MATHEMATICAL ENGINEERING
TECHNICAL REPORTS

Bayesian prediction and model selection for locally asymptotically mixed normal models

Tomonari SEI and Fumiyasu KOMAKI

METR2004-25 May 2004

DEPARTMENT OF MATHEMATICAL INFORMATICS
GRADUATE SCHOOL OF INFORMATION SCIENCE AND TECHNOLOGY
THE UNIVERSITY OF TOKYO
BUNKYO-KU, TOKYO 113-8656, JAPAN

WWW page: http://www.i.u-tokyo.ac.jp/mi/mi-e.htm

Bayesian prediction and model selection for locally asymptotically mixed normal models

Tomonari SEI (sei@stat.t.u-tokyo.ac.jp)
Department of Mathematical Informatics, Graduate School of Information Science and Technology, the University of Tokyo.

Fumiyasu KOMAKI (komaki@mist.i.u-tokyo.ac.jp)
Department of Mathematical Informatics, Graduate School of Information Science and Technology, the University of Tokyo.

Abstract
An information criterion for models with the local asymptotic mixed normality (LAMN) is proposed. Since the widely known Akaike's Information Criterion (AIC) is derived based on the local asymptotic normality (LAN), it cannot be directly used to model selection of LAMN models and a criterion for them is required. The proposed criterion for LAMN models is an asymptotically unbiased estimator of the Kullback-Leibler loss of Bayesian prediction. Simulation studies for a mixed normal model, a discretely observed diffusion model and a partially explosive Gaussian AR(2) model are given.

Keywords
Bayesian prediction, information criterion, Kullback-Leibler divergence, local asymptotic mixed normality, model selection.

1 Introduction

Consider a model $\mathcal{P}_n = \{p_n(\cdot|\theta) \mid \theta \in \Theta \subset \mathbb{R}^k\}$ $(n = 1, 2, \cdots)$ on a sequence of measure spaces $(\Omega_n, \mathcal{P}_n, \mu_n)$, where $p_n(\cdot|\theta)$ is a probability density with a parameter θ. The LAMN property defined below is regarded as a key concept throughout this paper.

Definition 1 (Jeganathan 1982) Let $\theta \in \Theta$. A model $(\mathcal{P}_n)_{n=1}^{\infty}$ is called *locally asymptotically mixed normal (LAMN)* at θ if there exist a sequence of matrices $\gamma_n = \gamma_{n,\theta} \in \mathbb{R}^{k\times k}$, a random matrix $J = J_\theta$ and a random vector ξ with $\xi|J \sim N(0, J^{-1})$ such that for any $h \in \mathbb{R}^k$ and any convergent sequence $h_n \to h$

$$\begin{aligned} \log \frac{p_n(x|\theta+\gamma_n h_n)}{p_n(x|\theta)} &= h' J_{n,\theta}\xi_{n,\theta} - \frac{1}{2}h' J_{n,\theta} h + \mathrm{o}_{p_\theta}(1), \\ (\xi_{n,\theta}, J_{n,\theta}) &\overset{\text{Law}}{\to} (\xi, J). \end{aligned}$$

Here $'$ denotes transpose of a vector. In particular, $(\mathcal{P}_n)_{n=1}^{\infty}$ is *locally asymptotically normal (LAN)* if J is deterministic.

Example 1. We give a trivial example (LeCam & Yang 2000, p.121). Let $x_1, \cdots, x_\nu$ be an independently and identically distributed (i.i.d.) sequence subject to the probability density $p(x_1|\theta)$ and ν be a random variable independent of x_i's. If ν/n weakly converges to a non-degenerate random variable c and $p(x_1|\theta)$ satisfies some mild conditions, the model has the LAMN property with $\gamma_n = 1/\sqrt{n}$ and $J = cJ_0$, where J_0 is the Fisher information matrix of $p(x_1|\theta)$.

Example 2 (Discretely observed diffusion models). Let X be the solution of the following 1-dimensional diffusion process

$$\mathrm{d}X_t \;=\; a(X_t,\theta)\mathrm{d}W_t + b(X_t,\theta)\mathrm{d}t, \quad X_0 \;=\; x_0, \quad t \in [0,1],$$

where x_0 is the fixed initial value of X, a and b are smooth bounded functions and W is a standard Wiener process. When θ is estimated from the discretely observed data X_{t_i}, where $t_i = i/n$ for $i = 1, \cdots, n$, it is known that the model has the LAMN property with $\gamma_n = 1/\sqrt{n}$ (Dohnal 1987, Genon-Catalot & Jacod 1994). The random Fisher information matrix is

$$J \;=\; 2\int_0^1 \left[\frac{\partial}{\partial\theta}\log a(X_t,\theta)\right]\left[\frac{\partial}{\partial\theta'}\log a(X_t,\theta)\right]\mathrm{d}t.$$

Example 3 (Partially explosive Gaussian AR models). Let us consider the Gaussian AR(2) model with known variance

$$\begin{aligned} X_t &= \beta_1 X_{t-1} + \beta_2 X_{t-2} + \varepsilon_t, \quad \varepsilon_t \overset{\text{i.i.d.}}{\sim} N(0,1), \quad t \in \{1, \cdots, n\}, \\ X_0 &= X_{-1} \;=\; 0. \end{aligned}$$

Let θ_1 and θ_2 be two roots of the characteristic equation $\theta^2 - \beta_1\theta - \beta_2 = 0$. Assume that $\theta_1 > 1 > |\theta_2|$. We use (θ_1, θ_2) as the parameter. Then the model is LAMN with the normalization matrix $\gamma_{n,\theta} = \mathrm{diag}(\theta_1^{-n}, n^{-1/2})$. The random Fisher information matrix is

$$J \;=\; \mathrm{diag}\left[\frac{\chi_1^2}{1-\theta_1^{-2}}, \frac{1}{1-\theta_2^2}\right],$$

where χ_1^2 is a random variable subject to the chi-square distribution with one degree of freedom. This result is generalized to any Gaussian AR(k) model for $k \geq 1$ (Jeganathan 1988, Theorem 16).

For examples other than described above, branching processes (See e.g. van der Vaart 1998) and some class of semimartingale models (Luschgy 1992) are LAMN.

The LAMN property implies the convergence of the likelihood ratio to that of the corresponding mixed normal model (van der Vaart 1998, Theorem 9.8). Therefore it allows us to reduce statistical problems to those of the mixed normal model at least formally. Several rigorous results including the convolution theorem and the local asymptotic minimax theorem (Jeganathan 1982, 1983) were obtained from this point of view.

The importance of statistical inference for LAMN models have been recognized in recent years because the LAMN property of the discretely observed diffusion model described in Example 2 will be extended to many models for time-series analysis and spatial statistics. The LAMN property of the discretely and randomly observed multivariate diffusion models with the commutative diffusion coefficients is proved by Genon-Catalot & Jacod (1994). Discretely observed models of multivariate diffusion processes with noncommutative diffusion coefficients, non-Markov processes and multiparameter stochastic processes may have the LAMN property. For example, the LAMN property of a model of multiparameter stochastic processes which is transformation of the Brownian sheet by a parametric function is proved by Sei (2004).

We propose an information criterion for LAMN models by studying the corresponding mixed normal model. Since the Akaike's Information Criterion (AIC) is derived based on the LAN property (Akaike 1974), it cannot be directly used to model selection of LAMN models. The proposed criterion *Bayes-LAMN-IC* for LAMN models is defined as an asymptotically unbiased estimator of the loss of Bayesian prediction. The loss function we adopt is equivalent to the Kullback-Leibler divergence. Here the Bayesian prediction is used since it dominates the plug-in predictive distribution as given in Section 3. We also give several other criteria based on other predictive distributions for comparison.

Some notations and assumptions are prepared in Section 2. For the mixed normal model, the Bayesian and some other predictive distributions are compared in Section 3. In Section 4, Bayes-LAMN-IC is defined for the mixed normal model. The criterion for non-limit models is given in Section 5. Simulation studies for the (not asymptotically) mixed normal model, the discretely observed diffusion model and the partially explosive Gaussian AR model are given in Section 6.

2 Notations and assumptions

We fix a full LAMN model $\{p_n(x|\theta) \mid \theta \in \Theta \subset \mathbb{R}^k\}$ and focus on its submodels. The corresponding full limit model is $\{p(\xi, J|h) = p(\xi|h, J)p(J) \mid h \in \mathbb{R}^k\}$, where h, ξ and J are defined in Definition 1. The conditional density $p(\xi|h, J)$ is $\phi(\xi|h, J^{-1})$, where $\phi(x|\mu, \Sigma)$ is the density of normal distribution with the mean vector μ and the covariance matrix Σ. The marginal density $p(J)$ of the random Fisher information matrix J does not depend on h from the definition. We use symbols indicated in Table 1.

Table 1: The symbols used in the paper.

	full model	submodel $\alpha \in A$
non-limit model	$\{p_n(x\|\theta) \mid \theta \in \Theta \subset \mathbb{R}^k\}$	$\{p_n(x\|\theta) \mid \theta \in \Theta_\alpha\}$
limit model	$\{p(\xi\|h, J) \mid h \in \mathbb{R}^k\}$	$\{p(\xi\|h, J) \mid h \in H_\alpha\}$

In the table, A is the index set of submodels. For each $\alpha \in A$, Θ_α is a k_α-dimensional subset in Θ, where $0 \leq k_\alpha \leq k$. Let θ_α be a smooth embedding map from $\mathbb{R}^{k_\alpha}$ to Θ_α and $B_\alpha \in \mathbb{R}^{k \times k_\alpha}$ is the derivative matrix of θ_α. The subspace corresponding to α in the limit model is denoted by $H_\alpha = \{h = B_\alpha u \mid u \in \mathbb{R}^{k_\alpha}\}$.

We denote $E = E_k$ as the identity matrix of size k. We put $J_\alpha^- = B_\alpha(B_\alpha' J B_\alpha)^{-1} B_\alpha'$ and $\pi_\alpha = J_\alpha^- J$. The matrix π_α is a (random) projection operator from $\mathbb{R}^k$ to H_α. A relation $\pi_\alpha J^{-1} \pi_\alpha' = J_\alpha^- J J_\alpha^- = J_\alpha^-$ holds.

We assume that the true parameter h of the limit model is an arbitrary point in $\mathbb{R}^k$. This corresponds to a local alternate in hypothesis testing. For each submodel $\alpha \in A$, we put $h_\alpha = \pi_\alpha h$ and $\xi_\alpha = \pi_\alpha \xi$ for the true parameter h and an observation ξ. The quantity ξ_α is the maximum likelihood estimator for the subspace H_α, whose conditional mean and variance are $\mathrm{E}[\xi_\alpha|J] = h_\alpha$ and $\mathrm{Var}[\xi_\alpha|J] = J_\alpha^-$, respectively. The random variable h_α is considered as "the true parameter in H_α" because it gives the nearest distribution in H_α to the true one. The phenomenon that the true parameter is random does not appear in the LAN situation.

We assume that the prior distribution $P_\alpha(\mathrm{d}h)$ for α is the uniform distribution on H_α. Use of the uniform prior for the limit model is natural in the sense that any smooth prior density for the non-limit model is locally approximated by the uniform prior density. The posterior distribution $P_\alpha(\mathrm{d}h|\xi, J)$ is the degenerate normal distribution

with mean ξ_α and variance J_α^-, since its characteristic function is

$$\begin{aligned}\psi_{h|\xi,J}(\lambda) &:= \frac{\int \exp(\mathrm{i}\lambda' h)\, p(\xi|h,J)\, P_\alpha(\mathrm{d}h)}{\int p(\xi|h,J)\, P_\alpha(\mathrm{d}h)} \\ &= \frac{\int_{\mathbb{R}^{k_\alpha}} \exp\left[\mathrm{i}\lambda' B_\alpha u - \frac{1}{2}(\xi - B_\alpha u)' J (\xi - B_\alpha u)\right] \mathrm{d}u}{\int_{\mathbb{R}^{k_\alpha}} \exp\left[-\frac{1}{2}(\xi - B_\alpha u)' J (\xi - B_\alpha u)\right] \mathrm{d}u} \\ &= \exp\left[\mathrm{i}\lambda'\xi_\alpha - \frac{1}{2}\lambda' J_\alpha^- \lambda\right].\end{aligned}$$

3 Risk of prediction

In this section and the next section, we consider prediction problem for limit models. The problem is prediction of $(\eta, \tilde{J})$ from an observation (ξ, J), where $(\eta, \tilde{J})$ and (ξ, J) are independently and identically distributed with true parameter $h \in \mathbb{R}^k$. Since the distributions of the random information matrices J and $\tilde{J}$ are independent of h, they are considered as ancillary statistics. Thus the prediction problem is reduced to that of η from ξ conditionally on J and $\tilde{J}$. When J and $\tilde{J}$ are conditioned, the arguments are usually abbreviated, for example, $q(\eta|\xi) = q(\eta|\xi, J, \tilde{J})$. Expectations are taken conditionally on J and $\tilde{J}$ unless otherwise stated.

The loss of a predictive distribution $q(\eta|\xi)$ is defined by

$$l(q(\cdot|\xi)) = -2\int p(\eta|h) \log q(\eta|\xi)\, \mathrm{d}\eta,$$

which is equivalent to the Kullback-Leibler divergence $\int p(\eta|h) \log(p(\eta|h)/q(\eta|\xi))\mathrm{d}\eta$. The risk is denoted by $r(q) = \int p(\xi|h) l(q(\cdot|\xi))\, \mathrm{d}\xi$.

We construct four predictive distributions by classifying Bayesian or plug-in, and LAMN or LAN.

Definition 2 The *Bayes-LAMN, plugin-LAMN, Bayes-LAN and plugin-LAN distributions* are defined by

$$\begin{aligned} q_\alpha^{\mathrm{B}}(\eta|\xi) &= \int p(\eta|h,\tilde{J})\, P_\alpha(\mathrm{d}h|\xi,J), \\ q_\alpha^{\mathrm{p}}(\eta|\xi) &= p(\eta|\xi_\alpha,\tilde{J}), \\ q_\alpha^{\mathrm{BN}}(\eta|\xi) &= \int p(\eta|h,J)\, P_\alpha(\mathrm{d}h|\xi,J), \\ q_\alpha^{\mathrm{pN}}(\eta|\xi) &= p(\eta|\xi_\alpha,J), \end{aligned}$$

respectively.

Lemma 1 *The predictive distributions defined in Definition 2 are expressed explicitly by*

$$\begin{aligned} q_\alpha^{\mathrm{B}}(\eta|\xi) &= \phi(\eta|\xi_\alpha, \tilde{J}^{-1}+J_\alpha^-), \\ q_\alpha^{\mathrm{p}}(\eta|\xi) &= \phi(\eta|\xi_\alpha, \tilde{J}^{-1}), \\ q_\alpha^{\mathrm{BN}}(\eta|\xi) &= \phi(\eta|\xi_\alpha, J^{-1}+J_\alpha^-), \\ q_\alpha^{\mathrm{pN}}(\eta|\xi) &= \phi(\eta|\xi_\alpha, J^{-1}), \end{aligned}$$

respectively.

Proof. The first expression is obtained by using the characteristic function

$$\begin{aligned} \psi_{\eta|\xi}^{\mathrm{B}}(\lambda) &:= \int \exp(\mathrm{i}\lambda'\eta)\, q_\alpha^{\mathrm{B}}(\eta|\xi)\, \mathrm{d}\eta \\ &= \iint \exp(\mathrm{i}\lambda'\eta)\, p(\eta|h, \tilde{J})\, P_\alpha(\mathrm{d}h|\xi, J)\, \mathrm{d}\eta \\ &= \int \exp\left[\mathrm{i}\lambda' h - \frac{1}{2}\lambda'\tilde{J}^{-1}\lambda\right]\, P_\alpha(\mathrm{d}h|\xi, J) \\ &= \exp\left[\mathrm{i}\lambda'\xi_\alpha - \frac{1}{2}\lambda'(\tilde{J}^{-1}+J_\alpha^-)\lambda\right]. \end{aligned}$$

The other expressions are also easily obtained. □

We introduce a class of predictive distributions including the four predictive distributions considered above.

Definition 3 Let $\Sigma_\alpha = \Sigma_\alpha(J, \tilde{J})$ be a $k \times k$ positive definite matrix. Then the *Σ-predictive distribution* is defined by

$$q_\alpha^{\Sigma}(\eta|\xi) = \phi(\eta|\xi_\alpha, \Sigma_\alpha).$$

The Bayes-LAMN, plugin-LAMN, Bayes-LAN and plugin-LAN distributions are Σ-predictive distributions with

$$\begin{aligned} \Sigma_\alpha^{\mathrm{B}} &= \tilde{J}^{-1} + J_\alpha^-, \\ \Sigma_\alpha^{\mathrm{p}} &= \tilde{J}^{-1}, \\ \Sigma_\alpha^{\mathrm{BN}} &= J^{-1} + J_\alpha^-, \\ \Sigma_\alpha^{\mathrm{pN}} &= J^{-1}, \end{aligned}$$

respectively.

The next two lemmas about the loss and risk of the Σ-predictive distributions are obtained by an elementary calculation.

Lemma 2 *Let $h \in \mathbb{R}^k$. The loss of the predictive distribution q_α^Σ is*

$$l(q_\alpha^\Sigma(\cdot|\xi)) \;=\; (h-\xi_\alpha)'\Sigma_\alpha^{-1}(h-\xi_\alpha) + \mathrm{tr}[\Sigma_\alpha^{-1}\tilde{J}^{-1}] + \log\det\Sigma_\alpha.$$

Lemma 3 *Let $h \in \mathbb{R}^k$. The risk of the predictive distribution q_α^Σ is*

$$r(q_\alpha^\Sigma) \;=\; (h-h_\alpha)'\Sigma_\alpha^{-1}(h-h_\alpha) + \mathrm{tr}[\Sigma_\alpha^{-1}(\tilde{J}^{-1}+J_\alpha^-)] + \log\det\Sigma_\alpha.$$

The next theorem reveals superiority of the Bayes-LAMN prediction q_α^{B} in a certain sense. Therefore we use the Bayes-LAMN distribution for the prediction problem throughout the paper.

Theorem 1 *(i) Let $h \in \mathbb{R}^k$. Then*

$$r(q_\alpha^{\mathrm{B}}) \;<\; r(q_\alpha^{\mathrm{p}}).$$

(ii) Let $h \in H_\alpha$. Then

$$r(q_\alpha^{\mathrm{B}}) \;\le\; r(q_\alpha^\Sigma)$$

for any $k \times k$ positive definite matrix Σ_α. The equality holds if and only if $q_\alpha^\Sigma = q_\alpha^{\mathrm{B}}$.

Proof. Let $h \in \mathbb{R}^k$ and let $\Sigma = \Sigma_\alpha$ be any $k \times k$ positive definite matrix. By Lemma 1 and Lemma 3,

$$\begin{aligned} & r(q_\alpha^\Sigma) - r(q_\alpha^{\mathrm{B}}) \\ &= (h-h_\alpha)'(\Sigma^{-1}-(\tilde{J}^{-1}+J_\alpha^-)^{-1})(h-h_\alpha) \\ &\quad +\mathrm{tr}[\Sigma^{-1/2}(\tilde{J}^{-1}+J_\alpha^-)\Sigma^{-1/2}] - k - \log\det[\Sigma^{-1/2}(\tilde{J}^{-1}+J_\alpha^-)\Sigma^{-1/2}] \\ &\ge (h-h_\alpha)'(\Sigma^{-1}-(\tilde{J}^{-1}+J_\alpha^-)^{-1})(h-h_\alpha) \end{aligned}$$

because of an inequality

$$\mathrm{tr}\, C - k - \log\det C \ge 0$$

for any non-negative definite matrix C, where the equality holds if and only if $C = E$. If $\Sigma = \tilde{J}^{-1}$, then $(h-h_\alpha)'(\Sigma^{-1}-(\tilde{J}^{-1}+J_\alpha^-)^{-1})(h-h_\alpha) \ge 0$ for any $h \in \mathbb{R}^k$. Thus (i) holds. On the other hand, if $h \in H_\alpha$, then $h = h_\alpha$. Thus (ii) holds. □

Remark. The difference between $r(q_\alpha^{\mathrm{p}})$ and $r(q_\alpha^{\mathrm{B}})$ is quite large if J is close to zero relative to $\tilde{J}$. Let $h \in H_\alpha$ for simplicity. If the maximum eigenvalue of $C_\alpha = \tilde{J}^{1/2}(\tilde{J}^{-1}+J_\alpha^-)\tilde{J}^{1/2}$ is $\bar{\lambda} > 1$, then the difference is assessed as

$$\begin{aligned} r(q_\alpha^{\mathrm{p}}) - r(q_\alpha^{\mathrm{B}}) &= \mathrm{tr}C_\alpha - k - \log\det C_\alpha \\ &\ge \bar{\lambda} - 1 - \log\bar{\lambda}. \end{aligned}$$

On the other hand, the expectation of twice the Kullback-Leibler risk of q_α^{B} is

$$\begin{aligned} r(q_\alpha^{\mathrm{B}}) - r(p) &= 2 \iint p(\eta|h,\tilde{J}) \log \frac{p(\eta|h,\tilde{J})}{q_\alpha^{\mathrm{B}}(\eta|\xi)} \,\mathrm{d}\eta \,\mathrm{d}\xi \\ &= k + \log\det(\tilde{J}^{-1} + J_\alpha^-) - k - \log\det \tilde{J}^{-1} \\ &= \log\det C_\alpha \\ &\leq k \log \bar{\lambda}. \end{aligned}$$

Thus

$$\frac{r(q_\alpha^{\mathrm{p}}) - r(q_\alpha^{\mathrm{B}})}{r(q_\alpha^{\mathrm{B}}) - r(p)} \;\geq\; \frac{\bar{\lambda} - 1 - \log\bar{\lambda}}{k \log \bar{\lambda}} \;\to\; \infty$$

as $\bar{\lambda} \to \infty$. Similarly, the difference between $r(q_\alpha^{\Sigma})$ and $r(q_\alpha^{\mathrm{B}})$ is very large for any Σ if J is close to zero relative to $\tilde{J}$.

4 Proposed information criterion

We introduce an information criterion Bayes-LAMN-IC for limit models, which forms $-2\log q_\alpha^{\mathrm{B}}(\xi|\xi) + c_\alpha$ with some correcting term c_α. Expectations are taken conditionally on J and $\tilde{J}$ unless otherwise stated.

We first define criteria based on Σ-predictive distributions. Bayes-LAMN-IC is a special case of them.

Definition 4 Fix Σ_α. An information criterion *Σ-IC* is defined by

$$\begin{aligned} \Sigma\text{-IC} &= \Sigma\text{-IC}(\alpha) = \Sigma\text{-IC}(\alpha, \xi, J, \tilde{J}) \\ &:= (\xi - \xi_\alpha)' \Sigma_\alpha^{-1} (\xi - \xi_\alpha) + \log\det \Sigma_\alpha + \mathrm{tr}[\Sigma_\alpha^{-1}(\tilde{J}^{-1} - J^{-1} + 2J_\alpha^-)]. \end{aligned} \tag{1}$$

The selected model by the criterion is denoted by

$$\hat{\alpha}^{\Sigma} = \hat{\alpha}^{\Sigma}(\xi) = \hat{\alpha}^{\Sigma}(\xi, J, \tilde{J}) := \operatorname*{argmin}_{\alpha \in A} \Sigma\text{-IC}(\alpha).$$

In particular, for $i \in \{\mathrm{B}, \mathrm{p}, \mathrm{BN}, \mathrm{pN}\}$, *$\Sigma^i$-IC* is called *Bayes-LAMN-IC*, *plugin-LAMN-IC*, *Bayes-LAN-IC* and *plugin-LAN-IC*, respectively.

Theorem 2 *The information criterion* Σ-IC(α) *is the unique unbiased estimator of the risk* $r(q_\alpha^{\Sigma})$.

Proof. Put $c_\alpha = \mathrm{tr}[\Sigma_\alpha^{-1}(\tilde{J}^{-1} - J^{-1} + 2J_\alpha^-)]$. The expectation of Σ-IC(α) is

$$\begin{aligned}
&\int p(\xi|h,J)\,\Sigma\text{-IC}(\alpha)\,\mathrm{d}\xi \\
&= \int p(\xi|h,J)\left[(\xi-\xi_\alpha)'\Sigma_\alpha^{-1}(\xi-\xi_\alpha)\right]\,\mathrm{d}\xi + \log\det\Sigma_\alpha + c_\alpha \\
&= (h-h_\alpha)'\Sigma_\alpha^{-1}(h-h_\alpha) + \mathrm{tr}\left[\Sigma_\alpha^{-1}(J^{-1}-J_\alpha^-)\right] + \log\det\Sigma_\alpha + c_\alpha \\
&= (h-h_\alpha)'\Sigma_\alpha^{-1}(h-h_\alpha) + \mathrm{tr}\left[\Sigma_\alpha^{-1}(\tilde{J}^{-1}+J_\alpha^-)\right] + \log\det\Sigma_\alpha \\
&= r(q_\alpha^\Sigma)
\end{aligned}$$

by Lemma 3. Uniqueness holds due to the completeness of the statistic ξ (Lehmann & Casella 1998, p.42 and p.87). □

Proposition 3 *The information criterion* plugin-LAN-IC *is equivalent to* AIC.

Proof. By putting $\Sigma_\alpha = J^{-1}$ in (1), it is shown that

$$\text{plugin-LAN-IC}(\alpha) \;=\; -2\log q_\alpha^{\mathrm{pN}}(\xi|\xi) + 2k_\alpha + \mathrm{tr}[J(\tilde{J}^{-1}-J^{-1})].$$

Since the last term is independent of α, plugin-LAN-IC is equivalent to AIC. □

We adopt Bayes-LAMN-IC among Σ-IC's because it is compatible with the Bayes-LAMN prediction, which has the dominating property obtained in Theorem 1. It should be noted that the criterion does not coincide with AIC even if a LAN model is considered. For LAN models, both Bayes-LAMN-IC and Bayes-LAN-IC coincide with

$$\mathrm{PIC}_2 \;=\; -2\log q_\alpha^{\mathrm{BN}}(\xi|\xi) + k_\alpha, \tag{2}$$

in Kitagawa (1997) when the uniform prior is used. It is also found in Akaike (1980, eq. (3.8)). The performance of PIC_2 and AIC does not seem very different since J is deterministic. On the other hand, for LAMN models, the difference between Bayes-LAMN-IC and AIC is quite serious as remarked after Theorem 1.

We compare Σ-IC's for different Σ's in Section 6. The risk R of the model selection procedure based on Σ-IC is defined by the risk of the Bayesian prediction based on the model selection procedure using Σ-IC, that is,

$$R \;=\; R(\Sigma,h) \;=\; \mathrm{E}[r(q_{\hat{\alpha}^\Sigma}^{\mathrm{B}})], \tag{3}$$

where E denotes the expectation with respect to J and $\tilde{J}$. An information criterion Σ-IC whose risk R is small is a good criterion.

In practice, a model selection is implemented without use of $\tilde{J}$. For this purpose, we define an expectation version of Σ-IC by

$$\int \Sigma\text{-IC}(\alpha, \xi, J, \tilde{J}) p(\tilde{J}) \mathrm{d}\tilde{J}.$$

We compare only Σ-IC for the numerical examples in Section 6 since the performance of an information criterion is assessed by its performance of prediction.

5 Information criterion for non-limit models

In this section, the information criteria for limit models defined in the previous section are restored to those for the non-limit models. Only a heuristic definition is given here. The conditions for asymptotic properties such as contiguity of the selected predictive distribution are not discussed.

Let $\hat{\theta}$ and $\hat{\theta}_\alpha$ be the maximum likelihood estimators (or other asymptotically efficient estimators) for the full model and the submodel $\alpha \in A$, respectively. Our Σ-IC for the non-limit models is, by using Σ-IC for the limit models,

$$\Sigma\text{-IC}(\alpha)|_{J=J^{(n)}, \tilde{J}=\tilde{J}^{(n)}, (\xi-\xi_\alpha)=(\xi-\xi_\alpha)^{(n)}}, \tag{4}$$

where a matrix $J^{(n)}$ is defined by

$$J^{(n)} \quad := \quad -\frac{\partial^2}{\partial u \partial u'} \log p_n(x|\hat{\theta} + \gamma_n u)\Big|_{u=0}, \tag{5}$$

$\tilde{J}^{(n)}$ is given by replacing x in (5) with y and

$$(\xi - \xi_\alpha)^{(n)} \quad := \quad \gamma_n^{-1}(\hat{\theta} - \hat{\theta}_\alpha).$$

Under mild conditions, $J^{(n)}$, $\tilde{J}^{(n)}$ and $(\xi - \xi_\alpha)^{(n)}$ converge to J, $\tilde{J}$ and $\xi - \xi_\alpha$ as $n \to \infty$, respectively.

In general, $J^{(n)}$ and $\tilde{J}^{(n)}$ may not be positive definite. Therefore some modification is needed. For the discretely observed diffusion models, we can use a non-negative definite matrix $J^{\sharp(n)}$ instead of $J^{(n)}$ defined by

$$J^{\sharp(n)} \quad := \quad \sum_{i=1}^{n} \frac{2}{n} \left[\frac{\partial}{\partial \theta} \log a(X_{t_i}, \hat{\theta}) \frac{\partial}{\partial \theta'} \log a(X_{t_i}, \hat{\theta}) \right]. \tag{6}$$

The matrix $J^{\sharp(n)}$ is used at the numerical experiments in Subsection 6.2.

6 Examples

Three examples are considered here. In Subsection 6.1, some theoretical and experimental results for a special limit model are given. Subsection 6.2 is devoted to a numerical study of the discretely observed diffusion models. Subsection 6.3 deals with the partially explosive Gaussian AR model.

6.1 Scalar-randomness model

Let $J = cJ_0$ with a 1-dimensional positive random variable c and a deterministic matrix J_0 as Example 1 in Section 1. We call it a scalar-randomness model. If we consider nested submodels, the information criterion has a simple representation. Let $A = \{0, 1, \cdots, k\}$. Suppose that $H_0 = \{0\} \subset \mathbb{R}^k$ and H_α is an α-dimensional subspace of $\mathbb{R}^k$ including $H_{\alpha-1}$ for $1 \leq \alpha \leq k$. Put $\tilde{J} = \tilde{c}J_0$ where $\tilde{c}$ is a random variable independent of c and has the same distribution as c.

We assume that $J_0 = E$ and $H_\alpha = \{(a_1, \cdots, a_\alpha, 0, \cdots, 0) \mid a_1, \cdots, a_\alpha \in \mathbb{R}\}$ without loss of generality. The loss $l(q_\alpha^{\mathrm{B}}(\cdot|\xi))$ of the Bayes prediction is, by Lemma 2,

$$\begin{aligned} l(q_\alpha^{\mathrm{B}}(\cdot|\xi)) &= \sum_{i=1}^{\alpha} \left[(\tilde{c}^{-1}+c^{-1})^{-1}\{(h_i-\xi_i)^2+\tilde{c}^{-1}\} + \log(\tilde{c}^{-1}+c^{-1})\right] \\ &\quad + \sum_{i=\alpha+1}^{k} \left[\tilde{c}h_i^2 + 1 + \log \tilde{c}^{-1}\right], \end{aligned}$$

where ξ_i and h_i is the i-th component of ξ and h, respectively.

We consider only Σ-IC satisfying the condition that Σ_α is a diagonal matrix whose i-th diagonal component σ_i is $s = s(c, \tilde{c})$ if $i \leq \alpha$ and $t = t(c, \tilde{c})$ otherwise, where s and t are common in all α. The four criteria (Bayes-LAMN, plugin-LAMN, Bayes-LAN and plugin-LAN) satisfy the condition as indicated in Table 2. The expression of Σ-IC is

$$\begin{aligned} \Sigma\text{-IC}(\alpha) &= \sum_{i=1}^{\alpha} \left[s^{-1}(\tilde{c}^{-1}+c^{-1}) + \log s\right] + \sum_{i=\alpha+1}^{k} \left[t^{-1}\xi_i^2 + t^{-1}(\tilde{c}^{-1}-c^{-1}) + \log t\right] \\ &= \sum_{i=1}^{k} \left[s^{-1}(\tilde{c}^{-1}+c^{-1}) + \log s\right] + t^{-1} \sum_{i=\alpha+1}^{k} (\xi_i^2 - \bar{\xi}^2), \end{aligned}$$

where

$$\bar{\xi}^2 = t\left[s^{-1}(\tilde{c}^{-1}+c^{-1}) + \log s - t^{-1}(\tilde{c}^{-1}-c^{-1}) - \log t\right].$$

For fixed $\alpha \in A$, the set Ξ_α of ξ such that $\hat{\alpha}^\Sigma(\xi) = \alpha$ is given by

$$\Xi_\alpha = L_\alpha \cap G_\alpha,$$

where

$$L_\alpha \;=\; \{\xi \mid \forall j \le \alpha,\ \sum_{i=j+1}^{\alpha} (\xi_i^2 - \bar{\xi}^2) > 0\}$$

and

$$G_\alpha \;=\; \{\xi \mid \alpha < \forall j \le k,\ \sum_{i=\alpha+1}^{j} (\xi_i^2 - \bar{\xi}^2) < 0\}.$$

The quantities s, t and $\bar{\xi}^2$ corresponding to the four criteria are summarized in Table 2, where we put $r = \tilde{c}/c$.

Table 2: The quantities s, t and $\bar{\xi}^2$ corresponding to the four predictive distributions. $(r = \tilde{c}/c)$

prediction	s	t	ξ^2
Bayes-LAMN	$\tilde{c}^{-1} + c^{-1}$	$\tilde{c}^{-1}$	$c^{-1}(r^{-1}\log(1+r)+1)$
plugin-LAMN	$\tilde{c}^{-1}$	$\tilde{c}^{-1}$	$2c^{-1}$
Bayes-LAN	$2c^{-1}$	c^{-1}	$c^{-1}(\log 2 + \frac{3}{2} - \frac{1}{2r})$
plugin-LAN	c^{-1}	c^{-1}	$2c^{-1}$

If $s = t$, then $\bar{\xi}^2 = 2c^{-1}$, which is independent of s. Thus the next proposition holds.

Proposition 4 *Suppose that the model is a scalar-randomness model and Σ_α is common in all $\alpha \in A$. Then* Σ-IC *is equivalent to* AIC.

Consider the scalar-randomness model with dimension $k = 10$. Assume that c takes only two values $\sqrt{10}$ and $1/\sqrt{10}$ with the same probability. We numerically evaluate the risk R of the model selection procedures (eq. (3)).

Figure 1 indicates R of FULL (which always selects the full model: $\bar{\xi}^2 = 0$), Bayes-LAMN-IC, Bayes-LAN-IC, AIC and BOUND (which is a lower bound based on the "best selection" $\hat{\alpha} = \mathrm{argmin}\ l(q_\alpha^{\mathrm{B}})$ using the true h), respectively. The true parameter h takes its value in $D_i = \{d_i e_i \mid d_i \in [0, 10]\}$ for $i \in \{1, \cdots, 10\}$, where $e_i = (0, \cdots, 0, 1, 0, \cdots, 0)$ is the i-th unit vector in $\mathbb{R}^{10}$. The horizontal axis denotes d_i such that $h = d_i e_i$.

The figure shows that Bayes-LAMN-IC is better than AIC especially for $h \in D_i$ $(i = 4, \cdots, 10)$. The minimax criterion is FULL. However, the difference between risks of Bayes-LAMN-IC and BOUND is stable throughout the parameter space compared

to that of FULL and BOUND. This kind of stability is considered important from the view point of model selection. Thus Bayes-LAMN-IC has a good performance in the example.

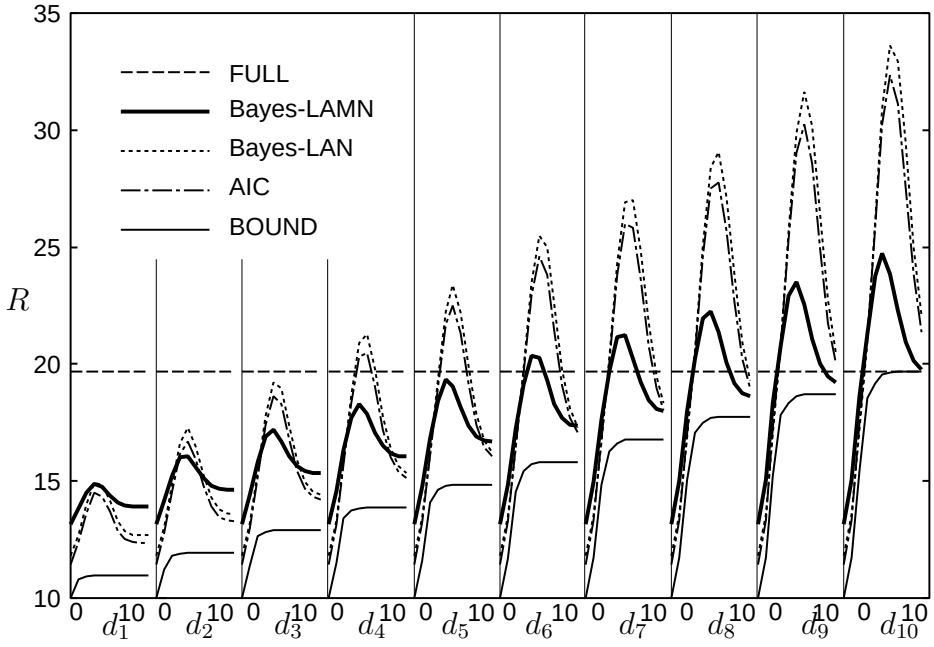

Figure 1: The risk R of the model selection procedures for the scalar-randomness model. The true parameter h takes its value in $D_i = \{(0, \cdots, 0, d_i, 0, \cdots, 0) \in \mathbb{R}^{10} \mid d_i \in [0, 10]\}$ for each $i \in \{1, \cdots, 10\}$, where $(0, \cdots, 0, d_i, 0, \cdots, 0)$ denotes the vector whose i-th coordinate is d_i. The horizontal axis denotes d_i such that $h = (0, \cdots, 0, d_i, 0, \cdots, 0)$.

6.2 Discretely observed diffusion models

Let us consider the discretely observed diffusion model stated in Example 2 of Section 1. The example used here is

$$\mathrm{d}X_t = \frac{1 + \theta X_t^2}{1 + X_t^2}\mathrm{d}W_t, \quad X_0 = 0, \quad \theta > 0, \tag{7}$$

which satisfies the regularity condition in Genon-Catalot & Jacod (1994). Two submodels $\Theta_{\mathrm{I}} = \{\theta \mid \theta = 1\}$ and $\Theta_{\mathrm{II}} = \{\theta \mid \theta > 0\}$ are compared, where $A = \{\mathrm{I}, \mathrm{II}\}$ is the index set. If $\theta = 1$, $X_t = W_t$.

Figure 2 gives a numerical result about the risk R of the model selection procedure based on Bayes-LAMN-IC and AIC. The simulation algorithm for Σ-IC is as follows.

1. Fix integers L and $\tilde{L}$. For each $l = 1, 2, \cdots, L$,

(a) Generate a path $\{X_t(l) \mid t \in \{\frac{1}{n}, \cdots, \frac{n}{n}\}\}$ according to the true parameter θ. Calculate the maximum likelihood estimator $\hat{\theta}(l)$ for the full model and the random Fisher information $J^{\sharp(n)}(l)$ by the formula (6).

(b) For $\tilde{l} = 1, 2, \cdots, \tilde{L}$, generate $\tilde{L}$ paths $\{Y_t(l, \tilde{l}) \mid t \in \{\frac{1}{n}, \cdots, \frac{n}{n}\}\}$ according to the estimated parameter $\hat{\theta}(l)$. Calculate the random Fisher information $\tilde{J}^{\sharp(n)}(l, \tilde{l})$.

(c) Select one of the submodels according to Σ-IC determined by eq. (4) and calculate the loss $\ell(l, \tilde{l})$ of selected predictive distribution, where the loss is also calculated by LAMN approximation for simplicity.

2. Calculate $R = (L\tilde{L})^{-1} \sum_{l=1}^{L} \sum_{\tilde{l}=1}^{\tilde{L}} \ell(l, \tilde{l})$.

The number of sampling points is $n = 100$. The number of loops is $L = \tilde{L} = 1000$ for each true $\theta \in \{0.25, 0.50, \cdots, 3.00\}$. In the example, Bayes-LAMN-IC is better than AIC in the minimax sense.

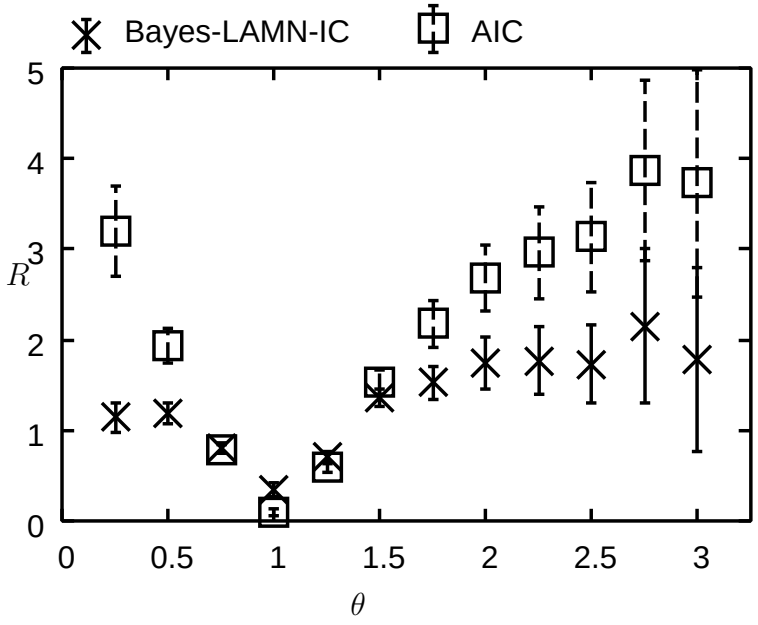

Figure 2: The risk R of the model selection procedures for the discretely observed diffusion model (eq. (7)). The confidence interval is based on 3 times of the standard deviation.

6.3 The partially explosive Gaussian AR model

Let us consider the partially explosive Gaussian AR(2) model stated in Example 3 of Section 1. Two submodels $\Theta_{\mathrm{I}} = \{\theta \mid \theta_1 > 1,\ \theta_2 = 0\}$ and $\Theta_{\mathrm{II}} = \{\theta \mid \theta_1 >$

1, $|\theta_2| < 1\}$ are compared, where $A = \{\mathrm{I}, \mathrm{II}\}$ is the index set. Let $J = \mathrm{diag}(J_{11}, J_{22})$, $\tilde{J} = \mathrm{diag}(\tilde{J}_{11}, \tilde{J}_{22})$ and $\xi = (\xi_1, \xi_2)'$.

Since $\tilde{J}_{22} = J_{22}$, Bayes-LAMN-IC's for the two submodels are

$$\begin{aligned} \text{Bayes-LAMN-IC(I)} &= \xi_2^2(2J_{22}^{-1})^{-1} + \log(\tilde{J}_{11}^{-1} + J_{11}^{-1}) + \log(J_{22}^{-1}) + 1, \\ \text{Bayes-LAMN-IC(II)} &= \log(\tilde{J}_{11}^{-1} + J_{11}^{-1}) + \log(2J_{22}^{-1}) + 2. \end{aligned}$$

Their difference is $\xi_2^2 J_{22}/2 - (\log 2 + 1)$. On the other hand,

$$\begin{aligned} \text{Bayes-LAN-IC(I)} &= \xi_2^2(2J_{22}^{-1})^{-1} + \log(2J_{11}^{-1}) + \log(J_{22}^{-1}) + 1, \\ \text{Bayes-LAN-IC(II)} &= \log(2J_{11}^{-1}) + \log(2J_{22}^{-1}) + 2. \end{aligned}$$

Their difference is $\xi_2^2 J_{22}/2 - (\log 2 + 1)$. Therefore both Bayes-LAMN-IC and Bayes-LAN-IC are equivalent to PIC_2 (eq. (2)). In particular, $\tilde{J}$ is not needed in order to calculate them. Similarly, both plugin-LAMN-IC and plugin-LAN-IC are equivalent to AIC. These properties hold for any AR(k) model if we consider only submodels where some of the stationary components of θ are restricted to zero.

We now compare Bayes-LAMN-IC and AIC by finite-sample experiments. Figure 3 gives a numerical result about the risk R of the model selection procedure based on Bayes-LAMN-IC and AIC. The simulation algorithm for Bayes-LAMN-IC is as follows. A similar algorithm is used for AIC.

1. Fix L and $\tilde{L}$. For each $l = 1, \cdots, L$,
   (a) Generate a path $\{X_t(l) \mid t \in \{1, \cdots, n\}\}$ according to the true parameter θ. Calculate the maximum likelihood estimator $\hat{\theta}_\alpha(l)$ and Bayes-LAMN-IC(α) for each model $\alpha \in A$.
   (b) Calculate the loss $\ell(l)$ by the Monte-Carlo method, that is, generate $\tilde{L}$ paths $\{Y_t(l, \tilde{l}) \mid t \in \{1, \cdots, n\}\}$ $(\tilde{l} = 1, \cdots, \tilde{L})$ according to the true parameter θ and take the sample mean: $\ell(l) = \tilde{L}^{-1} \sum_{\tilde{l}=1}^{\tilde{L}} 2 \log\{p_n(Y|\theta)/q_n^{\mathrm{B}}(Y|X)\}$, where $q_n^{\mathrm{B}}(Y|X)$ is the selected predictive distribution by Bayes-LAMN-IC.
2. Calculate $R = L^{-1} \sum_{l=1}^{L} \ell(l)$.

The number of sampling points is $n = 100$. The number of loops is $L = \tilde{L} = 1000$ for each $\theta = (\theta_1, \theta_2) \in \{1.03\} \times \{0.00, 0.05, 0.10, \cdots, 1.00\}$. In the example, Bayes-LAMN-IC is slightly better than AIC in the minimax sense.

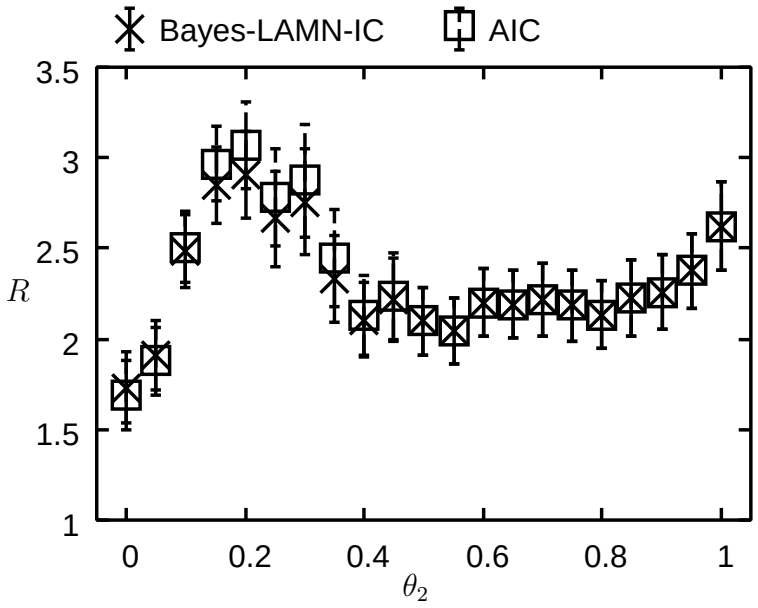

Figure 3: The risk R of the model selection procedures for the partially explosive Gaussian AR(2) model. The horizontal axis denotes θ_2. The value of θ_1 is fixed to 1.03. The confidence interval is based on 3 times of the standard deviation.

7 Conclusion and future works

We proposed an information criterion Bayes-LAMN-IC for LAMN models. It is the unique unbiased estimator of the risk of the Bayesian prediction. We numerically compared it with other criteria including AIC. For the scalar-randomness model, the risk of the model selection procedures based on Bayes-LAMN-IC was relatively stable over the true parameter space. For the discretely observed diffusion model and the partially explosive Gaussian AR model, the maximum risk of the model selection procedure based on Bayes-LAMN-IC was less than the maximum risk of the procedure based on the other criteria. These numerical results show that Bayes-LAMN-IC is better than the other criteria.

The remaining tasks are to give many numerical experiments, real data analysis, theoretical evaluation of the risk and characterization of Bayes-LAMN-IC. Another future work is to construct a version of Bayes-LAMN-IC like Takeuchi's information criterion (Burnham & Anderson 2002, p. 65). It is naturally constructed for the i.i.d. models with random number of samples. We believe that it is also available for the discretely observed diffusion models.

References

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

[2] Akaike, H. (1980). On the use of the predictive likelihood of a Gaussian model, *Ann. Inst. Statist. Math.*, **32**, 311-324.

[3] Burnham, K. P. and Anderson D. R. (2002). *Model selection and multimodel inference - A practical information-theoretic approach*, 2nd ed., Springer-Verlag, New York.

[4] Dohnal, G. (1987). On estimating the diffusion coefficient, *J. Appl. Probab.*, **24**, 105-114.

[5] Genon-Catalot, V. and Jacod, J. (1994). Estimation of the diffusion coefficient for diffusion processes: random sampling, *Scand. J. Statist.*, **21**, 193-221.

[6] Jeganathan, P. (1982). On the asymptotic theory of estimation when the limit of the log-likelihood ratios is mixed normal, *Sankhyā, Series A*, **44**, 173-212.

[7] Jeganathan, P. (1983). Some asymptotic properties of risk functions when the limit of the experiment is mixed normal, *Sankhyā, Series A*, **45**, 66-87.

[8] Jeganathan, P. (1988). On the strong approximation of the distributions of estimators in linear stochastic models, I and II: stationary and explosive AR models, *Ann. Statist.*, **16**, 1283-1314.

[9] Kitagawa, G. (1997). Information criteria for the predictive evaluation of Bayesian models, *Commun. Statist. - Theory Meth.*, **26**, 2223-2246.

[10] LeCam, L. and Yang, G. L. (2000), *Asymptotics in statistics*, 2nd ed., Springer-Verlag, New York.

[11] Lehmann, E. L. and Casella, G. (1998). *Theory of point estimation*, 2nd ed., Springer-Verlag, New York.

[12] Luschgy, H. (1992). Local asymptotic mixed normality for semimartingale experiments, *Probab. Theory Related Fields*, **92**, 151-176.

[13] Sei, T. (2004). Local asymptotic mixed normality for a model of discretely observed multiparameter stochastic processes, In preparation.

[14] A. W. van der Vaart, (1998). *Asymptotic statistics*, Cambridge University Press, Cambridge.