AIC for ergodic diffusion processes from discrete observations

Masayuki Uchida (Kyushu University)
Nakahiro Yoshida (University of Tokyo)

Outline

1. Introduction: concepts of information criteria

2. Akaike's information criteria (AIC) for ergodic diffusions

3. Simulation studies on approximate log likelihoods

4. Example of model selection based on AIC

1. Introduction : Concepts of information criteria

$\mathbf{X}_n \sim g(x_n)$: true density

$\mathbf{X}_n = x_n$: observations

$\{f(x_n, \theta); \theta \in \Theta\}$: statistical model

$g(x_n) = f(x_n, \theta_0) \Longrightarrow \{f(x_n, \theta); \theta \in \Theta\}$: specified parametric model

Problem

Suppose that there are two statistical models, $\{f_1(x_n, \theta_1); \theta_1 \in \Theta_1\}$ and $\{f_2(x_n, \theta_2); \theta_2 \in \Theta_2\}$.

Which model should we select, f_1 or f_2?

We want to obtain a criterion for selecting the best statistical model among a set of competing models.

Note that "best" means the best of the competing models.

It does not mean the best of all statistical models.

Akaike's information criterion (AIC)

<u>Step 1</u>

$\{f(x_n, \theta); \theta \in \Theta\}$: statistical model

$\theta \longleftarrow \widehat{\theta}(\mathbf{X}_n)$: an estimator obtained from the observations $\mathbf{X}_n$

<u>Step 2</u>

$Z_n \sim g(\cdot)$: a future observation

$g(z_n) \longleftarrow f(z_n, \widehat{\theta}(\mathbf{X}_n))$: predictive distribution (statistical model)

Step 3

We would like to assess the closeness of the predictive distribution $f(z_n, \widehat{\theta}(\mathbf{X}_n))$ to the true density $g(z_n)$.

As a measure of the divergence of $f(z_n, \widehat{\theta}(\mathbf{X}_n))$ from $g(z_n)$, we use the estimated Kullback-Leibler information

$$I\{g(z_n); f(z_n, \widehat{\theta}(\mathbf{X}_n))\} \;:=\; E_{\mathbf{Z}}\left[\log \frac{g(\mathbf{Z}_n)}{f(\mathbf{Z}_n, \widehat{\theta}(\mathbf{X}_n))}\right].$$

The rule of model selection

Choose a statistical model $\widehat{f}$ among competing models $\{f_1, \ldots, f_m\}$ such that

$$I\{g; \widehat{f}\} = \min_i I\{g(z_n); f_i(z_n, \widehat{\theta}_i(\mathbf{X}_n))\}.$$

The estimated Kullback-Leibler information

$$\begin{aligned} & I\{g(z_n); f(z_n, \widehat{\theta}(\mathbf{X}_n))\} \\ = & \int g(z_n) \log g(z_n) dz_n - \int g(z_n) \log f(z_n, \widehat{\theta}(\mathbf{X}_n)) dz_n. \end{aligned} \tag{1}$$

The first term in the right hand side of (1) does not depend on the statistical model $f(z_n, \widehat{\theta}(\mathbf{X}_n))$

The second term $\eta(\widehat{\theta}(\mathbf{X}_n)) := \int g(z_n) \log f(z_n, \widehat{\theta}(\mathbf{X}_n)) dz_n$ depends on it,

$\eta(\widehat{\theta}(\mathbf{X}_n))$: the expected log likelihood

Minimizing the estimated KLI $I\{g(z_n); f(z_n, \widehat{\theta}(\mathbf{X}_n))\}$ is equivalent to maximizing the expected log likelihood $\eta(\widehat{\theta}(\mathbf{X}_n))$.

Therefore, the rule of model selection is to choose a model which is maximizing the expected log likelihood $\eta(\widehat{\theta}(\mathbf{X}_n))$ among competing models.

However, since the expected log likelihood

$$\eta(\widehat{\theta}(\mathbf{X}_n)) = \int g(z_n) \log f(z_n, \widehat{\theta}(\mathbf{X}_n)) dz_n$$

depends on the true density $g(z_n) = f(z_n, \theta_0)$, we need to estimate it.

Step 4

$\eta(\widehat{\theta}(\mathbf{X}_n)) \longleftarrow \log f(\mathbf{X}_n, \widehat{\theta}(\mathbf{X}_n))$: simple estimator

The bias of $\log f(\mathbf{X}_n, \widehat{\theta}(\mathbf{X}_n))$ in the estimation of $\eta(\widehat{\theta}(\mathbf{X}_n))$ is given by

$$\text{bias} = E_{\mathbf{X}}\left[\log f(\mathbf{X}_n, \widehat{\theta}(\mathbf{X}_n)) - \int g(z_n) \log f(z_n, \widehat{\theta}(\mathbf{X}_n)) dz_n\right].$$

If $\widehat{\theta}$ is the MLE ($\widehat{\theta}^{(ML)}$), under some regularity conditions, as $n \to \infty$,

$$\text{bias} = \text{dim}(\Theta) + o(1),$$

where dim(Θ) denotes the dimension of a parameter space Θ.

The bias corrected log likelihood is given by

$$\log f(\mathbf{X}_n, \widehat{\theta}^{(ML)}(\mathbf{X}_n)) - \text{dim}(\Theta).$$

Thus, Akaike (1973, 1974) proposed

$$\text{AIC}(\mathbf{X}_n, \widehat{\theta}^{(ML)}(\mathbf{X}_n)) = -2 \log f(\mathbf{X}_n, \widehat{\theta}^{(ML)}(\mathbf{X}_n)) + 2\text{dim}(\Theta). \tag{2}$$

Note that we choose a statistical model for which the value of AIC is minimizing among a set of competing models.

2. Notation and assumptions

Consider a family of one-dimensional diffusion processes defined by the stochastic differential equations

$$\begin{aligned} dX_t &= b(X_t, \alpha)dt + \sigma(X_t, \beta)dw_t, \ t \in [0, T], \\ X_0 &= x_0, \end{aligned} \tag{3}$$

where

$\theta = (\alpha, \beta) \in \Theta_\alpha \times \Theta_\beta = \Theta$.

Θ_α, Θ_β : compact convex subsets of $\mathbf{R}^p$ and $\mathbf{R}^q$.

b : $\mathbf{R} \times \Theta_\alpha \to \mathbf{R}$.

σ : $\mathbf{R} \times \Theta_\beta \to \mathbf{R}$.

w : one-dimensional standard Wiener process.

We assume that the drift b and the diffusion coefficient σ are known apart from the parameters α and β.

The data : $\mathbf{X}_n = (X_{t_k^n})_{0 \le k \le n}$ with $t_k^n = kh_n$, where
h_n : the discretization step.
The asymptotics : $h_n \to 0$, $nh_n \to \infty$ and $nh_n^2 \to 0$ as $n \to \infty$.

We introduce the notation used in this paper.
1. α_0, β_0, θ_0 denote the true values of α, β and θ, respectively.
2. For a function $f(x, \theta)$, define that
$\delta_{\theta_i} f(x,\theta) = \frac{\partial}{\partial \theta_i} f(x,\theta)$, $f'(x,\theta) = \frac{\partial}{\partial x} f(x,\theta)$, $\delta_\theta f(x,\theta) = \Big(\delta_{\theta_i} f(x,\theta)\Big)_{i=1,\dots,p}$
and $\delta_\theta^2 f(x,\theta) = \left(\frac{\partial^2}{\partial \theta_i \partial \theta_j} f(x,\theta)\right)_{i,j=1,\dots,p+q}$.
3. E denotes the state space of X, $E \subseteq \mathbf{R}$.
4. When the distribution of X_t given $X_0 = x$ has a strictly positive density with respect to the Lebesgue measure on the state space E, we denote it by $y \longmapsto p(t, x, y, \theta)$, $y \in E$.

Moreover, we define the following functions.

$$
\begin{aligned}
s(x,\beta) &= \int_0^x \frac{du}{\sigma(u,\beta)}, \\
B(x,\theta) &= \frac{b(x,\alpha)}{\sigma(x,\beta)} - \frac{1}{2}\sigma'(x,\beta), \\
\tilde{B}(x,\theta) &= B(s^{-1}(x,\beta),\theta), \\
\tilde{h}(x,\theta) &= \tilde{B}^2(x,\theta) + \tilde{B}'(x,\theta).
\end{aligned}
$$

We make three sets of assumptions as follows.

Assumption 1 *(i) Equation (3) has a unique strong solution on* $[0, T]$*.*
(ii) $\inf_{x,\beta} \sigma^2(x, \beta) > 0$.
(iii) The process X *is ergodic for every* θ *with invariant probability measure* μ_θ*. All polynomial moments of* μ_θ *are finite.*
(iv) For all $m \geq 0$ *and for all* θ*,* $\sup_t E_\theta[|X_t|^m] < \infty$*.*
(v) For every θ*, the functions* $b(x, \alpha)$ *and* $\sigma(x, \beta)$ *are twice continuously differentiable with respect to* x *and the derivatives are of polynomial growth in* x*, uniformly in* θ*.*
(vi) The functions $b(x, \alpha)$ *and* $\sigma(x, \beta)$ *and all their partial* x*-derivatives up to order 2 are three times differentiable with respect to* θ *for all* x *in the state space. All these derivatives with respect to* θ *are of polynomial growth in* x*, uniformly in* θ*.*

Assumption 2 *(i)* $\tilde{h}(x,\theta) = O(|x|^2)$ *as* $x \to \infty$.
(ii) $\inf_x \tilde{h}(x,\theta) > -\infty$ *for all* θ.
(iii) $\sup_\theta \sup_x |\tilde{h}^3(x,\theta)| \leq M < \infty$.
(iv) There exists $\gamma > 0$ *such that for every* θ *and* $j = 1, 2$, $|\tilde{B}^j(x,\theta)| = O(|\tilde{B}|^\gamma(x,\theta))$ *as* $|x| \to \infty$.

Assumption 3

$$b(x,\alpha) = b(x,\alpha_0) \quad \text{for } \mu_{\theta_0} \text{ a.s. all } x \;\Rightarrow\; \alpha = \alpha_0,$$
$$\sigma(x,\beta) = \sigma(x,\beta_0) \quad \text{for } \mu_{\theta_0} \text{ a.s. all } x \;\Rightarrow\; \beta = \beta_0.$$

3. Information criterion

The log likelihood function of $\mathbf{X}_n$ is

$$l_n(\mathbf{X}_n, \theta) = \sum_{k=1}^{n} l(h_n, X_{t_{k-1}^n}, X_{t_k^n}, \theta),$$

where $l(t, x, y, \theta) = \log p(t, x, y, \theta)$.
Define the maximum likelihood estimator

$$\widehat{\theta}_n^{(ML)} = \arg\sup_{\theta} l_n(\mathbf{X}_n, \theta).$$

Then, Akaike's information criterion is as follows:

$$AIC = -2 l_n(\mathbf{X}_n, \widehat{\theta}_n^{(ML)}) + 2 dim(\Theta).$$

However, since the transition density p of the diffusion process X does not generally have an explicit form, we cannot directly obtain

$$\underline{\text{the log likelihood function } l_n}$$

and

$$\underline{\text{the maximum likelihood estimator } \widehat{\theta}_n^{(ML)}}.$$

That is why we need to obtain both

$$\underline{\text{an approximation of the log-likelihood function } l_n}$$

and

$$\underline{\text{an asymptotically efficient estimator } \widehat{\theta}_n}$$

in order to construct AIC type of information criteria for diffusion processes.

an approximation of l_n (Dacunha-Castelle and Florens-Zmirou (1986))

$$
\begin{aligned}
u_n(\mathbf{X}_n, \theta) &= \sum_{k=1}^{n} u(h_n, X_{t_{k-1}^n}, X_{t_k^n}, \theta), \\
u(t, x, y, \theta) &= -\frac{1}{2}\log(2\pi t) - \log\sigma(y, \beta) \\
&\quad -\frac{[S(x, y, \beta)]^2}{2t} + H(x, y, \theta) + t\tilde{g}(x, y, \theta), \qquad (4)
\end{aligned}
$$

where

$$
\begin{aligned}
S(x, y, \beta) &= \int_x^y \frac{du}{\sigma(u, \beta)}, \\
H(x, y, \theta) &= \int_x^y \left\{ \frac{b(u, \alpha)}{\sigma^2(u, \beta)} - \frac{1}{2}\frac{\sigma'(u, \beta)}{\sigma(u, \beta)} \right\} du, \\
\tilde{g}(x, y, \theta) &= -\frac{1}{2}\left\{ C(x, \theta) + C(y, \theta) + \frac{1}{3}B(x, \theta)B(y, \theta) \right\}, \\
B(x, \theta) &= \frac{b(x, \alpha)}{\sigma(x, \beta)} - \frac{1}{2}\sigma'(x, \beta), \\
C(x, \theta) &= \frac{1}{3}[B(x, \theta)]^2 + \frac{1}{2}[B(x, \theta)]'\sigma(x, \beta).
\end{aligned}
$$

an asymptotically efficient estimator

We use the contrast function based on locally Gaussian approximation as follows:

$$g_n(\mathbf{X}_n, \theta) = \sum_{k=1}^{n} g(h_n, X_{t_{k-1}^n}, X_{t_k^n}, \theta),$$

where

$$g(t, x, y, \theta) = -\frac{1}{2}\log(2\pi t) - \log\sigma(x, \beta) - \frac{[y - x - tb(x, \alpha)]^2}{2t\sigma^2(x, \beta)}. \tag{5}$$

We then define the maximum contrast estimator as

$$\hat{\theta}_n^{(C)} = \arg\sup_{\theta} g_n(\mathbf{X}_n, \theta).$$

For a process $\mathbf{Z}_n$ which is independent of (but has the same distribution as) the observed process $\mathbf{X}_n$,

$$
\begin{aligned}
& u_n(\mathbf{X}_n, \widehat{\theta}_n^{(C)}(\mathbf{X}_n)) - \int_\Omega l_n(\mathbf{Z}_n(\tilde{\omega}), \widehat{\theta}_n^{(C)}(\mathbf{X}_n)) dP(\tilde{\omega}) \\
= \; & u_n(\mathbf{X}_n, \widehat{\theta}_n^{(C)}(\mathbf{X}_n)) - u_n(\mathbf{X}_n, \theta_0) && (6) \\
+ \; & u_n(\mathbf{X}_n, \theta_0) - \int_\Omega l_n(\mathbf{Z}_n(\tilde{\omega}), \theta_0) dP(\tilde{\omega}) && (7) \\
+ \; & \int_\Omega l_n(\mathbf{Z}_n(\tilde{\omega}), \theta_0) dP(\tilde{\omega}) - \int_\Omega l_n(\mathbf{Z}_n(\tilde{\omega}), \widehat{\theta}_n^{(C)}(\mathbf{X}_n)) dP(\tilde{\omega}). && (8)
\end{aligned}
$$

Under the regularity conditions, one has

$$
\begin{aligned}
(6) \;\; = \;\; & [D^{1/2}\delta_\theta u_n(\mathbf{X}_n,\theta_0)]^T D^{-1/2}(\widehat{\theta}_n^{(C)}(\mathbf{X}_n)-\theta_0) \\
& +\frac{1}{2}[D^{-1/2}(\widehat{\theta}_n^{(C)}(\mathbf{X}_n)-\theta_0)]^T[D^{1/2}\delta_\theta^2 u_n(\mathbf{X}_n,\theta_0)D^{1/2}] \\
& \times D^{-1/2}(\widehat{\theta}_n^{(C)}(\mathbf{X}_n)-\theta_0)+o_p(1), \\
(8) \;\; = \;\; & -\left[\int_\Omega D^{1/2}\delta_\theta l_n(\mathbf{Z}_n(\tilde{\omega}),\theta_0)dP(\tilde{\omega})\right]^T D^{-1/2}(\widehat{\theta}_n^{(C)}(\mathbf{X}_n)-\theta_0) \\
& -\frac{1}{2}[D^{-1/2}(\widehat{\theta}_n^{(C)}(\mathbf{X}_n)-\theta_0)]^T\int_\Omega D^{1/2}\delta_\theta^2 l_n(\mathbf{Z}_n(\tilde{\omega}),\theta_0)D^{1/2}dP(\tilde{\omega}) \\
& \times D^{-1/2}(\widehat{\theta}_n^{(C)}(\mathbf{X}_n)-\theta_0)+o_p(1),
\end{aligned}
$$

where A^T is the transpose of A for a vector A, D is the following $(p+q)\times(p+q)$ matrix

$$
D=\begin{pmatrix} \frac{1}{nh_n}I_p & 0 \\ 0 & \frac{1}{n}I_q \end{pmatrix},
$$

and I_p is the $p\times p$ identity matrix.

Let $I(\theta_0)$ denote the Fisher information matrix as follows:

$$I(\theta_0) = \begin{pmatrix} (I_b^{ij}(\theta_0))_{i,j=1,\dots,p} & 0 \\ 0 & (I_\sigma^{ij}(\theta_0))_{i,j=1,\dots,q} \end{pmatrix},$$

where

$$I_b^{ij}(\theta_0) = \int_{\mathbf{R}} \frac{\delta_{\alpha_i} b(x,\alpha_0)\delta_{\alpha_j} b(x,\alpha_0)}{\sigma^2(x,\beta_0)} \mu_{\theta_0}(dx),$$

$$I_\sigma^{ij}(\theta_0) = 2\int_{\mathbf{R}} \frac{\delta_{\beta_i} \sigma(x,\beta_0)\delta_{\beta_j} \sigma(x,\beta_0)}{\sigma^2(x,\beta_0)} \mu_{\theta_0}(dx).$$

In order to obtain our main result, we need the following four lemmas.

Lemma 1 *Suppose that Assumptions 1 and 2 hold true. Then, as* $nh_n^2 \to 0$,

$$E_{\theta_0}[u_n(\mathbf{X}_n, \theta_0) - l_n(\mathbf{X}_n, \theta_0)] = o(1).$$

Lemma 2 *Suppose that Assumptions 1 and 2 hold true. Then, as* $nh_n^2 \to 0$,

$$D^{1/2}[\delta_\theta u_n(\mathbf{X}_n, \theta_0) - \delta_\theta g_n(\mathbf{X}_n, \theta_0)] = o_p(1).$$

Lemma 3 *Suppose that Assumptions 2 and 3 hold true. Then, as* $nh_n^2 \to 0$,
(i) $D^{-1/2}(\widehat{\theta}_n^{(C)} - \theta_0) = I^{-1}(\theta_0) D^{1/2}(\delta_\theta g_n)(\mathbf{X}_n, \theta_0) + o_p(1)$,
(ii) $D^{1/2}(\delta_\theta g_n)(\mathbf{X}_n, \theta_0) \xrightarrow{d} N(0, I(\theta_0))$.

Lemma 4 *Suppose that Assumptions 1 and 2 hold true. Then, as* $nh_n^2 \to 0$,

$$D^{1/2}(\delta_\theta^2 u_n)(\mathbf{X}_n, \theta_0)D^{1/2} \xrightarrow{p} -I(\theta_0).$$

The main result is as follows.

Theorem 1 *Suppose that Assumptions 1, 2 and 3 hold true. Then, as* $nh_n^2 \to 0$,

$$E_{\theta_0}\left[u_n(\mathbf{X}_n, \hat{\theta}_n^{(C)}) - \int_\Omega l_n(\mathbf{Z}_n(\tilde{w}), \hat{\theta}_n^{(C)})dP(\tilde{w})\right] = dim(\Theta) + o(1).$$

Remark 1 *By theorem 1, AIC type of information criterion for diffusion processes is*

$$AIC = -2u_n(\mathbf{X}_n, \hat{\theta}_n^{(C)}) + 2dim(\Theta).$$

4. Simulation studies on approximate log likelihoods

We examine the asymptotic behaviours of

$$E_{\theta_0}[g_n(\mathbf{X}_n, \theta_0) - l_n(\mathbf{X}_n, \theta_0)]$$

and

$$E_{\theta_0}[u_n(\mathbf{X}_n, \theta_0) - l_n(\mathbf{X}_n, \theta_0)]$$

through simulations, which are done for each $T = nh_n = 10, 30, 50$ and $h_n = 1/100, 1/1000$.

For a true parameter value θ_0 and an initial value x_0, 5000 independent sample paths are generated by the Milstein scheme. For the Milstein scheme, see Kloeden and Platen (1992).

4.1 The Ornstein-Uhlenbeck model

$$dX_t \;=\; -\alpha X_t dt + \beta dw_t, \qquad X_0 = x_0,$$

where $\alpha > 0$ and $\beta > 0$ are unknown parameters.

$$\begin{aligned}
g(t,x,y,\theta) &= -\frac{1}{2}\log(2\pi t) - \frac{1}{2}\log(\beta^2) - \frac{(y - x + t\alpha x)^2}{2t\beta^2}, \\
u(t,x,y,\theta) &= -\frac{1}{2}\log(2\pi t) - \frac{1}{2}\log(\beta^2) - \frac{(y-x)^2}{2t\beta^2} - \frac{\alpha}{2\beta^2}(y^2 - x^2) \\
&\quad -\frac{t}{2}\left\{\frac{\alpha^2}{3\beta^2}(x^2 + y^2 + xy) - \alpha\right\}, \\
l(t,x,y,\theta) &= \log p(t,x,y,\theta), \\
p(t,x,y,\theta) &= \frac{1}{\sqrt{\pi\beta^2(1 - \exp\{-2\alpha t\})/\alpha}} \exp\left[\frac{-(y - \exp\{-\alpha t\}x)^2}{\beta^2(1 - \exp\{-2\alpha t\})/\alpha}\right].
\end{aligned}$$

T	h_n	$E[g_n - l_n]$	$E[u_n - l_n]$
10	1/100	-0.02436	0.00826
	1/1000	-0.00227	0.00083
30	1/100	-0.06979	0.02488
	1/1000	-0.00790	0.00249
50	1/100	-0.11241	0.04151
	1/1000	-0.01186	0.00416

Table 1: Means of $g_n - l_n$ and $u_n - l_n$ for 5000 independent simulated sample paths with $\alpha_0 = 1$, $\beta_0 = 2$ and $x_0 = 10$.

4.2 The Radial Ornstein-Uhlenbeck process

$$dX_t \;=\; (\theta X_t^{-1} - X_t)dt + dw_t, \qquad X_0 = x_0,$$

where $\theta > 0$ is an unknown parameter.

$$\begin{aligned}
g(t,x,y,\theta) &= -\frac{1}{2}\log(2\pi t) - \frac{(y - x - t(\theta x^{-1} - x))^2}{2t}, \\
u(t,x,y,\theta) &= -\frac{1}{2}\log(2\pi t) - \frac{(y-x)^2}{2t} + \theta \log\left(\frac{y}{x}\right) - \frac{1}{2}(y^2 - x^2) \\
&\quad -\frac{t}{2}\left\{\frac{1}{3}\left(\frac{\theta}{x} - x\right)^2 - \frac{1}{2}\left(\frac{\theta}{x^2} + 1\right) + \frac{1}{3}\left(\frac{\theta}{y} - y\right)^2 - \frac{1}{2}\left(\frac{\theta}{y^2} + 1\right)\right. \\
&\quad \left. + \frac{1}{3}\left(\frac{\theta}{x} - x\right)\left(\frac{\theta}{y} - y\right)\right\}.
\end{aligned}$$

$$
\begin{aligned}
l(t,x,y,\theta) &= \log p(t,x,y), \\
p(t,x,y,\theta) &= \frac{(y/x)^{\theta}\sqrt{xy}\,\exp\{-y^2+(\theta+\frac{1}{2})t\}}{\sinh(t)} \exp\left[\frac{-(x^2+y^2)}{\exp\{2t\}-1}\right] \\
&\quad \times I_{\theta-\frac{1}{2}}\left(\frac{xy}{\sinh(t)}\right),
\end{aligned}
$$

where I_ν is a modified Bessel function with index ν.

T	h_n	$E[g_n - l_n]$	$E[u_n - l_n]$
10	1/100	-0.24315	0.05307
	1/1000	-0.10851	0.00544
30	1/100	-0.86191	0.18446
	1/1000	-0.37434	0.01892
50	1/100	-1.45814	0.31463
	1/1000	-0.64074	0.03222

Table 2: Means of $g_n - l_n$ and $u_n - l_n$ for 5000 independent simulated sample paths with $\theta_0 = 2$ and $x_0 = 10$.

4.3 The Cox-Ingersoll-Ross process

$$dX_t = -\alpha_1(X_t - \alpha_2)dt + \beta\sqrt{X_t}dw_t, \quad X_0 = x_0,$$

where $\alpha_1 > 0$, $\alpha_2 > 0$ and $\beta > 0$ are unknown parameters.

$$g(t,x,y,\theta) = -\frac{1}{2}\log(2\pi t) - \frac{1}{2}\log(\beta^2 x) - \frac{(y - x + t\alpha_1(x-\alpha_2))^2}{2t\beta^2 x}, \quad (9)$$

$$\begin{aligned} u(t,x,y,\theta) &= -\frac{1}{2}\log(2\pi t) - \frac{1}{2}\log(\beta^2 y) - \frac{2(\sqrt{y}-\sqrt{x})^2}{t\beta^2} \\ &\quad -\frac{\alpha_1(y-x)}{\beta^2} + \left(\frac{\alpha_1\alpha_2}{\beta^2} - \frac{1}{4}\right)\log\left(\frac{y}{x}\right) \\ &\quad -\frac{t}{2}\left[C(x,\theta) + C(y,\theta) + \frac{1}{3}B(x,\theta)B(y,\theta)\right], \end{aligned} \quad (10)$$

where

$$\begin{aligned} B(x,\theta) &= -\frac{\alpha_1}{\beta}\sqrt{x} + \left(\frac{\alpha_1\alpha_2}{\beta} - \frac{\beta}{4}\right)\frac{1}{\sqrt{x}}, \\ C(x,\theta) &= \frac{1}{3}\{B(x,\theta)\}^2 + \frac{1}{2}\left\{-\frac{\alpha_1}{2} - \frac{1}{2}\left(\alpha_1\alpha_2 - \frac{\beta^2}{4}\right)\frac{1}{x}\right\}. \end{aligned}$$

$$l(t,x,y,\theta) = \log p(t,x,y,\theta),$$

$$p(t,x,y,\theta) = \frac{\gamma(y/x)^{\frac{1}{2}\nu}\exp\{\frac{1}{2}\alpha_1\nu t - \gamma y\}}{1-\exp\{-\alpha_1 t\}}\exp\left[\frac{-\gamma(x+y)}{\exp\{\alpha_1 t\}-1}\right] I_\nu\left(\frac{\gamma\sqrt{xy}}{\sinh(\frac{1}{2}\alpha_1 t)}\right),$$

where $\gamma = 2\alpha_1\beta^{-2}$, $\nu = \gamma\alpha_2 - 1$.

T	h_n	$E[g_n - l_n]$	$E[u_n - l_n]$
10	1/100	-5.63288	0.01009
	1/1000	-20.92659	0.00100
30	1/100	-17.06086	0.03066
	1/1000	-63.01867	0.00305
50	1/100	-28.42232	0.05122
	1/1000	-105.04559	0.00510

Table 3: Means of $g_n - l_n$ and $u_n - l_n$ for 5000 independent simulated sample paths with $\alpha_{1,0} = 1$, $\alpha_{2,0} = 10$, $\beta_0 = 2$ and $x_0 = 10$.

5. Example of model selection based on AIC

The true model is

$$dX_t = -(X_t - 10)dt + 2\sqrt{X_t}dw_t,$$

where $X_0 = 10$ and $t \in [0, T]$. We consider the following three statistical models:

$$\begin{aligned} dX_t &= -\alpha_1(X_t - \alpha_2)dt + \beta\sqrt{X_t}dw_t, && (11) \\ dX_t &= -\alpha_1(X_t - \alpha_2)dt + \sqrt{\beta_1 + \beta_2 X_t}dw_t, && (12) \\ dX_t &= -\alpha_1(X_t - \alpha_2)dt + (\beta_1 + \beta_2 X_t)^{\beta_3}dw_t, && (13) \end{aligned}$$

where $\alpha_1 > 0$, $\alpha_2 > 0$, $\beta > 0$, $\beta_1 \geq 0$, $\beta_2 > 0$ and $\beta_3 \geq 0$.

It follows from (5) that the contrast functions for the models (11), (12) and (13) are

$$
\begin{aligned}
g^{(1)}(t,x,y,\theta) &= -\frac{1}{2}\log(2\pi t) - \frac{1}{2}\log(\beta^2 x) \\
&\quad -\frac{(y-x+t\alpha_1(x-\alpha_2))^2}{2t\beta^2 x}, \qquad (14)\\
g^{(2)}(t,x,y,\theta) &= -\frac{1}{2}\log(2\pi t) - \frac{1}{2}\log(\beta_1 + \beta_2 x) \\
&\quad -\frac{(y-x+t\alpha_1(x-\alpha_2))^2}{2t(\beta_1 x + \beta_2)}, \qquad (15)\\
g^{(3)}(t,x,y,\theta) &= -\frac{1}{2}\log(2\pi t) - \beta_3\log(\beta_1 + \beta_2 x) \\
&\quad -\frac{(y-x+t\alpha_1(x-\alpha_2))^2}{2t(\beta_1 x + \beta_2)^{2\beta_3}}, \qquad (16)
\end{aligned}
$$

respectively.

By (4), the approximate log-likelihood functions of the models (11), (12) and (13) are

$$\begin{aligned} u^{(1)}(t,x,y,\theta) &= -\frac{1}{2}\log(2\pi t) - \frac{1}{2}\log(\beta^2 y) - \frac{2(\sqrt{y}-\sqrt{x})^2}{t\beta^2} \\ &\quad -\frac{\alpha_1(y-x)}{\beta^2} + \left(\frac{\alpha_1\alpha_2}{\beta^2} - \frac{1}{4}\right)\log\left(\frac{y}{x}\right) \\ &\quad -\frac{t}{2}\left[C(x,\theta) + C(y,\theta) + \frac{1}{3}B(x,\theta)B(y,\theta)\right], \qquad (17) \end{aligned}$$

where

$$\begin{aligned} B(x,\theta) &= -\frac{\alpha_1}{\beta}\sqrt{x} + \left(\frac{\alpha_1\alpha_2}{\beta} - \frac{\beta}{4}\right)\frac{1}{\sqrt{x}}, \\ C(x,\theta) &= \frac{1}{3}\{B(x,\theta)\}^2 + \frac{1}{2}\left\{-\frac{\alpha_1}{2} - \frac{1}{2}\left(\alpha_1\alpha_2 - \frac{\beta^2}{4}\right)\frac{1}{x}\right\}, \end{aligned}$$

$$
\begin{aligned}
u^{(2)}(t,x,y,\theta) &= -\frac{1}{2}\log(2\pi t) - \frac{1}{2}\log(\beta_1+\beta_2 y) \\
&\quad -\frac{2(\sqrt{\beta_1+\beta_2 y}-\sqrt{\beta_1+\beta_2 x})^2}{t\beta_2^2} \\
&\quad -\frac{\alpha_1(y-x)}{\beta_2} + \left(\frac{\alpha_1\alpha_2}{\beta_2}+\frac{\alpha_1\beta_1}{\beta_2^2}-\frac{1}{4}\right)\log\left(\frac{\beta_1+\beta_2 y}{\beta_1+\beta_2 x}\right) \\
&\quad -\frac{t}{2}\left[C(x,\theta)+C(y,\theta)+\frac{1}{3}B(x,\theta)B(y,\theta)\right], \qquad (18)
\end{aligned}
$$

where

$$
\begin{aligned}
B(x,\theta) &= \frac{-\alpha_1 x+\alpha_1\alpha_2-\beta_2/4}{\sqrt{\beta_1+\beta_2 x}}, \\
C(x,\theta) &= \frac{1}{3}\left\{\frac{-\alpha_1 x+\alpha_1\alpha_2-\beta_2/4}{\sqrt{\beta_1+\beta_2 x}}\right\}^2 \\
&\quad +\frac{1}{2}\left\{\frac{-\alpha_1\beta_2 x/2-\alpha_1\beta_1-\alpha_1\alpha_2\beta_2/2+\beta_2^2/8}{\beta_1+\beta_2 x}\right\},
\end{aligned}
$$

$$
\begin{aligned}
u^{(3)}(t,x,y,\theta) \;=\; & -\frac{1}{2}\log(2\pi t) - \beta_3 \log(\beta_1+\beta_2 y) \\
& -\frac{\left\{(\beta_1+\beta_2 y)^{1-\beta_3} - (\beta_1+\beta_2 x)^{1-\beta_3}\right\}^2}{2t(1-\beta_3)^2\beta_2^2} \\
& -\alpha_1\left[\frac{(-\beta_1+\beta_2 y(1-2\beta_3) - 2\alpha_2\beta_2(1-\beta_3))}{2\beta_2^2(1-\beta_3)(1-2\beta_3)(\beta_1+\beta_2 y)^{2\beta_3-1}}\right. \\
& \left. -\frac{(-\beta_1+\beta_2 x(1-2\beta_3) - 2\alpha_2\beta_2(1-\beta_3))}{2\beta_2^2(1-\beta_3)(1-2\beta_3)(\beta_1+\beta_2 x)^{2\beta_3-1}}\right] \\
& -\frac{\beta_3}{2}\log\left(\frac{\beta_1+\beta_2 y}{\beta_1+\beta_2 x}\right) \\
& -\frac{t}{2}\left[C(x,\theta)+C(y,\theta)+\frac{1}{3}B(x,\theta)B(y,\theta)\right], \qquad (19)
\end{aligned}
$$

where

$$
\begin{aligned}
B(x,\theta) &= \frac{-\alpha_1(x-\alpha_2)}{(\beta_1+\beta_2 x)^{\beta_3}} - \frac{\beta_2\beta_3}{2(\beta_1+\beta_2 x)^{1-\beta_3}}, \\
C(x,\theta) &= \frac{1}{3}\left\{\frac{-\alpha_1(x-\alpha_2)}{(\beta_1+\beta_2 x)^{\beta_3}} - \frac{\beta_2\beta_3}{2(\beta_1+\beta_2 x)^{1-\beta_3}}\right\}^2 \\
&\quad + \frac{1}{2}\left\{-\alpha_1 + \frac{\alpha_1(x-\alpha_2)\beta_2\beta_3}{\beta_1+\beta_2 x} - \frac{\beta_2^2\beta_3(\beta_3-1)}{2(\beta_1+\beta_2 x)^{2-2\beta_3}}\right\},
\end{aligned}
$$

respectively. Note that $u^{(3)}(t,x,y,\theta)$ is obtained under the assumption $\beta_3 \neq 1/2$. When $\beta_3 = 1/2$, it suffices to consider the model (12).

Therefore, AIC for each model (11), (12), (13) is as follows.

$$\begin{aligned} AIC_1(\mathbf{X}_n, \widehat{\theta}_n^{(1)}) &= -2u^{(1)}(\mathbf{X}_n, \widehat{\theta}_n^{(1)}) + 2 \times 3, \\ AIC_2(\mathbf{X}_n, \widehat{\theta}_n^{(2)}) &= -2u^{(2)}(\mathbf{X}_n, \widehat{\theta}_n^{(2)}) + 2 \times 4, \\ AIC_3(\mathbf{X}_n, \widehat{\theta}_n^{(3)}) &= -2u^{(3)}(\mathbf{X}_n, \widehat{\theta}_n^{(3)}) + 2 \times 5, \end{aligned}$$

where $\widehat{\theta}_n^{(i)}$ is obtained from the contrast function $g_n^{(i)}$ for $i = 1, 2, 3$.

Simulation result

We examine the number of models selected by AIC among competing models (11), (12), (13) through simulations for 1000 independent sample paths generated by the Milstein scheme.

The simulations are done for each $T = 10, 30$ and $h_n = 1/100$.

$$
\begin{aligned}
\text{(true model)} \quad dX_t &= -(X_t - 10)dt + 2\sqrt{X_t}dw_t, \\
\text{(model 1)} \quad dX_t &= -\alpha_1(X_t - \alpha_2)dt + \beta\sqrt{X_t}dw_t, \\
\text{(model 2)} \quad dX_t &= -\alpha_1(X_t - \alpha_2)dt + \sqrt{\beta_1 + \beta_2 X_t}dw_t, \\
\text{(model 3)} \quad dX_t &= -\alpha_1(X_t - \alpha_2)dt + (\beta_1 + \beta_2 X_t)^{\beta_3}dw_t,
\end{aligned}
$$

where $X_0 = 10$, $t \in [0, T]$.

T	h_n	model 1	model 2	model 3
10	1/100	761	185	54
30	1/100	803	185	12

Table 4: The number of models selected by AIC for 1000 independent simulated sample paths.

6. Conclusion

1. In order to obtain an efficient estimator, we use the locally Gaussian approximation g_n.

2. For an approximation of log likelihood l_n, it is better to use the approximate log likelihood u_n based on Dacunha-Castelle and Florens-Zmirou (1986) than the locally Gaussian approximation g_n.

3. In special, for the CIR model, we cannot use g_n as an approximation of l_n because g_n has a considerable bias.