MEITMA International Workshop on Spatio-Temporal Modelling
Santiago de Compostela
30 Jun - 2 Jul 2010

Some Asymptotics for Geostatistical Model Selection

Zhi-Hao Zhang[1], Hsin-Cheng Huang[2,*] and Ching-Kang Ing[2]

[1] *1001 Ta Hsueh Road, Institute of Statistics, Hsinchu 300, Taiwan; jhow@stat.sinica.edu.tw*
[2] *128 Section 2 Academia Road, Institute of Statistical Science, Academia Sinica, Taipei 115, Taiwan; hchuang@stat.sinica.edu.tw, cking@stat.sinica.edu.tw*
[*] *Corresponding author*

Abstract. *Information criteria such as AIC and BIC are often used for model selection. However, their asymptotic behaviors in geostatistical model selection have not been well studied. In this article, we provide some asymptotic results for the generalized information criterion, including both AIC and BIC.*

Keywords. *Akaike information criterion; Asymptotic efficiency; Bayesian information criterion; Consistency; Variable selection.*

1 Geostatistical Models

Consider a spatial process $\{S(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ of interest defined over a d-dimensional region $D \subset \mathbb{R}^d$. Suppose that we observe data $\boldsymbol{Z} = (Z(\boldsymbol{s}_1), \ldots, Z(\boldsymbol{s}_n))'$ at locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n \in D$, according to the measurement equation:

$$Z(\boldsymbol{s}_i) = S(\boldsymbol{s}_i) + \varepsilon(\boldsymbol{s}_i); \quad i = 1, \ldots, n,$$

where $\varepsilon(\boldsymbol{s}_1), \ldots, \varepsilon(\boldsymbol{s}_n) \sim N(0, \sigma_\varepsilon^2)$ are white noise variables representing measurement errors, and are independent of the process $S(\cdot)$. In addition, we observe p explanatory variables, $\boldsymbol{x}_i = (x_1(\boldsymbol{s}_i), \ldots, x_p(\boldsymbol{s}_i))'$ for $i = 1, \ldots, n$. We model $Z(\boldsymbol{s}_i)$ in terms of a linear combination of $\boldsymbol{x}_i$ by considering the following geostatistical regression model:

$$\begin{aligned} Z(\boldsymbol{s}_i) &= \mu(\boldsymbol{s}_i) + \eta(\boldsymbol{s}_i) + \varepsilon(\boldsymbol{s}_i) \\ &= \beta_0 + \sum_{j=1}^{p} \beta_j x_j(\boldsymbol{s}_i) + \eta(\boldsymbol{s}_i) + \varepsilon(\boldsymbol{s}_i); \quad i = 1, \ldots, n, \end{aligned} \tag{1}$$

where $\mu(\cdot) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j(\cdot)$ is the deterministic mean process, β_j's are unknown regression coefficients, and $\eta(\cdot)$ is a zero-mean, L_2-continuous spatially dependent Gaussian process.

Instead of using all p variables, it is sometimes preferable to select only a subset of important variables. By doing this, we are able to trade off some bias for smaller variance. We consider a class of candidate models indexed by $\alpha \in \mathcal{A} \subset 2^{\{1,\ldots,p\}}$ with each α corresponding to a subset of p variables. Then the geostatistical regression model corresponding to α can be written as:

$$\boldsymbol{Z} = \boldsymbol{X}_\alpha \boldsymbol{\beta}_\alpha + \boldsymbol{\eta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\beta}_\alpha$ is the parameter vector consisting of β_0 and $\{\beta_j : j \in \alpha\}$, $\boldsymbol{X}_\alpha$ is the $n \times (p_\alpha + 1)$ design matrix corresponding to α with p_α being the number of elements in α, $\boldsymbol{\eta} = (\eta(\boldsymbol{s}_1), \ldots, \eta(\boldsymbol{s}_n))'$, and $\boldsymbol{\varepsilon} = (\varepsilon(\boldsymbol{s}_1), \ldots, \varepsilon(\boldsymbol{s}_n))'$. Let $\alpha^c = \{j : |\beta_j| > 0\}$, which is the smallest correct model, and let $\mathcal{A}^c = \{\alpha \in \mathcal{A} : \alpha^c \subset \alpha\}$ be the set of all correct models.

2 Generalized Information Criterion

Suppose that $\boldsymbol{\Sigma} = \mathrm{var}(\boldsymbol{Z})$ is known. Then the log-likelihood function of $\boldsymbol{Z}$ is

$$l(\boldsymbol{\beta}_\alpha; \boldsymbol{Z}) = \text{constant} - \frac{1}{2}(Z - \boldsymbol{X}_\alpha \boldsymbol{\beta}_\alpha)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{Z} - \boldsymbol{X}_\alpha \boldsymbol{\beta}_\alpha).$$

The maximum likelihood estimate of $\boldsymbol{\beta}_\alpha$ can be written as $\hat{\boldsymbol{\beta}}_\alpha = \left(\boldsymbol{X}_\alpha' \boldsymbol{\Sigma}^{-1} \boldsymbol{X}_\alpha\right)^{-1} \boldsymbol{X}_\alpha' \boldsymbol{\Sigma}^{-1} \boldsymbol{Z}$. A commonly used loss function for the parameters fitted by model α is the Kullback-Leibler loss, which satisfies

$$L(\alpha) = \frac{1}{2}\left(\hat{\boldsymbol{\mu}}_\alpha - \boldsymbol{\mu}\right)' \boldsymbol{\Sigma}^{-1} \left(\hat{\boldsymbol{\mu}}_\alpha - \boldsymbol{\mu}\right),$$

where $\boldsymbol{\mu} = E(\boldsymbol{Z})$ and $\hat{\boldsymbol{\mu}}_\alpha = \boldsymbol{X}_\alpha \hat{\boldsymbol{\beta}}_\alpha$. We consider the generalized information criterion (GIC) [2]:

$$\mathrm{GIC}_\lambda(\alpha) = \text{constant} + (\boldsymbol{Z} - \hat{\boldsymbol{\mu}}_\alpha)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{Z} - \hat{\boldsymbol{\mu}}_\alpha) + \lambda p_\alpha,$$

where λ is a penalty parameter with a smaller λ corresponding to a larger model, and vice versa. The criterion includes the commonly used Akaike information criterion (AIC) [1] corresponding to $\lambda = 2$, and the Bayesian information criterion (BIC) [3] corresponding to $\lambda = \log(n)$ as special cases.

Let $\hat{\alpha}_2 = \arg\min_{\alpha \in \mathcal{A}} \mathrm{GIC}_2(\alpha)$ be the model selected by AIC. The following theorem provides sufficient conditions under which AIC is asymptotically efficient.

Theorem 1 *Suppose that* $\min_{\alpha \in \mathcal{A} \setminus \mathcal{A}^c} E(L(\alpha)) \to \infty$, *as* $n \to \infty$.

(i) If $\mathcal{A}^c = \emptyset$, *then* $L(\hat{\alpha}_2) \big/ \min_{\alpha \in \mathcal{A}} L(\alpha) \xrightarrow{p} 1$, *as* $n \to \infty$.

(ii) If $\mathcal{A}^c = \{\alpha^c\}$, *then* $P(\hat{\alpha}_2 = \alpha^c) \to 1$, *as* $n \to \infty$.

However, AIC is not able to distinguish among $\mathcal{A}^c$ when $|\mathcal{A}^c| \geq 2$, and hence may select an over-fitted model. To avoid over-fitting, it is natural to consider a larger penalty. Let $\hat{\alpha}_\lambda = \arg\min_{\alpha\in\mathcal{A}} \mathrm{GIC}_\lambda(\alpha)$ be the model selected by GIC_λ. The following theorem provides some results when λ is large.

Theorem 2 *Suppose that* $\lambda \to \infty$ *and* $\min_{\alpha\in\mathcal{A}\setminus\mathcal{A}^c} \lambda^{-1}E(L(\alpha)) \to \infty$, *as* $n \to \infty$. *Then*

$$L(\hat{\alpha}_\lambda)\big/\min_{\alpha\in\mathcal{A}} L(\alpha) \xrightarrow{p} 1, \quad \text{as } n \to \infty.$$

In addition, if $\mathcal{A}^c \neq \emptyset$, *then* $P(\hat{\alpha}_\lambda = \alpha^c) \to 1$, *as* $n \to \infty$.

In what follows, we provide an example in the one-dimensional space. Suppose that the data $\{\boldsymbol{x}_i : i = 1,\ldots,n\}$ and $\boldsymbol{Z}$ are sampled at $s_i = in^{\delta-1} \in D = [0, n^\delta]$; $i = 1,\ldots,n$, for some $0 \leq \delta \leq 1$, according to (1) with p fixed, where $\mathrm{cov}(\eta(s),\eta(s+h)) = \sigma_\eta^2 \exp(-\kappa_\eta|h|)$. Note that different δ values correspond to different asymptotic frameworks. When $\delta = 0$, increasingly dense observations are sampled in a bounded fixed region $D = [0,1]$, corresponding to the fixed-domain asymptotic framework. On the other hand, when $0 < \delta \leq 1$, the region D increases with the sample size, corresponding to the increasing-domain asymptotic framework. Suppose that $x_j(\cdot)$'s are independently generated from zero-mean Gaussian processes with $\mathrm{cov}(\eta(s),\eta(s+h)) = \sigma_j^2 \exp(-\kappa_j|h|)$, for $j = 1,\ldots,p$, where $\sigma_j^2 > 0$ and $\kappa_j > 0$. We have the following results showing whether selection consistency is satisfied may depend on which asymptotic framework is chosen.

Theorem 3 *Under the setup above, suppose that* $\kappa_\eta > 0$, $\sigma_\eta^2 > 0$, *and* $\sigma_\varepsilon^2 > 0$ *are known. If* $\mathcal{A}^c \neq \emptyset$, *and* $\lambda \to \infty$ *and* $\lambda n^{-(1+\delta)/2} \to 0$, *as* $n \to \infty$, *then* $P(\hat{\alpha}_\lambda = \alpha^c) \to 1$, *as* $n \to \infty$.

Corollary 1 *Under the conditions of Theorem 3, consider two sampling schemes with* δ *given by* δ_1 *and* δ_2, *where* $\delta_1 < \delta_2$. *Suppose that* $\lambda = n^{(2+\delta_1+\delta_2)/4}$. *Then* $\limsup_{n\to\infty} P(\hat{\alpha}_\lambda = \alpha^c) < 1$ *for* $\delta = \delta_1$, *and* $\lim_{n\to\infty} P(\hat{\alpha}_\lambda = \alpha^c) = 1$ *for* $\delta = \delta_2$.

References

[1] Akaike, H. (1973). Information theory and the maximum likelihood principle. *International Symposium on Information Theory* (V. Petrov and F. Csáki eds.), Akademiai Kiádo, Budapest, 267–281.

[2] Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics* **12**, 758–765.

[3] Schwartz, G. (1978). Estimating the dimensions of a model. *The Annals of Statistics* **6**, 461–464.