C > 0$ for a universal positive constant which may vary at each appearance, and $a_n \lesssim b_n$ for possibly random nonnegative sequences (a_n) and (b_n) if $a_n \le C b_n$ a.s. holds for every n large enough. The density of the Gaussian distribution $N_d(\mu, \Sigma)$ will be denoted by $\phi_d(x; \mu, \Sigma)$.

The basic setting is as follows. We denote by $(\Omega, \mathcal{F}, P)$ the underlying probability space and by E the associated expectation operator. For notational convenience, instead of (2) we look at a single model

$$dX_t = c(X_{t-}, \gamma) dZ_t + a(X_t, \alpha) dt, \tag{4}$$

where $\gamma = (\gamma_k) \in \Theta_\gamma \subset \mathbb{R}^{p_\gamma}$ and $\alpha = (\alpha_l) \in \Theta_\alpha \subset \mathbb{R}^{p_\alpha}$, both parameter spaces being bounded convex domains. Let $p := p_\alpha + p_\gamma$. Let $\Delta_j Y := Y_{t_j} - Y_{t_{j-1}}$ for a process Y, and $f_{j-1}(\theta) := f(X_{t_{j-1}}, \theta)$ for any measurable function on $f : \mathbb{R}^d \times \Theta$. The symbols $\xrightarrow{p}$ and $\xrightarrow{\mathcal{L}}$ denote the convergence in probability and distribution, respectively. In Sects. 2, 3 and 3, we assume that coefficients c and a are correctly specified in the sense that there exist $\gamma_0 \in \Theta_\gamma$ and $\alpha_0 \in \Theta_\alpha$ such that $c(\cdot, \gamma_0) = C(\cdot)$ and $a(\cdot, \alpha_0) = A(\cdot)$.

2.2 Two-stage Gaussian quasi-likelihood estimation

Write $S(x, \gamma) = c(x, \gamma)^{\otimes 2}$ for the scale matrix, which will play the role of diffusion matrix in the diffusion context. Let $\nu(dz)$ denote the Lévy measure of Z, and then for

$i_1, \dots, i_m \in \{1, \dots, r\}$ with $m \geq 3$ we write $\nu(m)$ the tensor consisting of all the mth-mixed moments of ν:

$$\nu(m) = \{\nu_{i_1 \dots i_m}(m)\}_{i_1,\dots,i_m} := \left\{ \int z_{i_1} \dots z_{i_m} \nu(dz) \right\}_{i_1,\dots,i_m}.$$

Denote by $\lambda_{\min}\{S(x,\gamma)\}$ the minimum eigenvalue of $S(x,\gamma)$. The symbol $\mathcal{C}_{\sharp}^{k,l}$ for non-negative integers k and l denotes the function space consisting of all measurable $f : \mathbb{R}^d \times \overline{\Theta} \to \mathbb{R}$ such that:

- $f(\cdot, \theta)$ is globally Lipschitz uniformly in $\theta \in \overline{\Theta}$;
- $f(x, \theta)$ is k-times (resp. l-times) continuously differentiable in x (resp. in θ), respectively, all the partial derivatives are continuous over $\overline{\Theta}$ for each x, and the estimates

$$\max_{i \leq k} \max_{j \leq l} \sup_{\theta \in \overline{\Theta}} |\partial_\theta^j \partial_x^i f(x,\theta)| \lesssim 1 + |x|^{C_{k,l}}$$

holds for some constant $C_{k,l} \geq 0$.

We need some regularity conditions on the process (X, Z) to ensure the asymptotic normality and the uniform tail-probability estimate.

The following conditions are standard in the literature, essentially borrowed from Masuda (2013) and Masuda and Uehara (2017).

Assumption 1 (Moments) $E[Z_1] = 0$, $E[Z_1^{\otimes 2}] = I_r$, and $E[|Z_1|^q] < \infty$ for all $q > 0$.

Assumption 2 (Smoothness and non-degeneracy) The components of a and c belong to the class $\mathcal{C}_{\sharp}^{2,4}$, and

$$\sup_{\gamma \in \overline{\Theta}_\gamma} \lambda_{\min}\{S(x,\gamma)\}^{-1} \lesssim 1 + |x|^{C_0}$$

for some constant $C_0 \geq 0$.

Assumption 3 (Stability) There exists a probability measure $\pi = \pi_{\theta_0}$ such that for every $q > 0$ we can find positive constant a for which

$$\sup_{t \in \mathbb{R}_+} e^{at} \sup_{f:\, |f| \leq g} \left| \int f(y) P_t(x, dy) - \int f(y) \pi(dy) \right| \lesssim g(x), \qquad x \in \mathbb{R}^d,$$

where $g(x) := 1 + \|x\|^q$ and $P_t(x, dy) := P(X_t \in dy | X_0 = x)$. Further, for every $q > 0$,

$$\sup_{t \in \mathbb{R}_+} E[|X_t|^q] < \infty. \tag{5}$$

It follows from Assumptions 2 and 3 that

$$\frac{1}{n}\sum_{j=1}^{n} g(X_{t_{j-1}},\theta)\xrightarrow{p}\int g(x,\theta)\pi(dx),\qquad n\to\infty,$$

uniformly in θ for sufficiently smooth function $g(x,\theta)$ whose partial derivative with respect to x are of at most polynomial growth in x uniformly in θ. This can be seen in the standard moment estimates and the tightness argument: see (Masuda 2013, p.1598 and Section 4.1.1). Also to be noted is that the seemingly stringent moment condition (5) could be removed in compensation for the boundedness of the coefficients and the uniform non-degeneracy of S: see (Masuda 2013, Theorem 2.9).

The Euler approximation for (4) under P_θ is given by

$$X_{t_j}\approx X_{t_{j-1}}+a_{j-1}(\alpha)h+c_{j-1}(\gamma)\Delta_j Z. \tag{6}$$

Taking the small-time Gaussian approximation

$$\mathcal{L}(X_{t_j}|X_{t_{j-1}}=x)\approx N_d(x+a(x,\alpha)h,\,hS(x,\gamma)) \tag{7}$$

into account, we are led to the joint GQLF $\mathbb{H}_n(\theta)=\mathbb{H}_n(\boldsymbol{X}_n,\theta)$:

$$\begin{aligned}\mathbb{H}_n(\theta)&:=\sum_{j=1}^{n}\log\phi_d\Big(X_{t_j};X_{t_{j-1}}+a_{j-1}(\alpha)h,\,hS_{j-1}(\gamma)\Big)\\&=-\frac{1}{2}\sum_{j=1}^{n}\Big(\log\Big|2\pi hS_{j-1}(\gamma)\Big|+\frac{1}{h}S_{j-1}^{-1}(\gamma)\Big[\big(\Delta_jX-ha_{j-1}(\alpha)\big)^{\otimes2}\Big]\Big).\end{aligned} \tag{8}$$

In the present study, although the model of interest is described by (4), we will not consider the associated exact likelihood. Instead, we will regard (7) as our statistical model and deal with the explicit GQLF (8) based on the theoretically incorrect (7) for inference purposes; in this sense, our statistical model is misspecified. Still, it is possible to estimate the true coefficients when they are correctly specified (see Theorem 1 below).

We can write $\mathbb{H}_n(\theta)=\mathbb{H}_{1,n}(\gamma)+\mathbb{H}_{2,n}(\theta)$ where

$$\begin{aligned}\mathbb{H}_{1,n}(\gamma;\boldsymbol{X}_n)=\ \mathbb{H}_{1,n}(\gamma)\ &:=\sum_{j=1}^{n}\log\phi_d\Big(X_{t_j};X_{t_{j-1}},\,hS_{j-1}(\gamma)\Big),\\\mathbb{H}_{2,n}(\theta;\boldsymbol{X}_n)=\ \mathbb{H}_{2,n}(\theta)\ &:=\sum_{j=1}^{n}\Big(S_{j-1}^{-1}(\gamma)\big[\Delta_jX,\,a_{j-1}(\alpha)\big]-\frac{h}{2}S_{j-1}^{-1}(\gamma)\Big[a_{j-1}^{\otimes2}(\alpha)\Big]\Big).\end{aligned} \tag{9}$$

The joint GQLF $\mathbb{H}_n(\theta)$ has two different "resolutions" (see (13) below), which comes from the fact that the last term $c_{j-1}(\gamma)\Delta_jZ$ in the right-hand side of (6) is stochastically dominant compared with the second one $a_{j-1}(\alpha)h$. It was seen in Masuda and Uehara (2017) that under suitable conditions including the ergodicity of X that both $n^{-1}\mathbb{H}_{1,n}(\gamma)$ and $T_n^{-1}\mathbb{H}_{2,n}(\alpha,\gamma)$ have non-trivial limits (of the ergodic theorem) for each θ, in particular the former limits depending on γ only. Building on this observation, the following two-stage estimation strategy is suggested: *first*, we estimate γ by

$\hat{\gamma}_n \in \operatorname{argmax}_\gamma \mathbb{H}_{1,n}(\gamma)$; *and then*, estimate α by $\hat{\alpha}_n \in \operatorname{argmax}_\alpha \mathbb{H}_{2,n}(\alpha)$ where, with a slight abuse of notation,

$$\mathbb{H}_{2,n}(\alpha) := \mathbb{H}_{2,n}(\alpha, \hat{\gamma}_n).$$

Note that maximizing $\alpha \mapsto \mathbb{H}_{2,n}(\alpha, \gamma)$ *given a value of* γ amounts to maximizing the discrete-time approximation of the log-likelihood function corresponding to the continuous-time observation (see (Liptser and Shiryaev 2001, Chapter 7) for details), and also to maximizing

$$\mathbb{H}^*_{2,n}(\theta) := \sum_{j=1}^{n} \log \phi_d\Big(X_{t_j}; X_{t_{j-1}} + a_{j-1}(\alpha)h,\, hS_{j-1}(\gamma)\Big). \tag{10}$$

Therefore, in either case, the second stage itself may be recognized as a GQLF. We denote this two-stage *Gaussian quasi-maximum likelihood estimator (GQMLE)* by $\hat{\theta}_n = (\hat{\alpha}_n, \hat{\gamma}_n)$, which is essentially the same as in the one considered in Masuda and Uehara (2017).

Assumption 4 (Identifiability) There exist positive constants χ_γ and χ_α such that

$$-\frac{1}{2}\int \left\{ \operatorname{trace}\left(S(x,\gamma)^{-1}S(x,\gamma_0) - I_d\right) + \log \frac{|S(x,\gamma)|}{|S(x,\gamma_0)|} \right\} \pi(dx) \le -\chi_\gamma |\gamma - \gamma_0|^2,$$
$$-\frac{1}{2}\int S^{-1}(x,\gamma_0)\Big[\big(a(x,\alpha) - a(x,\alpha_0)\big)^{\otimes 2}\Big]\pi(dx) \le -\chi_\alpha |\alpha - \alpha_0|^2,$$

for every γ and α.

The two integrals in the left-hand sides in Assumption 4 correspond to the Kullback-Leibler divergences associated with $\mathbb{H}_{1,n}$ and $\mathbb{H}_{2,n}$, respectively.

To state the result we need to introduce the matrix $V(\theta_0) = \Gamma(\theta_0)^{-1}\Sigma(\theta_0)\Gamma(\theta_0)^{-1}$ where

$$\Sigma(\theta_0) = (\Sigma^{(kl)}(\theta_0))_{k,l} := \begin{pmatrix} \Gamma_\alpha(\theta_0) & W_{\alpha,\gamma}(\theta_0) \\ W_{\alpha,\gamma}(\theta_0)^\top & W_\gamma(\gamma_0) \end{pmatrix},$$
$$\Gamma(\theta_0) = (\Gamma^{(kl)}(\theta_0))_{k,l} := \operatorname{diag}\{\Gamma_\alpha(\theta_0), \Gamma_\gamma(\gamma_0)\},$$

with, letting $\Psi = (\Psi^{(kl)})_{k,l} := \partial_\gamma(S^{-1}) = -S^{-1}(\partial S)S^{-1}$ where $\Psi^{(kl)} \in \mathbb{R}^{p_\gamma}$,

$$\Gamma_\alpha^{(kl)}(\theta_0) := \int S^{-1}(x,\beta_0)\big[\partial_{\alpha_k} a(x,\alpha_0), \partial_{\alpha_l} a(x,\alpha_0)\big]\pi(dx),$$
$$\Gamma_\gamma^{(kl)}(\gamma_0) := \frac{1}{2}\int \operatorname{trace}\big[\big\{(S^{-1}\partial_{\gamma_k}S)(S^{-1}\partial_{\gamma_l}S)\big\}(x,\gamma_0)\big]\pi(dx),$$
$$W_{\alpha,\gamma}^{(qr)}(\theta_0) := -\frac{1}{2}\sum_{k,l,k',l',s,t,t'} \nu_{stt'}(3)\int \big((\Psi^{(kl)})^{(q)}(\partial_{\alpha_r}a^{(k')})c^{(ks)}c^{(lt)}c^{(l't')}(S^{-1})^{(k't')}\big)(x,\theta_0)\pi(dx),$$
$$W_\gamma^{(qr)}(\gamma_0) := \frac{1}{2}\sum_{k,l,k',l',s,t,s',t'} \nu_{sts't'}(4)\int \big((\Psi^{(kl)})^{(q)}(\Psi^{(k'l')})^{(r)}c^{(ks)}c^{(lt)}c^{(k's')}c^{(l't')}\big)(x,\gamma_0)\pi(dx).$$

Under Assumption 4, the matrix $\Gamma(\theta_0)$ is positive definite. We introduce the following empirical counterparts: let $\hat{V}_n := \hat{\Gamma}_n^{-1}\hat{\Sigma}_n\hat{\Gamma}_n^{-1}$, where

$$\hat{\Sigma}_n := \begin{pmatrix} \hat{\Gamma}_{\alpha,n} & \hat{W}_{\alpha,\gamma,n} \\ \hat{W}^{\top}_{\alpha,\gamma,n} & \hat{W}_{\gamma,n} \end{pmatrix}, \qquad \hat{\Gamma}_n := \mathrm{diag}\{\hat{\Gamma}_{\alpha,n}, \hat{\Gamma}_{\gamma,n}\},$$

with, using the shorthand $\partial_\theta \hat{f}_{j-1}$ for $(\partial_\theta f_{j-1})(\hat{\theta}_n)$ and so on, the entries of $\hat{\Gamma}_{\alpha,n}$ and $\hat{\Gamma}_{\gamma,n}$ being given by

$$\hat{\Gamma}^{(kl)}_{\alpha,n} := \frac{1}{n}\sum_{j=1}^{n} \hat{S}^{-1}_{j-1}\left[\partial_{\alpha_k}\hat{a}_{j-1}, \partial_{\alpha_l}\hat{a}_{j-1}\right],$$

$$\hat{\Gamma}^{(kl)}_{\gamma,n} := \frac{1}{2n}\sum_{j=1}^{n} \mathrm{trace}\left[(\hat{S}^{-1}_{j-1}\partial_{\gamma_k}\hat{S}_{j-1})(\hat{S}^{-1}_{j-1}\partial_{\gamma_l}\hat{S}_{j-1})\right],$$

and also those of $\hat{W}_{\alpha,\gamma,n}$ and $\hat{W}_{\gamma,n}$ by, writing $\hat{\chi}_j = \Delta_j X - h\hat{a}_{j-1}$,

$$\begin{aligned}
\hat{W}^{(qr)}_{\alpha,\gamma,n} &:= \frac{1}{2T_n}\sum_{j=1}^{n} \left\{\left(\hat{S}^{-1}_{j-1}(\partial_{\gamma_q}\hat{S}_{j-1})\hat{S}^{-1}_{j-1}\right)[\hat{\chi}_j^{\otimes 2}]\right\}\left\{(\hat{S}^{-1}_{j-1})[\hat{\chi}_j, \partial_{\alpha_r}\hat{a}_{j-1}]\right\}, \\
\hat{W}^{(qr)}_{\gamma,n} &:= \frac{1}{4T_n}\sum_{j=1}^{n} \left\{\left(\hat{S}^{-1}_{j-1}(\partial_{\gamma_q}\hat{S}_{j-1})\hat{S}^{-1}_{j-1}\right)[\hat{\chi}_j^{\otimes 2}]\right\}\left\{\left(\hat{S}^{-1}_{j-1}(\partial_{\gamma_r}\hat{S}_{j-1})\hat{S}^{-1}_{j-1}\right)[\hat{\chi}_j^{\otimes 2}]\right\}.
\end{aligned} \tag{11}$$

Theorem 1 *Suppose that Assumptions 1, 2, 3, and 4 hold true. Then, for any continuous function $f : \mathbb{R}^p \to \mathbb{R}$ of at most polynomial growth, we have the convergence of moments*

$$E\left[f\left(\sqrt{T_n}(\hat{\theta}_n - \theta_0)\right)\right] \to \int f(u)\phi(u;0, V(\theta_0))du;$$

in particular, we have

$$\hat{u}_n := \sqrt{T_n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} N_p\left(0, V(\theta_0)\right),$$

$E(\hat{u}_n) \to 0$, $E(\hat{u}_n^{\otimes 2}) \to V(\theta_0)$, and also $\sup_n E(|\hat{u}_n|^q) < \infty$ for every $q > 0$. Further, we have $\hat{\Sigma}_n \xrightarrow{p} \Sigma(\theta_0)$ and $\hat{\Gamma}_n \xrightarrow{p} \Gamma(\theta_0)$, followed by $\hat{V}_n \xrightarrow{p} V(\theta_0)$, so that

$$\hat{V}_n^{-1/2}\sqrt{T_n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} N_p(0, I_p), \tag{12}$$

as soon as $\Sigma(\theta_0)$ is positive definite (hence so is $V(\theta_0)$).

Comparing (Masuda 2013, Theorem 2.7) and Theorem 1 shows that the two GQMLEs have the same asymptotic distribution, so that there is no loss of asymptotic efficiency when using the two-stage procedure instead of the joint one. For convenience, we give a sketch of the proof in Sect. 7.1. The positive definiteness of Σ_0

seems not straightforward to be verified. Yet, if $\mathcal{L}(Z_1)$ is assumed to be symmetric from the beginning so that $\nu(3) = 0$, then $\Sigma(\theta_0) = \mathrm{diag}\{\Gamma_\alpha(\theta_0), W_\gamma(\gamma_0)\}$ and the assumption reduce to the positive definiteness of $W_\gamma(\gamma_0)$.

Remark 1 (Asymptotic normality of the joint GQMLE) Although (Masuda 2013) used a Z-estimation framework, we may follow an M-estimation one; typically, regularity conditions in the latter case are less restrictive. Recalling that $\mathbb{H}_n(\theta) = \mathbb{H}_{1,n}(\gamma) + \mathbb{H}_{2,n}(\theta)$, we have the mixed-rates expression

$$\begin{aligned}\mathbb{H}_n(\theta_0 + T_n^{-1/2}u) - \mathbb{H}_n(\theta_0) &= \frac{1}{h}\big\{h\big(\mathbb{H}_{1,n}(\gamma_0 + T_n^{-1/2}u_\gamma) - \mathbb{H}_{1,n}(\gamma_0)\big)\big\} \\ &\quad + \big(\mathbb{H}_{2,n}(\theta_0 + T_n^{-1/2}u) - \mathbb{H}_{2,n}(\theta_0)\big) \\ &=: \frac{1}{h}\log \mathbb{Z}_{1,n}(u_\gamma) + \log \mathbb{Z}_{2,n}(u),\end{aligned} \tag{13}$$

where $u = (u_\alpha, u_\gamma) \in \mathbb{R}^{p_\alpha} \times \mathbb{R}^{p_\gamma}$. Importantly, it can be shown that both of the random functions $u_\gamma \mapsto \log \mathbb{Z}_{1,n}(u_\gamma)$ and $u \mapsto \log \mathbb{Z}_{2,n}(u)$ are locally asymptotically quadratic with

$$\begin{aligned}\log \mathbb{Z}_{1,n}(u_\gamma) &\xrightarrow{\mathcal{L}} \Delta_1(\gamma_0)[u_\gamma] - \frac{1}{2}\Gamma_\gamma(\gamma_0)[u_\gamma^{\otimes 2}] =: f_0(u_\gamma), \\ \log \mathbb{Z}_{2,n}(u) &\xrightarrow{\mathcal{L}} \Delta_{21}(\theta_0)[u_\alpha] - \frac{1}{2}\Gamma_\alpha(\theta_0)[u_\alpha^{\otimes 2}] - \frac{1}{2}\Gamma_{2,\gamma}(\theta_0)[u_\gamma^{\otimes 2}] =: g_0(u),\end{aligned}$$

where $\Delta_1(\gamma_0)$ and $\Delta_{21}(\theta_0)$ are weak limits of $T_n^{-1/2}h\partial_\gamma \mathbb{H}_{1,n}(\gamma_0)$ and $T_n^{-1/2}\partial_\alpha \mathbb{H}_{2,n}(\theta_0)$, respectively, where $\Gamma_{2,\gamma}(\theta_0)$ is the limit in probability of $-(1/2)n^{-1}\sum_{j=1}^n \partial_\gamma^2 (S^{-1})_{j-1}(\gamma_0)[a_{j-1}(\alpha_0)^{\otimes 2}]$, and where $\mathrm{argmax}_{u_\gamma} f_0(u_\gamma) = \{\Gamma_\gamma(\gamma_0)^{-1}\Delta_1(\gamma_0)\} =: \{\hat{u}_{\gamma,0}\}$ and $\mathrm{argmax}_{u_\alpha} g_0(u_\alpha, \hat{u}_{\gamma,0}) = \{\Gamma_\alpha(\alpha_0)^{-1}\Delta_{21}(\theta_0)\} =: \{\hat{u}_{\alpha,0}\}$ a.s. Note that in the limit we have no "cross term" involving both u_α and u_γ, entailing that $\hat{u}_0 := (\hat{u}_{\alpha,0}, \hat{u}_{\gamma,0}) \sim N_p(0, V(\theta_0))$. Now, by means of (Radchenko 2008, Theorem 1), it is possible to deduce the asymptotic normality $\sqrt{T_n}(\hat{\theta}_n' - \theta_0) \xrightarrow{\mathcal{L}} N_p\big(0, V(\theta_0)\big)$ of the joint GQMLE $\hat{\theta}_n' \in \mathrm{argmax}_\theta \mathbb{H}_n(\theta)$.

Remark 2 (Univariate case) If $d = r = 1$, then the asymptotic covariance matrix $V(\theta_0)$ takes the following much simpler forms:

$$\begin{aligned}\Gamma_\alpha(\theta_0) &= \int \left(\frac{\partial_\alpha a(x, \alpha_0)}{c(x, \gamma_0)}\right)^{\otimes 2} \pi(dx), \qquad \Gamma_\gamma(\gamma_0) = \frac{1}{2}\int \left(\frac{\partial_\gamma S(x, \gamma_0)}{S(x, \gamma_0)}\right)^{\otimes 2} \pi(dx), \\ W_{\alpha,\gamma}(\theta_0) &= \frac{\nu(3)}{2}\int \frac{(\partial_\gamma S) \otimes (\partial_\alpha a)}{c^3}(x, \theta_0)\pi(dx), \\ W_\gamma(\gamma_0) &= \frac{\nu(4)}{4}\int \left(\frac{\partial_\gamma S(x, \gamma_0)}{S(x, \gamma_0)}\right)^{\otimes 2} \pi(dx) = \frac{\nu(4)}{2}\Gamma_\gamma(\gamma_0).\end{aligned}$$

Remark 3 We are primarily interested in cases of a driving Lévy process Z having a non-null jump part. The studentization (asymptotic standard normality) (12) is, however, valid as it is even for the case of diffusion where $Z = w$, an r-dimensional

standard Wiener process. This implies that the factor $\hat{V}_n$ automatically distinguishes whether or not the driving noise is Gaussian or not. To see this, we recall the well-known fact $D_n(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} N_p(0, \Gamma(\theta_0)^{-1})$ with the different (partly faster) rate of convergence $D_n := \mathrm{diag}(\sqrt{T_n} I_{p_\alpha}, \sqrt{n} I_{p_\gamma})$, see e.g. Uchida and Yoshida (2012). By writing

$$\hat{V}_n^{-1/2} \sqrt{T_n} (\hat{\theta}_n - \theta_0) = \left(\sqrt{T_n} \hat{V}_n D_n^{-1} \right) D_n (\hat{\theta}_n - \theta_0)$$

and then taking into account the difference of the orders of the conditional moments of $\Delta_j X$ given $\mathcal{F}_{t_{j-1}}$ (see e.g. (Masuda 2013, Lemma 4.5)), it is possible to deduce that $\sqrt{T_n} \hat{V}_n D_n^{-1} \xrightarrow{p} \Gamma(\theta_0)^{1/2}$, hence (12); the proof is standard and omitted. Concerned with the rate of convergence, if Z has a diffusion component and compound-Poisson jumps, the same rate as in the diffusion case can be achieved: $\sqrt{T_n}$ for the drift (and jump) component and $\sqrt{n}$ for the diffusion one (see (Shimizu and Yoshida 2006) and Ogihara and Yoshida (2011)). If Z is a locally β-stable pure-jump Lévy process, the drift parameter has the rate of convergence $\sqrt{n} h^{1-1/\beta}$ (see (Clément and Gloter 2020; Masuda 2010, 2019), and Masuda (2023)). It should be noted that the rates of convergence affect the regularization term in the quasi-BIC type statistics. See Remark 7.

3 Gaussian quasi-AIC

Building on Theorem 1, we turn to AIC-type model selection. Before proceeding, let us briefly describe the classical Akaike paradigm in a general setting: for the unknown true distribution $g(x)\mu(x)$ of a sample $\boldsymbol{X}_n$, we are given a statistical model, say $\{f(\cdot;\theta) : \theta \in \Theta\}$, and a divergence $\mathcal{D}(f;g)$ measuring deviation from g to f. We will follow the standard route by taking the Kullback-Leibler divergence: we estimate g by $\hat{f}_n(\cdot) = f(\cdot;\hat{\theta}_n)$ for some estimator $\hat{\theta}_n = \hat{\theta}_n(\boldsymbol{X}_n)$, and look at the random quantity $\mathcal{D}(\hat{f}_n;g)$, where

$$\mathcal{D}(f;g) = \int \log \left(\frac{g}{f} \right) g d\mu,$$

which we want to minimize over given candidate models. It amounts to minimizing the relative entropy $\mathcal{E}(\hat{f}_n;g)$ where $\mathcal{E}(f;g) := -\int (\log f) g d\mu$. As g is unknown, we substitute the empirical counterpart $\mathcal{E}(\hat{f}_n;\delta_{\boldsymbol{X}_n})$ for $\mathcal{E}(\hat{f}_n;g)$. Then, by removing the randomness by integrating out $\boldsymbol{X}_n$ with respect to g, it is desired to derive a computable corrector $\hat{\mathfrak{b}}_n$ such that

$$E\left[\mathcal{E}(\hat{f}_n;g) - \left(\mathcal{E}(\hat{f}_n;\delta_{\boldsymbol{X}_n}) + \hat{\mathfrak{b}}_n \right) \right] = o(1), \qquad n \to \infty. \tag{14}$$

The routine way is to first compute the "leading" term(s) of

$$\mathfrak{b}_n := E\big[\mathcal{E}(\hat{f}_n;g) - \mathcal{E}(\hat{f}_n;\delta_{\boldsymbol{X}_n})\big], \tag{15}$$

and then construct an asymptotically unbiased estimator $\hat{\mathfrak{b}}_n = \hat{\mathfrak{b}}_n(\boldsymbol{X}_n)$, namely $E[\hat{\mathfrak{b}}_n - \mathfrak{b}_n] = o(1)$. The stochastic order of $\hat{\mathfrak{b}}_n$ should be strictly smaller than that of $\mathcal{E}(\hat{f}_n;\delta_{\boldsymbol{X}_n})$; For regular models, very often $\hat{\mathfrak{b}}_n$ equals the number of the unknown parameters involved in the model, although it may not be the case, depending on each situation; see (Akaike 1973) and Konishi and Kitagawa (1996) with the references therein for details.

The above strategy remains the same and makes sense, whatever the probabilistic structure of $\mathcal{L}(\boldsymbol{X}_n)$ is and when $\mathcal{D}(f;g)$ is replaced by other divergences. Moreover, since we are considering a relative comparison among the candidate models, the above claim is satisfied as long as the suitable asymptotic properties of estimators are given even if the model $\{f(\cdot;\theta) : \theta \in \Theta\}$ is misspecified. In other words, what is essential is that we can find a suitable corrector $\hat{\mathfrak{b}}_n$ satisfying the property (14); again, note that our statistical models are all misspecified since the intractable true log-likelihood function was replaced by the fake Gaussian ones. Still, the candidate coefficients $c_{m_1}(x, \gamma_{m_1})$ and $a_{m_2}(x, \alpha_{m_2})$ are estimable as specified in Theorem 1.

3.1 Expansion of moments: joint case

Turning to our setup, we keep considering the SDE model (4). We first observe what will occur if we look at the AIC associated not with the two-stage GQLF $(\mathbb{H}_{1,n}, \mathbb{H}_{2,n})$, but with the *joint* GQLF $\mathbb{H}_n(\theta) = \mathbb{H}_n(\theta;\boldsymbol{X}_n) = \mathbb{H}_{1,n}(\gamma) + \mathbb{H}_{2,n}(\theta)$ of (8); most of the computations will be of direct use in the stepwise case of our primary interest as well (Sect. 3.2).

In what follows, we will omit a large portion of the technical details, for they are essentially based on and quite analogous to the basic computations in (Masuda 2013, Sect. 4).

Denote by $\tilde{\boldsymbol{X}}_n$ the independent copy of $\boldsymbol{X}_n = (X_{t_j})_{j=0}^n$, and by $\tilde{E}$ the expectation operator with respect to $\mathcal{L}(\tilde{\boldsymbol{X}}_n)$. Let $\hat{u}_{\gamma,n} := \sqrt{T_n}(\hat{\gamma}_n - \gamma_0)$ and $\hat{u}_{\alpha,n} := \sqrt{T_n}(\hat{\alpha}_n - \alpha_0)$, so that $\hat{u}_n = (\hat{u}_{\alpha,n}, \hat{u}_{\gamma,n})$; by Theorem 1, both quantities are $L^q(P)$-bounded for every $q > 0$. As was mentioned in (15), we want to compute (approximate) the quantity

$$\begin{aligned}\mathfrak{b}_n &:= E\big[\mathbb{H}_n(\hat{\theta}_n(\boldsymbol{X}_n);\boldsymbol{X}_n) - \tilde{E}\big[\mathbb{H}_n(\hat{\theta}_n(\boldsymbol{X}_n);\tilde{\boldsymbol{X}}_n)\big]\big] \\ &= E\tilde{E}\big[\big(\mathbb{H}_n(\hat{\theta}_n) - \mathbb{H}_n(\theta_0)\big) - \big(\tilde{\mathbb{H}}_n(\hat{\theta}_n) - \tilde{\mathbb{H}}_n(\theta_0)\big)\big].\end{aligned}$$

Concerned with the first term $\mathbb{H}_n(\hat{\theta}_n) - \mathbb{H}_n(\theta_0)$ inside the sign $E\tilde{E}$, we note that

$$\forall k \geq 1\, \forall l \geq 0, \quad \partial_\alpha^k \partial_\gamma^l \mathbb{H}_n(\theta) = \partial_\alpha^k \partial_\gamma^l \mathbb{H}_{2,n}(\theta).$$

Let $\Delta_{1,n}(\gamma_0) := h\, T_n^{-1/2} \partial_\gamma \mathbb{H}_{1,n}(\gamma_0)$ and $\Delta_{2,n}(\theta_0) := T_n^{-1/2} \partial_\alpha \mathbb{H}_{2,n}(\theta_0)$. We have

$$\begin{aligned}\Gamma_{\gamma,n}(\theta_0) &:= -\frac{1}{n}\partial_\gamma^2 \mathbb{H}_n(\theta_0) \xrightarrow{p} \Gamma_\gamma(\gamma_0),\\ \Gamma_{\alpha,n}(\theta_0) &:= -\frac{1}{T_n}\partial_\alpha^2 \mathbb{H}_{2,n}(\theta_0) \xrightarrow{p} \Gamma_\alpha(\theta_0).\end{aligned} \tag{16}$$

Then, using the third-order Taylor expansion, we can derive the following key asymptotically quadratic structure having two different resolutions for α and γ: for some $\check{\theta}_n = \check{\theta}_n(s) = s\theta_0 + (1-s)\hat{\theta}_n$ with random $s \in [0,1]$,

$$\begin{aligned}\mathbb{H}_n(\hat{\theta}_n) - \mathbb{H}_n(\theta_0) &= \frac{1}{\sqrt{T_n}}\partial_\theta \mathbb{H}_n(\theta_0)[\hat{u}_n] - \frac{1}{2}\left(-\frac{1}{T_n}\partial_\theta^2 \mathbb{H}_n(\theta_0)\right)[\hat{u}_n^{\otimes 2}]\\ &\quad + \frac{1}{\sqrt{T_n}}\left(\frac{1}{6T_n}\partial_\theta^3 \mathbb{H}_n(\check{\theta}_n)\right)[\hat{u}_n^{\otimes 3}]\\ &=: \frac{1}{h}\mathcal{Q}_{1,n}(\gamma_0) + \mathcal{Q}_{2,n}(\theta_0),\end{aligned} \tag{17}$$

where

$$\mathcal{Q}_{1,n}(\gamma_0) := \Delta_{1,n}(\gamma_0)[\hat{u}_{\gamma,n}] - \frac{1}{2}\Gamma_\gamma(\gamma_0)[\hat{u}_{\gamma,n}^{\otimes 2}] + \frac{1}{\sqrt{T_n}}R_{1,n}(\hat{\theta}_n;\theta_0), \tag{18}$$

$$\mathcal{Q}_{2,n}(\theta_0) := \Delta_{2,n}(\theta_0)[\hat{u}_{\alpha,n}] - \frac{1}{2}\Gamma_\alpha(\theta_0)[\hat{u}_{\alpha,n}^{\otimes 2}] + \frac{1}{\sqrt{T_n}}R_{2,n}(\hat{\theta}_n;\theta_0). \tag{19}$$

Here the "remainder" terms are given as follows:

$$\begin{aligned}R_{1,n}(\hat{\theta}_n;\theta_0) &= nh^2\left(\frac{1}{T_n}\partial_\gamma \mathbb{H}_{2,n}(\theta_0)\right)[\hat{u}_{\gamma,n}] - \frac{1}{2}\left(\sqrt{T_n}\big(\Gamma_{n,\gamma}(\gamma_0) - \Gamma_\gamma(\gamma_0)\big)\right)[\hat{u}_{\gamma,n}^{\otimes 2}]\\ &\quad + \left(\frac{h}{6T_n}\partial_\gamma^3 \mathbb{H}_n(\check{\theta}_n)\right)[\hat{u}_{\gamma,n}^{\otimes 3}],\\ R_{2,n}(\hat{\theta}_n;\theta_0) &= \left(\frac{1}{\sqrt{T_n}}\partial_\alpha\partial_\gamma \mathbb{H}_{2,n}(\theta_0)\right)[\hat{u}_{\alpha,n}, \hat{u}_{\gamma,n}] - \frac{1}{2}\left(\sqrt{T_n}\big(\Gamma_{n,\alpha}(\theta_0) - \Gamma_\alpha(\theta_0)\big)\right)[\hat{u}_{\alpha,n}^{\otimes 2}]\\ &\quad + \sum_{r\in\mathbb{Z}_+^p;\,|r|=3,\,|r_\alpha|\geq 1} \frac{1}{6T_n}\partial_\theta^r \mathbb{H}_{2,n}(\check{\theta}_n)\hat{u}_n^r,\end{aligned}$$

where the standard multi-index notation is used for the summation sign in the latter ($\mathbb{Z}_+ := \{0,1,2,\dots\}$). Following the proofs of the basic lemmas in (Masuda 2013, Sections 4.1.2 and 4.1.3), we can deduce that all of the random sequences $\{\Delta_{1,n}(\gamma_0)\}_n$, $\{\Delta_{2,n}(\theta_0)\}_n$, $\{R_{1,n}(\hat{\theta}_n;\theta_0)\}_n$, and $\{R_{2,n}(\hat{\theta}_n;\theta_0)\}_n$ are $L^q(P)$-bounded for every $q > 0$. Moreover, we have (see (Masuda 2013, Lemma 4.6))

$$\big(\Delta_{2,n}(\theta_0),\ \Delta_{1,n}(\gamma_0)\big) \xrightarrow{\mathcal{L}} N_p(0, \Sigma(\theta_0)).$$

Completely analogously to (17), with replacing $\mathbb{H}_n$ by $\tilde{\mathbb{H}}_n$ and so on we obtain the expression

$$\tilde{\mathbb{H}}_n(\hat{\theta}_n) - \tilde{\mathbb{H}}_n(\theta_0) = \frac{1}{h}\tilde{\mathcal{Q}}_{1,n}(\gamma_0) + \tilde{\mathcal{Q}}_{2,n}(\theta_0)$$

with similar structures to (18) and (19). Therefore

$$\mathfrak{b}_n = \frac{1}{h}E\tilde{E}\big[\mathcal{Q}_{1,n}(\gamma_0) - \tilde{\mathcal{Q}}_{1,n}(\gamma_0)\big] + E\tilde{E}\big[\mathcal{Q}_{2,n}(\theta_0) - \tilde{\mathcal{Q}}_{2,n}(\theta_0)\big]. \tag{20}$$

Concerning $\mathcal{Q}_{1,n}(\gamma_0) - \tilde{\mathcal{Q}}_{1,n}(\gamma_0)$, we apply the standard Taylor-expansion argument to obtain

$$\hat{u}_{\gamma,n} = \Gamma_\gamma(\gamma_0)^{-1}\Delta_{1,n}(\gamma_0) + \frac{1}{\sqrt{T_n}}\delta_{\gamma,n}, \tag{21}$$

with $\{\delta_{\gamma,n}\}_n$ being $L^q(P)$-bounded for every $q > 0$. From the stochastic expansions (18) and (21), it follows that

$$\begin{aligned} E\tilde{E}\big[\mathcal{Q}_{1,n}(\theta_0)\big] &= \text{trace}\left\{\Gamma_\gamma(\gamma_0)^{-1}E[\Delta_{1,n}(\gamma_0)^{\otimes 2}] + o(1)\right\} - \frac{1}{2}\Gamma_\gamma(\gamma_0)\Big[E[\hat{u}_{\gamma,n}^{\otimes 2}]\Big] + O(T_n^{-1/2}) \\ &= \text{trace}\left\{\Gamma_\gamma(\gamma_0)^{-1}W_\gamma(\gamma_0)\right\} - \frac{1}{2}\Gamma_\gamma(\gamma_0)\big[\Gamma_\gamma(\gamma_0)^{-1}W_\gamma(\gamma_0)\Gamma_\gamma(\gamma_0)^{-1}\big] + O(T_n^{-1/2}). \end{aligned} \tag{22}$$

Likewise, we can show that $E\tilde{E}[\hat{\mathcal{Q}}_{1,n}(\theta_0)]$ admits the same expansion as in (22) except that the first term on the right-hand side is replaced by $o(h\sqrt{T_n}) = o(1)$: indeed, using the independence between $\boldsymbol{X}_n$ and $\tilde{\boldsymbol{X}}_n$, Burkholder's inequality, $E[\hat{u}_{\gamma,n}] = o(1)$, and also the obvious notation with tilde, we have

$$E\tilde{E}[\tilde{\Delta}_{1,n}(\gamma_0)[\hat{u}_{\gamma,n}]] = \tilde{E}[\tilde{\Delta}_{1,n}(\gamma_0)]\big[E[\hat{u}_{\gamma,n}]\big] = O(h\sqrt{T_n})\cdot o(1) = o(nh^3) = o(1).$$

By subtraction and recalling that $nh^2 \to 0$, we conclude that

$$E\tilde{E}\big[\mathcal{Q}_{1,n}(\gamma_0) - \tilde{\mathcal{Q}}_{1,n}(\gamma_0)\big] = \text{trace}\left\{\Gamma_\gamma(\gamma_0)^{-1}W_\gamma(\gamma_0)\right\} + O(T_n^{-1/2}). \tag{23}$$

Remark 4 In case of $d = r = 1$ we have $W_\gamma(\gamma_0) = \nu(4)\Gamma_\gamma(\gamma_0)/2$ (Remark 2) the first term on the right-hand side of (23) becomes $p_\gamma\nu(4)/2$.

We can handle $\mathcal{Q}_{2,n}(\theta_0) - \tilde{\mathcal{Q}}_{2,n}(\theta_0)$ similarly. As in (21) we can derive the stochastic expansion

$$\hat{u}_{\alpha,n} = \Gamma_\alpha(\theta_0)^{-1}\Delta_{2,n}(\theta_0) + \frac{1}{\sqrt{T_n}}\delta_{\alpha,n}, \tag{24}$$

with $\{\delta_{\alpha,n}\}_n$ being $L^q(P)$-bounded for every $q > 0$. Substituting (24) in (19) and proceeding as in the case of $\mathcal{Q}_{1,n}(\gamma_0) - \tilde{\mathcal{Q}}_{1,n}(\gamma_0)$, we obtain

$$E\tilde{E}\big[\mathcal{Q}_{2,n}(\theta_0) - \tilde{\mathcal{Q}}_{2,n}(\theta_0)\big] = \text{trace}\left\{\Gamma_\alpha(\theta_0)^{-1}\Gamma_\alpha(\theta_0)\right\} + O(T_n^{-1/2}) = p_\alpha + O(T_n^{-1/2}). \tag{25}$$

We have derived the bias expressions for the γ and α parts separately. Combining (20) with (23) and (25), we obtain the following proposition.

Proposition 1 *Suppose that Assumptions* 1, 2, 3, *and* 4 *hold. Then*,

$$\mathfrak{b}_n = \frac{1}{h}\big(\text{trace}\left\{\Gamma_\gamma(\gamma_0)^{-1}W_\gamma(\gamma_0)\right\} + O(T_n^{-1/2})\big) + p_\alpha + O(T_n^{-1/2}). \tag{26}$$

Note that $h^{-1}T_n^{-1/2} \to \infty$. This means that the above form of $\mathfrak{b}_n$ is inconvenient, for the residual term in the γ part becomes stochastically larger than the leading term of the α part. To make the expansion fully explicit in decreasing order, we thus have to further expand $\mathbb{H}_{1,n}(\gamma)$. Rather roughly, recalling (17) and the subsequent paragraphs, we may formally write

$$\mathbb{H}_n(\hat{\theta}_n) - \mathbb{H}_n(\theta_0) = \sum_{k\in\mathbb{N}} \frac{1}{k!}\Big(\frac{1}{n}\partial_\theta^k \mathbb{H}_n(\theta_0)\Big)[\hat{u}_n^{\otimes k}]\, n\, T_n^{-k/2}$$

with each $n^{-1}\partial_\theta^k \mathbb{H}_n(\theta_0)$ being asymptotically non-null (as in (16)). This implies that we need to pick up the terms up to the order k_0 which is the minimal integer such that $n\, T_n^{-k_0/2} \to 0$; since we are assuming that $nh^2 \to 0$, the number k_0 is necessarily greater than or equal to 5.

In sum, because of the mixed-rates structure, the direct evaluation of $\mathfrak{b}_n$ based on the *joint* GQLF $\mathbb{H}_n(\theta)$ necessitates the higher-order derivatives of $\mathbb{H}_n$, resulting in rather complicated expressions. In the next section, we are going to take a different route through a *stepwise* manner to bypass this annoying point.

3.2 Stepwise bias corrections

Building on the observations in the previous section, we can expect that the stepwise AIC procedure will work, making a simple formula of the bias correction for the scale coefficient. Recall the stepwise GQMLE $\hat{\gamma}_n \in \text{argmax}_\gamma\, \mathbb{H}_{1,n}(\gamma)$ and $\hat{\alpha}_n \in \text{argmax}_\alpha\, \mathbb{H}_{2,n}(\alpha)$ introduced in Sect. 2.2.

First, we focus on the relative comparison of the scale coefficient, looking at the quasi-likelihood $\mathbb{H}_{1,n}(\gamma)$. By inspecting the derivation of (26) in Sect. 3.1, we see that the bias

$$\mathfrak{b}_{\gamma,n} := hE\big[\mathbb{H}_{1,n}(\hat{\gamma}_n(\boldsymbol{X}_n);\boldsymbol{X}_n) - \tilde{E}\big[\mathbb{H}_{1,n}(\hat{\gamma}_n(\boldsymbol{X}_n);\tilde{\boldsymbol{X}}_n)\big]\big]$$

admits the expression

$$\mathfrak{b}_{\gamma,n} = \text{trace}\left\{\Gamma_\gamma(\gamma_0)^{-1}W_\gamma(\gamma_0)\right\} + O(T_n^{-1/2}).$$

Hence it is natural to define (dividing by h)

$$\mathrm{GQAIC}_{1,n} := -2\,\mathbb{H}_{1,n}(\hat{\gamma}_n) + \frac{2}{h}\,\mathrm{trace}\left(\hat{\Gamma}_{\gamma,n}^{-1}\hat{W}_{\gamma,n}\right) \tag{27}$$

as the first-stage GQAIC; recall that $\hat{\Sigma}_n \xrightarrow{p} \Sigma(\theta_0)$ and $\hat{\Gamma}_n \xrightarrow{p} \Gamma(\theta_0)$ (Theorem 1). If in particular $d = r = 1$, then we may define (Remark 4)

$$\mathrm{GQAIC}_{1,n} = -2\,\mathbb{H}_{1,n}(\hat{\gamma}_n) + \frac{p_\gamma}{h}\hat{\nu}_n(4), \tag{28}$$

where $\hat{\nu}_n(4)$ is a suitable consistent estimator of $\nu(4)$; it can be conveniently estimated by

$$\hat{\nu}_n(4) := \frac{1}{T_n}\sum_{j=1}^{n}\left(\frac{\Delta_j X}{c_{j-1}(\hat{\gamma}_n)}\right)^4 \xrightarrow{p} \nu(4).$$

Note that this convergence does hold in $L^1(P)$. The penalty term in (27), hence in (28) as well, is stochastically divergent at the non-standard order $1/h$, which is in sharp contrast to the classical AIC and also to the CIC of Uchida (2010). Still, it can be seen that the first term $-2\,\mathbb{H}_{1,n}(\hat{\gamma}_n)$ is the leading one: $-2hn^{-1}\,\mathbb{H}_{1,n}(\hat{\gamma}_n)$ has a non-trivial constant limit in probability, while $2(nh)^{-1}\,\mathrm{trace}(\hat{\Gamma}_{\gamma,n}^{-1}\hat{W}_{\gamma,n}) = O_p(T_n^{-1}) = o_p(1)$.

Having the estimate $\hat{\gamma}_n$ in hand, we proceed to the bias evaluation concerning the drift coefficient. Again inspecting the derivation of (26) in Sect. 3.1, we see that

$$\begin{aligned}
\mathfrak{b}_{\alpha,n} &:= E\big[\mathbb{H}_{2,n}(\hat{\alpha}_n,\hat{\gamma}_n) - \tilde{E}\big[\tilde{\mathbb{H}}_{2,n}(\hat{\alpha}_n,\hat{\gamma}_n)\big]\big] \\
&= E\big[\mathbb{H}_{2,n}(\hat{\alpha}_n,\hat{\gamma}_n) - \mathbb{H}_{2,n}(\alpha_0,\hat{\gamma}_n)\big] - E\tilde{E}\big[\tilde{\mathbb{H}}_{2,n}(\hat{\alpha}_n,\hat{\gamma}_n) - \tilde{\mathbb{H}}_{2,n}(\alpha_0,\hat{\gamma}_n)\big] \\
&\quad + E\tilde{E}\big[\mathbb{H}_{2,n}(\alpha_0,\hat{\gamma}_n) - \tilde{\mathbb{H}}_{2,n}(\alpha_0,\hat{\gamma}_n)\big] \\
&=: \mathfrak{b}_{\alpha,n}^{(1)} - \mathfrak{b}_{\alpha,n}^{(2)} + \mathfrak{b}_{A,n}.
\end{aligned} \tag{29}$$

Using the same devices as in Sect. 3.1 together with the results in (Masuda 2013, Sections 4.1.2 and 4.1.3), we can deduce that

$$\begin{aligned}
\mathfrak{b}_{\alpha,n}^{(1)} &= -E\tilde{E}\left[-\frac{1}{\sqrt{T_n}}\partial_\alpha\mathbb{H}_{2,n}(\hat{\alpha}_n,\hat{\gamma}_n)[\hat{u}_{\alpha,n}] + \frac{1}{2}\left(-\frac{1}{T_n}\partial_\alpha^2\mathbb{H}_{2,n}(\check{\alpha}_n,\hat{\gamma}_n)[\hat{u}_{\alpha,n}^{\otimes 2}]\right)\right] \\
&= o(1) + \frac{1}{2}\,\mathrm{trace}\left(\Gamma_\alpha(\theta_0)E[\hat{u}_{\alpha,n}^{\otimes 2}]\right) = o(1) + \frac{1}{2}p_\alpha,
\end{aligned} \tag{30}$$

and that for some $\check{\alpha}_n' = \check{\alpha}_n'(s') = s'\alpha_0 + (1-s')\hat{\alpha}_n$ with random $s' \in [0,1]$,

$$\begin{aligned}
\mathfrak{b}_{\alpha,n}^{(2)} &= -E\tilde{E}\left[\frac{1}{\sqrt{T_n}}\partial_\alpha\tilde{\mathbb{H}}_{2,n}(\hat{\alpha}_n,\hat{\gamma}_n)[\hat{u}_{\alpha,n}] - \frac{1}{2}\left(-\frac{1}{T_n}\partial_\alpha^2\tilde{\mathbb{H}}_{2,n}(\check{\alpha}_n',\hat{\gamma}_n)[\hat{u}_{\alpha,n}^{\otimes 2}]\right)\right] \\
&= o(1) - \frac{1}{2}\,\mathrm{trace}\left(\Gamma_\alpha(\theta_0)E[\hat{u}_{\alpha,n}^{\otimes 2}]\right) = o(1) - \frac{1}{2}p_\alpha;
\end{aligned} \tag{31}$$

in part, we used the facts that $P[\partial_\alpha\tilde{\mathbb{H}}_{2,n}(\hat{\alpha}_n,\hat{\gamma}_n) = 0] \to 1$ and that $\{T_n^{-1/2}\partial_\alpha\tilde{\mathbb{H}}_{2,n}(\hat{\alpha}_n,\hat{\gamma}_n)[\hat{u}_{\alpha,n}]\}_n$ is L^q bounded for any $q > 0$. Thus we have obtained

$$\mathfrak{b}_{\alpha,n} = p_\alpha + \mathfrak{b}_{A,n} + o(1).$$

Recall that we are assuming that scale and drift coefficients are correctly specified. Since $\mathfrak{b}_{A,n} = E\tilde{E}[\mathbb{H}_{2,n}(\alpha_0, \hat{\gamma}_n) - \tilde{\mathbb{H}}_{2,n}(\alpha_0, \hat{\gamma}_n)]$ is independent of the drift estimator $\hat{\alpha}_n$ and, given a $\hat{\gamma}_n$, is common to all the candidates, we may and do ignore $\mathfrak{b}_{A,n}$ in relative model comparison. Therefore, we define the second-stage GQAIC as the usual form

$$\mathrm{GQAIC}_{2,n} := -2\,\mathbb{H}_{2,n}(\hat{\alpha}_n) + 2p_\alpha. \tag{32}$$

Summarizing the above observations yields the following result.

Theorem 2 *Suppose that Assumptions 1, 2, 3, and 4 hold true, and that*

$$E\left[\mathrm{trace}\left(\hat{\Gamma}_{\gamma,n}^{-1}\hat{W}_{\gamma,n}\right)\right] \to \mathrm{trace}\left\{\Gamma_\gamma(\gamma_0)^{-1}W_\gamma(\gamma_0)\right\}. \tag{33}$$

Then, we have $\mathfrak{b}_n = h^{-1}\mathfrak{b}_{\gamma,n} + \mathfrak{b}_{\alpha,n}$, *where*

$$\mathfrak{b}_{\gamma,n} = \mathrm{trace}\left\{\Gamma_\gamma(\gamma_0)^{-1}W_\gamma(\gamma_0)\right\} + O(T_n^{-1/2}), \tag{34}$$

$$\mathfrak{b}_{\alpha,n} = p_\alpha + \mathfrak{b}_{A,n} + o(1). \tag{35}$$

The equations (23), (25), and (34) are derived similarly, while (35) is derived by dividing the bias into three parts and expanding them. The point here is that, by considering $\mathfrak{b}_{\gamma,n}$ and $\mathfrak{b}_{\alpha,n}$ separately, we can bypass the problem of the residual term in the γ part being stochastically larger than the leading term in the α part; recall the expression (26).

Here is a simple sufficient condition for (33).

Proposition 2 *Under the assumptions in Theorem 2, (33) is implied by*

$$\exists\delta > 0,\ \exists N \in \mathbb{N},\quad \sup_{n\geq N} E\left[\lambda_{\min}^{-(1+\delta)}(\hat{\Gamma}_{\gamma,n})\right] < \infty. \tag{36}$$

Unfortunately, verification of (36) may not be technically trivial. Naively, if the off-diagonal elements of $\hat{\Gamma}_{\gamma,n}$ are small enough in magnitude, a simple sufficient condition for (36) can be given through the Gerschgorin circle theorem, which ensures that $|\lambda_i| \geq |a_{ii}| - \sum_{j\neq i}|a_{ij}|$ for any eigenvalue λ_i of a square matrix $A = (a_{ij})$: writing $M(x,\gamma) = [M^{(k)}(x,\gamma)]_{k=1}^{p_\gamma}$ with $M^{(k)}(x,\gamma) := (S^{-1}\partial_{\gamma_k}S)(x,\gamma)$, we have

$$\begin{aligned}\lambda_{\min}(\hat{\Gamma}_{\gamma,n}) &\geq \min_{1\leq k\leq d}\left(\hat{\Gamma}_{\gamma,n}^{(kk)} - \sum_{l\neq k}|\hat{\Gamma}_{\gamma,n}^{(kl)}|\right)\\ &= \min_{1\leq k\leq d}\frac{1}{2n}\sum_{j=1}^{n}\left(\mathrm{trace}\left[(\hat{M}_{j-1}^{(k)})^2\right] - \sum_{l\neq k}\left|\mathrm{trace}\left[\hat{M}_{j-1}^{(k)}\hat{M}_{j-1}^{(l)}\right]\right|\right).\end{aligned}$$

Then, (36) holds if

$$\inf_{\gamma} \min_{1\le k\le d} \left(\mathrm{trace}\left[M^{(k)}(x,\gamma)^2\right] - \sum_{l\neq k} \left|\mathrm{trace}\left[M^{(k)}(x,\gamma)\, M^{(l)}(x,\gamma)\right]\right| \right) \gtrsim (1+|x|^C)^{-1}.$$

Things become simpler if, for example, we assume the spectral representation $S(x,\gamma) = \sum_k \lambda_k(x,\gamma)\Pi_k(x)$ for some positive functions $\lambda_k(x,\gamma)$ and some projection-valued ones $\Pi_k(x)$. We will get a little bit further into the issue of bounding inverse moments (36) in Sect. 3.3.

For the convergence of moments (33), we could bypass the non-trivial bound (36) by suitably truncating the minimum eigenvalue $\lambda_{\min}(\hat{\Gamma}_{\gamma,n})$. Specifically, define

$$\hat{\Gamma}_{\gamma,n}^{-1}(b_n) := \hat{\Gamma}_{\gamma,n}^{-1} I\big(\lambda_{\min}(\hat{\Gamma}_{\gamma,n}) \ge b_n\big)$$

for some positive sequence (b_n) such that $b_n \to 0$ and that

$$\exists \kappa > 0, \quad b_n \gtrsim T_n^{-(1-\kappa)/2}. \tag{37}$$

Let $\lambda_n := \lambda_{\min}(\hat{\Gamma}_{\gamma,n})$ and $\lambda_0 := \lambda_{\min}(\Gamma_\gamma(\gamma_0)) > 0$ for brevity. Then, since $\lambda_n \xrightarrow{p} \lambda_0$, we have

$$\hat{\Gamma}_{\gamma,n}^{-1}(b_n) \xrightarrow{p} \Gamma_\gamma(\gamma_0)^{-1}.$$

Observe that $\lambda_n = \inf_{|u|=1} u^\top \hat{\Gamma}_{\gamma,n} u \ge \lambda_0 - |\hat{\Gamma}_{\gamma,n} - \Gamma_\gamma(\gamma_0)|$. Also, $\sup_n E[|\sqrt{T_n}(\hat{\Gamma}_{\gamma,n} - \Gamma_\gamma(\gamma_0))|^q] < \infty$ for any $q > 0$ under the present assumptions; see (Masuda 2013) for details. Building on the above observations, for $K > 1$ and $q > 0$ and for n large enough,

$$\begin{aligned}
E\left[\left|\hat{\Gamma}_{\gamma,n}^{-1}(b_n)\right|^K\right] &= E\left[\left|\hat{\Gamma}_{\gamma,n}^{-1}\right|^K; \lambda_n \ge b_n\right] \lesssim E\left[\lambda_n^{-K}; \lambda_n \ge b_n\right] \\
&= \int_0^\infty P\left[\lambda_n^{-K} I(\lambda_n \ge b_n) \ge x\right] dx \\
&= \int_0^\infty P\left[b_n \le \lambda_n \le x^{-1/K}\right] dx \\
&\le 1 + \int_1^{b_n^{-K}} P\left[\lambda_n \le x^{-1/K}\right] dx \\
&\le 1 + \int_1^{b_n^{-K}} P\left[\lambda_0 \le T_n^{-1/2} |\sqrt{T_n}(\hat{\Gamma}_{\gamma,n} - \Gamma_\gamma(\gamma_0))| + x^{-1/K}\right] dx \\
&\lesssim 1 + \int_1^{b_n^{-K}} \left(P\left[\left|\sqrt{T_n}(\hat{\Gamma}_{\gamma,n} - \Gamma_\gamma(\gamma_0))\right| \ge \sqrt{T_n} x^{-1/K}\right] + P\left[\lambda_0 \le 2x^{-1/K}\right] \right) dx \\
&\lesssim 1 + \int_1^{b_n^{-K}} P\left[\left|\sqrt{T_n}(\hat{\Gamma}_{\gamma,n} - \Gamma_\gamma(\gamma_0))\right| \ge \sqrt{T_n} x^{-1/K}\right] dx \\
&\lesssim 1 + \left(\sup_n E\left[\left|\sqrt{T_n}(\hat{\Gamma}_{\gamma,n} - \Gamma_\gamma(\gamma_0))\right|^q\right] \right) T_n^{-q/2} \int_1^{b_n^{-K}} x^{q/K} ds \\
&\lesssim 1 + T_n^{-q/2} b_n^{-q-K}.
\end{aligned}$$

The rightmost side is bounded in n if $b_n \gtrsim T_n^{-(1-K/(q+K))/2}$, which holds under (37). In sum, under the additional ad-hoc tuning (37), we could remove the requirement (33) by adopting

$$\mathrm{GQAIC}^{\flat}_{1,n} := -2\,\mathbb{H}_{1,n}(\hat{\gamma}_n) + \frac{2}{h}\,\mathrm{trace}\left(\hat{\Gamma}^{-1}_{\gamma,n}(b_n)\hat{W}_{\gamma,n}\right) \tag{38}$$

instead of (27); a similar modification can be applied to (40) below as well.

Remark 5 Inspecting the derivation it is trivial that the same bias corrections as in Theorem 2 hold even if we replace $\mathbb{H}_{2,n}(\theta)$ of (9) by $\mathbb{H}^*_{2,n}(\theta)$ of (10) all through bias evaluation at the second step.

Remark 6 (GQAIC for diffusion) In relation to Remark 3, it is worth mentioning the case of a diffusion process where Z is an r-dimensional standard Wiener process w.

(1) Then, the first and second GQAICs become $-2\mathbb{H}_{1,n}(\hat{\gamma}_n) + 2p_\gamma$ and $-2\mathbb{H}_{2,n}(\hat{\alpha}_n) + 2p_\alpha$, respectively. Moreover, in the case of the joint estimation through the GQLF $\mathbb{H}_n(\theta)$ of (8), the GQAIC is given by $-2\mathbb{H}_n(\hat{\theta}_n) + 2(p_\alpha + p_\gamma)$, showing that the GQAIC takes the same form as in the CIC of Uchida (2010), the contrast information criterion. We omit the technical details of these observations, for they can be derived in an analogous way to the Lévy SDE case, with an essential difference that, although in this case the rates of convergence are different for the diffusion and drift parameters, we can simultaneously normalize the associated random field by a single matrix to conclude the locally asymptotically quadratic structure (see (Yoshida 2011, Sect. 6) for details):

$$\begin{aligned} u = (u_\alpha, u_\gamma) \mapsto\ & \mathbb{H}_n(\alpha_0 + T_n^{-1/2}u_\alpha, \gamma_0 + n^{-1/2}u_\gamma) - \mathbb{H}_n(\theta_0) \\ =\ & \Delta_n(\theta_0)[u] - \frac{1}{2}\Gamma(\theta_0)[u,u] + o_p(1), \qquad u \in \mathbb{R}^p. \end{aligned}$$

This implies that the associated statistical random field does not have the mixed-rates structure of such as (13). The difference between the Gaussian and the non-Gaussian cases comes from the fact that the random sequence $(h^{-1/2}w_h)_{h>0}$ is $L^K(P)$-bounded for any $K > 0$ whereas it is not the case for $(h^{-1/2}Z_h)_{h>0}$ with non-Gaussian Lévy process Z; more specifically, it is known that in the one-dimensional case,

$$\lim_{h\to 0}\frac{1}{h}E[|Z_h|^K] = \int |z|^K \nu(dz)$$

for $K > 2$ if $E[Z_1] = 0$ and $E[|Z_1|^K] < \infty$ (see (Asmussen and Rosiński 2001, Lemma 3.1)). This is also why the matrix rate of convergence D_n mentioned in Remark 3 emerges in the diffusion case. We refer to (Masuda 2013, Section 4.1.1) for related remarks.

(2) Write $o_p^*(1)$ for a random sequence $(\zeta_n)_n$ such that $E(|\zeta_n|^q) \to 0$ for any $q > 0$. It can be shown that $\hat{\Gamma}^{(kl)}_{\gamma,n} \xrightarrow{p} \Gamma^{(kl)}_{\gamma}(\gamma_0)$ and that (recall (11)), through repeated compensations,

$$
\begin{aligned}
\frac{2}{h}\hat{W}^{(qr)}_{\gamma,n} &= \frac{1}{2n}\sum_{j=1}^{n}\left\{\left(\hat{S}_{j-1}^{-1}(\partial_{\gamma_q}\hat{S}_{j-1})\hat{S}_{j-1}^{-1}\right)\left[\left(\frac{\hat{\chi}_j}{\sqrt{h}}\right)^{\otimes 2}\right]\right\} \\
&\quad\times\left\{\left(\hat{S}_{j-1}^{-1}(\partial_{\gamma_r}\hat{S}_{j-1})\hat{S}_{j-1}^{-1}\right)\left[\left(\frac{\hat{\chi}_j}{\sqrt{h}}\right)^{\otimes 2}\right]\right\} \\
&= \frac{1}{2n}\sum_{j=1}^{n}\left\{\left(S_{j-1}^{-1}(\partial_{\gamma_q}S_{j-1})\right)\left[\left(\frac{\Delta_j w}{\sqrt{h}}\right)^{\otimes 2}\right]\right\} \\
&\quad\times\left\{\left(S_{j-1}^{-1}(\partial_{\gamma_r}S_{j-1})\right)\left[\left(\frac{\Delta_j w}{\sqrt{h}}\right)^{\otimes 2}\right]\right\} + o_p^*(1) \\
&= \frac{1}{2n}\sum_{j=1}^{n}E^{j-1}\left[\left\{\left(S_{j-1}^{-1}(\partial_{\gamma_q}S_{j-1})\right)\left[\left(\frac{\Delta_j w}{\sqrt{h}}\right)^{\otimes 2}\right]\right\}\right. \\
&\quad\left.\times\left\{\left(S_{j-1}^{-1}(\partial_{\gamma_r}S_{j-1})\right)\left[\left(\frac{\Delta_j w}{\sqrt{h}}\right)^{\otimes 2}\right]\right\}\right] + o_p^*(1) \\
&= \frac{1}{2n}\sum_{j=1}^{n}\left\{2\,\mathrm{trace}\left(S_{j-1}^{-1}(\partial_{\gamma_q}S_{j-1})S_{j-1}^{-1}(\partial_{\gamma_r}S_{j-1})\right)\right. \\
&\quad\left.+\,\mathrm{trace}\left(\left(S_{j-1}^{-1}(\partial_{\gamma_q}S_{j-1})\right)\otimes\left(S_{j-1}^{-1}(\partial_{\gamma_r}S_{j-1})\right)\right)\right\} + o_p^*(1) \\
&= 2\Gamma^{(qr)}_{\gamma}(\gamma_0) + \frac{1}{2n}\sum_{j=1}^{n}\mathrm{trace}\left(\left(\hat{S}_{j-1}^{-1}(\partial_{\gamma_q}\hat{S}_{j-1})\right)\otimes\left(\hat{S}_{j-1}^{-1}(\partial_{\gamma_r}\hat{S}_{j-1})\right)\right) + o_p^*(1),
\end{aligned}
\tag{39}
$$

where we used the identity (Magnus and Neudecker 1979, Theorem 4.2(i)) for the fourth equality. Write $\hat{A}^{(qr)}_{\gamma,n}$ for the second term in (39), and let $\hat{A}_{\gamma,n} := (\hat{A}^{(qr)}_{\gamma,n})_{q,r}$; obviously, we have $\hat{A}_{\gamma,n} = O_p^*(1)$. The identity (39) suggests us use the following modified version

$$
\mathrm{GQAIC}_{1,n} = -2\,\mathbb{H}_{1,n}(\hat{\gamma}_n) + \mathrm{trace}\left\{\hat{\Gamma}^{-1}_{\gamma,n}\left(\frac{2}{h}\hat{W}_{\gamma,n} - \hat{A}_{\gamma,n}\right)\right\} \tag{40}
$$

instead of (27) as an alternative that can be used for both diffusion and Lévy driven SDE in common, in exchange for a slight additional computational cost. In particular for $d = 1$, instead of (28) we could use

$$
\mathrm{GQAIC}_{1,n} = -2\,\mathbb{H}_{1,n}(\hat{\gamma}_n) + p_\gamma\left(\frac{1}{h}\hat{\nu}_n(4) - \hat{\nu}_n(2)^2\right),
$$

since $h^{-1}\hat{\nu}_n(4) \xrightarrow{p} 3$ and $\hat{\nu}_n(2) := T_n^{-1}\sum_{j=1}^{n} c_{j-1}(\hat{\gamma}_n)^{-2}(\Delta_j X)^2 \xrightarrow{p} 1$, with the latter holding for both diffusion and Lévy driven SDE while the former only for diffusion; or, more simply we could use $\mathrm{GQAIC}_{1,n} = -2\,\mathbb{H}_{1,n}(\hat{\gamma}_n) + p_\gamma\left(h^{-1}\hat{\nu}_n(4) - 1\right)$ in common.

3.3 Inverse-moment bound

In this section, we revisit (36). For notational simplicity, we write

$$\zeta(x,\gamma) = (\partial_\gamma \log|S|)(x,\gamma)$$

and $\mathbb{S} := \{u \in \mathbb{R}^{p_\gamma} : |u| = 1\}$. We will prove the following criterion.

Lemma 1 *Suppose that the assumptions given in Sect. 2.2 hold. Moreover, suppose that there exist positive constants ρ and C' such that for each $\epsilon \in (0,1]$,*

$$\sup_n \sup_{2\le j\le n} \sup_{u\in\mathbb{S}} \sup_\gamma P\left[\left|\zeta(X_{t_{j-1}},\gamma)[u]\right| < \epsilon \middle| \mathcal{F}_{t_{j-2}}\right] \le C'\epsilon^\rho \quad \text{a.s.} \tag{41}$$

Then (36) holds.

Note that the bound (41) is implied by

$$\sup_n \sup_{2\le j\le n} \sup_\gamma P\left[\lambda_{\min}\left(\zeta(X_{t_{j-1}},\gamma)^{\otimes 2}\right) < \epsilon^2 \middle| \mathcal{F}_{t_{j-2}}\right] \le C'\epsilon^\rho \quad \text{a.s.}$$

The proof of Lemma 1 utilizes the technique which dates back to Bhansali and Papangelou (1991), later improved and generalized by Findley and Wei (2002) and Chan and Ing (2011): For the sake of reference, we will give the proof in an almost self-contained form. Write $q = 1+\delta$ in what follows. We recall the following lemma for later reference.

Lemma 2 (Lemma A.3 in Findley and Wei (2002)) *For $\delta \in (0,1)$, there exists a finite subset $\mathbb{S}(\delta) \subset \mathbb{S}$ such that*:

(1) *$\mathbb{S}(\delta)$ has at most $\lfloor C_{p_\gamma}\delta^{-(p_\gamma-1)}\rfloor$ elements for some constant C_{p_γ} only depending on p_γ;*
(2) *For each $u \in \mathbb{S}$ there exists an element $v \in \mathbb{S}(\delta)$ for which $|u-v| < \delta$.*

Here is a direct corollary to Lemma 1.

Corollary 1 *Suppose that the assumptions given in Sect. 2.2 hold. Further, suppose that there exists a nonnegative measurable function $\underline{\lambda}(x)$ for which*

$$\inf_\gamma \lambda_{\min}\left(\zeta(x,\gamma)^{\otimes 2}\right) \ge \underline{\lambda}(x),$$

and for every $\Delta > 0$ small enough,

$$\sup_{t\ge 0} P\left(\underline{\lambda}(X_{t+\Delta}) \le \epsilon \,\middle|\, \mathcal{F}_t\right) \lesssim \epsilon^\rho \quad \text{a.s.}$$

for $\epsilon \in (0, 1]$. *Then* (36) *holds for any N large enough.*

We remark that the "tuning" parameter k in the proof of Lemma 1 plays a role to relieve possible high concentration probability of $\lambda_{\min}(\hat{\Gamma}_{\gamma,n})$ around the origin; obviously, it is redundant under the present assumptions if the stronger non-degeneracy condition of $\zeta(x, \gamma)$ holds:

Corollary 2 *Suppose that the assumptions given in Sect.* 2.2 *hold. Further, suppose that*

$$\inf_{\gamma} \lambda_{\min}\big(\zeta(x,\gamma)^{\otimes 2}\big) \gtrsim (1+|x|)^{-C}. \tag{42}$$

Then (36) *holds for any N large enough.*

4 Gaussian quasi-BIC

In this section, we consider a two-stage Schwarz's type Bayesian information criterion, termed Gaussian quasi-BIC (GQBIC), through the GQLF. We keep using the notation introduced in Sect. 2. Suppose that Assumptions 1, 2, 3, and 4 hold true. In addition, we consider the prior densities $\pi_1(\gamma)$ and $\pi_2(\alpha)$ for α and γ, respectively. We assume that both π_1 and π_2 are continuous and bounded in $\overline{\Theta}_\gamma$ and $\overline{\Theta}_\alpha$ respectively, and moreover that $\pi_1(\gamma_0) > 0$ and $\pi_2(\alpha_0) > 0$. Moreover, for a technical reason, throughout this section we assume that there exists a constant $c_1 \in (0, 1)$ for which

$$T_n \gtrsim n^{c_1}. \tag{43}$$

This is a real restriction in addition to $nh^2 \to 0$ as is seen by the example $h = n^{-1}\log n$.

4.1 Scale

We introduce the stochastic expansion of the *free energy at the inverse temperature* $\mathfrak{b} > 0$, which is defined using the negative normalized logarithmic partition function (we refer to Watanabe (2013) for relevant backgrounds):

$$\mathfrak{F}_{1,n}(\mathfrak{b}) := -\frac{1}{n\mathfrak{b}} \log \left(\int_{\Theta_\gamma} \exp\{\mathfrak{b}\, \mathbb{H}_{1,n}(\gamma)\} \pi_1(\gamma) d\gamma \right).$$

Here, the terminology "normalized" means that $\mathfrak{F}_{1,n}(\mathfrak{b})$ has non-trivial limit (in probability) for each $\mathfrak{b} > 0$. The normalized marginal quasi-log likelihood corresponds to $\mathfrak{F}_{1,n}(1)$, and the classical BIC methodology is based on a stochastic expansion of $\mathfrak{F}_{1,n}(1)$. See (Eguchi and Masuda 2018) and the references therein.

We will prove the following expansions in Sect. 7.6.

Theorem 3 *We have the following stochastic expansions*:

$$\mathfrak{F}_{1,n}(1) = -\frac{1}{n}\mathbb{H}_{1,n}(\hat{\gamma}_n) + \frac{p_\gamma}{2n}\log n + O_p\Big(\frac{1}{n}\Big), \tag{44}$$

$$\mathfrak{F}_{1,n}(h) = -\frac{1}{n}\mathbb{H}_{1,n}(\hat{\gamma}_n) + \frac{p_\gamma}{2T_n}\log T_n + O_p\Big(\frac{1}{T_n}\Big). \tag{45}$$

The first one (44) was previously given in Eguchi and Uehara (2021), based on which the authors introduced the GQBIC for the scale by

$$\mathrm{GQBIC}^{\sharp}_{1,n} := -2\mathbb{H}_{1,n}(\hat{\gamma}_n) + p_\gamma \log n.$$

Theorem 6 below revises the incorrect part of (Eguchi and Uehara 2021, Theorem 3.2), showing that $\mathrm{GQBIC}^{\sharp}_{1,n}$ does *not* bring about the model-selection consistency.

Instead, building on (45), we propose to use

$$\mathrm{GQBIC}_{1,n} := -2\,\mathbb{H}_{1,n}(\hat{\gamma}_n) + \frac{p_\gamma}{h}\log T_n. \tag{46}$$

In Theorem 7 below, it will show that this form has the model-selection consistency. This implies that in the present Lévy driven SDE setting, we need to "heat up" the quasi-likelihood $\mathbb{H}_{1,n}$ by multiplying h^{-1}.

4.2 Drift

Different from the previous scale case, the second stage QBIC corresponding to $\mathbb{H}_{2,n}(\alpha)$ is standard: we do not need to heat up the second-stage quasi-likelihood $\mathbb{H}_{1,n}(\alpha)$. We can directly look at the normalized marginal quasi-log likelihood

$$\mathfrak{F}_{2,n} = \mathfrak{F}_{2,n}(1) := -\frac{1}{T_n}\log\left(\int_{\Theta_\alpha}\exp\{\mathbb{H}_{2,n}(\alpha)\}\pi_2(\alpha)d\alpha\right).$$

We have the following stochastic expansion:

Theorem 4

$$\mathfrak{F}_{2,n}(1) = -\frac{1}{T_n}\mathbb{H}_{2,n}(\hat{\alpha}_n) + \frac{p_\alpha}{2T_n}\log T_n + O_p\Big(\frac{1}{T_n}\Big). \tag{47}$$

Theorem 4 can be proved similarly to the proof of (45) in Theorem 3. The proof of the stochastic expansion (47) is much simpler, and we omit the proof.

Ignoring the vanishing term $O_p(T_n^{-1})$ of $\mathfrak{F}_{2,n}$ (just as in Eguchi and Masuda (2018)), we introduce the GQBIC for the drift in the same form as in Eguchi and Uehara (2021):

$$\mathrm{GQBIC}_{2,n} = -2\,\mathbb{H}_{2,n}(\hat{\alpha}_n) + p_\alpha \log T_n. \tag{48}$$

In the next section, we will formulate a two-step selection procedure for both GQAIC and GQBIC.

Remark 7 The proposed two-stage methodology itself is simple enough, and we believe that, in principle, it can be applied to other types of quasi-likelihoods such as the non-Gaussian stable one ((Clément and Gloter 2020; Jasra et al. 2019; Masuda 2019), and Masuda (2023) for details). As long as considering the ergodic case, the derivation of the AIC-type statistics remains valid since what is essential therein is the convergence of moments of the asymptotically normally distributed estimator. The case of BIC-type statistics is easier to handle, for it is only based on the rate of convergence of the estimator; of more interest is that different from the AIC type, it is not essential for the quasi-BIC statistics that the model is ergodic (see (Eguchi and Masuda 2018) and also (Eguchi and Masuda 2019, Appendix)). Whatever the case, careful consideration and calculation are needed in terms of models. We would like to leave these issues to future tasks.

5 Model comparison and asymptotic probability of relative model selection

In this section, we consider relative (pairwise) model selection probabilities of the GQAIC and GQBIC. Below, we assume that $0 < \sharp\mathfrak{M}_1 < M_1$ and $0 < \sharp\mathfrak{M}_2 < M_2$, where $\sharp\mathfrak{M}_1$ and $\sharp\mathfrak{M}_2$ denote the numbers of elements of $\mathfrak{M}_1$ and $\mathfrak{M}_2$, respectively, with

$$\mathfrak{M}_1 := \{m_1 \in \left\{1, \dots, M_1\right\} : \text{there exists a } \gamma_{m_1,0} \in \Theta_{\gamma_{m_1}} \text{ such that } c_{m_1}(\cdot, \gamma_{m_1,0}) = C(\cdot)\Big\},$$
$$\mathfrak{M}_2 := \{m_2 \in \left\{1, \dots, M_2\right\} : \text{there exists a } \alpha_{m_2,0} \in \Theta_{\alpha_{m_2}} \text{ such that } a_{m_2}(\cdot, \alpha_{m_2,0}) = A(\cdot)\Big\}.$$

This means that the candidate coefficients $c_1, \dots, c_{M_1}$ and $a_1, \dots, a_{M_2}$ contain both correctly specified coefficients and misspecified coefficients. In cases of Lévy driven SDEs where either or both of the drift and scale coefficients are misspecified, the asymptotic properties of estimators are shown in Uehara (2019) under suitable conditions. Also, we formally use the $\mathrm{GQAIC}_{1,n}$ and $\mathrm{GQAIC}_{2,n}$ even for the possibly misspecified coefficients, although the assumptions of Theorem 2 may not hold. Using the GQAIC, the stepwise model comparison is performed as follows.

(i) We compute $\mathrm{GQAIC}_{1,n}$ for each candidate scale coefficient, say $\mathrm{GQAIC}_{1,n}^{(1)}, \dots, \mathrm{GQAIC}_{1,n}^{(M_1)}$, and select the best scale coefficient $c_{\hat{m}_{1,n}}$ having the minimum $\mathrm{GQAIC}_{1,n}$-value:

$$\{\hat{m}_{1,n}\} = \underset{1 \le m_1 \le M_1}{\operatorname{argmin}} \, \mathrm{GQAIC}_{1,n}^{(m_1)}.$$

(ii) Under the result of (i), we choose the best drift coefficient with index $\hat{m}_{2,n}$ such that

$$\{\hat{m}_{2,n}\} = \underset{1\le m_2\le M_2}{\operatorname{argmin}} \mathrm{GQAIC}_{2,n}^{(m_2|\hat{m}_{1,n})},$$

where $\mathrm{GQAIC}_{2,n}^{(m_2|m_{1,n})}$ corresponds to (32) with $c_{m_{1,n}}$ and $\hat{\gamma}_{m_{1,n},n}$.

The total number of comparisons in this procedure is $M_1 + M_2$, and we can obtain the model $\mathcal{M}_{\hat{m}_{1,n},\hat{m}_{2,n}}$ as the final best model among the candidates. When we use GQBIC for model comparison, the best model is selected by a similar procedure.

Let the functions $\mathbb{H}_{1,n}^{(m_1)}$ and $\mathbb{H}_{2,n}^{(m_2|m_1)}$ denote $\mathbb{H}_{1,n}$ and $\mathbb{H}_{2,n}$ in each candidate model $\mathcal{M}_{m_1,m_2}$, respectively. Then, we have

$$\begin{aligned}\frac{1}{n}\mathbb{H}_{1,n}^{(m_1)}(\gamma_{m_1}) &\xrightarrow{p} -\frac{1}{2}\int_{\mathbb{R}^d}\left\{\mathrm{trace}\left(S(x,\gamma_{m_1})^{-1}S(x)\right) + \log\left|S(x,\gamma_{m_1})\right|\right\}\pi(dx)\\ &=: \mathbb{H}_{1,0}^{(m_1)}(\gamma_{m_1}),\end{aligned}$$

where $S(x) = C(x)^{\otimes 2}$. We assume that the optimal scale parameter $\gamma_{m_1}^*$ and scale index set $\mathfrak{M}_1^*$ are defined as

$$\begin{aligned}\{\gamma_{m_1}^*\} &= \underset{\gamma_{m_1}}{\operatorname{argmax}}\, \mathbb{H}_{1,0}^{(m_1)}(\gamma_{m_1}),\\ \mathfrak{M}_1^* &= \underset{m_1\in\mathfrak{M}_1}{\operatorname{argmin}} \dim(\Theta_{\gamma_{m_1}}),\end{aligned}$$

respectively. For any $m_1 \in \mathfrak{M}_1$, $\gamma_{m_1}^* = \gamma_{m_1,0}$.

Next, for any fixed $m_1 \in \{1,\dots,M_1\}$,

$$\begin{aligned}\frac{1}{T_n}\mathbb{H}_{2,n}^{(m_2|m_1)}(\alpha_{m_2}) &\xrightarrow{p} -\frac{1}{2}\int_{\mathbb{R}^d} S^{-1}(x,\gamma_{m_1}^*)\left[\left(a_{m_2}(x,\alpha_{m_2}) - A(x)\right)^{\otimes 2}\right]\pi(dx)\\ &=: \mathbb{H}_{2,0}^{(m_2|m_1)}(\alpha_{m_2}),\end{aligned}$$

and assume that the optimal drift parameter $\alpha_{m_2}^*$ is given by maximizing $\mathbb{H}_{2,0}^{(m_2|m_1)}$:

$$\{\alpha_{m_2}^*\} = \underset{\alpha_{m_2}}{\operatorname{argmax}}\, \mathbb{H}_{2,0}^{(m_2|m_1)}(\alpha_{m_2}).$$

When m_2 is included in $\mathfrak{M}_2$, $\alpha_{m_2}^* = \alpha_{m_2,0}$. We also suppose that the drift index set $\mathfrak{M}_2^*$ is defined as

$$\mathfrak{M}_2^* = \underset{m_2\in\mathfrak{M}_2}{\operatorname{argmin}} \dim(\Theta_{\alpha_{m_2}}).$$

From the assumptions and definitions of $\mathbb{H}_{1,0}^{(m_1)}$ and $\mathbb{H}_{2,0}^{(m_2|m_1)}$, $\mathfrak{M}_1 = \operatorname{argmax}_{m_1} \mathbb{H}_{1,0}^{(m_1)}(\gamma_{m_1}^*)$ and $\mathfrak{M}_2 = \operatorname{argmax}_{m_2} \mathbb{H}_{2,0}^{(m_2|m_1)}(\alpha_{m_2}^*)$ hold.

Let $\Theta_{\gamma_{i_1}} \times \Theta_{\alpha_{i_2}} \subset \mathbb{R}^{p_{\gamma_{i_1}}} \times \mathbb{R}^{p_{\alpha_{i_2}}}$ and $\Theta_{\gamma_{j_1}} \times \Theta_{\alpha_{j_2}} \subset \mathbb{R}^{p_{\gamma_{j_1}}} \times \mathbb{R}^{p_{\alpha_{j_2}}}$ be the parameter space associated with model $\mathcal{M}_{i_1,i_2}$ and $\mathcal{M}_{j_1,j_2}$, respectively. If $p_{\gamma_{i_1}} < p_{\gamma_{j_1}}$ and there

exists a matrix $F_1 \in \mathbb{R}^{p_{\gamma_{j_1}} \times p_{\gamma_{i_1}}}$ with $F_1^\top F_1 = I_{p_{\gamma_{i_1}} \times p_{\gamma_{i_1}}}$ as well as a $\mathbb{H}_{1,n}^{(i_1)}(\gamma_{i_1}) = \mathbb{H}_{1,n}^{(j_1)}(F_1\gamma_{i_1} + c_1)$ for all $\gamma_{i_1} \in \Theta_{\gamma_{i_1}}$, we say $\Theta_{\gamma_{i_1}}$ is nested in $\Theta_{\gamma_{j_1}}$. It is defined in a similar manner that $\Theta_{\alpha_{i_2}}$ is nested in $\Theta_{\alpha_{j_2}}$.

Now we are in a position to state the results. The theoretical properties of the GQAIC are given in Theorem 5, and those of the GQBIC in Theorems 6 and 7. For convenience, we give the summaries before the statements:

- Theorem 5 1(i) and 2(i) reveal that the probability of relative selection is asymptotically characterized by the non-central chi-squared distribution; in general, this happens when an estimator under consideration is asymptotically normally distributed with the asymptotic covariance matrix being of the sandwich form (see (Kent 1982)). Further, Theorem 5 1(ii) and 2(ii) indicate that the probability that GQAIC chooses the misspecified coefficients tends to 0 as $n \to \infty$.
- Theorem 6 1(i) shows that, when comparing correctly specified models, the probability that $\mathrm{GQBIC}_{1,n}^{\sharp}$ selects a larger model tends to 1. Moreover, Theorem 7 means that the GQBIC proposed by (46) and (48) has the model selection consistency.

Let $\mathrm{GQBIC}_{1,n}^{(m_1)}$ and $\mathrm{GQBIC}_{1,n}^{\sharp(m_1)}$ denote the $\mathrm{GQBIC}_{1,n}$ and $\mathrm{GQBIC}_{1,n}^{\sharp}$ of the m_1-th candidate scale coefficient, respectively. Also, let $\mathrm{GQBIC}_{2,n}^{(m_2|m_{1,n})}$ correspond to (48) associated with $c_{m_{1,n}}$ and $\hat{\gamma}_{m_{1,n},n}$.

Theorem 5 *Suppose that the assumptions of Theorem 2 hold for all candidate coefficients which are included in $\mathfrak{M}_1$ and $\mathfrak{M}_2$. We also assume that indexes m_1^* and m_2^* satisfy $m_1^* \in \mathfrak{M}_1^*$ and $m_2^* \in \mathfrak{M}_2^*$, respectively.*

1. (i) *Let $m_1 \in \mathfrak{M}_1 \backslash \{m_1^*\}$. If $\Theta_{\gamma_{m_1^*}}$ is nested in $\Theta_{\gamma_{m_1}}$ with map F_1, then*

$$\begin{aligned}
&\lim_{n\to\infty} P\left(\mathrm{GQAIC}_{1,n}^{(m_1^*)} - \mathrm{GQAIC}_{1,n}^{(m_1)} > 0\right) \\
&= P\left[\sum_{j=1}^{p_{\gamma_{m_1}}} \lambda_j \chi_j^2 > 2\,\mathrm{trace}\left\{\Gamma_{\gamma_{m_1}}(\gamma_{m_1,0})^{-1} W_{\gamma_{m_1}}(\gamma_{m_1,0})\right\}\right. \\
&\quad \left. - 2\,\mathrm{trace}\left\{\Gamma_{\gamma_{m_1^*}}(\gamma_{m_1^*,0})^{-1} W_{\gamma_{m_1^*}}(\gamma_{m_1^*,0})\right\}\right] \\
&\quad > 0,
\end{aligned}$$

   where

$$G_{\gamma_{m_1}}(\gamma_{m_1,0}) = \Gamma_{\gamma_{m_1}}(\gamma_{m_1,0})^{-1} - F_1\left(F_1^\top \Gamma_{\gamma_{m_1}}(\gamma_{m_1,0}) F_1\right)^{-1} F_1^\top,$$

(χ^2_j) is a sequence of independent χ^2 random variables with one degree of freedom, and $\lambda_1, \lambda_2, \ldots, \lambda_{p_{\gamma_{m_1}}}$ are the eigenvalues of $W_{\gamma_{m_1}}(\gamma_{m_1,0})^{1/2} G_{\gamma_{m_1}}(\gamma_{m_1,0}) W_{\gamma_{m_1}}(\gamma_{m_1,0})^{1/2}$.

(ii) *If* $m_1 \in \{1, \ldots, M_1\} \backslash \mathfrak{M}_1$, *then*

$$\lim_{n\to\infty} P\left(\mathrm{GQAIC}^{(m_1^*)}_{1,n} < \mathrm{GQAIC}^{(m_1)}_{1,n}\right) = 1.$$

2. (i) *Let* $m_2 \in \mathfrak{M}_2 \backslash \{m_2^*\}$. *If* $\Theta_{\alpha_{m_2^*}}$ *is nested in* $\Theta_{\alpha_{m_2}}$ *with map* F_2, *then*

$$\lim_{n\to\infty} P\left(\mathrm{GQAIC}^{(m_2^*|\hat{m}_{1,n})}_{2,n} - \mathrm{GQAIC}^{(m_2|\hat{m}_{1,n})}_{2,n} > 0\right)$$
$$= P\left[\sum_{j=1}^{p_{\alpha_{m_2}}} \lambda'_j \chi^2_j > 2(p_{\alpha_{m_2}} - p_{\alpha_{m_2^*}})\right] > 0,$$

where

$$G_{\alpha_{m_2}}(\alpha_{m_2,0}, \gamma_{\hat{m}_{1,n},0}) = \Gamma_{\alpha_{m_2}}(\alpha_{m_2,0}, \gamma_{\hat{m}_{1,n},0})^{-1} - F_2\left(F_2^\top \Gamma_{\alpha_{m_2}}(\alpha_{m_2,0}, \gamma_{\hat{m}_{1,n},0}) F_2\right)^{-1} F_2^\top$$

and $\lambda'_1, \lambda'_2, \ldots, \lambda'_{p_{\alpha_{m_2}}}$ are the eigenvalues of $\Gamma_{\alpha_{m_2}}(\alpha_{m_2,0}, \gamma_{\hat{m}_{1,n},0})^{1/2} G_{\alpha_{m_2}}(\alpha_{m_2,0}, \gamma_{\hat{m}_{1,n},0})$ $\Gamma_{\alpha_{m_2}}(\alpha_{m_2,0}, \gamma_{\hat{m}_{1,n},0})^{1/2}$.

(ii) *If* $m_2 \in \{1, \ldots, M_2\} \backslash \mathfrak{M}_2$, *then*

$$\lim_{n\to\infty} P\left(\mathrm{GQAIC}^{(m_2^*|\hat{m}_{1,n})}_{2,n} < \mathrm{GQAIC}^{(m_2|\hat{m}_{1,n})}_{2,n}\right) = 1.$$

Theorem 6 *Suppose that the assumptions of Theorem 2 hold for all candidate coefficients which are included in* $\mathfrak{M}_1$. *We also assume that index* m_1^* *satisfies* $m_1^* \in \mathfrak{M}_1^*$.

(i) *Let* $m_1 \in \mathfrak{M}_1 \backslash \{m_1^*\}$. *If* $\Theta_{\gamma_{m_1^*}}$ *is nested in* $\Theta_{\gamma_{m_1}}$, *then*

$$\lim_{n\to\infty} P\left(\mathrm{GQBIC}^{\sharp(m_1^*)}_{1,n} > \mathrm{GQBIC}^{\sharp(m_1)}_{1,n}\right) = 1.$$

(ii) *If* $m_1 \in \{1, \ldots, M_1\} \backslash \mathfrak{M}_1$, *then*

$$\lim_{n\to\infty} P\left(\mathrm{GQBIC}^{\sharp(m_1^*)}_{1,n} < \mathrm{GQBIC}^{\sharp(m_1)}_{1,n}\right) = 1.$$

Theorem 7 *Suppose that the assumptions of Theorem 2 hold for all candidate coefficients which are included in* $\mathfrak{M}_1$ *and* $\mathfrak{M}_2$. *We also assume that indexes* m_1^* *and* m_2^* *satisfy* $m_1^* \in \mathfrak{M}_1^*$ *and* $m_2^* \in \mathfrak{M}_2^*$, *respectively*.

1. (i) *Let* $m_1 \in \mathfrak{M}_1 \backslash \{m_1^*\}$. *If* $\Theta_{\gamma_{m_1^*}}$ *is nested in* $\Theta_{\gamma_{m_1}}$, *then*

$$\lim_{n\to\infty} P\left(\mathrm{GQBIC}^{(m_1^*)}_{1,n} < \mathrm{GQBIC}^{(m_1)}_{1,n}\right) = 1.$$

(ii) *If* $m_1 \in \{1, \dots, M_1\} \backslash \mathfrak{M}_1$, *then*

$$\lim_{n\to\infty} P\Big(\mathrm{GQBIC}_{1,n}^{(m_1^*)} < \mathrm{GQBIC}_{1,n}^{(m_1)}\Big) = 1.$$

2. (i) *Let* $m_2 \in \mathfrak{M}_2 \backslash \{m_2^*\}$. *If* $\Theta_{\alpha_{m_2^*}}$ *is nested in* $\Theta_{\alpha_{m_2}}$, *then*

$$\lim_{n\to\infty} P\Big(\mathrm{GQBIC}_{2,n}^{(m_2^*|\hat{m}_{1,n})} < \mathrm{GQBIC}_{2,n}^{(m_2|\hat{m}_{1,n})}\Big) = 1.$$

(ii) *If* $m_2 \in \{1, \dots, M_2\} \backslash \mathfrak{M}_2$, *then*

$$\lim_{n\to\infty} P\Big(\mathrm{GQBIC}_{2,n}^{(m_2^*|\hat{m}_{1,n})} < \mathrm{GQBIC}_{2,n}^{(m_2|\hat{m}_{1,n})}\Big) = 1.$$

6 Numerical experiments

In this section, we present simulation results to observe finite-sample performance of the proposed GQBIC and GQAIC. We use the `yuima` package on R (see (Brouste et al. 2014)) for generating data. All the Monte Carlo trials are based on 1000 independent sample paths, and the simulations are done for $(h, T_n) = (0.01, 10)$, $(0.005, 10)$, $(0.01, 50)$, and $(0.005, 50)$ (hence in each case, $n = 1000, 2000, 5000$, and 10000).

The sample data $\boldsymbol{X}_n = (X_{t_j})_{j=0}^n$ with $t_j = jh$ is obtained from

$$dX_t = -\frac{1}{2}X_t dt + \frac{3}{1+X_{t-}^2} dZ_t, \quad t \in [0, T_n], \quad X_0 = 0,$$

where $T_n = nh_n$. The numerical experiments are conducted in three situations:

(i) $\mathcal{L}(Z_t) = NIG(10, 0, 10t, 0)$,
(ii) $\mathcal{L}(Z_t) = bGamma(t, \sqrt{2}, t, \sqrt{2})$,
(iii) $\mathcal{L}(Z_t) = NIG\Big(\frac{25}{3}, \frac{20}{3}, \frac{9}{5}t, -\frac{12}{5}t\Big)$.

Here *NIG* and *bGamma* refer to the normal inverse-Gaussian and bilateral gamma distributions, respectively (see (Iacus and Yoshida 2018) for the definitions). In this example, we consider the following candidate scale (Scale) and drift (Drift) coefficients:

$$\textbf{Scale 1}: c_1(x, \gamma_1) = \gamma_1; \ \textbf{Scale 2}: c_2(x, \gamma_2) = \frac{\gamma_2}{1+x^2};$$

$$\textbf{Scale 3}: c_3(x, \gamma_3) = \frac{\gamma_{3,1} + \gamma_{3,2}x^2}{1+x^2}; \ \textbf{Scale 4}: c_4(x, \gamma_4) = \frac{\gamma_{4,1} + \gamma_{4,2}x + \gamma_{4,3}x^2}{1+x^2},$$

and

Drift 1 : $a_1(x,\alpha_1) = -\alpha_1$; **Drift 2** : $a_2(x,\alpha_2) = -\alpha_2 x$; **Drift 3** : $a_3(x,\alpha_3) = -\alpha_{3,1}x - \alpha_{3,2}$.

Each candidate model is given by a combination of the scale and drift coefficients, and the stochastic differential equation models based on local Gaussian approximate models of Lévy stochastic differential equations are assumed as misspecified candidate models. For example, in the case of Scale 1 and Drift 1, the statistical model is a stochastic differential equation model given by

$$\mathcal{L}(X_{t_j}|X_{t_{j-1}} = x) = N\big(x + \alpha_1 h, h\gamma_1^2\big).$$

Then, the Scale 2 and Drift 2 with $(\gamma_2,\alpha_2) = (3,\frac{1}{2})$ are the true coefficients, and the coefficients Scale 3, 4, and Drift 3 include the true coefficient.

We compare model selection frequency through GQAIC, GQBIC, and GQBIC$^\sharp$. Also, we formally use classical AIC, say formal AIC(fAIC), and compare the model selection results with those of the proposed criteria. The fAIC for scale and drift are given by

$$\begin{aligned}\mathrm{fAIC}_{1,n} &= -2\mathbb{H}_{1,n}(\hat{\gamma}_n) + 2p_\gamma,\\ \mathrm{fAIC}_{2,n} &= -2\mathbb{H}_{2,n}(\hat{\alpha}_n) + 2p_\alpha,\end{aligned}$$

respectively. Tables 1, 2, and 3 summarize the comparison results of model selection frequency. The GQAIC and GQBIC select the true coefficients Scale 2 and Drift 2 with high frequency in all cases, while the fAIC selects the true coefficients with less frequency in (ii) and (iii) cases. Moreover, we can observe the frequencies of selecting the true coefficient by GQBIC become larger as sample size n increases. In the (ii) and (iii) cases, also observed is that the frequencies that the misspecified coefficients Scale 1 and Drift 1 are chosen by GQAIC become lower as n increases.

7 Proofs

7.1 Proof of Theorem 1

The proofs are essentially the same as in those of (Masuda 2013, Theorem2.7), (Masuda and Uehara 2017, Theorem 3.4), making use of the general machinery (Yoshida 2011). Hence we only mention the formal difference, omitting the further details: The only difference to be mentioned is that the proofs are based on the two-stage procedure for M-estimators as in (Yoshida 2011, Section 6), where the first-stage random field is

$$u_\gamma \mapsto \log \mathbb{Z}_{1,n}(u_\gamma) := h\big(\mathbb{H}_{1,n}(\gamma_0 + T_n^{-1/2}u_\gamma) - \mathbb{H}_{1,n}(\gamma_0)\big),$$

and the second-stage one (depending on $\hat{\gamma}_n$) is the rescaled

$$u_\alpha \mapsto \log \mathbb{Z}_{2,n}(u_\alpha) := \mathbb{H}_{2,n}(\alpha_0 + T_n^{-1/2}u_\alpha) - \mathbb{H}_{2,n}(\alpha_0).$$

Table 1 Computation results of (i) case. Model selection frequencies for various situations are shown. The true model consists of Scale 2 and Drift 2

fAIC	T_n	h		Scale 1	Scale 2*	Scale 3	Scale 4
	10	0.01	Drift 1	0	1	0	0
	($n = 1000$)		Drift 2*	0	**536**	124	199
			Drift 3	0	74	27	39
	10	0.005	Drift 1	0	0	1	0
	($n = 2000$)		Drift 2*	0	**458**	110	285
			Drift 3	0	67	25	54
	50	0.01	Drift 1	0	0	0	0
	($n = 5000$)		Drift 2*	0	**466**	202	231
			Drift 3	0	51	28	32
	50	0.005	Drift 1	0	0	0	0
	($n = 10000$)		Drift 2*	0	**402**	150	352
			Drift 3	0	51	13	32
GQAIC	T_n	h_n		Scale 1	Scale 2*	Scale 3	Scale 4
	10	0.01	Drift 1	0	1	0	0
	($n = 1000$)		Drift 2*	0	**714**	77	64
			Drift 3	0	110	20	14
	10	0.005	Drift 1	0	1	0	0
	($n = 2000$)		Drift 2*	0	**733**	62	56
			Drift 3	0	117	17	14
	50	0.01	Drift 1	0	0	0	0
	($n = 5000$)		Drift 2*	0	**713**	122	64
			Drift 3	0	83	11	7
	50	0.005	Drift 1	0	0	0	0
	($n = 10000$)		Drift 2*	0	**765**	79	59
			Drift 3	0	88	5	4
GQBIC	T_n	h_n		Scale 1	Scale 2*	Scale 3	Scale 4
	10	0.01	Drift 1	0	1	0	0
	($n = 1000$)		Drift 2*	0	**861**	0	0
			Drift 3	0	138	0	0
	10	0.005	Drift 1	0	1	0	0
	($n = 2000$)		Drift 2*	0	**866**	0	0
			Drift 3	0	133	0	0
	50	0.01	Drift 1	0	0	0	0
	($n = 5000$)		Drift 2*	0	**965**	0	0
			Drift 3	0	35	0	0
	50	0.005	Drift 1	0	0	0	0
	($n = 10000$)		Drift 2*	0	**964**	0	0
			Drift 3	0	36	0	0
GQBIC$^\sharp$	T_n	h_n		Scale 1	Scale 2*	Scale 3	Scale 4
	10	0.01	Drift 1	0	1	0	0

Table 1 (continued)

GQBIC$^\sharp$	T_n	h_n		Scale 1	Scale 2*	Scale 3	Scale 4
	($n = 1000$)		Drift 2*	0	**788**	49	30
			Drift 3	0	119	7	6
	10	0.005	Drift 1	0	1	0	0
	($n = 2000$)		Drift 2*	0	**759**	69	45
			Drift 3	0	107	9	10
	50	0.01	Drift 1	0	0	0	0
	($n = 5000$)		Drift 2*	0	**882**	64	19
			Drift 3	0	32	2	1
	50	0.005	Drift 1	0	0	0	0
	($n = 10000$)		Drift 2*	0	**862**	68	34
			Drift 3	0	32	2	2

Note the resolution in handling the scale coefficient in the first stage is corrected by the multiplicative factor "h". The Studentization (12) can be verified exactly as in (Masuda 2013, Corollary 2.8). □

7.2 Proof of proposition 2

Let us recall the expression (11) for $\hat{W}_{\gamma,n} = (\hat{W}^{(qr)}_{\gamma,n})$. Under the integrability conditions, the sequence $(\hat{W}_{\gamma,n})_n$ is $L^q(P)$-bounded for any $q > 0$. To see this, let $\chi_j := \Delta_j X - h a_{j-1}(\alpha_0)$, and write $O^*_p(1)$ for a random sequence $(\zeta_n)_n$ such that $\sup_n E(|\zeta_n|^q) < \infty$ for any $q > 0$. Write E^{j-1} for the expectation conditional on $\mathcal{F}_{t_{j-1}}$, where $(\mathcal{F}_t)$ denotes the underlying filtration to which all the stochastic processes are adapted. Then, by compensation and Burkholder's inequality, we have

$$
\begin{aligned}
|\hat{W}_{\gamma,n}| &\lesssim \frac{1}{T_n}\sum_{j=1}^n (1 + |X_{t_{j-1}}|)^C |\hat{\chi}_j|^4 \\
&\lesssim \frac{1}{T_n}\sum_{j=1}^n (1 + |X_{t_{j-1}}|)^C |\chi_j|^4 + O^*_p(1) \\
&\lesssim \frac{1}{T_n}\sum_{j=1}^n (1 + |X_{t_{j-1}}|)^C E^{j-1}\left[|\chi_j|^4\right] + O^*_p(T_n^{-1/2}) + O^*_p(1) \\
&\lesssim \frac{1}{n}\sum_{j=1}^n (1 + |X_{t_{j-1}}|)^C + O^*_p(T_n^{-1/2}) + O^*_p(1) = O^*_p(1).
\end{aligned}
$$

Since $|\hat{\Gamma}^{-1}_{\gamma,n}|$ is bounded by a universal-constant multiple of $\lambda^{-1}_{\min}(\hat{\Gamma}_{\gamma,n})$, we can apply Hölder' inequality to ensure that (36) is sufficient for (33). □

Table 2 Computation results of (ii) case. Model selection frequencies for various situations are shown. The true model consists of Scale 2 and Drift 2

fAIC	T_n	h		Scale 1	Scale 2*	Scale 3	Scale 4
	10	0.01	Drift 1	0	0	0	9
	($n = 1000$)		Drift 2*	5	**106**	26	767
			Drift 3	0	10	1	76
	10	0.005	Drift 1	0	0	0	9
	($n = 2000$)		Drift 2*	2	**83**	24	793
			Drift 3	0	6	2	81
	50	0.01	Drift 1	0	0	0	0
	($n = 5000$)		Drift 2*	2	**85**	47	782
			Drift 3	0	6	4	74
	50	0.005	Drift 1	0	0	0	0
	($n = 10000$)		Drift 2*	0	**68**	25	826
			Drift 3	0	4	3	74
GQAIC	T_n	h_n		Scale 1	Scale 2*	Scale 3	Scale 4
	10	0.01	Drift 1	1	0	0	4
	($n = 1000$)		Drift 2*	97	**591**	10	186
			Drift 3	15	86	0	10
	10	0.005	Drift 1	1	0	0	5
	($n = 2000$)		Drift 2*	99	**584**	7	191
			Drift 3	15	88	0	10
	50	0.01	Drift 1	0	0	0	0
	($n = 5000$)		Drift 2*	29	**741**	36	100
			Drift 3	4	73	5	12
	50	0.005	Drift 1	0	0	0	0
	($n = 10000$)		Drift 2*	27	**747**	27	104
			Drift 3	4	74	5	12
GQBIC	T_n	h_n		Scale 1	Scale 2*	Scale 3	Scale 4
	10	0.01	Drift 1	3	0	0	1
	($n = 1000$)		Drift 2*	142	**700**	2	36
			Drift 3	14	101	0	1
	10	0.005	Drift 1	3	0	0	2
	($n = 2000$)		Drift 2*	142	**700**	1	37
			Drift 3	14	100	0	1
	50	0.01	Drift 1	0	0	0	0
	($n = 5000$)		Drift 2*	42	**890**	9	22
			Drift 3	3	33	0	1
	50	0.005	Drift 1	0	0	0	0
	($n = 10000$)		Drift 2*	38	**894**	6	23
			Drift 3	3	35	0	1
GQBIC$^\sharp$	T_n	h_n		Scale 1	Scale 2*	Scale 3	Scale 4
	10	0.01	Drift 1	0	0	0	10

Table 2 (continued)

GQBIC$^\sharp$	T_n	h_n		Scale 1	Scale 2*	Scale 3	Scale 4
	($n = 1000$)		Drift 2*	15	**189**	32	685
			Drift 3	2	14	2	51
	10	0.005	Drift 1	0	0	0	10
	($n = 2000$)		Drift 2*	10	**130**	35	743
			Drift 3	1	12	2	57
	50	0.01	Drift 1	0	0	0	0
	($n = 5000$)		Drift 2*	3	**200**	84	686
			Drift 3	0	4	2	21
	50	0.005	Drift 1	0	0	0	0
	($n = 10000$)		Drift 2*	1	**159**	49	763
			Drift 3	0	3	2	23

7.3 Proof of Lemma 1

To begin with, let $k \ll n$ be a positive integer not depending on n, and let $m := \lfloor n/k \rfloor$; without loss of generality, we set $k \le n/2$. Then, for $u \in \mathbb{R}^{p_\gamma}$,

$$\begin{aligned}
\hat{\Gamma}_{\gamma,n}[u^{\otimes 2}] &= \frac{1}{2n}\sum_{j=1}^{n} \operatorname{trace}\left\{\left(\hat{S}_{j-1}^{-1}(\partial_\gamma \hat{S}_{j-1})[u]\right)^2\right\} \\
&\ge \frac{1}{2n}\sum_{j=1}^{n} \frac{1}{d}\left\{\operatorname{trace}\left((\hat{S}_{j-1}^{-1}\partial_\gamma \hat{S}_{j-1})[u]\right)\right\}^2 \\
&= \frac{1}{n}\sum_{j=1}^{n} \frac{1}{2d}\left\{\zeta_{j-1}(\hat{\gamma}_n)[u]\right\}^2 \\
&\gtrsim \frac{1}{mk}\sum_{j=1}^{mk} \left\{\zeta_{j-1}(\hat{\gamma}_n)[u]\right\}^2 =: \frac{1}{k}\cdot\frac{1}{m}\sum_{i=1}^{m} V_i(\hat{\gamma}_n)[u^{\otimes 2}],
\end{aligned} \tag{49}$$

where $V_i(\gamma) := \sum_{j=(i-1)k+1}^{ik} \zeta_{j-1}(\gamma)^{\otimes 2}$; the first inequality is due to the Cauchy-Schwarz inequality: $\operatorname{trace}(A^2) \ge d^{-1}\operatorname{trace}(A)^2$ for any real square matrix A with real eigenvalues. Observe that by (49) and the Jensen inequality,

$$\begin{aligned}
\lambda_{\min}^{-q}(\hat{\Gamma}_{\gamma,n}) &\le \left(\inf_{u\in\mathbb{S}} \hat{\Gamma}_{\gamma,n}[u^{\otimes 2}]\right)^{-q} \\
&\lesssim \left(\frac{1}{m}\sum_{i=1}^{m} \inf_{u\in\mathbb{S}} \inf_{\gamma} V_i(\gamma)[u^{\otimes 2}]\right)^{-q} \\
&\lesssim \frac{1}{m}\sum_{i=1}^{m} \left(\inf_{u\in\mathbb{S}} \inf_{\gamma} V_i(\gamma)[u^{\otimes 2}]\right)^{-q}.
\end{aligned}$$

It suffices for (36) to have

Table 3 Computation results of (iii) case. Model selection frequencies for various situations are shown. The true model consists of Scale 2 and Drift 2

fAIC	T_n	h		Scale 1	Scale 2*	Scale 3	Scale 4
	10	0.01	Drift 1	0	0	0	0
	($n = 1000$)		Drift 2*	0	**128**	52	600
			Drift 3	0	30	30	160
	10	0.005	Drift 1	0	0	0	1
	($n = 2000$)		Drift 2*	0	**88**	37	653
			Drift 3	0	25	16	180
	50	0.01	Drift 1	0	0	0	0
	($n = 5000$)		Drift 2*	0	**88**	75	731
			Drift 3	0	12	4	9
	50	0.005	Drift 1	0	0	0	0
	($n = 10000$)		Drift 2*	0	**70**	39	785
			Drift 3	0	6	1	99
GQAIC	T_n	h_n		Scale 1	Scale 2*	Scale 3	Scale 4
	10	0.01	Drift 1	0	0	0	0
	($n = 1000$)		Drift 2*	18	**589**	33	155
			Drift 3	32	125	33	15
	10	0.005	Drift 1	0	0	0	1
	($n = 2000$)		Drift 2*	15	**602**	30	149
			Drift 3	32	122	32	17
	50	0.01	Drift 1	0	0	0	0
	($n = 5000$)		Drift 2*	0	**672**	102	122
			Drift 3	0	80	8	16
	50	0.005	Drift 1	0	0	0	0
	($n = 10000$)		Drift 2*	0	**710**	77	110
			Drift 3	0	78	4	21
GQBIC	T_n	h_n		Scale 1	Scale 2*	Scale 3	Scale 4
	10	0.01	Drift 1	0	0	0	0
	($n = 1000$)		Drift 2*	24	**798**	5	5
			Drift 3	35	133	0	0
	10	0.005	Drift 1	0	0	0	0
	($n = 2000$)		Drift 2*	25	**795**	4	4
			Drift 3	37	135	0	0
	50	0.01	Drift 1	0	0	0	0
	($n = 5000$)		Drift 2*	0	**943**	28	0
			Drift 3	0	28	1	0
	50	0.005	Drift 1	0	0	0	0
	($n = 10000$)		Drift 2*	0	**957**	16	0
			Drift 3	0	26	1	0
GQBIC$^\sharp$	T_n	h_n		Scale 1	Scale 2*	Scale 3	Scale 4
	10	0.01	Drift 1	0	0	0	0

Table 3 (continued)

GQBIC$^\sharp$	T_n	h_n		Scale 1	Scale 2*	Scale 3	Scale 4
	($n = 1000$)		Drift 2*	0	**264**	64	477
			Drift 3	5	37	45	108
	10	0.005	Drift 1	0	0	0	1
	($n = 2000$)		Drift 2*	0	**181**	48	565
			Drift 3	2	35	29	139
	50	0.01	Drift 1	0	0	0	0
	($n = 5000$)		Drift 2*	0	**210**	136	619
			Drift 3	0	7	1	27
	50	0.005	Drift 1	0	0	0	0
	($n = 10000$)		Drift 2*	0	**166**	96	708
			Drift 3	0	4	1	25

$$\sup_n \sup_{i\in\mathbb{N}} E\left[\left(\inf_{u\in\mathbb{S}} \inf_{\gamma} V_i(\gamma)[u^{\otimes 2}]\right)^{-q}\right] < \infty. \tag{50}$$

Fix a constant $r > 0$ in the sequel. The expectation in (50) equals

$$\begin{aligned}
&\int_0^\infty P\left[\left(\inf_{u\in\mathbb{S}} \inf_{\gamma} V_i(\gamma)[u^{\otimes 2}]\right)^{-q} > s\right] ds \\
&\le 1 + \int_1^\infty P\left[\inf_{u\in\mathbb{S}} \inf_{\gamma} \sum_{j=(i-1)k+1}^{ik} \left(\zeta_{j-1}(\gamma)[u]\right)^2 < s^{-1/q}\right] ds \\
&\le 1 + \int_1^\infty \left(P\left[\sum_{j=(i-1)k+1}^{ik} \sup_{\gamma} |\zeta_{j-1}(\gamma)|^2 \ge s^{r/q}\right] \right. \\
&\qquad \left. + P\left[\inf_{u\in\mathbb{S}} \inf_{\gamma} \sum_{j=(i-1)k+1}^{ik} \left(\zeta_{j-1}(\gamma)[u]\right)^2 < s^{-1/q}, \sum_{j=(i-1)k+1}^{ik} \sup_{\gamma} |\zeta_{j-1}(\gamma)|^2 \le s^{r/q}\right] \right) ds \\
&=: 1 + \int_1^\infty \left(I'_{i,k,q}(s) + I''_{i,k,q}(s) \right) ds.
\end{aligned}$$

Since k is fixed and

$$I'_{i,k,q}(s) \lesssim s^{-2} \sup_t E\left[\sup_{\gamma} \left|\zeta(X_t, \gamma)\right|^{4q/r}\right] \lesssim s^{-2}\left(1 + \sup_t E[|X_t|^C]\right) \lesssim s^{-2}$$

whatever $r > 0$ is under the present assumptions, it remains to be shown that

$$\int_1^\infty I''_{i,k,q}(s) ds \lesssim 1. \tag{51}$$

First, to handle the infimum for u, we will apply Lemma 2. Let $r' := (r+1)/2$. Taking $\delta = s^{-r'/q}$ in Lemma 2, we can pick some elements $u'_1, \ldots, u'_{D(s)}$ with the integer

$$D(s) = O\big(s^{r'(p_\gamma - 1)/q}\big) \qquad s \uparrow \infty,$$

which is constant for $p_\gamma = 1$. Write $\mathbb{S}_l := \{u \in \mathbb{S} : |u - u'_l| < s^{-r'/q}\}$. Since $\inf_{u\in\mathbb{S}_l} \inf_\gamma \big|\zeta_{j-1}(\gamma)[u]\big| \geq \inf_\gamma \big|\zeta_{j-1}(\gamma)[u'_l]\big| - s^{-r'/q} \sup_\gamma |\zeta_{j-1}(\gamma)|$ for each $u \in \mathbb{S}_l$, we have

$$\begin{aligned} I''_{i,k,q}(s) &\leq \sum_{l=1}^{D(s)} P\left[\bigcap_{j=(i-1)k+1}^{ik} \left\{ \inf_{u\in\mathbb{S}_l} \inf_\gamma \big|\zeta_{j-1}(\gamma)[u]\big| < s^{-1/(2q)},\ \sup_\gamma |\zeta_{j-1}(\gamma)| \leq s^{r/(2q)} \right\}\right] \\ &\leq \sum_{l=1}^{D(s)} P\left[\bigcap_{j=(i-1)k+1}^{ik} \left\{ \inf_\gamma \big|\zeta_{j-1}(\gamma)[u'_l]\big| < 2s^{-1/(2q)},\ \sup_\gamma |\zeta_{j-1}(\gamma)| \leq s^{r/(2q)} \right\}\right] \\ &\leq \sum_{l=1}^{D(s)} P\left[\bigcap_{j=(i-1)k+1}^{ik} \left\{ \inf_\gamma \big|\zeta_{j-1}(\gamma)[u'_l]\big| < 2s^{-1/(2q)} \right\}\right]. \end{aligned} \tag{52}$$

Next, we get rid of the infimum with respect to γ. Since $\overline{\Theta}_\gamma \subset \mathbb{R}^{p_\gamma}$ is compact, we can cover it by finitely many hypercubes $\mathcal{U}_1, \dots, \mathcal{U}_{H(s)}$, each with side length $s^{-1/(2q)}$ and the number

$$H(s) = O\big(s^{p_\gamma/(2q)}\big) \qquad s \uparrow \infty.$$

Pick elements $\gamma'_b \in \mathcal{U}_b$ arbitrarily $(b = 1, \dots, H(s))$. We also have $\sup_\gamma |\partial_\gamma \zeta_{j-1}(\gamma)| \lesssim 1 + |X_{t_{j-1}}|^C$. With these observations, for some nonnegative function F such that $F(x) \lesssim 1 + |x|^C$ we can continue (52) as follows:

$$\begin{aligned} I''_{i,k,q}(s) &\leq \sum_{l=1}^{D(s)} \sum_{b=1}^{H(s)} P\left[\bigcap_{j=(i-1)k+1}^{ik} \left\{ \inf_{\gamma\in\mathcal{U}_b} \big|\zeta_{j-1}(\gamma)[u'_l]\big| < 2s^{-1/(2q)} \right\}\right] \\ &\leq \sum_{l=1}^{D(s)} \sum_{b=1}^{H(s)} P\left[\bigcap_{j=(i-1)k+1}^{ik} \left\{ \big|\zeta_{j-1}(\gamma'_b)[u'_l]\big| \leq s^{-1/(2q)} F_{j-1} \right\}\right] \\ &\leq \sum_{l=1}^{D(s)} \sum_{b=1}^{H(s)} P\left[\bigcap_{j=(i-1)k+1}^{ik} \left\{ \big|\zeta_{j-1}(\gamma'_b)[u'_l]\big| \leq s^{-1/(4q)} \right\}\right] + \sum_{l=1}^{D(s)} \sum_{b=1}^{H(s)} P\big[F_{j-1} \geq s^{1/(4q)}\big]. \end{aligned} \tag{53}$$

Whatever $r > 0$ is, the second term in (53) can be bounded by Cs^{-2} through the Markov inequality with sufficiently high-order moments. Finally, by iterative conditioning, the first term in (53) is a.s. bounded by $C \sum_{l=1}^{D(s)} \sum_{b=1}^{H(s)} (s^{-1/(4q)})^{k\rho}$ with $\rho > 0$ of (41); here is the only place where we used the condition (41). Given p_γ, $q = 1 + \delta$, $r > 0$, and $\rho > 0$, we can pick a sufficiently large $k \in \mathbb{N}$ to ensure that $C \sum_{l=1}^{D(s)} \sum_{b=1}^{H(s)} (s^{-1/(4q)})^{k\rho} \lesssim s^{-2}$, thus concluding (51). The proof is complete. □

7.4 Proof of Corollary 1

By the iterative conditioning with taking k large enough, we can bound the first term in (53) from above by $\sum_{l=1}^{D(s)} \sum_{b=1}^{H(s)} s^{-\rho k/(2q)} \lesssim s^{-2}$. Hence (51). □

7.5 Proof of Corollary 2

(50) readily follows from (42); in this case we may set $k = 1$ in the proof of Lemma 1. □

7.6 Proof of Theorem 3

Roughly, (44) will be proved by expanding $\mathbb{H}_{1,n}(\gamma)$ around $\hat{\gamma}_n$ with vicinity size of order $n^{-1/2}$, while (45) around γ_0 with vicinity size of order $T_n^{-1/2}$. □

7.6.1 Proof of (44)

Our proof is achieved in an analogous way to the derivation of the Bernstein-von Mises theorem given in (Jasra et al. 2019, Theorem 1), which dealt with a more complicated two-step non-ergodic setting.[1]

By the change of variable $\gamma = \hat{\gamma}_n + n^{-1/2}u$, we have

$$\mathfrak{F}_{1,n}(1) = -\frac{1}{n}\mathbb{H}_{1,n}(\hat{\gamma}_n) + \frac{p_\gamma}{2n}\log n - \frac{1}{n}\log \hat{\mathbb{Z}}^*_{1,n},$$

where $\hat{\mathbb{Z}}^*_{1,n} = \int_{\hat{U}_{1,n}} \hat{\mathbb{Z}}_{1,n}(u)\pi_1(\hat{\gamma}_n + n^{-1/2}u)du$ with $\hat{U}_{1,n} := \{v \in \mathbb{R}^{p_\gamma} : \hat{\gamma}_n + n^{-1/2}v \in \Theta_\gamma\}$ and $\hat{\mathbb{Z}}_{1,n}(u) := \exp\{\mathbb{H}_{1,n}(\hat{\gamma}_n + n^{-1/2}u) - \mathbb{H}_{1,n}(\hat{\gamma}_n)\}$. It suffices to show that $\log \hat{\mathbb{Z}}^*_{1,n} = O_p(1)$.

We need some notation and preliminary remarks. Let $\mathbb{Y}_{1,n}(\gamma) := n^{-1}\{\mathbb{H}_{1,n}(\gamma) - \mathbb{H}_{1,n}(\gamma_0)\}$ and $\mathbb{Y}_{1,0}(\gamma) := -(1/2)\int\{\mathrm{trace}\left(S(x,\gamma)^{-1}S(x,\gamma_0) - I_d\right) + \log(|S(x,\gamma)|/|S(x,\gamma_0)|)\}\pi(dx)$. Then, Assumption 4 ensures that we can find a constant $\chi_\gamma > 0$ such that

$$\sup_{\gamma:\,|\gamma-\gamma_0|\geq\delta} \mathbb{Y}_{1,0}(\gamma) \leq -\chi_\gamma\delta^2 \tag{54}$$

for every $\delta > 0$. By (Masuda 2013, Lemma 4.3) we know that $\sup_\theta \sqrt{T_n}|\mathbb{Y}_{1,n}(\gamma) - \mathbb{Y}_{1,0}(\gamma)| = O_p(1)$. Moreover, note that $\Delta_{1,n}(\hat{\gamma}_n) = h\,T_n^{-1/2}\partial_\gamma\mathbb{H}_{1,n}(\hat{\gamma}_n) = 0$ if $\hat{\gamma}_n \in \Theta_\gamma$, that $\Gamma_{\gamma,n}(\theta_0) \xrightarrow{p} \Gamma_\gamma(\gamma_0)$ (see (16)), and that $n^{-1}\sup_\gamma|\partial_\theta^3\mathbb{H}_{1,n}(\gamma)| = O_p(1)$. We pick a constant (recall (43))

$$0 < c_0 < \frac{c_1}{4}. \tag{55}$$

[1] Therefore, it should be remarked that without any essential change the proof below could be easily extended to the non-ergodic framework, where the matrices $\Sigma(\theta_0)$ and $\Gamma(\theta_0)$ are random.

Put $\epsilon_n = n^{-c_0}$ in the sequel.

We introduce the following auxiliary event for constants $M, \lambda > 0$ (recall $\hat{u}_{\gamma,n} := \sqrt{T_n}(\hat{\gamma}_n - \gamma_0)$):

$$G_{1,n}(M,\lambda) := \Big\{ \hat{\gamma}_n \in \Theta_\gamma, \quad \Big|\Gamma_{\gamma,n}(\hat{\gamma}_n) - \Gamma_\gamma(\hat{\gamma}_n)\Big| \le \lambda, \quad \lambda_{\min}(\Gamma_\gamma(\hat{\gamma}_n)) \ge 4\lambda,$$
$$\Big|\sqrt{T_n}\sup_\theta(\mathbb{Y}_{1,n}(\gamma) - \mathbb{Y}_1(\gamma))\Big| \vee |\hat{u}_{\gamma,n}| < M, \quad \frac{1}{n}\sup_\gamma \Big|\partial_\theta^3 \mathbb{H}_{1,n}(\gamma)\Big| \le \frac{3\lambda}{\epsilon_n}\Big\}.$$

Fix any $\epsilon > 0$. Then, we can find a pair (M_1, λ_1) and an $N_1 \in \mathbb{N}$ such that

$$\sup_{n \ge N_1} P\big\{G_{1,n}(M,\lambda)^c\big\} < \epsilon$$

holds for every $M \ge M_1$ and $\lambda \in (0, \lambda_1]$. Therefore, to deduce $\log \hat{\mathbb{Z}}^*_{1,n} = O_p(1)$, we may and do focus on the event $G_{1,n}(M,\lambda)$ with $M = M(\epsilon)$ and $\lambda = \lambda(\epsilon)$ being sufficiently large and small, respectively.

Let $A_{1,n} := \{u \in \mathbb{R}^{p_\gamma} : |u| \le \epsilon_n\sqrt{n}\}$. For $u \in A_{1,n}$ and on $G_{1,n}(M,\lambda)$, we apply the third-order Taylor expansion to conclude that

$$\log \hat{\mathbb{Z}}_{1,n}(u) \le -\lambda|u|^2,$$

by noting that, for some random point $\tilde{\gamma}_n$ on the segment joining $\hat{\gamma}_n$ and γ_0, we have $\log \hat{\mathbb{Z}}_{1,n}(u) = n^{-1/2}\partial_\gamma \mathbb{H}_{1,n}(\hat{\gamma}_n)[u] - (1/2)\{\Gamma_\gamma(\hat{\gamma}_n) + (\Gamma_{\gamma,n}(\hat{\gamma}_n) - \Gamma_\gamma(\hat{\gamma}_n))$. This entails that $\sup_n \sup_{u \in A_{1,n} \cap \hat{U}_{1,n}} \hat{\mathbb{Z}}_{1,n}(u)\pi_1(\hat{\gamma}_n + n^{-1/2}u) \lesssim \exp(-|u|^2)$, followed by

$$\begin{aligned}
&\left|\int_{A_{1,n}\cap\hat{U}_{1,n}} \hat{\mathbb{Z}}_{1,n}(u)\{\pi_1(\hat{\gamma}_n + n^{-1/2}u) - \pi_1(\hat{\gamma}_n)\}du\right| \\
&\le \sup_{u \in A_{1,n}\cap\hat{U}_{1,n}} \Big|\pi_1(\hat{\gamma}_n + n^{-1/2}u) - \pi_1(\hat{\gamma}_n)\Big| \int_{A_{1,n}\cap\hat{U}_{1,n}} \hat{\mathbb{Z}}_{1,n}(u)du \\
&\lesssim \sup_{|v_n| \le \epsilon_n} |\pi_1(\hat{\gamma}_n + v_n) - \pi_1(\hat{\gamma}_n)| \xrightarrow{p} 0.
\end{aligned}$$

For $\mathbb{Z}^0_1(u) := \exp\{-(1/2)\Gamma_\gamma(\gamma_0)[u^{\otimes 2}]\}$, we can deduce that $\int_{A_{1,n}\cap\hat{U}_{1,n}} \hat{\mathbb{Z}}_{1,n}(u)du = \int_{A_{1,n}\cap\hat{U}_{1,n}} \mathbb{Z}^0_1(u)du + o_p(1)$ (on $G_{1,n}(M,\lambda)$) by using the subsequence argument in much the same way as in Jasra et al. (2019). Since $\int_{\mathbb{R}^{p_\gamma}} \mathbb{Z}^0_1(u)du = (2\pi)^{p_\gamma/2}|\Gamma_\gamma(\gamma_0)|^{-1/2} + o(1)$, we conclude that $\log\{\int_{A_{1,n}\cap\hat{U}_{1,n}} \hat{\mathbb{Z}}_{1,n}(u)\pi_1(\hat{\gamma}_n + n^{-1/2}u)du\} = \log \pi(\gamma_0) + (p_\gamma/2)\log(2\pi) - (1/2)\log|\Gamma_\gamma(\gamma_0)| + o_p(1) = O_p(1)$.

We are left to proving $\int_{A^c_{1,n}\cap\hat{U}_{1,n}} \hat{\mathbb{Z}}_{1,n}(u)\pi_1(\hat{\gamma}_n + n^{-1/2}u)du = o_p(1)$; since π_1 is bounded, it suffices to show that $\int_{A^c_{1,n}\cap\hat{U}_{1,n}} \hat{\mathbb{Z}}_{1,n}(u)du = o_p(1)$. We have on $G_{1,n}(M,\lambda)$,

$$\begin{aligned}\sup_{u\in A_{1,n}^c\cap\hat{U}_{1,n}}\log\hat{\mathbb{Z}}_{1,n}(u) &\le n\sup_{u\in A_{1,n}^c\cap\hat{U}_{1,n}}\left\{\mathbb{Y}_{1,n}\left(\hat{\gamma}_n+\frac{u}{\sqrt{n}}\right)-\mathbb{Y}_{1,0}\left(\hat{\gamma}_n+\frac{u}{\sqrt{n}}\right)\right\}\\ &\quad+\mathbb{Y}_{1,0}\left(\hat{\gamma}_n+\frac{u}{\sqrt{n}}\right)\\ &\le n\left(\sup_{\gamma}|\mathbb{Y}_{1,n}(\gamma)-\mathbb{Y}_{1,0}(\gamma)|+\sup_{u\in A_{1,n}^c\cap\hat{U}_{1,n}}\mathbb{Y}_{1,0}\left(\hat{\gamma}_n+\frac{u}{\sqrt{n}}\right)\right).\\ &\le n\left(Mn^{-c_1/2}+\sup_{u\in A_{1,n}^c\cap\hat{U}_{1,n}}\mathbb{Y}_{1,0}\left(\hat{\gamma}_n+\frac{u}{\sqrt{n}}\right)\right)\end{aligned}\tag{56}$$

Observe that

$$\inf_{u\in A_{1,n}^c}\left|\left(\hat{\gamma}_n+\frac{u}{\sqrt{n}}\right)-\gamma_0\right|\ge\inf_{u\in A_{1,n}^c}\frac{|u|}{\sqrt{n}}-\frac{|\hat{u}_{\gamma,n}|}{\sqrt{T_n}}\ge n^{-c_0}\left(1-Mn^{c_0-c_1/2}\right)\ge\frac{1}{2}n^{-c_0}$$

for every n large enough. Recalling (54) and (55), we can continue the estimate (56) as follows:

$$\sup_{u\in A_{1,n}^c\cap\hat{U}_{1,n}}\log\hat{\mathbb{Z}}_{1,n}(u)\lesssim-n^{1-2c_0}\left(1-n^{2c_0-c_1/2}\right)\lesssim-n^{1-2c_0}\downarrow-\infty.$$

Thus $\int_{A_{1,n}^c\cap\hat{U}_{1,n}}\hat{\mathbb{Z}}_{1,n}(u)du\lesssim\exp(-Cn^{1-2c_0})\int_{|u|\le C\sqrt{n}}du\lesssim n^{p_\gamma/2}\exp(-Cn^{1-2c_0})\to 0$, concluding that $\int_{A_{1,n}^c\cap\hat{U}_{1,n}}\hat{\mathbb{Z}}_{1,n}(u)du=o_p(1)$. The proof is complete.

7.6.2 Proof of (45)

The proof is similar to the proof of (44) and much closer to that of (Jasra et al. 2019, Theorem 1).

Let $\mathbb{Z}_{1,n}(u):=\exp\{h(\mathbb{H}_{1,n}(\gamma_0+T_n^{-1/2}u)-\mathbb{H}_{1,n}(\gamma_0))\}$. The change of variable $\gamma=\gamma_0+T_n^{-1/2}u$ yields

$$\begin{aligned}\mathfrak{F}_{1,n}(h)&=-\frac{1}{T_n}h\mathbb{H}_{1,n}(\gamma_0)+\frac{p_\gamma}{2T_n}\log T_n-\frac{1}{T_n}\log\left(\int_{U_{1,n}}\mathbb{Z}_{1,n}(u)\pi_1(\gamma_0+T_N^{-1/2}u)du\right)\\ &=:-\frac{1}{T_n}h\mathbb{H}_{1,n}(\gamma_0)+\frac{p_\gamma}{2T_n}\log T_n-\frac{1}{T_n}\log\overline{\mathbb{Z}}_{1,n},\end{aligned}$$

where $U_{1,n}:=\{v\in\mathbb{R}^{p_\gamma}:\gamma_0+T_n^{-1/2}v\in\Theta_\gamma\}$; a relevant form already appeared in Remark 1. By Theorem 1 (and its proof), it is easily seen that $h\mathbb{H}_{1,n}(\gamma_0)=h\mathbb{H}_{1,n}(\hat{\gamma}_n)+O_p(1)$, so that it suffices to show $\log\overline{\mathbb{Z}}_{1,n}=O_p(1)$.

Recall that $\Delta_{1,n}(\gamma_0) := h\, T_n^{-1/2} \partial_\gamma \mathbb{H}_{1,n}(\gamma_0) = O_p(1)$. In the present case, the auxiliary event is given as follows (we keep using $\epsilon_n = n^{-c_0}$ such that (55) holds): for constants $\lambda \in (0, \lambda_{\min}(\Gamma_\gamma(\gamma_0))/4)$ and $M > 0$,

$$G_{1,n}(M,\lambda) := \left\{ |\Delta_{1,n}(\gamma_0)| \le M, \quad \left|\Gamma_{\gamma,n}(\theta_0) - \Gamma_\gamma(\gamma_0)\right| < \lambda, \right.$$
$$\left. \sqrt{T_n} \sup_\theta |\mathbb{Y}_{1,n}(\gamma) - \mathbb{Y}_1(\gamma)| < M, \quad \frac{1}{n} \sup_\gamma \left|\partial_\theta^3 \mathbb{H}_{1,n}(\gamma)\right| \le \frac{3\lambda}{\epsilon_n} \right\}.$$

With this $G_{1,n}(M,\lambda)$, the remaining arguments are almost identical to those of Jasra et al. (2019), hence omitted. □

7.7 Proof of Theorem 5

Under the assumptions, the coefficients $c_{m_1^*}$ and $a_{m_2^*}$ are correctly specified, $\gamma^*_{m_1^*} = \gamma_{m_1^*,0}$, and $\alpha^*_{m_2^*} = \alpha_{m_2^*,0}$.

7.7.1 Proof of 1

(i) In this case, c_{m_1} is correctly specified, and $\gamma^*_{m_1} = \gamma_{m_1,0}$. Furthermore, the equation $\mathbb{H}_{1,0}^{(m_1^*)}(\gamma_{m_1^*,0}) = \mathbb{H}_{1,0}^{(m_1)}(\gamma_{m_1,0})$ holds. Define the map $f_1 : \Theta_{\gamma_{m_1^*}} \to \Theta_{\gamma_{m_1}}$ by $f_1(\gamma_{m_1^*}) = F_1 \gamma_{m_1^*} + c_1$, where F_1 and c_1 satisfy the equation $\mathbb{H}_{1,n}^{(m_1^*)}(\gamma_{m_1^*}) = \mathbb{H}_{1,n}^{(m_1)}\big(f_1(\gamma_{m_1^*})\big)$ for any $\gamma_{m_1^*} \in \Theta_{\gamma_{m_1^*}}$. When $f_1(\gamma_{m_1^*,0}) \neq \gamma_{m_1,0}$, $\mathbb{H}_{1,0}^{(m_1^*)}(\gamma_{m_1^*,0}) = \mathbb{H}_{1,0}^{(m_1)}\big(f_1(\gamma_{m_1^*,0})\big) < \mathbb{H}_{1,0}^{(m_1)}(\gamma_{m_1,0})$. Hence, we have $f_1(\gamma_{m_1^*,0}) = \gamma_{m_1,0}$.

Using the fact that $P\{\partial_\gamma \mathbb{H}_{1,n}^{(m_1)}(\hat\gamma_{m_1,n}) = 0\} \to 1$ and Taylor expansion of $\mathbb{H}_{1,n}^{(m_1)}$,

$$\begin{aligned}\mathbb{H}_{1,n}^{(m_1^*)}(\hat\gamma_{m_1^*,n}) &= \mathbb{H}_{1,n}^{(m_1)}\left(f_1(\hat\gamma_{m_1^*,n})\right) \\ &= \mathbb{H}_{1,n}^{(m_1)}(\hat\gamma_{m_1,n}) - \frac{1}{2}\left(-\partial^2_{\gamma_{m_1}} \mathbb{H}_{1,n}^{(m_1)}(\tilde\gamma_{m_1,n})\right)\left[\left\{\hat\gamma_{m_1,n} - f_1(\hat\gamma_{m_1^*,n})\right\}^{\otimes 2}\right]\end{aligned}$$

where $\tilde\gamma_{m_1,n} = \hat\gamma_{m_1,n} - \xi_1\left\{f_1(\hat\gamma_{m_1^*,n}) - \hat\gamma_{m_1,n}\right\}$ for some $0 < \xi_1 < 1$ and $\tilde\gamma_{m_1,n} \xrightarrow{p} \gamma_{m_1,0}$ as $n \to \infty$. Therefore, the difference between $\mathrm{GQAIC}_{1,n}^{(m_1^*)}$ and $\mathrm{GQAIC}_{1,n}^{(m_1)}$ is given by

$$\begin{aligned}
\mathrm{GQAIC}^{(m_1^*)}_{1,n} - \mathrm{GQAIC}^{(m_1)}_{1,n} &= \left(-\partial^2_{\gamma_{m_1}} \mathbb{H}^{(m_1)}_{1,n}(\tilde{\gamma}_{m_1,n})\right)\left[\left\{\hat{\gamma}_{m_1,n} - f_1(\hat{\gamma}_{m_1^*,n})\right\}^{\otimes 2}\right] \\
&\quad + \frac{2}{h}\,\mathrm{trace}\left(\hat{\Gamma}^{-1}_{\gamma_{m_1^*},n}\hat{W}_{\gamma_{m_1^*},n}\right) - \frac{2}{h}\,\mathrm{trace}\left(\hat{\Gamma}^{-1}_{\gamma_{m_1},n}\hat{W}_{\gamma_{m_1},n}\right) \\
&= \frac{1}{h}\left(-\frac{1}{n}\partial^2_{\gamma_{m_1}} \mathbb{H}^{(m_1)}_{1,n}(\tilde{\gamma}_{m_1,n})\right)\left[\left\{\sqrt{T_n}\left(\hat{\gamma}_{m_1,n} - f_1(\hat{\gamma}_{m_1^*,n})\right)\right\}^{\otimes 2}\right] \\
&\quad + \frac{2}{h}\,\mathrm{trace}\left(\hat{\Gamma}^{-1}_{\gamma_{m_1^*},n}\hat{W}_{\gamma_{m_1^*},n}\right) - \frac{2}{h}\,\mathrm{trace}\left(\hat{\Gamma}^{-1}_{\gamma_{m_1},n}\hat{W}_{\gamma_{m_1},n}\right).
\end{aligned}$$

Since the chain rule gives $\partial_{\gamma_{m_1^*}} \mathbb{H}^{(m_1^*)}_{1,n}(\gamma_{m_1^*,0}) = F_1^\top \partial_{\gamma_{m_1}} \mathbb{H}^{(m_1)}_{1,n}(\gamma_{m_1,0})$ and $\partial^2_{\gamma_{m_1^*}} \mathbb{H}^{(m_1^*)}_{1,n}(\gamma_{m_1^*,0}) = F_1^\top \partial^2_{\gamma_{m_1}} \mathbb{H}^{(m_1)}_{1,n}(\gamma_{m_1,0}) F_1$,

$$\begin{aligned}
\sqrt{T_n}\left\{f_1(\hat{\gamma}_{m_1^*,n}) - \gamma_{m_1,0}\right\} &= \sqrt{T_n}\left\{f_1(\hat{\gamma}_{m_1^*,n}) - f_1(\gamma_{m_1^*,0})\right\} \\
&= F_1\sqrt{T_n}(\hat{\gamma}_{m_1^*,n} - \gamma_{m_1^*,0}) \\
&= F_1\left(-\frac{1}{n}\partial^2_{\gamma_{m_1^*}} \mathbb{H}^{(m_1^*)}_{1,n}(\gamma_{m_1^*,0})\right)^{-1}\left(\sqrt{\frac{h}{n}}\partial_{\gamma_{m_1^*}} \mathbb{H}^{(m_1^*)}_{1,n}(\gamma_{m_1^*,0})\right) \\
&\quad + O_p(T_n^{-1/2}) \\
&= F_1\left[F_1^\top\left\{-\frac{1}{n}\partial^2_{\gamma_{m_1}} \mathbb{H}^{(m_1)}_{1,n}(\gamma_{m_1,0})\right\}F_1\right]^{-1} F_1^\top\left(\sqrt{\frac{h}{n}}\partial_{\gamma_{m_1}} \mathbb{H}^{(m_1)}_{1,n}(\gamma_{m_1,0})\right) \\
&\quad + O_p(T_n^{-1/2}) \xrightarrow{\mathcal{L}} F_1\left(F_1^\top \Gamma_{\gamma_{m_1}}(\gamma_{m_1,0})F_1\right)^{-1} F_1^\top W_{\gamma_{m_1}}(\gamma_{m_1,0})^{1/2}\mathbf{N}_{m_1},
\end{aligned}$$

where $\mathbf{N}_{m_1} \sim N_{p_{\gamma_{m_1}}}(0, I_{p_{\gamma_{m_1}}})$ (under P without loss of generality). Moreover, we have

$$\begin{aligned}
\sqrt{T_n}\left(\hat{\gamma}_{m_1,n} - f_1(\hat{\gamma}_{m_1^*,n})\right) &= \sqrt{T_n}\left\{(\hat{\gamma}_{m_1,n} - \gamma_{m_1,0}) - \left(f_1(\hat{\gamma}_{m_1^*,n}) - \gamma_{m_1,0}\right)\right\} \\
&\xrightarrow{\mathcal{L}} \left\{\Gamma_{\gamma_{m_1}}(\gamma_{m_1,0})^{-1} - F_1\left(F_1^\top \Gamma_{\gamma_{m_1}}(\gamma_{m_1,0})F_1\right)^{-1} F_1^\top\right\} \\
&\quad W_{\gamma_{m_1}}(\gamma_{m_1,0})^{1/2}\mathbf{N}_{m_1} \\
&= G_{\gamma_{m_1}}(\gamma_{m_1,0}) W_{\gamma_{m_1}}(\gamma_{m_1,0})^{1/2}\mathbf{N}_{m_1}.
\end{aligned}$$

Hence,

$$
\begin{aligned}
&P\Big(\mathrm{GQAIC}_{1,n}^{(m_1^*)} - \mathrm{GQAIC}_{1,n}^{(m_1)} > 0\Big) \\
&= P\left[\frac{1}{h}\Big(-\frac{1}{n}\partial^2_{\gamma_{m_1}} \mathbb{H}_{1,n}^{(m_1)}(\tilde{\gamma}_{m_1,n})\Big)\left[\Big(G_{\gamma_{m_1}}(\gamma_{m_1,0})W_{\gamma_{m_1}}(\gamma_{m_1,0})^{1/2}\mathbf{N}_{m_1}\Big)^{\otimes 2}\right]\right. \\
&\qquad \left. + \frac{2}{h}\,\mathrm{trace}\left(\hat{\Gamma}^{-1}_{\gamma_{m_1^*},n}\hat{W}_{\gamma_{m_1^*},n}\right) - \frac{2}{h}\,\mathrm{trace}\left(\hat{\Gamma}^{-1}_{\gamma_{m_1},n}\hat{W}_{\gamma_{m_1},n}\right) > 0\right] \\
&= P\left[\Big(-\frac{1}{n}\partial^2_{\gamma_{m_1}} \mathbb{H}_{1,n}^{(m_1)}(\tilde{\gamma}_{m_1,n})\Big)\left[\Big(G_{\gamma_{m_1}}(\gamma_{m_1,0})W_{\gamma_{m_1}}(\gamma_{m_1,0})^{1/2}\mathbf{N}_{m_1}\Big)^{\otimes 2}\right] > \right. \\
&\qquad \left. 2\,\mathrm{trace}\left(\hat{\Gamma}^{-1}_{\gamma_{m_1},n}\hat{W}_{\gamma_{m_1},n}\right) - 2\,\mathrm{trace}\left(\hat{\Gamma}^{-1}_{\gamma_{m_1^*},n}\hat{W}_{\gamma_{m_1^*},n}\right)\right] \\
&\to P\left[\Gamma_{\gamma_{m_1}}(\gamma_{m_1,0})\left[\Big(G_{\gamma_{m_1}}(\gamma_{m_1,0})W_{\gamma_{m_1}}(\gamma_{m_1,0})^{1/2}\mathbf{N}_{m_1}\Big)^{\otimes 2}\right]\right. \\
&\qquad \left. > 2\,\mathrm{trace}\left\{\Gamma_{\gamma_{m_1}}(\gamma_{m_1,0})^{-1}W_{\gamma_{m_1}}(\gamma_{m_1,0})\right\} - 2\,\mathrm{trace}\left\{\Gamma_{\gamma_{m_1^*}}(\gamma_{m_1^*,0})^{-1}W_{\gamma_{m_1^*}}(\gamma_{m_1^*,0})\right\}\right] \\
&\to P\left[\mathbf{N}_{m_1}^\top W_{\gamma_{m_1}}(\gamma_{m_1,0})^{1/2}G_{\gamma_{m_1}}(\gamma_{m_1,0})W_{\gamma_{m_1}}(\gamma_{m_1,0})^{1/2}\mathbf{N}_{m_1}\right. \\
&\qquad \left. > 2\,\mathrm{trace}\left\{\Gamma_{\gamma_{m_1}}(\gamma_{m_1,0})^{-1}W_{\gamma_{m_1}}(\gamma_{m_1,0})\right\} - 2\,\mathrm{trace}\left\{\Gamma_{\gamma_{m_1^*}}(\gamma_{m_1^*,0})^{-1}W_{\gamma_{m_1^*}}(\gamma_{m_1^*,0})\right\}\right]
\end{aligned}
$$

as $n \to \infty$. Since $W_{\gamma_{m_1}}(\gamma_{m_1,0})^{1/2}G_{\gamma_{m_1}}(\gamma_{m_1,0})W_{\gamma_{m_1}}(\gamma_{m_1,0})^{1/2}$ is the symmetric matrix, there exists an orthogonal matrix M such that

$$
\begin{aligned}
&M^\top\Big(W_{\gamma_{m_1}}(\gamma_{m_1,0})^{1/2}G_{\gamma_{m_1}}(\gamma_{m_1,0})W_{\gamma_{m_1}}(\gamma_{m_1,0})^{1/2}\Big)M \\
&= \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{p_{\gamma_{m_1}}}).
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\mathbf{N}_{m_1}^\top W_{\gamma_{m_1}}(\gamma_{m_1,0})^{1/2}G_{\gamma_{m_1}}(\gamma_{m_1,0})W_{\gamma_{m_1}}(\gamma_{m_1,0})^{1/2}\mathbf{N}_{m_1} \\
&= \mathbf{N}_{m_1}^\top M\,\mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{p_{\gamma_{m_1}}})M^\top\mathbf{N}_{m_1} \\
&= \sum_{j=1}^{p_{\gamma_{m_1}}} \lambda_j \chi_j^2
\end{aligned}
$$

in distribution.

(ii) Since $\mathfrak{M}_1 = \mathrm{argmax}_{m_1}\, \mathbb{H}_{1,0}^{(m_1)}(\gamma_{m_1}^*)$ holds, the inequality $\mathbb{H}_{1,0}^{(m_1)}(\gamma_{m_1}^*) < \mathbb{H}_{1,0}^{(m_1^*)}(\gamma_{m_1^*}^*)\big(= \mathbb{H}_{1,0}^{(m_1^*)}(\gamma_{m_1^*,0})\big)$ is satisfied. Further, we have

$$
\begin{aligned}
\frac{1}{n}\mathbb{H}_{1,n}^{(m_1)}(\hat{\gamma}_{m_1,n}) &= \mathbb{H}_{1,0}^{(m_1)}(\gamma_{m_1}^*) + o_p(1), \\
\frac{1}{n}\mathbb{H}_{1,n}^{(m_1^*)}(\hat{\gamma}_{m_1^*,n}) &= \mathbb{H}_{1,0}^{(m_1^*)}(\gamma_{m_1^*,0}) + o_p(1).
\end{aligned}
$$

Hence,

$$
\begin{aligned}
&P\big(\mathrm{GQAIC}^{(m_1^*)} - \mathrm{GQAIC}^{(m_1)} > 0\big) \\
&\quad = P\Big[-\frac{2}{n}\Big(\mathbb{H}_{1,n}^{(m_1^*)}(\hat{\gamma}_{m_1^*,n}) - \mathbb{H}_{1,n}^{(m_1)}(\hat{\gamma}_{m_1,n})\Big) \\
&\quad > \frac{2}{n}\Big\{\mathrm{trace}\Big(\hat{\Gamma}_{\gamma_{m_1},n}^{-1}\hat{W}_{\gamma_{m_1},n}\Big) - \mathrm{trace}\Big(\hat{\Gamma}_{\gamma_{m_1^*},n}^{-1}\hat{W}_{\gamma_{m_1^*},n}\Big)\Big\}\Big] \\
&\quad = P\Big\{-2\Big(\mathbb{H}_{1,0}^{(m_1^*)}(\gamma_{m_1^*,0}) - \mathbb{H}_{1,0}^{(m_1)}(\gamma_{m_1}^*)\Big)o_p(1)\Big\} \to 0
\end{aligned}
$$

as $n \to \infty$.

7.7.2 Proof of 2

(i) The claims of 1 of this theorem mean that $P(\hat{m}_{1,n} \in \mathfrak{M}_1) \to 1$. For any $m_2 \in \mathfrak{M}_2 \backslash \{m_2^*\}$, we have

$$
\begin{aligned}
&P\Big(\mathrm{GQAIC}_{2,n}^{(m_2^*|\hat{m}_{1,n})} - \mathrm{GQAIC}_{2,n}^{(m_2|\hat{m}_{1,n})} > 0\Big) \\
&= P\Big(\mathrm{GQAIC}_{2,n}^{(m_2^*|\hat{m}_{1,n})} - \mathrm{GQAIC}_{2,n}^{(m_2|\hat{m}_{1,n})} > 0 \mid \hat{m}_{1,n} \in \mathfrak{M}_1\Big)P\big(\hat{m}_{1,n} \in \mathfrak{M}_1\big) \\
&\qquad + P\Big(\mathrm{GQAIC}_{2,n}^{(m_2^*|\hat{m}_{1,n})} - \mathrm{GQAIC}_{2,n}^{(m_2|\hat{m}_{1,n})} > 0 \mid \hat{m}_{1,n} \notin \mathfrak{M}_1\Big)P\big(\hat{m}_{1,n} \notin \mathfrak{M}_1\big) \\
&= P\Big(\mathrm{GQAIC}_{2,n}^{(m_2^*|\hat{m}_{1,n})} - \mathrm{GQAIC}_{2,n}^{(m_2|\hat{m}_{1,n})} > 0 \mid \hat{m}_{1,n} \in \mathfrak{M}_1\Big) + o(1).
\end{aligned}
$$

Below, we focus on the cases where $\hat{m}_{1,n} \in \mathfrak{M}_1$. Then, $\hat{\gamma}_{\hat{m}_{1,n},n}$ converges to $\gamma_{\hat{m}_{1,n},0}$ in probability, and $c_{\hat{m}_{1,n}}(\cdot, \gamma_{\hat{m}_{1,n},0}) = C(\cdot)$.

Because of assumptions, a_{m_2} is correctly specified, and $\alpha_{m_2}^* = \alpha_{m_2,0}$. Define the map $f_2 : \Theta_{\alpha_{m_2^*}} \to \Theta_{\alpha_{m_2}}$ by $f_2(\alpha_{m_2^*}) = F_2\alpha_{m_2^*} + c_2$, where F_2 and c_2 satisfy the equation $\mathbb{H}_{2,n}^{(m_2^*|m_1)}(\alpha_{m_2^*}) = \mathbb{H}_{2,n}^{(m_2|m_1)}\big(f_2(\alpha_{m_2^*})\big)$ for any $\alpha_{m_2^*} \in \Theta_{\alpha_{m_2^*}}$. If $f_2(\alpha_{m_2^*,0}) = \alpha_{m_2,0}$, then the inequality $\mathbb{H}_{2,0}^{(m_2^*|m_1)}(\alpha_{m_2^*,0}) = \mathbb{H}_{2,0}^{(m_2|m_1)}\big(f_2(\alpha_{m_2^*,0})\big) < \mathbb{H}_{2,0}^{(m_2|m_1)}(\alpha_{m_2,0})$ holds, and the assumption of $\mathfrak{M}_2$ is not satisfied. Hence, we have $f_2(\alpha_{m_2^*,0}) = \alpha_{m_2,0}$.

Considering the fact that $P\{\partial_\alpha \mathbb{H}_{2,n}^{(m_2|\hat{m}_{1,n})}(\hat{\alpha}_{m_2,n}) = 0\} \to 1$ and Taylor expansion of $\mathbb{H}_{2,n}^{(m_2|\hat{m}_{1,n})}$,

$$
\begin{aligned}
\mathbb{H}_{2,n}^{(m_2^*|\hat{m}_{1,n})}(\hat{\alpha}_{m_2^*,n}) &= \mathbb{H}_{2,n}^{(m_2|\hat{m}_{1,n})}\big(f_2(\hat{\alpha}_{m_2^*,n})\big) \\
&= \mathbb{H}_{2,n}^{(m_2|\hat{m}_{1,n})}(\hat{\alpha}_{m_2,n}) \\
&\quad - \frac{1}{2}\Big(-\frac{1}{T_n}\partial_{\alpha_{m_2}}^2 \mathbb{H}_{2,n}^{(m_2|\hat{m}_{1,n})}(\tilde{\alpha}_{m_2,n})\Big)\Big[\Big\{\sqrt{T_n}\big(\hat{\alpha}_{m_2,n} - f_2(\hat{\alpha}_{m_2^*,n})\big)\Big\}^{\otimes 2}\Big],
\end{aligned}
$$

where $\tilde{\alpha}_{m_2,n} = \hat{\alpha}_{m_2,n} - \xi_2 \left\{ f_2(\hat{\alpha}_{m_2^*,n}) - \hat{\alpha}_{m_2,n} \right\}$ for some $0 < \xi_2 < 1$ and $\tilde{\alpha}_{m_2,n} \xrightarrow{p} \alpha_{m_2,0}$ as $n \to \infty$. Moreover,

$$
\begin{aligned}
&\sqrt{T_n}(\hat{\alpha}_{m_2,n} - \alpha_{m_2,0}) \\
&= \left(-\frac{1}{T_n} \partial^2_{\alpha_{m_2}} \mathbb{H}^{(m_2|\hat{m}_{1,n})}_{2,n}(\alpha_{m_2,0}, \hat{\gamma}_{\hat{m}_{1,n},n}) \right)^{-1} \left(\frac{1}{\sqrt{T_n}} \partial_{\alpha_{m_2}} \mathbb{H}^{(m_2|\hat{m}_{1,n})}_{2,n}(\alpha_{m_2,0}, \hat{\gamma}_{\hat{m}_{1,n},n}) \right) + O_p(T_n^{-1/2}) \\
&= \left(-\frac{1}{T_n} \partial^2_{\alpha_{m_2}} \mathbb{H}^{(m_2|\hat{m}_{1,n})}_{2,n}(\alpha_{m_2,0}, \gamma_{\hat{m}_{1,n},0}) \right)^{-1} \left(\frac{1}{\sqrt{T_n}} \partial_{\alpha_{m_2}} \mathbb{H}^{(m_2|\hat{m}_{1,n})}_{2,n}(\alpha_{m_2,0}, \gamma_{\hat{m}_{1,n},0}) \right. \\
&\qquad \left. -\frac{1}{T_n} \partial_{\alpha_{m_2}} \partial_{\gamma_{\hat{m}_{1,n}}} \mathbb{H}^{(m_2|\hat{m}_{1,n})}_{2,n}(\alpha_{m_2,0}, \gamma_{\hat{m}_{1,n},0}) \left[\sqrt{T_n}(\hat{\gamma}_{\hat{m}_{1,n},n} - \gamma_{\hat{m}_{1,n},0}) \right] \right) + O_p(T_n^{-1/2}) \\
&\xrightarrow{\mathcal{L}} \Gamma_{\alpha_{m_2}}(\alpha_{m_2,0}, \gamma_{\hat{m}_{1,n},0})^{-1/2} \mathbf{N}_{m_2},
\end{aligned}
$$

where $\mathbf{N}_{m_2} \sim N_{p_{\alpha_{m_2}}}\left(0, I_{p_{\alpha_{m_2}}}\right)$. By the above and chain rule,

$$
\begin{aligned}
&\sqrt{T_n}\left\{ \hat{\alpha}_{m_2,n} - f_2(\hat{\alpha}_{m_2^*,n}) \right\} \\
&= \sqrt{T_n}\left\{ (\hat{\alpha}_{m_2,n} - \alpha_{m_2,0}) - \left(f_2(\hat{\alpha}_{m_2^*,n}) - f_2(\alpha_{m_2^*,0}) \right) \right\} \\
&= \sqrt{T_n}(\hat{\alpha}_{m_2,n} - \alpha_{m_2,0}) - F_2 \sqrt{T_n}(\hat{\alpha}_{m_2^*,n} - \alpha_{m_2^*,0}) \\
&= \sqrt{T_n}(\hat{\alpha}_{m_2,n} - \alpha_{m_2,0}) - F_2 \left(-\frac{1}{T_n} \partial^2_{\alpha_{m_2^*}} \mathbb{H}^{(m_2^*|\hat{m}_{1,n})}_{2,n}(\alpha_{m_2^*,0}, \gamma_{\hat{m}_{1,n},0}) \right)^{-1} \\
&\qquad \times \left(\frac{1}{\sqrt{T_n}} \partial_{\alpha_{m_2^*}} \mathbb{H}^{(m_2^*|\hat{m}_{1,n})}_{2,n}(\alpha_{m_2^*,0}, \gamma_{\hat{m}_{1,n},0}) \right) + O_p(T_n^{-1/2}) \\
&= \sqrt{T_n}(\hat{\alpha}_{m_2,n} - \alpha_{m_2,0}) - F_2 \left[F_2^\top \left\{ -\frac{1}{T_n} \partial^2_{\alpha_{m_2}} \mathbb{H}^{(m_2|\hat{m}_{1,n})}_{2,n}\left(\alpha_{m_2,0}, \gamma_{\hat{m}_{1,n},0}\right) \right\} F_2 \right]^{-1} \\
&\qquad \times F_2^\top \left(\frac{1}{\sqrt{T_n}} \partial_{\alpha_{m_2}} \mathbb{H}^{(m_2|\hat{m}_{1,n})}_{2,n}(\alpha_{m_2,0}, \gamma_{\hat{m}_{1,n},0}) \right) + O_p(T_n^{-1/2}) \\
&\xrightarrow{\mathcal{L}} \left\{ \Gamma_{\alpha_{m_2}}(\alpha_{m_2,0}, \gamma_{\hat{m}_{1,n},0})^{-1} - F_2 \left(F_2^\top \Gamma_{\alpha_{m_2}}(\alpha_{m_2,0}, \gamma_{\hat{m}_{1,n},0}) F_2 \right)^{-1} F_2^\top \right\} \\
&\quad \Gamma_{\alpha_{m_2}}(\alpha_{m_2,0}, \gamma_{\hat{m}_{1,n},0})^{1/2} \mathbf{N}_{m_2} \\
&= G_{\alpha_{m_2}}(\alpha_{m_2,0}, \gamma_{\hat{m}_{1,n},0}) \Gamma_{\alpha_{m_2}}(\alpha_{m_2,0}, \gamma_{\hat{m}_{1,n},0})^{1/2} \mathbf{N}_{m_2}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&P\Big(\mathrm{GQAIC}_{2,n}^{(m_2^*|\hat{m}_{1,n})} - \mathrm{GQAIC}_{2,n}^{(m_2|\hat{m}_{1,n})} > 0\Big) \\
&= P\Big(\mathrm{GQAIC}_{2,n}^{(m_2^*|\hat{m}_{1,n})} - \mathrm{GQAIC}_{2,n}^{(m_2|\hat{m}_{1,n})} > 0 \mid \hat{m}_{1,n} \in \mathfrak{M}_1\Big) + o(1) \\
&= P\left[\left(-\frac{1}{T_n}\partial^2_{\alpha_{m_2}} \mathbb{H}_{2,n}^{(m_2|\hat{m}_{1,n})}(\tilde{\alpha}_{m_2,n})\right)\left[\left\{\sqrt{T_n}\big(\hat{\alpha}_{m_2,n} - f_2(\hat{\alpha}_{m_2^*,n})\big)\right\}^{\otimes 2}\right]\right. \\
&\qquad \left. + 2p_{\alpha_{m_2^*}} - 2p_{\alpha_{m_2}} > 0 \mid \hat{m}_{1,n} \in \mathfrak{M}_1\right] + o(1) \\
&\to P\left[\Gamma_{\alpha_{m_2}}(\alpha_{m_2,0}, \gamma_{\hat{m}_{1,n},0})\left[\left(G_{\alpha_{m_2}}(\alpha_{m_2,0}, \gamma_{\hat{m}_{1,n},0})\Gamma_{\alpha_{m_2}}(\alpha_{m_2,0}, \gamma_{\hat{m}_{1,n},0})^{1/2}\mathbf{N}_{m_2}\right)^{\otimes 2}\right]\right. \\
&> s2(p_{\alpha_{m_2}} - p_{\alpha_{m_2^*}})\Big] \\
&= P\left[\mathbf{N}_{m_2}^\top \Gamma_{\alpha_{m_2}}(\alpha_{m_2,0}, \gamma_{\hat{m}_{1,n},0})^{1/2} G_{\alpha_{m_2}}(\alpha_{m_2,0}, \gamma_{\hat{m}_{1,n},0})\Gamma_{\alpha_{m_2}}(\alpha_{m_2,0}, \gamma_{\hat{m}_{1,n},0})^{1/2}\mathbf{N}_{m_2}\right. \\
&> 2(p_{\alpha_{m_2}} - p_{\alpha_{m_2^*}})\Big]
\end{aligned}
$$

as $n \to \infty$. In a similar way as (i) of Sect. 7.7.2, we can show that

$$
\mathbf{N}_{m_2}^\top \Gamma_{\alpha_{m_2}}(\alpha_{m_2,0}, \gamma_{\hat{m}_{1,n},0})^{1/2} G_{\alpha_{m_2}}(\alpha_{m_2,0}, \gamma_{\hat{m}_{1,n},0})\Gamma_{\alpha_{m_2}}(\alpha_{m_2,0}, \gamma_{\hat{m}_{1,n},0})^{1/2}\mathbf{N}_{m_2} = \sum_{j=1}^{p_{\alpha_{m_2}}} \lambda_j' \chi_j^2
$$

in distribution.

(ii) The set $\mathfrak{M}_2$ satisfies that $\mathfrak{M}_2 = \mathrm{argmax}_{m_2} \mathbb{H}_{2,0}^{(m_2|\hat{m}_{1,n})}(\alpha_{m_2}^*)$, so that the inequality $\mathbb{H}_{2,0}^{(m_2|\hat{m}_{1,n})}(\alpha_{m_2}^*) < \mathbb{H}_{2,0}^{(m_2^*|\hat{m}_{1,n})}(\alpha_{m_2^*}^*)\big(= \mathbb{H}_{2,0}^{(m_2^*|\hat{m}_{1,n})}(\alpha_{m_2^*,0})\big)$ holds. Since

$$
\begin{aligned}
\frac{1}{T_n}\mathbb{H}_{2,n}^{(m_2|\hat{m}_{1,n})}(\hat{\alpha}_{m_2,n}) &= \mathbb{H}_{2,0}^{(m_2|\hat{m}_{1,n})}(\alpha_{m_2}^*) + o_p(1), \\
\frac{1}{T_n}\mathbb{H}_{2,n}^{(m_2^*|\hat{m}_{1,n})}(\hat{\alpha}_{m_2^*,n}) &= \mathbb{H}_{2,0}^{(m_2^*|\hat{m}_{1,n})}(\alpha_{m_2^*,0}) + o_p(1),
\end{aligned}
$$

we have

$$
\begin{aligned}
&P\Big(\mathrm{GQAIC}_{2,n}^{(m_2^*|\hat{m}_{1,n})} - \mathrm{GQAIC}_{2,n}^{(m_2|\hat{m}_{1,n})} > 0\Big) \\
&= P\left\{-\frac{2}{T_n}\left(\mathbb{H}_{2,n}^{(m_2^*|\hat{m}_{1,n})}(\hat{\alpha}_{m_2^*,n}) - \frac{1}{T_n}\mathbb{H}_{2,n}^{(m_2|\hat{m}_{1,n})}(\hat{\alpha}_{m_2,n})\right) > \frac{2}{T_n}(p_{\alpha_{m_2}} - p_{\alpha_{m_2^*}})\right\} \\
&= P\left\{-2\Big(\mathbb{H}_{2,0}^{(m_2^*|\hat{m}_{1,n})}(\alpha_{m_2^*,0}) - \mathbb{H}_{2,0}^{(m_2|\hat{m}_{1,n})}(\alpha_{m_2}^*) > o_p(1)\Big)\right\} \\
&\to 0
\end{aligned}
$$

as $n \to \infty$. □

7.8 Proof of Theorem 6

(i) Since $h \log n \to 0$, in a similar way as the proof of Theorem 5 1(i), we obtain

$$\begin{aligned}
&P\Big(\mathrm{GQBIC}_{1,n}^{\sharp(m_1^*)} - \mathrm{GQBIC}_{1,n}^{\sharp(m_1)} > 0\Big) \\
&= P\left[\left(-\frac{1}{n}\partial_{\gamma_{m_1}}^2 \mathbb{H}_{1,n}^{(m_1)}(\tilde{\gamma}_{m_1,n})\right)\left[\left(G_{\gamma_{m_1}}(\gamma_{m_1,0})W_{\gamma_{m_1}}(\gamma_{m_1,0})^{1/2}\mathbf{N}_{m_1}\right)^{\otimes 2}\right] > \left(\gamma_{m_1} - \gamma_{m_1^*}\right)h\log n\right] \\
&\to P\left[\Gamma_{\gamma_{m_1}}(\gamma_{m_1,0})\left[\left(G_{\gamma_{m_1}}(\gamma_{m_1,0})W_{\gamma_{m_1}}(\gamma_{m_1,0})^{1/2}\mathbf{N}_{m_1}\right)^{\otimes 2}\right] > 0\right] \\
&= 1
\end{aligned}$$

as $n \to \infty$.

(ii) As with Theorem 5 1(ii), we can get

$$\begin{aligned}
&P\Big(\mathrm{GQBIC}_{1,n}^{\sharp(m_1^*)} - \mathrm{GQBIC}_{1,n}^{\sharp(m_1)} > 0\Big) \\
&= P\left\{-2\left(\mathbb{H}_{1,0}^{(m_1^*)}(\gamma_{m_1^*,0}) - \mathbb{H}_{1,0}^{(m_1)}(\gamma_{m_1}^*)\right)\right. \\
&\left.> \left(\gamma_{m_1} - \gamma_{m_1^*}\right)\frac{\log n}{n}\right\} \\
&\to 0
\end{aligned}$$

as $n \to \infty$. □

7.9 Proof of Theorem 7

Since the proof of claim 2 can be handled analogously as claim 1 and Theorem 5 2, we only prove claim 1.

(i) In a similar way as the proof of Theorem 5 1(i), we have

$$\begin{aligned}
&P\Big(\mathrm{GQBIC}_{1,n}^{(m_1^*)} - \mathrm{GQBIC}_{1,n}^{(m_1)} > 0\Big) \\
&= P\left[\left(-\frac{1}{n}\partial_{\gamma_{m_1}}^2 \mathbb{H}_{1,n}^{(m_1)}(\tilde{\gamma}_{m_1,n})\right)\left[\left(G_{\gamma_{m_1}}(\gamma_{m_1,0})W_{\gamma_{m_1}}(\gamma_{m_1,0})^{1/2}\mathbf{N}_{m_1}\right)^{\otimes 2}\right]\right. \\
&\left.> \left(\gamma_{m_1} - \gamma_{m_1^*}\right)\log T_n\right] \\
&\to 0,
\end{aligned}$$

as $n \to \infty$.

(ii) As with Theorem 5 1(ii), we can deduce that

$$\begin{aligned}
P\Big(\mathrm{GQBIC}_{1,n}^{(m_1^*)} - \mathrm{GQBIC}_{1,n}^{(m_1)} > 0\Big) &= P\left\{-2\left(\mathbb{H}_{1,0}^{(m_1^*)}(\gamma_{m_1^*,0}) - \mathbb{H}_{1,0}^{(m_1)}(\gamma_{m_1}^*)\right)\right. \\
&\left.> \left(\gamma_{m_1} - \gamma_{m_1^*}\right)\frac{\log T_n}{T_n}\right\} \\
&\to 0
\end{aligned}$$

as $n \to \infty$. □

Acknowledgements The authors are grateful to the anonymous reviewers for their valuable comments, which led to significant improvements in the first version. This work was partially supported by JST CREST Grant Numbers JPMJCR14D7 and JPMJCR2115, and by JSPS KAKENHI Grant Numbers JP19K14593 and 22H01139.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Second international symposium on information theory (Tsahkadsor, 1971)* 267–281, Budapest: Akadémiai Kiadó.

Asmussen, S., Rosiński, J. (2001). Approximations of small jumps of Lévy processes with a view towards simulation. *Journal of Applied Probability, 38*(2), 482–493.

Bhansali, R. J., Papangelou, F. (1991). Convergence of moments of least squares estimators for the coefficients of an autoregressive process of unknown order. *The Annals of Statistics, 19*(3), 1155–1162.

Brouste, A., Fukasawa, M., Hino, H., Iacus, S. M., Kamatani, K., Koike, Y., Masuda, H., Nomura, R., Ogihara, T., Shimizu, Y., Uchida, M., Yoshida, N. (2014). The yuima project: A computational framework for simulation and inference of stochastic differential equations. *Journal of Statistical Software, 57*(4), 1–51.

Chan, N. H., Ing, C.-K. (2011). Uniform moment bounds of Fisher's information with applications to time series. *The Annals of Statistics, 39*(3), 1526–1550.

Clément, E., Gloter, A. (2020). Joint estimation for SDE driven by locally stable Lévy processes. *Electronic Journal of Statistics, 14*(2), 2922–2956.

Eguchi, S., Masuda, H. (2018). Schwarz type model comparison for LAQ models. *Bernoulli, 24*(3), 2278–2327.

Eguchi, S., Masuda, H. (2019). Data driven time scale in Gaussian quasi-likelihood inference. *Statistical Inference for Stochastic Processes, 22*(3), 383–430.

Eguchi, S., Uehara, Y. (2021). Schwartz-type model selection for ergodic stochastic differential equation models Scandinavian Journal of Statistics. *Theory and Applications, 48*(3), 950–968.

Findley, D. F., Wei, C.-Z. (2002). AIC, overfitting principles, and the boundedness of moments of inverse matrices for vector autoregressions and related models. *Journal of Multivariate Analysis, 83*(2), 415–450.

Iacus, S.M. and Yoshida, N. (2018). Simulation and inference for stochastic processes with YUIMA. Use R! Springer, Cham. A comprehensive R framework for SDEs and other stochastic processes.

Jasra, A., Kamatani, K., Masuda, H. (2019). Bayesian inference for stable Lévy-driven stochastic differential equations with high-frequency data. *Scandinavian Journal of Statistics., 46*(2), 545–574.

Kent, J. T. (1982). Robust properties of likelihood ratio tests. *Biometrika, 69*(1), 19–27.

Konishi, S., Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika, 83*(4), 875–890.

Liptser, R. S., Shiryaev, A. N. (2001). *Statistics of random processes. I*, expanded ed., Applications of Mathematics, Volume 5, New York, Berlin: Springer-Verlag.

Magnus, J. R., Neudecker, H. (1979). The commutation matrix: some properties and applications. *The Annals of Statistics, 7*(2), 381–394.

Masuda, H. (2010). Approximate self-weighted LAD estimation of discretely observed ergodic Ornstein-Uhlenbeck processes. *Electronic Journal of Statistics, 4*, 525–565.

Masuda, H. (2013). Convergence of Gaussian quasi-likelihood random fields for ergodic Lévy driven SDE observed at high frequency. *The Annals of Statistics, 41*(3), 1593–1641.

Masuda, H. (2019). Non-Gaussian quasi-likelihood estimation of SDE driven by locally stable Lévy process. *Stochastic Processes and their Applications, 129*(3), 1013–1059.

Masuda, H. (2023). Optimal stable Ornstein-Uhlenbeck regression. *Japanese Journal of Statistics and Data Science, accepted*.

Masuda, H., Uehara, Y. (2017). On stepwise estimation of Lévy driven stochastic differential equation (Japanese). *Proceedings of the Institute of Statistical Mathematics, 65*(1), 21–38.

Ogihara, T., Yoshida, N. (2011). Quasi-likelihood analysis for the stochastic differential equation with jumps. *Statistical Inference for Stochastic Processes, 14*(3), 189–229.
Radchenko, P. (2008). Mixed-rates asymptotics. *The Annals of Statistics, 36*(1), 287–309.
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464.
Shimizu, Y., Yoshida, N. (2006). Estimation of parameters for diffusion processes with jumps from discrete observations. *Statistical Inference and Stochastic Processes, 9*(3), 227–277.
Uchida, M. (2010). Contrast-based information criterion for ergodic diffusion processes from discrete observations. *Annals of the Institute of Statistical Mathematics, 62*(1), 161–187.
Uchida, M., Yoshida, N. (2012). Adaptive estimation of an ergodic diffusion process based on sampled data. *Stochastic Processes and their Applications, 122*(8), 2885–2924.
Uehara, Y. (2019). Statistical inference for misspecified ergodic Lévy driven stochastic differential equation models. *Stochastic Processes and their Applications, 129*(10), 4051–4081.
Watanabe, S. (2013). A widely applicable Bayesian information criterion. *Journal of Machine Learning Research, 14*, 867–897.
Yoshida, N. (2011). Polynomial type large deviation inequalities and quasi-likelihood analysis for stochastic differential equations. *Annals of the Institute of Statistical Mathematics, 63*(3), 431–479.