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

\( X_n \sim g(x_n) \): true density

\( X_n = x_n \): observations

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

\( g(x_n) = f(x_n, \theta_0) \implies \{f(x_n, \theta); \theta \in \Theta\} \): 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)

Step 1

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

\( \theta \leftarrow \hat{\theta}(X_n) \) : an estimator obtained from the observations \( X_n \)

Step 2

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

\( g(z_n) \leftarrow f(z_n, \hat{\theta}(X_n)) \) : predictive distribution (statistical model)
Step 3

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

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

$$I\{g(z_n); f(z_n, \hat{\theta}(X_n))\} := E_Z \left[ \log \frac{g(Z_n)}{f(Z_n, \hat{\theta}(X_n))} \right].$$

The rule of model selection

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

$$I\{g; \hat{f}\} = \min_i I\{g(z_n); f_i(z_n, \hat{\theta}_i(X_n))\}.$$
The estimated Kullback-Leibler information

\[ I\{g(z_n); f(z_n, \hat{\theta}(X_n))\} = \int g(z_n) \log g(z_n) dz_n - \int g(z_n) \log f(z_n, \hat{\theta}(X_n)) dz_n. \]  

The first term in the right hand side of (1) does not depend on the statistical model \( f(z_n, \hat{\theta}(X_n)) \)

The second term \( \eta(\hat{\theta}(X_n)) := \int g(z_n) \log f(z_n, \hat{\theta}(X_n)) dz_n \) depends on it,

\( \eta(\hat{\theta}(X_n)) : \text{the expected log likelihood} \)

Minimizing the estimated KLI \( I\{g(z_n); f(z_n, \hat{\theta}(X_n))\} \) is equivalent to maximizing the expected log likelihood \( \eta(\hat{\theta}(X_n)) \).
Therefore, the rule of model selection is to choose a model which is maximizing the expected log likelihood $\eta(\hat{\theta}(X_n))$ among competing models.

However, since the expected log likelihood

$$\eta(\hat{\theta}(X_n)) = \int g(z_n) \log f(z_n, \hat{\theta}(X_n)) dz_n$$

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

**Step 4**

$\eta(\hat{\theta}(X_n)) \leftarrow \log f(X_n, \hat{\theta}(X_n)) : \text{simple estimator}$

The bias of $\log f(X_n, \hat{\theta}(X_n))$ in the estimation of $\eta(\hat{\theta}(X_n))$ is given by

$$\text{bias} = E_X \left[ \log f(X_n, \hat{\theta}(X_n)) - \int g(z_n) \log f(z_n, \hat{\theta}(X_n)) dz_n \right].$$
If $\hat{\theta}$ is the MLE ($\hat{\theta}^{(ML)}$), under some regularity conditions, as $n \to \infty$,  
\[ \text{bias} = \text{dim}(\Theta) + o(1), \]
where dim($\Theta$) denotes the dimension of a parameter space $\Theta$.

The bias corrected log likelihood is given by  
\[ \log f(X_n, \hat{\theta}^{(ML)}(X_n)) - \text{dim}(\Theta). \]

Thus, Akaike (1973, 1974) proposed  
\[ \text{AIC}(X_n, \hat{\theta}^{(ML)}(X_n)) = -2 \log f(X_n, \hat{\theta}^{(ML)}(X_n)) + 2\text{dim}(\Theta). \tag{2} \]

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

\[ dX_t = b(X_t, \alpha)dt + \sigma(X_t, \beta)d\omega_t, \quad t \in [0, T], \]

\[ X_0 = x_0, \]

where

\[ \theta = (\alpha, \beta) \in \Theta_{\alpha} \times \Theta_{\beta} = \Theta. \]

\( \Theta_{\alpha}, \Theta_{\beta} \) : compact convex subsets of \( \mathbb{R}^p \) and \( \mathbb{R}^q \).

\( b : \mathbb{R} \times \Theta_{\alpha} \rightarrow \mathbb{R} \).

\( \sigma : \mathbb{R} \times \Theta_{\beta} \rightarrow \mathbb{R} \).

\( w \) : one-dimensional standard standard Wiener process.

We assume that the drift \( b \) and the diffusion coefficient \( \sigma \) are known apart from the parameters \( \alpha \) and \( \beta \).
The data: \( X_n = (X_{t_k}^n)_{0 \leq k \leq n} \) 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. \( \alpha_0, \beta_0, \theta_0 \) denote the true values of \( \alpha, \beta \) and \( \theta \), respectively.
2. For a function \( f(x, \theta) \), define that
   \( \delta_{\theta_i} f(x, \theta) = \frac{\partial}{\partial \theta_i} f(x, \theta) \),
   \( f'(x, \theta) = \frac{\partial}{\partial x} f(x, \theta) \),
   \( \delta_{\theta} f(x, \theta) = \left( \delta_{\theta_i} f(x, \theta) \right)_{i=1, \ldots, p} \)
   and \( \delta_{\theta}^2 f(x, \theta) = \left( \frac{\partial^2}{\partial \theta_i \partial \theta_j} f(x, \theta) \right)_{i,j=1, \ldots, p+q} \).
3. \( E \) denotes the state space of \( X, \; E \subseteq \mathbb{R} \).
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 \( y \mapsto \mathbf{p}(t, x, y, \theta), \; y \in E. \)
Moreover, we define the following functions.

\[ s(x, \beta) = \int_0^x \frac{du}{\sigma(u, \beta)}, \]
\[ B(x, \theta) = \frac{b(x, \alpha)}{\sigma(x, \beta)} - \frac{1}{2} \sigma'(x, \beta), \]
\[ \tilde{B}(x, \theta) = B(s^{-1}(x, \beta), \theta), \]
\[ \tilde{h}(x, \theta) = \tilde{B}^2(x, \theta) + \tilde{B}'(x, \theta). \]
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 $\theta$ with invariant probability measure $\mu_\theta$. All polynomial moments of $\mu_\theta$ are finite.  
(iv) For all $m \geq 0$ and for all $\theta$, $\sup_t E_\theta[|X_t|^m] < \infty$.  
(v) For every $\theta$, 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 $\theta$.  
(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 $\theta$ for all $x$ in the state space. All these derivatives with respect to $\theta$ are of polynomial growth in $x$, uniformly in $\theta$. 
Assumption 2  (i) $\tilde{h}(x, \theta) = O(|x|^2)$ as $x \to \infty$.
(ii) $\inf_x \tilde{h}(x, \theta) > -\infty$ for all $\theta$.
(iii) $\sup_{\theta} \sup_x |\tilde{h}^3(x, \theta)| \leq M < \infty$.
(iv) There exists $\gamma > 0$ such that for every $\theta$ 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

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

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

Define the maximum likelihood estimator

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

Then, Akaike’s information criterion is as follows:

$$AIC = -2l_n(\mathbf{X}_n, \hat{\theta}_n^{(ML)}) + 2 \dim(\Theta).$$
However, since the transition density $p$ of the diffusion process $X$ does not generally have an explicit form, we cannot directly obtain

the log likelihood function $l_n$

and

the maximum likelihood estimator $\hat{\theta}_n^{(ML)}$.

That is why we need to obtain both

an approximation of the log-likelihood function $l_n$

and

an asymptotically efficient estimator $\hat{\theta}_n$

in order to construct AIC type of information criteria for diffusion processes.
an approximation of $l_n$ (Dacunha-Castelle and Florens-Zmirou (1986))

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

where

$$
\begin{align*}
  S(x, y, \beta) &= \int_{x}^{y} \frac{du}{\sigma(u, \beta)}, \\
  H(x, y, \theta) &= \int_{x}^{y} \left\{ \frac{b(u, \alpha)}{\sigma^2(u, \beta)} - \frac{1}{2} \frac{\sigma'(u, \beta)}{\sigma(u, \beta)} \right\} du, \\
  \tilde{g}(x, y, \theta) &= -\frac{1}{2} \left\{ C(x, \theta) + C(y, \theta) + \frac{1}{3} B(x, \theta)B(y, \theta) \right\}, \\
  B(x, \theta) &= \frac{b(x, \alpha)}{\sigma(x, \beta)} - \frac{1}{2} \sigma'(x, \beta), \\
  C(x, \theta) &= \frac{1}{3} [B(x, \theta)]^2 + \frac{1}{2} [B(x, \theta)]' \sigma(x, \beta).
\end{align*}
$$
an asymptotically efficient estimator
We use the contrast function based on locally Gaussian approximation as follows:

\[ g_n(X_n, \theta) = \sum_{k=1}^{n} g(h_n, X_{tn}^{k-1}, X_{tn}^{k}, \theta), \]

where

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

We then define the maximum contrast estimator as

\[ \hat{\theta}_n^{(C)} = \arg \sup_{\theta} g_n(X_n, \theta). \]
For a process $Z_n$ which is independent of (but has the same distribution as) the observed process $X_n$,

$$u_n(X_n, \hat{\theta}_n^{(C)}(X_n)) - \int_\Omega l_n(Z_n(\tilde{\omega}), \hat{\theta}_n^{(C)}(X_n))dP(\tilde{\omega})$$

$$= u_n(X_n, \hat{\theta}_n^{(C)}(X_n)) - u_n(X_n, \theta_0)$$

$$+ u_n(X_n, \theta_0) - \int_\Omega l_n(Z_n(\tilde{\omega}), \theta_0)dP(\tilde{\omega})$$

$$+ \int_\Omega l_n(Z_n(\tilde{\omega}), \theta_0)dP(\tilde{\omega}) - \int_\Omega l_n(Z_n(\tilde{\omega}), \hat{\theta}_n^{(C)}(X_n))dP(\tilde{\omega}).$$
Under the regularity conditions, one has

\[
(6) = \left[ D^{1/2} \delta \mu_n(\mathbf{X}_n, \theta_0) \right]^T \left[ D^{-1/2} \left( \hat{\theta}_n^C \right)(\mathbf{X}_n) - \theta_0 \right] \\
+ \frac{1}{2} \left[ D^{-1/2} \left( \hat{\theta}_n^C \right)(\mathbf{X}_n) - \theta_0 \right]^T \left[ D^{1/2} \delta^2 \mu_n(\mathbf{X}_n, \theta_0) D^{1/2} \right] \\
\times D^{-1/2} \left( \hat{\theta}_n^C \right)(\mathbf{X}_n) - \theta_0 + o_p(1),
\]

\[
(8) = - \left[ \int_\Omega D^{1/2} \delta \ell_n(\mathbf{Z}_n(\tilde{\omega}), \theta_0) dP(\tilde{\omega}) \right]^T \left[ D^{-1/2} \left( \hat{\theta}_n^C \right)(\mathbf{X}_n) - \theta_0 \right] \\
- \frac{1}{2} \left[ D^{-1/2} \left( \hat{\theta}_n^C \right)(\mathbf{X}_n) - \theta_0 \right]^T \int_\Omega D^{1/2} \delta^2 \ell_n(\mathbf{Z}_n(\tilde{\omega}), \theta_0) D^{1/2} dP(\tilde{\omega}) \\
\times D^{-1/2} \left( \hat{\theta}_n^C \right)(\mathbf{X}_n) - \theta_0 + o_p(1),
\]

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:

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

where

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

$$I^{ij}_\sigma(\theta_0) = 2 \int_\mathbb{R} \frac{\delta_{\beta_i} \sigma(x, \beta_0) \delta_{\beta_j} \sigma(x, \beta_0)}{\sigma^2(x, \beta_0)} \mu_{\theta_0}(dx).$$
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 $nh_n^2 \to 0$,

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

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

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

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

(i) $D^{-1/2}(\hat{\theta}_n^{(C)} - \theta_0) = I^{-1}(\theta_0)D^{1/2}(\delta_{\theta} g_n)(X_n, \theta_0) + o_p(1)$,

(ii) $D^{1/2}(\delta_{\theta} g_n)(X_n, \theta_0) \overset{d}{\to} N(0, I(\theta_0))$. 

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

$$D^{1/2}(\delta^2 u_n)(X_n, \theta_0) D^{1/2} p \to -I(\theta_0).$$

The main result is as follows.

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

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

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

$$AIC = -2u_n(X_n, \hat{\theta}_n^{(C)}) + 2\text{dim}(\Theta).$$
4. Simulation studies on approximate log likelihoods

We examine the asymptotic behaviours of

$$E_{\theta_0}[g_n(X_n, \theta_0) - l_n(X_n, \theta_0)]$$

and

$$E_{\theta_0}[u_n(X_n, \theta_0) - l_n(X_n, \theta_0)]$$

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 $\theta_0$ 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

\[
dX_t = -\alpha X_t dt + \beta dw_t, \quad X_0 = x_0,
\]
where \(\alpha > 0\) and \(\beta > 0\) are unknown parameters.

\[
g(t, x, y, \theta) = -\frac{1}{2} \log(2\pi t) - \frac{1}{2} \log(\beta^2) - \frac{(y - x + t\alpha x)^2}{2t\beta^2},
\]

\[
u(t, x, y, \theta) = -\frac{1}{2} \log(2\pi t) - \frac{1}{2} \log(\beta^2) - \frac{(y - x)^2}{2t\beta^2} - \frac{\alpha}{2\beta^2}(y^2 - x^2)
\]

\[
-\frac{t}{2} \left\{ \alpha^2 \left( x^2 + y^2 + xy \right) - \alpha \right\},
\]

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

\[
p(t, x, y, \theta) = \frac{1}{\sqrt{\pi\beta^2(1 - \exp\{-2\alpha t\})/\alpha}} \exp \left[ \frac{-(y - \exp\{-\alpha t\} x)^2}{\beta^2(1 - \exp\{-2\alpha t\})/\alpha} \right].
\]
<table>
<thead>
<tr>
<th>$T$</th>
<th>$h_n$</th>
<th>$E[g_n - l_n]$</th>
<th>$E[u_n - l_n]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1/100</td>
<td>-0.02436</td>
<td>0.00826</td>
</tr>
<tr>
<td></td>
<td>1/1000</td>
<td>-0.00227</td>
<td>0.00083</td>
</tr>
<tr>
<td>30</td>
<td>1/100</td>
<td>-0.06979</td>
<td>0.02488</td>
</tr>
<tr>
<td></td>
<td>1/1000</td>
<td>-0.00790</td>
<td>0.00249</td>
</tr>
<tr>
<td>50</td>
<td>1/100</td>
<td>-0.11241</td>
<td>0.04151</td>
</tr>
<tr>
<td></td>
<td>1/1000</td>
<td>-0.01186</td>
<td>0.00416</td>
</tr>
</tbody>
</table>

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

\[ dX_t = (\theta X_t^{-1} - X_t)dt + dw_t, \quad X_0 = x_0, \]

where \( \theta > 0 \) is an unknown parameter.

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

where \( I_\nu \) is a modified Bessel function with index \( \nu \).

<table>
<thead>
<tr>
<th>( T )</th>
<th>( h_n )</th>
<th>( E[g_n - l_n] )</th>
<th>( E[u_n - l_n] )</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1/100</td>
<td>-0.24315</td>
<td>0.05307</td>
</tr>
<tr>
<td></td>
<td>1/1000</td>
<td>-0.10851</td>
<td>0.00544</td>
</tr>
<tr>
<td>30</td>
<td>1/100</td>
<td>-0.86191</td>
<td>0.18446</td>
</tr>
<tr>
<td></td>
<td>1/1000</td>
<td>-0.37434</td>
<td>0.01892</td>
</tr>
<tr>
<td>50</td>
<td>1/100</td>
<td>-1.45814</td>
<td>0.31463</td>
</tr>
<tr>
<td></td>
<td>1/1000</td>
<td>-0.64074</td>
<td>0.03222</td>
</tr>
</tbody>
</table>

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.

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

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

where

\[ B(x, \theta) = -\frac{\alpha_1}{\beta} \sqrt{x} + \left( \frac{\alpha_1 \alpha_2}{\beta} - \frac{\beta}{4} \right) \frac{1}{\sqrt{x}}, \]
\[ C(x, \theta) = \frac{1}{3} \{ B(x, \theta) \}^2 + \frac{1}{2} \left\{ -\frac{\alpha_1}{2} - \frac{1}{2} \left( \frac{\alpha_1 \alpha_2 - \frac{\beta^2}{4}}{\frac{\beta^2}{4}} \right) \frac{1}{x} \right\}. \]
\[
\begin{align*}
l(t, x, y, \theta) &= \log p(t, x, y, \theta), \\
p(t, x, y, \theta) &= \frac{\gamma (y/x)\frac{1}{2}\nu \exp\{\frac{1}{2} \alpha_1 \nu t - \gamma y\}}{1 - \exp\{-\alpha_1 t\}} \exp \left[ \frac{-\gamma (x + y)}{\exp\{\alpha_1 t\} - 1} \right] I_{\nu} \left( \frac{\gamma \sqrt{xy}}{\sinh(\frac{1}{2} \alpha_1 t)} \right),
\end{align*}
\]

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

<table>
<thead>
<tr>
<th>( T )</th>
<th>( h_n )</th>
<th>( E[g_n - l_n] )</th>
<th>( E[u_n - l_n] )</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1/100</td>
<td>-5.63288</td>
<td>0.01009</td>
</tr>
<tr>
<td></td>
<td>1/1000</td>
<td>-20.92659</td>
<td>0.00100</td>
</tr>
<tr>
<td>30</td>
<td>1/100</td>
<td>-17.06086</td>
<td>0.03066</td>
</tr>
<tr>
<td></td>
<td>1/1000</td>
<td>-63.01867</td>
<td>0.00305</td>
</tr>
<tr>
<td>50</td>
<td>1/100</td>
<td>-28.42232</td>
<td>0.05122</td>
</tr>
<tr>
<td></td>
<td>1/1000</td>
<td>-105.04559</td>
<td>0.00510</td>
</tr>
</tbody>
</table>

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:

\[
\begin{align*}
    dX_t &= -\alpha_1 (X_t - \alpha_2) dt + \beta \sqrt{X_t} dw_t, \quad (11) \\
    dX_t &= -\alpha_1 (X_t - \alpha_2) dt + \sqrt{\beta_1 + \beta_2 X_t} dw_t, \quad (12) \\
    dX_t &= -\alpha_1 (X_t - \alpha_2) dt + (\beta_1 + \beta_2 X_t)^{\beta_3} dw_t, \quad (13)
\end{align*}
\]

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
\[
g^{(1)}(t, x, y, \theta) = -\frac{1}{2} \log(2\pi t) - \frac{1}{2} \log(\beta^2 x) - \frac{(y - x + t\alpha_1(x - \alpha_2))^2}{2t\beta^2 x},
\]
\[
g^{(2)}(t, x, y, \theta) = -\frac{1}{2} \log(2\pi t) - \frac{1}{2} \log(\beta_1 + \beta_2 x) - \frac{(y - x + t\alpha_1(x - \alpha_2))^2}{2t(\beta_1 x + \beta_2)},
\]
\[
g^{(3)}(t, x, y, \theta) = -\frac{1}{2} \log(2\pi t) - \beta_3 \log(\beta_1 + \beta_2 x) - \frac{(y - x + t\alpha_1(x - \alpha_2))^2}{2t(\beta_1 x + \beta_2)^2\beta_3},
\]
respectively.
By (4), the approximate log-likelihood functions of the models (11), (12) and (13) are

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

(17)

where

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


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

where
\[ B(x, \theta) = \frac{-\alpha_1 x + \alpha_1 \alpha_2 - \beta_2/4}{\sqrt{\beta_1 + \beta_2 x}}, \]
\[ C(x, \theta) = \frac{1}{3} \left\{ \frac{-\alpha_1 x + \alpha_1 \alpha_2 - \beta_2/4}{\sqrt{\beta_1 + \beta_2 x}} \right\}^2 \]
\[ + \frac{1}{2} \left\{ \frac{-\alpha_1 \beta_2 x/2 - \alpha_1 \beta_1 - \alpha_1 \alpha_2 \beta_2 /2 + \beta_2^2/8}{\beta_1 + \beta_2 x} \right\}, \]
\[ u^{(3)}(t, x, y, \theta) = -\frac{1}{2} \log(2\pi t) - \beta_3 \log(\beta_1 + \beta_2 y) \]

\[ -\left\{(\beta_1 + \beta_2 y)^{1-\beta_3} - (\beta_1 + \beta_2 x)^{1-\beta_3}\right\}^2 \]

\[ 2t(1 - \beta_3)^2 \beta_2^2 \]

\[ -\alpha_1 \frac{(-\beta_1 + \beta_2 y(1 - 2\beta_3) - 2\alpha_2 \beta_2 (1 - \beta_3))}{2\beta_2^2(1 - \beta_3)(1 - 2\beta_3)(\beta_1 + \beta_2 y)^{2\beta_3 - 1}} \]

\[ -\alpha_1 \frac{(-\beta_1 + \beta_2 x(1 - 2\beta_3) - 2\alpha_2 \beta_2 (1 - \beta_3))}{2\beta_2^2(1 - \beta_3)(1 - 2\beta_3)(\beta_1 + \beta_2 x)^{2\beta_3 - 1}} \]

\[ -\frac{\beta_3}{2} \log \left(\frac{\beta_1 + \beta_2 y}{\beta_1 + \beta_2 x}\right) \]

\[ -\frac{t}{2} \left[ C(x, \theta) + C(y, \theta) + \frac{1}{3} B(x, \theta) B(y, \theta) \right], \quad (19) \]
where

\[
B(x, \theta) = \frac{-\alpha_1 (x - \alpha_2)}{(\beta_1 + \beta_2 x)^{\beta_3}} - \frac{\beta_2 \beta_3}{2(\beta_1 + \beta_2 x)^{1-\beta_3}},
\]

\[
C(x, \theta) = \frac{1}{3} \left\{ \frac{-\alpha_1 (x - \alpha_2)}{(\beta_1 + \beta_2 x)^{\beta_3}} - \frac{\beta_2 \beta_3}{2(\beta_1 + \beta_2 x)^{1-\beta_3}} \right\}^2
\]

\[
+ \frac{1}{2} \left\{ -\alpha_1 + \frac{\alpha_1 (x - \alpha_2) \beta_2 \beta_3}{\beta_1 + \beta_2 x} - \frac{\beta_2^2 \beta_3 (\beta_3 - 1)}{2(\beta_1 + \beta_2 x)^{2-2\beta_3}} \right\},
\]

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.

\[
\begin{align*}
AIC_1(X_n, \hat{\theta}_n^{(1)}) &= -2u^{(1)}(X_n, \hat{\theta}_n^{(1)}) + 2 \times 3, \\
AIC_2(X_n, \hat{\theta}_n^{(2)}) &= -2u^{(2)}(X_n, \hat{\theta}_n^{(2)}) + 2 \times 4, \\
AIC_3(X_n, \hat{\theta}_n^{(3)}) &= -2u^{(3)}(X_n, \hat{\theta}_n^{(3)}) + 2 \times 5,
\end{align*}
\]

where \( \hat{\theta}_n^{(i)} \) 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 \( h_n = 1/100 \).
(true model) \[ dX_t = - (X_t - 10) dt + 2 \sqrt{X_t} dw_t, \]

(model 1) \[ dX_t = - \alpha_1 (X_t - \alpha_2) dt + \beta \sqrt{X_t} dw_t, \]

(model 2) \[ dX_t = - \alpha_1 (X_t - \alpha_2) dt + \sqrt{\beta_1 + \beta_2 X_t} dw_t, \]

(model 3) \[ dX_t = - \alpha_1 (X_t - \alpha_2) dt + (\beta_1 + \beta_2 X_t)^{\beta_3} dw_t, \]

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

<table>
<thead>
<tr>
<th>( T )</th>
<th>( h_n )</th>
<th>model 1</th>
<th>model 2</th>
<th>model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1/100</td>
<td>761</td>
<td>185</td>
<td>54</td>
</tr>
<tr>
<td>30</td>
<td>1/100</td>
<td>803</td>
<td>185</td>
<td>12</td>
</tr>
</tbody>
</table>

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.