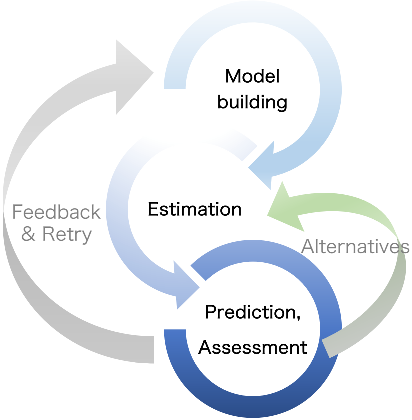
Gaussian quasi-likelihood inference for ergodic Lévy driven SDE

Hiroki Masuda

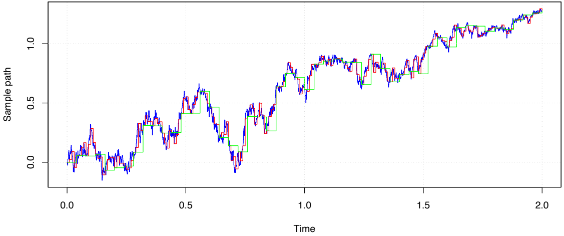
Graduate School of Mathematical Sciences
University of Tokyo, Japan

EFFI Spring School, Le Mans
May 27 & 28, 2024

Inference for an ergodic Lévy driven stochastic differential equation (SDE)



$$X_{t+dt} \leftarrow X_t + (\text{Trend})dt + (\text{Scale}) \cdot d(\text{Noise})$$



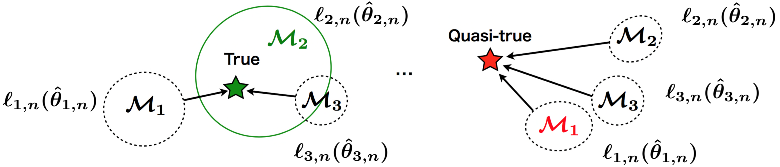
Contents

Eguchi, Shoichi and Masuda, Hiroki (2024).
Gaussian quasi-information criteria for ergodic Lévy driven SDE.
Ann. Inst. Statist. Math., 76(1):111–157. (arXiv:2203.04039v3 (22 May, 2023))

AIC and BIC in general: Twin jewels of information criteria

- AIC and BIC are based on different philosophies, neither is absolutely better than the other:
  - AIC is designed for **predictive (generalization) performance**, while
  - BIC is for **model-description** performance.

How to relatively measure the discrepancy from the model to the (virtual) truth?



Avoid **over-fitting** $\Longrightarrow$ $\underbrace{-2\ell_n(\hat{\theta}_n)}_{\text{Goodness-of-fit}} + (\text{Complexity/Redundancy regularization})$

Discrete-time setting, as a prelude

- Location-scale time series:

$$Y_j = a(Y_{j-1}, \alpha) + c(Y_{j-1}, \gamma)\epsilon_j, \qquad \epsilon_1, \epsilon_2, \dots \overset{\text{i.i.d.}}{\sim} (0, 1)$$

- Gaussian (logarithmic) quasi-likelihood for estimating $\theta = (\alpha, \gamma)$:

$$\theta \mapsto \ell_n(\theta) := \sum_{j=1}^{n} \log \phi\left(Y_j;\ a(Y_{j-1}, \alpha),\ c(Y_{j-1}, \gamma)^2\right)$$

$$\xrightarrow{\text{Maximize}} \hat{\theta}_n = (\hat{\alpha}_n, \hat{\gamma}_n) \in \text{argmax}\ell_n$$

$$\xrightarrow[\text{conditions}]{\text{Regularity}} \sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} N\left(0, \Gamma(\theta_0)^{-1}\Sigma(\theta_0)\Gamma(\theta_0)^{-1}\right)$$

$\Rightarrow$ Model assessment:
$$\begin{cases} \text{AIC} & -2\ell_n(\hat{\theta}_n) + 2\ \mathrm{trace}\left(\Gamma(\hat{\theta}_n)^{-1}\Sigma(\hat{\theta}_n)\right) \\ \text{BIC} & -2\ell_n(\hat{\theta}_n) + (\log n)\dim\theta \end{cases}$$

"The" continuous-time counterpart of the abovementioned machinery?

Objective

- Equally spaced high-frequency sample $\boldsymbol{X}_n = (X_{t_j^n})_{j=0}^n$ from **data-generating process**:

$$dX_t = C(X_{t-})dZ_t + A(X_t)dt \tag{1.1}$$

  - $t_j^n = t_j = jh, \quad T_n := nh \to \infty, \quad nh^2 \to 0 \ (n \to \infty)$
  - $A: \mathbb{R}^d \to \mathbb{R}^d$ and $C: \mathbb{R}^d \to \mathbb{R}^d \otimes \mathbb{R}^r$
  - $Z = (Z_t)_{t \in \mathbb{R}_+}$ is an r-dimensional **non-Gaussian** Lévy process independent of X_0.

- **Correctly specified** statistical models $\mathcal{M}_{m_1,m_2}$ for the unknown (1.1):

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

  - **Scale** candidates $c_1(x,\gamma_1), \dots, c_{M_1}(x, \gamma_{M_1}), \quad \gamma_{m_1} \in \Theta_{\gamma_{m_1}} \subset \mathbb{R}^{p_{\gamma_{m_1}}}$.
  - **Drift** candidates $a_1(x,\alpha_1), \dots, a_{M_2}(x, \alpha_{M_2}), \quad \alpha_{m_2} \in \Theta_{\alpha_{m_2}} \subset \mathbb{R}^{p_{\alpha_{m_2}}}$.

Fully explicit **two-stage** inference based on the **Gaussian quasi-likelihood function (GQLF)**

Regularity conditions

- Omitting the model indices, we look at a single model:

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

  - Unknown parameter $\theta := (\alpha,\gamma)$: $\gamma = (\gamma_k) \in \Theta_\gamma \subset \mathbb{R}^{p_\gamma}$ and $\alpha = (\alpha_l) \in \Theta_\alpha \subset \mathbb{R}^{p_\alpha}$
  - Let $\theta_0 = (\alpha_0,\gamma_0) \in \Theta := \Theta_\alpha \times \Theta_\gamma$ denote the true value of $\theta = (\alpha,\gamma)$.
  - Scale matrix $S(x,\gamma) := c(x,\gamma)^{\otimes 2}$

1. **Moments** of the driving noise
2. **Smoothness and non-degeneracy** of the coefficients
3. **Stability**: exponential ergodicity and boundednedss of moments
4. **Identifiability**

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

Assumption 1.1 (Moments of driving noise)

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

- For $Z=\sigma^{1/2}w+J$, the pure-jump part J satisfies

$$\frac{1}{h}E\left[J_h^{(i_1)}\cdots J_h^{(i_m)}\right]=\begin{cases}\nu_{i_1i_2i_3}(3) & m=3,\\ \nu_{i_1i_2i_3i_4}(4)+O(h) & m=4,\end{cases}$$

where, for $m\geq 3$ and $z=(z_1,\dots,z_r)$,

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

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

Assumption 1.2 (Smoothness and non-degeneracy)

The coefficient $(x, \theta) \mapsto (a(x, \alpha), c(x, \gamma))$ *is smooth enough:*

- *It is globally Lipschitz uniformly in* $\theta \in \overline{\Theta}$*;*
- *For some constant* $C_{k,l} \geq 0$*,*

$$\max_{i \leq d} \max_{j,k} \sup_{\theta \in \overline{\Theta}} \left(|\partial_\alpha^j \partial_x^i a(x, \alpha)| + |\partial_\gamma^k \partial_x^i c(x, \gamma)| \right) \lesssim 1 + |x|^{C_{k,l}}$$

and moreover, $\sup_{\gamma \in \overline{\Theta}_\gamma} \lambda_{\min}\{S(x, \gamma)\}^{-1} \lesssim 1 + |x|^{C_0}$ *for some constant* $C_0 \geq 0$*.*

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

Assumption 1.3 (Stability)

There exists a probability measure $\pi = \pi_{\theta_0}$ *s.t. for every* $q > 0$ *there exists a positive constant* a *s.t.*

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

where $g(x) := 1 + \|x\|^q$. *Further,* $\sup_{t \in \mathbb{R}_+} E[|X_t|^q] < \infty$ *for every* $q > 0$*,*

- Several sufficient conditions are available: among others,
  - Meyn-Tweedie-theory based [Masuda, 2007]
  - Local Doeblin condition based [Kulik, 2009]
  - Monograph [Kulik, 2018] with systematic and detailed descriptions
  - ...

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

Assumption 1.4 (Identifiability)

There exist positive constants χ_γ *and* χ_α *such that for every* γ *and* α*,*

$$-\frac{1}{2}\int \left\{ \mathrm{trace}\left(S(x,\gamma)^{-1}S(x,\gamma_0) - I_d\right) + \log\frac{|S(x,\gamma)|}{|S(x,\gamma_0)|} \right\}\pi(dx) \le -\chi_\gamma|\gamma-\gamma_0|^2,$$

$$-\frac{1}{2}\int S^{-1}(x,\gamma_0)\left[(a(x,\alpha)-a(x,\alpha_0))^{\otimes 2}\right]\pi(dx) \le -\chi_\alpha|\alpha-\alpha_0|^2.$$

- Different parameters should not lead to the same data-generating $\mathcal{L}(X)$.
- Essential in the **argmax** argument.

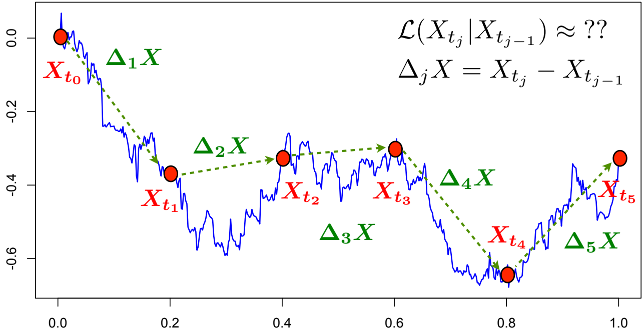
Advantage of **stepwise** estimation of the coefficients: scale → trend

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

Statistical problem, roughly

Find a resonable alternative to the unknown genuine **log-likelihood**, formally given by

$$\log p_h(X_{t_1}, \ldots, X_{t_n}; \theta) = \sum_{j=1}^{n} \log p_h(X_{t_{j-1}}, X_{t_j}; \theta)$$



Formulation of two-stage Gaussian quasi-likelihood estimation

- Write $\Delta_j Y = Y_{t_j} - Y_{t_{j-1}}$ and $f_{j-1}(\theta) = f(X_{t_{j-1}}, \theta)$.
- The Euler approximation for $dX_t = c(X_{t-}, \gamma)dZ_t + a(X_t, \alpha)dt$ under P_θ is given by

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

- Taking the small-time Gaussian approximation **(incorrect!)**:

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

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

$$\begin{aligned}\mathbb{H}_n(\theta) &:= \sum_{j=1}^{n} \log \phi_d\left(X_{t_j}; X_{t_{j-1}} + a_{j-1}(\alpha)h,\ hS_{j-1}(\gamma)\right) \\ &= -\frac{1}{2}\sum_{j=1}^{n}\left(\log|2\pi hS_{j-1}(\gamma)| + \frac{1}{h}S_{j-1}^{-1}(\gamma)\left[(\Delta_j X - ha_{j-1}(\alpha))^{\otimes 2}\right]\right). \qquad (2.1)\end{aligned}$$

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

$$\mathbb{H}_{1,n}(\gamma) := \sum_{j=1}^{n} \log \phi_d \left(X_{t_j};\, X_{t_{j-1}},\, hS_{j-1}(\gamma) \right),$$

$$\mathbb{H}_{2,n}(\theta) := \sum_{j=1}^{n} \left(S_{j-1}^{-1}(\gamma)\left[\Delta_j X,\, a_{j-1}(\alpha)\right] - \frac{h}{2} S_{j-1}^{-1}(\gamma)\left[a_{j-1}^{\otimes 2}(\alpha)\right] \right)$$

have two different resolutions: $n^{-1}\mathbb{H}_{1,n}(\gamma)$ and $T_n^{-1}\mathbb{H}_{2,n}(\theta)$ have proper LLN limits.
- The following **two-stage** estimation strategy is natural (Gaussian quasi-MLE: **GQMLE**):

1. Estimate γ by $\hat{\gamma}_n \in \mathrm{argmax}_\gamma \mathbb{H}_{1,n}(\gamma)$;
2. Estimate α by $\hat{\alpha}_n \in \mathrm{argmax}_\alpha \mathbb{H}_{2,n}(\alpha)$ where, with a slight abuse of notation:

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

Theorem 2.1 (Mighty convergence of the GQMLE [Eguchi and Masuda, 2024])

Under Assumptions 1.1, 1.2, 1.3, and 1.4, we have the convergence of moments for any continuous function $f : \mathbb{R}^p \to \mathbb{R}$ *of at most polynomial growth:*

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

with $V(\theta_0) = \Gamma(\theta_0)^{-1}\Sigma(\theta_0)\Gamma(\theta_0)^{-1}$ *explicitly given, hence in particular:*

- $\hat{u}_n := \sqrt{T_n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} N_p(0, V(\theta_0))$;
- $E[\hat{u}_n] \to 0$, $E[\hat{u}_n^{\otimes 2}] \to V(\theta_0)$, *and also* $\sup_n E[|\hat{u}_n|^q] < \infty$ *for every* $q > 0$.

Further, it is possible to construct an explicit consistent estimator $\hat{V}_n \xrightarrow{p} V(\theta_0)$, *so that*

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

Related remarks

- **Dividing estimation stages does not change asymptotic covariance.**

  Comparing Theorem 2.1 and [Masuda, 2013, Theorem 2.7] shows that the joint and stepwise GQMLE have the same asymptotic distribution.

$$\mathbb{H}_n(\theta) = \sum_{j=1}^{n} \log \phi_d \left(X_{t_j};\, X_{t_{j-1}} + a_{j-1}(\alpha)h,\, hS_{j-1}(\gamma) \right) = \mathbb{H}_{1,n}(\gamma) + \mathbb{H}_{2,n}(\theta),$$

$$\mathbb{H}_{1,n}(\gamma) = \sum_{j=1}^{n} \log \phi_d \left(X_{t_j};\, X_{t_{j-1}},\, hS_{j-1}(\gamma) \right),$$

$$\mathbb{H}_{2,n}(\theta) = \sum_{j=1}^{n} \left(S_{j-1}^{-1}(\gamma) \left[\Delta_j X,\, a_{j-1}(\alpha) \right] - \frac{h}{2} S_{j-1}^{-1}(\gamma) \left[a_{j-1}^{\otimes 2}(\alpha) \right] \right)$$

- **Simpler form for the univariate case** $(d = r = 1)$ for $V(\theta_0) = \Gamma(\theta_0)^{-1}\Sigma(\theta_0)\Gamma(\theta_0)^{-1}$:

$$\Sigma(\theta_0) = \begin{pmatrix} \Gamma_\alpha(\theta_0) & W_{\alpha,\gamma}(\theta_0) \\ W_{\alpha,\gamma}(\theta_0)^\top & W_\gamma(\gamma_0) \end{pmatrix}, \qquad \Gamma(\theta_0) = \mathrm{diag}\{\Gamma_\alpha(\theta_0), \Gamma_\gamma(\gamma_0)\}.$$

$$\Gamma_\alpha(\theta_0) = \int \left(\frac{\partial_\alpha a(x,\alpha_0)}{c(x,\gamma_0)} \right)^{\otimes 2} \pi(dx), \qquad \Gamma_\gamma(\gamma_0) = \frac{1}{2} \int \left(\frac{\partial_\gamma S(x,\gamma_0)}{S(x,\gamma_0)} \right)^{\otimes 2} \pi(dx).$$

$$W_{\alpha,\gamma}(\theta_0) = \frac{1}{2}\nu(3) \int \frac{(\partial_\gamma S) \otimes (\partial_\alpha a)}{c^3}(x,\theta_0)\pi(dx),$$

$$W_\gamma(\gamma_0) = \frac{1}{4}\nu(4) \int \left(\frac{\partial_\gamma S(x,\gamma_0)}{S(x,\gamma_0)} \right)^{\otimes 2} \pi(dx) = \frac{\nu(4)}{2}\Gamma_\gamma(\gamma_0).$$

  - Multivariate case involves many multil-linear forms, notationally messy.

- **The Studentization**

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

automatically distinguish "diffusion or Lévy with jumps".

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 [Kessler, 1997] and [Uchida and Yoshida, 2012]:

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

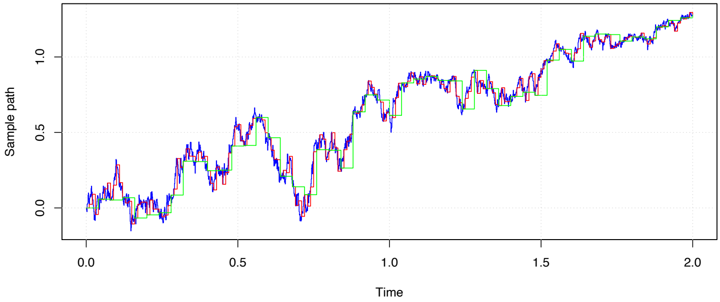
$$\sqrt{T_n}\hat{V}_n D_n^{-1} \xrightarrow{p} \Gamma(\theta_0)^{1/2},$$

resulting in

$$\left(\sqrt{T_n}(\hat{\alpha}_n - \alpha_0), \sqrt{n}(\hat{\gamma}_n - \gamma_0)\right) \xrightarrow{\mathcal{L}} N_p(0, \Gamma(\theta_0)^{-1}).$$

- Next section: Relative model comparison through AIC and BIC philosophies.

$$X_{t+dt} \leftarrow X_t + (\text{Trend})dt + (\text{Scale}) \cdot d(\text{Noise})$$



Summary: GQMLE for ergodic Lévy SDE

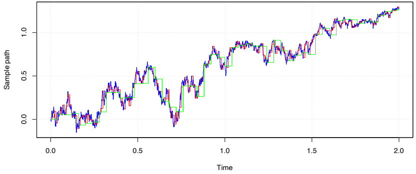
$$dX_t = C(X_{t-})dZ_t + A(X_t)dt$$

$$\boldsymbol{X}_n = (X_{t_j^n})_{j=0}^n$$

$$t_j^n = t_j = jh, \quad T_n := nh \to \infty, \quad nh^2 \to 0 \ (n \to \infty)$$

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

$X_{t+dt} \leftarrow X_t + (\mathrm{Trend})dt + (\mathrm{Scale}) \cdot d(\mathrm{Noise})$



Scale Estimation · Selection → **Trend (Drift)** Estimation · Selection → Noise Structure

$$\mathbb{H}_n(\theta) = \sum_{j=1}^n \log \phi_d \left(X_{t_j};\, X_{t_{j-1}} + a_{j-1}(\alpha)h,\, hS_{j-1}(\gamma)\right) = \mathbb{H}_{1,n}(\gamma) + \mathbb{H}_{2,n}(\theta)$$

$$\mathbb{H}_{1,n}(\gamma) = \sum_{j=1}^n \log \phi_d \left(X_{t_j};\, X_{t_{j-1}},\, hS_{j-1}(\gamma)\right),$$

$$\mathbb{H}_{2,n}(\theta) = \sum_{j=1}^n \left(S_{j-1}^{-1}(\gamma)\left[\Delta_j X,\, a_{j-1}(\alpha)\right] - \frac{h}{2} S_{j-1}^{-1}(\gamma)\left[a_{j-1}^{\otimes 2}(\alpha)\right]\right)$$

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

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

- **Correctly specified** statistical models $\mathcal{M}_{m_1,m_2}$ $(m_1 \le M_1;\ m_2 \le M_2)$:

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

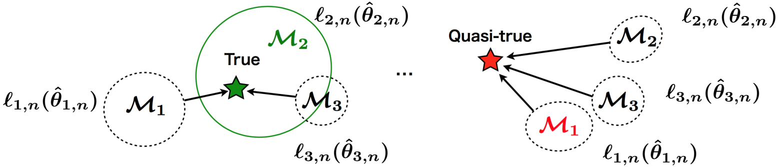
$$\text{e.g.} \quad dX_t = \exp\left(\frac{1}{2}\sum_{k=1}^{p_\gamma} \gamma_k S_k(X_{t-})\right) dZ_t + \sum_{l=1}^{p_\alpha} \alpha_l a_l(X_{t-})\, dt$$

- Akaike's AIC type and Schwarz's BIC type statistics for ergodic Lévy SDE?
- **Stepwise** selection strategy of the coefficients would be **essential**: scale $\to$ trend.

AIC and BIC in general: Twin jewels of information criteria (reprint)

- AIC and BIC are based on different philosophies, neither is absolutely better than the other:
  - AIC is designed for **predictive (generalization) performance**, while
  - BIC is for **model-description** performance.

How to relatively measure the discrepancy from the model to the (virtual) truth?



Avoid **over-fitting** $\Longrightarrow$ $\underbrace{-2\ell_n(\hat{\theta}_n)}_{\text{Goodness-of-fit}} + (\text{Complexity/Redundancy regularization})$

Preliminary general remarks on AIC paradigm

- True distribution $g(x)\mu(x)$ of a sample $\boldsymbol{X}_n$ $\leftarrow$ Statistical model $\{f(\cdot;\theta) : \theta \in \Theta\}$ wrt μ.
- Estimate g by $\hat{f}_n(\cdot) = f(\cdot;\hat{\theta}_n)$ for some $\hat{\theta}_n = \hat{\theta}_n(\boldsymbol{X}_n)$.
- Minimize the **Kullback-Leibler divergence**

$$\mathcal{D}(\hat{f}_n; g) = \left(\int g \log \left(\frac{g}{f_n} \right) g d\mu \right) \Bigg|_{f_n=\hat{f}_n}.$$

- 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

$$E\left[\mathcal{E}(\hat{f}_n; g) - \left(\mathcal{E}(\hat{f}_n; \delta_{\boldsymbol{X}_n}) + \hat{\mathfrak{b}}_n \right) \right] = o(1), \qquad n \to \infty. \tag{3.1}$$

  - Put simply: derivation of AIC amounts to managing this expectation.

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

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

and then construct an asymptotically unbiased estimator $\hat{\mathfrak{b}}_n = \hat{\mathfrak{b}}_n(\boldsymbol{X}_n)$, i.e. $E[\hat{\mathfrak{b}}_n - \mathfrak{b}_n] = o(1)$.

  - For models regular enough, we very often reach

$$\hat{\mathfrak{b}}_n = \mathrm{dim}(\text{Unknown parameters}).$$

The above strategy can be traced (formally makes sense)

- whatever the probabilistic structure of $\mathcal{L}(\boldsymbol{X}_n)$ is, and
- when the model $\{f(\cdot;\theta) : \theta \in \Theta\}$ is misspecified.

What's essential therein is that we can find a suitable corrector $\hat{\mathfrak{b}}_n$ satisfying the property (3.1).

Approximation of bias in joint case?

- Turning back to our setup, we keep considering the SDE model (index (m_1, m_2) omitted):

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

- If we try to evaluate the bias directly for the joint GQMLE $\mathbb{H}_n(\theta)$, as

$$\mathfrak{b}_n := E\left[\mathbb{H}_n(\hat{\theta}_n(\boldsymbol{X}_n); \boldsymbol{X}_n) - \tilde{E}\left[\tilde{\mathbb{H}}_n(\hat{\theta}_n(\boldsymbol{X}_n); \tilde{\boldsymbol{X}}_n)\right]\right] \qquad (\tilde{\boldsymbol{X}}_n \text{ is an indep. copy of } \boldsymbol{X}_n)$$
$$=: E\tilde{E}\left[\left(\mathbb{H}_n(\hat{\theta}_n) - \mathbb{H}_n(\theta_0)\right) - \left(\tilde{\mathbb{H}}_n(\hat{\theta}_n) - \tilde{\mathbb{H}}_n(\theta_0)\right)\right],$$

we need to look at the identity:

$$\mathbb{H}_n(\hat{\theta}_n) - \mathbb{H}_n(\theta_0) = \mathbb{H}_{1,n}(\hat{\gamma}_n) - \mathbb{H}_{1,n}(\gamma_0) + \mathbb{H}_{2,n}(\hat{\alpha}_n, \hat{\gamma}_n) - \mathbb{H}_{2,n}(\alpha_0, \gamma_0) = \frac{1}{h}\mathcal{Q}_{1,n}(\gamma_0) + \mathcal{Q}_{2,n}(\theta_0). \tag{3.2}$$

  - Importantly, both $\mathcal{Q}_{1,n}(\gamma_0)$ and $\mathcal{Q}_{2,n}(\theta_0)$ are the asymptotically non-trivial;
  - while $\mathbb{H}_n(\hat{\theta}_n) - \mathbb{H}_n(\theta_0) = O_p(1)$ in case of diffusions.

- Let $\hat{u}_n := (\hat{u}_{\alpha,n}, \hat{u}_{\gamma,n})$ where $\hat{u}_{\gamma,n} := \sqrt{T_n}(\hat{\gamma}_n - \gamma_0)$ and $\hat{u}_{\alpha,n} := \sqrt{T_n}(\hat{\alpha}_n - \alpha_0)$.
- Roughly, under suitable regularity conditions we may write

$$\mathbb{H}_n(\hat{\theta}_n) - \mathbb{H}_n(\theta_0) = \sum_{k\in\mathbb{N}} \frac{1}{k!}\left\{ \underbrace{\left(\frac{1}{n}\partial_\theta^k \mathbb{H}_n(\theta_0)\right)}_{\substack{=O_p(1)\ \ (\neq o_p(1)) \\ \text{for } \partial_\gamma^k\text{-components}}} [\hat{u}_n^{\otimes k}] \right\} n\, T_n^{-k/2}.$$

  - $n\, T_n^{-k/2} \to 0$ only for k large enough (will lead to messy and unpleasant terms!).
  - In case of diffusions, $\mathbb{H}_n(\hat{\theta}_n) - \mathbb{H}_n(\theta_0)$ itself is of the log-likelihood ratio type character.
- The **mixed-rates structure** necessitates the higher-order derivatives of $\mathbb{H}_n$ in the direct evaluation of $\mathfrak{b}_n$ based on the joint GQLF $\mathbb{H}_n(\theta)$, resulting in **rather complicated expressions**.

We can bypass this issue through the two-stage GQLF $\mathbb{H}_{1,n}(\gamma)$ and $\mathbb{H}_{2,n}(\alpha)$: in other words, we're going to **separately** look at the two terms $\mathcal{Q}_{1,n}(\gamma_0)$ and $\mathcal{Q}_{2,n}(\theta_0)$ in (3.2) **in this order**.

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

Stepwise bias corrections

1. **Scale**: $\mathcal{Q}_{1,n}(\gamma_0)$-part AIC-bias (w.r.t. γ) is

$$\mathfrak{b}_{\gamma,n} := hE\left[\mathbb{H}_{1,n}(\hat{\gamma}_n(\boldsymbol{X}_n);\boldsymbol{X}_n) - \tilde{E}\left[\tilde{\mathbb{H}}_{1,n}(\hat{\gamma}_n(\boldsymbol{X}_n);\tilde{\boldsymbol{X}}_n)\right]\right]$$
$$= \mathrm{trace}\left\{\Gamma_\gamma(\gamma_0)^{-1}W_\gamma(\gamma_0)\right\} + O(T_n^{-1/2}). \tag{3.3}$$

with both $\Gamma_\gamma(\gamma_0)$ and $W_\gamma(\gamma_0)$ being explicit: plugging-in their estimates,

$$\mathrm{GQAIC}_{1,n} := -2\,\mathbb{H}_{1,n}(\hat{\gamma}_n) + \frac{2}{h}\,\mathrm{trace}\left(\hat{\Gamma}_{\gamma,n}^{-1}\hat{W}_{\gamma,n}\right). \tag{3.4}$$

- The particular case of $d = r = 1$:

$$\mathrm{GQAIC}_{1,n} = -2\,\mathbb{H}_{1,n}(\hat{\gamma}_n) + \frac{1}{h}\hat{\nu}_n(4)p_\gamma, \qquad \hat{\nu}_n(4) := \frac{1}{T_n}\sum_{j=1}^{n}\left(\frac{\Delta_j X}{c_{j-1}(\hat{\gamma}_n)}\right)^4 \xrightarrow{p} \nu(4) := \int z^4\nu(dz).$$

A finite-sample correction is possible to make the formula consistent with the diffusion case.

- Having the estimate $\hat{\gamma}_n$ in hand, we proceed to the bias evaluation for the drift coefficient.

2. **Drift**: $\mathcal{Q}_{2,n}(\theta_0)$-part AIC-bias (w.r.t. α) turned out to be

$$\begin{aligned}\mathfrak{b}_{\alpha,n} &:= E\left[\mathbb{H}_{2,n}(\hat{\alpha}_n,\hat{\gamma}_n) - \tilde{E}\left[\tilde{\mathbb{H}}_{2,n}(\hat{\alpha}_n,\hat{\gamma}_n)\right]\right] \\ &= p_\alpha + \mathfrak{b}_{A,n} + o(1), \end{aligned} \tag{3.5}$$

with the term $\mathfrak{b}_{A,n}$ being common to all the drift-coefficient candidate, so that we may and do ignore $\mathfrak{b}_{A,n}$ in *relative* model comparison, for it is common to all the candidates:

$$\mathrm{GQAIC}_{2,n} := -2\,\mathbb{H}_{2,n}(\hat{\alpha}_n) + 2p_\alpha. \tag{3.6}$$

Theorem 3.1 (GQAIC formula for ergodic Lévy SDE: [Eguchi and Masuda, 2024])

Suppose additionally that $E\left[\mathrm{trace}\left(\hat{\Gamma}_{\gamma,n}^{-1}\hat{W}_{\gamma,n}\right)\right] \to \mathrm{trace}\left\{\Gamma_\gamma(\gamma_0)^{-1}W_\gamma(\gamma_0)\right\}$. *Then, the approximation* (3.3) *and* (3.5) *hold.*

- **Correctly specified** statistical models $\mathcal{M}_{m_1,m_2}$ $(m_1 \le M_1;\ m_2 \le M_2)$:

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

$$\mathbb{H}_{1,n}(\gamma) = \sum_{j=1}^{n} \log \phi_d\left(X_{t_j};\ X_{t_{j-1}},\ hS_{j-1}(\gamma)\right),$$

$$\mathbb{H}_{2,n}(\theta) = \sum_{j=1}^{n} \left(S_{j-1}^{-1}(\gamma)\left[\Delta_j X,\ a_{j-1}(\alpha)\right] - \frac{h}{2} S_{j-1}^{-1}(\gamma)\left[a_{j-1}^{\otimes 2}(\alpha)\right] \right)$$

Our proposal for how to select the scale coefficient through AIC

1. FIRST select the scale by (3.4): $\mathrm{GQAIC}_{1,n} := -2\,\mathbb{H}_{1,n}(\hat{\gamma}_n) + \dfrac{2}{h}\,\mathrm{trace}\left(\hat{\Gamma}_{\gamma,n}^{-1}\hat{W}_{\gamma,n}\right)$
2. THEN select the drift by (3.6): $\mathrm{GQAIC}_{2,n} := -2\,\mathbb{H}_{2,n}(\hat{\alpha}_n) + 2p_\alpha.$
   - For the GQAIC, **the prob. of "underestimating model" goes to** 0.

Remark 3.2 (GQAIC for diffusion)

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

- GQAIC become $-2\mathbb{H}_{1,n}(\hat{\gamma}_n) + 2p_\gamma$ for scale, and $-2\mathbb{H}_{2,n}(\hat{\alpha}_n) + 2p_\alpha$ for drift.
- Moreover, in case of the joint estimation through the GQLF $\mathbb{H}_n(\theta)$ of (2.1), the GQAIC is

$$-2\mathbb{H}_n(\hat{\theta}_n) + 2(p_\alpha + p_\gamma),$$

the same form as in the CIC of [Uchida, 2010], the contrast information criterion.

- The source of this essential difference is that, for diffusions, the statistical random fields is *jointly* LAQ.

Preliminary remarks on BIC paradigm

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

- BIC methodology is to select the model that approximately minimizes the minus logarithmic **model evidence** defined to be, with $\pi(\theta)$ denoting the joint prior density of (α, γ),

$$-\log\left(\int_\Theta \exp\{\mathbb{H}_n(\theta)\}\pi(\theta)d\theta\right)$$

  for the joint GQLF. As in the GQAIC case, we formulate a stepwise procedure.

- We keep using the two-stage GQLFs $\mathbb{H}_{1,n}(\gamma)$ and $\mathbb{H}_{2,n}(\alpha) = \mathbb{H}_{2,n}(\alpha, \hat{\gamma}_n)$.
- 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 $\pi_2(\alpha_0) > 0$.
- We additionally assume that there exists a constant $c_1 \in (0, 1)$ for which $T_n \gtrsim n^{c_1}$.

Stepwise stochastic expansions

1. **Scale**: **Free energy at the inverse temperature** $\mathfrak{b} > 0$, defined through the negative normalized logarithmic partition function:

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

   - 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 $\mathfrak{F}_{1,n}(1)$.
   - 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.

- $\mathfrak{F}_{1,n}(1)$ quantifies the minus model evidence for the $\mathcal{Q}_{1,n}(\gamma_0)$-part: a smaller $\mathfrak{F}_{1,n}(1)$ is **better in model description**.

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

Theorem 3.3 (Scale-GQBIC formula for ergodic Lévy SDE: [Eguchi and Masuda, 2024])

$$\mathfrak{F}_{1,n}(1) = -\frac{1}{n}\mathbb{H}_{1,n}(\hat{\gamma}_n) + \frac{p_\gamma}{2n}\log n + O_p\left(\frac{1}{n}\right), \tag{3.7}$$

$$\mathfrak{F}_{1,n}(h) = -\frac{1}{n}\mathbb{H}_{1,n}(\hat{\gamma}_n) + \frac{p_\gamma}{2T_n}\log T_n + O_p\left(\frac{1}{T_n}\right). \tag{3.8}$$

- The first one (3.7) was previously given in [Eguchi and Uehara, 2021], based on which the GQBIC for the scale was defined to be $-2\mathbb{H}_{1,n}(\hat{\gamma}_n) + p_\gamma \log n$.
- For the **model-selection consistency** to be in force, we propose to adopt (3.8):

$$\mathrm{GQBIC}_{1,n} = -2\,\mathbb{H}_{1,n}(\hat{\gamma}_n) + \frac{p_\gamma}{h}\log T_n. \tag{3.9}$$

2. **Drift**: We can directly look at the normalized marginal quasi-log likelihood

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

Theorem 3.4 (Drift-GQBIC formula for ergodic Lévy SDE: [Eguchi and Masuda, 2024])

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

- Ignoring $O_p(T_n^{-1})$ of $\mathfrak{F}_{2,n}$ (as in [Eguchi and Masuda, 2018]), we introduce the second GQBIC:

$$\mathrm{GQBIC}_{2,n} = -2\,\mathbb{H}_{2,n}(\hat{\alpha}_n) + p_\alpha \log T_n. \tag{3.10}$$

- **Correctly specified** statistical models $\mathcal{M}_{m_1,m_2}$ $(m_1 \le M_1;\ m_2 \le M_2)$:

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

$$\mathbb{H}_{1,n}(\gamma) = \sum_{j=1}^{n} \log \phi_d\left(X_{t_j};\ X_{t_{j-1}},\ hS_{j-1}(\gamma)\right),$$

$$\mathbb{H}_{2,n}(\theta) = \sum_{j=1}^{n} \left(S_{j-1}^{-1}(\gamma)\left[\Delta_j X,\ a_{j-1}(\alpha)\right] - \frac{h}{2} S_{j-1}^{-1}(\gamma)\left[a_{j-1}^{\otimes 2}(\alpha)\right] \right)$$

Our proposal for how to select the scale coefficient through BIC

1. FIRST select the scale by (3.9): $\mathrm{GQBIC}_{1,n} = -2\,\mathbb{H}_{1,n}(\hat{\gamma}_n) + \frac{p_\gamma}{h}\log T_n.$
2. THEN select the drift by (3.10): $\mathrm{GQBIC}_{2,n} = -2\,\mathbb{H}_{2,n}(\hat{\alpha}_n) + p_\alpha \log T_n.$

   - The **selection consistency** follows as in [Eguchi and Masuda, 2018] or [Eguchi and Masuda, 2019].

Simulation design

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

- Three different noises:
  - (i) Normal inverse Gaussian noise $\mathcal{L}(Z_t) = NIG(10, 0, 10t, 0)$,
  - (ii) Normal bilateral gamma noise $\mathcal{L}(Z_t) = bGamma(t, \sqrt{2}, t, \sqrt{2})$,
  - (iii) Skewed NIG noise $\mathcal{L}(Z_t) = NIG\left(\frac{25}{3}, \frac{20}{3}, \frac{9}{5}t, -\frac{12}{5}t\right)$.
- We use YUIMA R package [Brouste et al., 2014] for generating data.
- Monte Carlo trials are based on 1000 independent sample paths, done for

$$(h_n, T_n) = (0.01, 10), (0.005, 10), (0.01, 50), \text{ and } (0.005, 50),$$

hence in each case, $n = 1000, 2000, 5000,$ and 10000.

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

- Candidate scale (Scale) and drift (Drift) coefficients:
  - Scale 1 : $c_1(x, \gamma_1) = \gamma_1$
  - Scale 2 : $c_2(x, \gamma_2) = \frac{\gamma_2}{1+x^2}$
  - Scale 3 : $c_3(x, \gamma_3) = \frac{\gamma_{3,1} + \gamma_{3,2}x^2}{1+x^2}$
  - Scale 4 : $c_4(x, \gamma_4) = \frac{\gamma_{4,1} + \gamma_{4,2}x + \gamma_{4,3}x^2}{1+x^2}$
  - Drift 1 : $a_1(x, \alpha_1) = -\alpha_1$
  - Drift 2 : $a_2(x, \alpha_2) = -\alpha_2 x$
  - Drift 3 : $a_3(x, \alpha_3) = -\alpha_{3,1}x - \alpha_{3,2}$
- Then, the optimal (minimal true) model consists of Scale 2 and Drift 2 with $(\gamma_2, \alpha_2) = (3, \frac{1}{2})$.

- The proposed information criteria (with non-essential modification for $\mathrm{GQAIC}_{1,n}$):

$$\mathrm{CIC}_{1,n} = -2\mathbb{H}_{1,n}(\hat{\gamma}_n) + 2p_\gamma, \qquad \mathrm{CIC}_{2,n} = -2\mathbb{H}_{2,n}(\hat{\alpha}_n) + 2p_\alpha,$$

(Calculated just for comparison)

$$\mathrm{GQAIC}_{1,n} = -2\mathbb{H}_{1,n}(\hat{\gamma}_n) + p_\gamma \left\{ \frac{1}{hT_n} \sum_{j=1}^{n} \left(\frac{\Delta_j X}{c_{j-1}(\hat{\gamma}_n)} \right)^4 - \left(\frac{1}{T_n} \sum_{j=1}^{n} \left(\frac{\Delta_j X}{c_{j-1}(\hat{\gamma}_n)} \right)^2 \right)^2 \right\},$$

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

$$\mathrm{GQBIC}^\sharp_{1,n} = -2\mathbb{H}_{1,n}(\hat{\gamma}_n) + p_\gamma \log n, \qquad \mathrm{QBIC}_{1,n} = -2\mathbb{H}_{1,n}(\hat{\gamma}_n) + \frac{p_\gamma}{h_n} \log T_n,$$

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

Model-selection frequencies in (i) $\mathcal{L}(Z_t) = NIG(10, 0, 10t, 0)$

CIC	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	**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

QBIC	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
		($n = 1000$)	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

Model-selection frequencies in (ii) $\mathcal{L}(Z_t) = bGamma(t, \sqrt{2}, t, \sqrt{2})$

CIC	T_n	h_n		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

QBIC	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
	($n = 1000$)		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

Model-selection frequencies in (iii) $\mathcal{L}(Z_t) = NIG\left(\frac{25}{3}, \frac{20}{3}, \frac{9}{5}t, -\frac{12}{5}t\right)$ Back

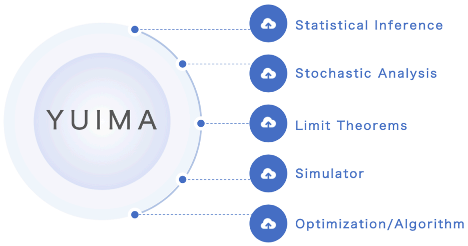
CIC	T_n	h_n		Scale 1	Scale 2*	Scale 3	Scale 4
	10	0.01	Drift 1	0	0	0	0
	($n = 1000$)		Drift 2*	0	**128**	52	600
			Drift 3	0	30	30	160
	10	0.005	Drift 1	0	0	0	1
	($n = 2000$)		Drift 2*	0	**88**	37	653
			Drift 3	0	25	16	180
	50	0.01	Drift 1	0	0	0	0
	($n = 5000$)		Drift 2*	0	**88**	75	731
			Drift 3	0	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

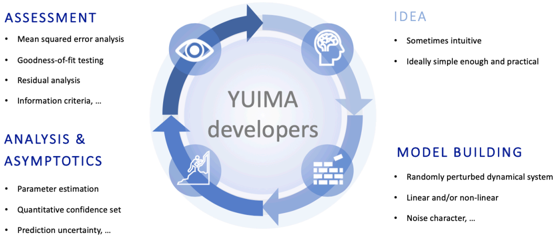
GQAIC	T_n	h_n		Scale 1	Scale 2*	Scale 3	Scale 4
	10	0.01	Drift 1	0	0	0	0
	($n = 1000$)		Drift 2*	18	**589**	33	155
			Drift 3	32	125	33	15
	10	0.005	Drift 1	0	0	0	1
	($n = 2000$)		Drift 2*	15	**602**	30	149
			Drift 3	32	122	32	17
	50	0.01	Drift 1	0	0	0	0
	($n = 5000$)		Drift 2*	0	**672**	102	122
			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

QBIC	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
	($n = 1000$)		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

YUIMA package





- Documents: [Brouste et al., 2014], [Iacus and Yoshida, 2018], ...
- Demonstrated also by e.g. *Quant Education*: https://www.youtube.com/watch?v=trWzgLj20XU
- https://r-forge.r-project.org/projects/yuima/

Summary: Noise character does significantly matter

- Inference based on the two-stage Gaussian quasi-likelihood:

$$\text{True data-generating process} \quad \leftarrow \quad \text{Statistical model } \mathcal{M}_{m_1,m_2} \ (m_1 \le M_1,\ m_2 \le M_2)$$

$$dX_t = C(X_{t-})dZ_t + A(X_t)dt \quad \leftarrow \quad dX_t = a_{m_2}(X_t, \alpha_{m_2})dt + c_{m_1}(X_{t-}, \gamma_{m_1})dZ_t$$

- Estimation of $\theta_{m_1,m_2} = (\alpha_{m_2}, \gamma_{m_1})$ by GQMLE:

$$E\left[f\left(\sqrt{T_n}(\hat{\theta}_n - \theta_0)\right)\right] \to \int f(u)\phi(u; 0, V(\theta_0))du, \quad \hat{V}_n^{-1/2}\sqrt{T_n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} N_p(0, I_p).$$

Select $\mathcal{M}_{\hat{m}_{1,n},\hat{m}_{2,n}}$ for GQX**IC type relative model selection (X←A,B)**

$$\{\hat{m}_{1,n}\} = \operatorname*{argmin}_{m_1} \mathrm{GQXIC}_{1,n}^{(m_1)}, \quad \{\hat{m}_{2,n}\} = \operatorname*{argmin}_{m_2} \mathrm{GQXIC}_{2,n}^{(m_2|\hat{m}_{1,n})}:$$

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

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

References I

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

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

Eguchi, S. and Masuda, H. (2019).
Data driven time scale in Gaussian quasi-likelihood inference.
Stat. Inference Stoch. Process., 22(3):383–430.

Eguchi, S. and Masuda, H. (2024).
Gaussian quasi-information criteria for ergodic Lévy driven SDE.
Ann. Inst. Statist. Math., 76(1):111–157.

References II

Eguchi, S. and Uehara, Y. (2021).
Schwartz-type model selection for ergodic stochastic differential equation models.
Scand. J. Stat., 48(3):950–968.

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.

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

Kessler, M. (1997).
Estimation of an ergodic diffusion from discrete observations.
Scand. J. Statist., 24(2):211–229.

References III

Kulik, A. (2018).
Ergodic behavior of Markov processes, volume 67 of *De Gruyter Studies in Mathematics*.
De Gruyter, Berlin.
With applications to limit theorems.

Kulik, A. M. (2009).
Exponential ergodicity of the solutions to SDE's with a jump noise.
Stochastic Process. Appl., 119(2):602–632.

Masuda, H. (2007).
Ergodicity and exponential β-mixing bounds for multidimensional diffusions with jumps.
Stochastic Process. Appl., 117(1):35–56.

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

References IV

Uchida, M. (2010).
Contrast-based information criterion for ergodic diffusion processes from discrete observations.
Ann. Inst. Statist. Math., 62(1):161–187.

Uchida, M. and Yoshida, N. (2012).
Adaptive estimation of an ergodic diffusion process based on sampled data.
Stochastic Process. Appl., 122(8):2885–2924.

Uehara, Y. (2019).
Statistical inference for misspecified ergodic lévy driven stochastic differential equation models.
Stochastic Process. Appl., 129(10):4051 – 4081.

Preliminary observations I

- 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

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

  - The candidate coefficients $c_1,\dots,c_{M_1}$ and $a_1,\dots,a_{M_2}$ contain both correctly specified coefficients and misspecified coefficients.
  - See [Uehara, 2019] for asymptotics for misspecified-coefficient case.

Preliminary observations II

- 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:
$$\{\hat{m}_{1,n}\} = \operatorname*{argmin}_{1\le m_1\le M_1} \mathrm{GQAIC}_{1,n}^{(m_1)}.$$
  (ii) Under the result of (i), we choose the best drift coefficient with index $\hat{m}_{2,n}$ such that
$$\{\hat{m}_{2,n}\} = \operatorname*{argmin}_{1\le m_2\le M_2} \mathrm{GQAIC}_{2,n}^{(m_2|\hat{m}_{1,n})},$$
  where $\mathrm{GQAIC}_{2,n}^{(m_2|m_{1,n})}$ corresponds to (3.6) 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.

Preliminary observations III

- 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

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

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 respectively defined as

$$\{\gamma_{m_1}^*\} = \mathrm{argmax}_{\gamma_{m_1}}\mathbb{H}_{1,0}^{(m_1)}(\gamma_{m_1}), \qquad \mathfrak{M}_1^* = \underset{m_1\in\mathfrak{M}_1}{\mathrm{argmin}}\,\dim(\Theta_{\gamma_{m_1}}).$$

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

Preliminary observations IV

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

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

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

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

- 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

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

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

Preliminary observations V

- 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 $\Theta_{\alpha_{j_2}}$.

- 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 (3.10) associated with $c_{m_{1,n}}$ and $\hat{\gamma}_{m_{1,n},n}$.

GQAIC is asymptotically weighed chi-square distributed

- Somewhat complicated description: [Eguchi and Masuda, 2024, Theorem 5]
- Summary
  - 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, the probability that GQAIC chooses the misspecified coefficients tends to 0 as $n \to \infty$.

GQBIC has selection consistency I

Theorem 6.1

Suppose that the assumptions of Theorem 3.1 hold for all candidate coefficients which are included in $\mathfrak{M}_1$. We also assume that index m_1^* satisfies $m_1^* \in \mathfrak{M}_1^*$.

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

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

2. *If $m_1 \in \{1, \ldots, M_1\} \backslash \mathfrak{M}_1$, then*

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

GQBIC has selection consistency II

Theorem 6.2

Suppose that the assumptions of Theorem 3.1 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. *If $\Theta_{\gamma_{m_1^*}}$ is nested in $\Theta_{\gamma_{m_1}}$ for $m_1 \in \mathfrak{M}_1 \backslash \{m_1^*\}$, or if $m_1 \in \{1, \dots, M_1\} \backslash \mathfrak{M}_1$, then*

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

2. *If $\Theta_{\alpha_{m_2^*}}$ is nested in $\Theta_{\alpha_{m_2}}$ for $m_2 \in \mathfrak{M}_2 \backslash \{m_2^*\}$, or if $m_2 \in \{1, \dots, M_2\} \backslash \mathfrak{M}_2$, then*

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

- Theorem 6.1.1 shows that, when comparing correctly specified models, the probability that $\mathrm{GQBIC}_{1,n}^{\sharp}$ selects a larger model tends to 1. Moreover, Theorem 6.2 means that the GQBIC proposed by (3.9) and (3.10) has the model selection consistency.