Environ Ecol Stat (2012) 19:23–35
DOI 10.1007/s10651-011-0171-2

Geostatistical model averaging based on conditional information criteria

Chun-Shu Chen · Hsin-Cheng Huang

Received: 11 March 2010 / Revised: 14 September 2010 / Published online: 11 April 2011

Abstract Variable selection in geostatistical regression is an important problem, but has not been well studied in the literature. In this paper, we focus on spatial prediction and consider a class of conditional information criteria indexed by a penalty parameter. Instead of applying a fixed criterion, which leads to an unstable predictor in the sense that it is discontinuous with respect to the response variables due to that a small change in the response may cause a different model to be selected, we further stabilize the predictor by local model averaging, resulting in a predictor that is not only continuous but also differentiable even after plugging-in estimated model parameters. Then Stein's unbiased risk estimate is applied to select the penalty parameter, leading to a data-dependent penalty that is adaptive to the underlying model. Some numerical experiments show superiority of the proposed model averaging method over some commonly used variable selection methods. In addition, the proposed method is applied to a mercury data set for lakes in Maine.

Keywords Conditional Akaike information criterion · Data perturbation · Spatial prediction · Stabilization · Stein's unbiased risk estimate · Variable selection

C.-S. Chen (✉)
Institute of Statistics and Information Science, National Changhua University of Education, No.1, Jin-De Road, 500 Changhua, Taiwan
e-mail: cschen@cc.ncue.edu.tw

H.-C. Huang
Institute of Statistical Science, Academia Sinica, 128 Academia Road, Section 2, Nankang, 115 Taipei, Taiwan
e-mail: hchuang@stat.sinica.edu.tw

H.-C. Huang
Institute of Statistics, National Chiao Tung University, Hsinchu, Taiwan

1 Introduction

In many geostatistical problems, the variable of interest is often observed along with other variables at some locations. Typically, a geostatistical regression model is applied by treating the variable of interest as the response and some other variables as explanatory variables. How to select a suitable subset of explanatory variables is crucial, because accuracy of selection directly affects estimation and prediction. Model selection and model averaging have been well studied for linear regression; see Shao (1997), Hoeting et al. (1999), and Burnham and Anderson (2002) for a review. For selection of geostatistical models, the Akaike information criterion (AIC) (Akaike 1973), the Bayesian information criterion (BIC) (Schwarz 1978), and the cross-validation method are often applied. Although some heuristic arguments for AIC were provided by Hoeting et al. (2006), as pointed out by Breiman (1996), these model selection procedures are unstable in the sense that a small change in the response variables may cause a different model to be selected, resulting in less accurate estimation and prediction.

In this paper, we focus on spatial prediction and attempt to make three proposals. First, we introduce a class of conditional information criteria indexed by a penalty parameter, which we believe is more suitable for geostatistical model selection comparing to commonly used unconditional information criteria, such as AIC and BIC. A particular criterion in this class, called the conditional AIC (CAIC), was proposed by Vaida and Blanchard (2005) for variable selection in linear mixed-effects models. Despite some tendency to select an over-complex model, this criterion has a desired property that it is an unbiased estimate of the sum of the mean squared prediction errors over data locations up to a constant. Second, due to that the spatial predictor obtained from a fixed conditional information criterion is also unstable, we propose to stabilize the spatial predictor by local model averaging using perturbed data, resulting in a stabilized predictor that can be shown to be differentiable with respect to the response variables even after plugging-in estimated model parameters. Third, utilizing the differentiable property of the stabilized predictor, we propose to apply Stein's unbiased risk estimate (Stein 1981) to select among a collection of conditional information criteria, leading to a data-dependent penalty with an adaptive feature such that it tends to selects a large penalty, and hence a small model (based on a small number of explanatory variables), when the underlying true model is small, and vice versa.

The rest of this article is organized as follows. Section 2 introduces the geostatistical regression model and the class of conditional information criteria for variable selection. Section 3 introduces the proposed stabilized predictor and Stein's unbiased risk estimate for selecting the penalty parameter. Some numerical results are shown in Sect. 4. An application of the proposed methodology for assessing mercury levels of fish in Maine lakes is presented in Sect. 5. Finally, a brief discussion is given in Sect. 6.

2 Geostatistical regression and variable selection

2.1 Geostatistical regression models

Consider a spatial process $\{S(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ defined over a region $D \subseteq \mathbb{R}^d$. Suppose that the spatial process can be decomposed into the following:

$$S(\boldsymbol{s}) = \mu(\boldsymbol{s}) + \eta(\boldsymbol{s}); \quad \boldsymbol{s} \in D, \tag{1}$$

where $\mu(\cdot)$ is a deterministic mean process corresponding to a large-scale mean structure, and $\eta(\cdot)$ is a zero-mean, L^2-continuous, spatial dependent process corresponding to a small-scale structure. Suppose that we observe response variable Z_i and a p-dimensional vector of explanatory variables, $(x_1(\boldsymbol{s}_i), \ldots, x_p(\boldsymbol{s}_i))'$, associated with Z_i at location $\boldsymbol{s}_i \in D$; $i = 1, \ldots, n$. The mean process $\mu(\boldsymbol{s})$ is usually modeled as $\beta_0 + \sum_{j=1}^{p} \beta_j x_j(\boldsymbol{s})$, where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ are regression parameters. Hence the geostatistical regression model can be written as:

$$\begin{aligned} Z_i &= S(\boldsymbol{s}_i) + \varepsilon(\boldsymbol{s}_i) \\ &= \beta_0 + \sum_{j=1}^{p} \beta_j x_j(\boldsymbol{s}_i) + \eta(\boldsymbol{s}_i) + \varepsilon(\boldsymbol{s}_i); \quad i = 1, \ldots, n, \end{aligned} \tag{2}$$

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 spatial process $\eta(\cdot)$. Here, the spatial dependent process $\eta(\cdot)$ is usually assumed to be stationary with some parametric covariance function class. A commonly used one is the isotropic Matérn model (Matérn 1986):

$$C(\boldsymbol{h}) = cov(\eta(\boldsymbol{s}+\boldsymbol{h}), \eta(\boldsymbol{s})) = \begin{cases} \dfrac{\sigma_\eta^2 (a^2||\boldsymbol{h}||/2)^\nu \, 2\,\kappa_\nu(a^2||\boldsymbol{h}||)}{\Gamma(\nu)}; & \text{if } \boldsymbol{h} \neq \boldsymbol{0}, \\ \sigma_\eta^2; & \text{if } \boldsymbol{h} = \boldsymbol{0}, \end{cases} \tag{3}$$

where $\|\cdot\|$ represents the Euclidean distance, $\kappa_\nu(\cdot)$ is the modified Bessel function of the second kind with order $\nu > 0$ (Abramowitz and Stegun 1965), ν indicates the smoothness of the process, $a > 0$ is a scaling parameter, and $\sigma_\eta^2 = \text{var}(\eta(\boldsymbol{s}))$. Note that (3) reduces to the exponential covariance model when $\nu = 0.5$.

In this paper, we consider selecting among p explanatory variables. Each candidate model corresponds to a subset of p variables and is indexed by $\gamma \in \Gamma$, where $\Gamma \subset 2^{\{1,\ldots,p\}}$ is the class of all candidate models. Then the geostatistical regression model corresponding to γ can be written as:

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

where $\boldsymbol{\beta}_\gamma$ is the parameter vector consisting of β_0 and $\{\beta_j : j \in \gamma\}$, $\boldsymbol{X}_\gamma$ is the $n \times (p_\gamma + 1)$ design matrix corresponding to γ with p_γ being the number of explanatory variables in γ, $\boldsymbol{Z} = (Z_1, \ldots, Z_n)'$, $\boldsymbol{\eta} = (\eta(\boldsymbol{s}_1), \ldots, \eta(\boldsymbol{s}_n))'$, and $\boldsymbol{\varepsilon} = (\varepsilon(\boldsymbol{s}_1), \ldots, \varepsilon(\boldsymbol{s}_n))'$.

Denote the covariance parameters of $\boldsymbol{Z}$ by $\boldsymbol{\theta} \equiv (\nu, a, \sigma_\eta^2, \sigma_\varepsilon^2)'$ and let $\boldsymbol{\Sigma}(\boldsymbol{\theta}) \equiv var(\boldsymbol{Z})$. For known $\boldsymbol{\theta}$, the best linear unbiased predictor (BLUP), typically called the universal kriging (UK) predictor, of $S(\boldsymbol{s})$ based on model γ satisfies

$$\hat{S}_\gamma(\boldsymbol{s}; \boldsymbol{\theta}) = \hat{\boldsymbol{\beta}}_\gamma' \boldsymbol{x}_\gamma(\boldsymbol{s}) + \text{cov}(\eta(\boldsymbol{s}), \boldsymbol{\eta}) \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} \big(\boldsymbol{Z} - \boldsymbol{X}_\gamma \hat{\boldsymbol{\beta}}_\gamma\big); \quad \boldsymbol{s} \in D,$$

where $\hat{\boldsymbol{\beta}}_\gamma = \left(\boldsymbol{X}_\gamma' \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} \boldsymbol{X}_\gamma\right)^{-1} \boldsymbol{X}_\gamma' \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} \boldsymbol{Z}$, and $\boldsymbol{x}_\gamma(\boldsymbol{s})$ consists of the constant 1 and the p_γ explanatory variables observed at location $\boldsymbol{s}$. Then the BLUP of $\boldsymbol{S} \equiv (S(\boldsymbol{s}_1), \ldots, S(\boldsymbol{s}_n))'$ based on model γ can be written as

$$\hat{\boldsymbol{S}}_\gamma(\boldsymbol{\theta}) \equiv (\hat{S}_\gamma(\boldsymbol{s}_1; \boldsymbol{\theta}), \ldots, \hat{S}_\gamma(\boldsymbol{s}_n; \boldsymbol{\theta}))' = \boldsymbol{H}_\gamma(\boldsymbol{\theta}) \boldsymbol{Z}, \tag{4}$$

in terms of some matrix $\boldsymbol{H}_\gamma(\boldsymbol{\theta})$.

In practice, the model parameters in (2) and (3) can be estimated by methods of moments, maximum likelihood (ML), restricted maximum likelihood (REML) (Patterson and Thompson 1971), or Bayesian methods. Note that after plugging-in an estimated parameter vector $\hat{\boldsymbol{\theta}}_\gamma$ of $\boldsymbol{\theta}$ based on model γ in (4), the resulting estimated UK predictor $\hat{\boldsymbol{S}}_\gamma(\hat{\boldsymbol{\theta}}_\gamma) = \boldsymbol{H}_\gamma(\hat{\boldsymbol{\theta}}_\gamma) \boldsymbol{Z}$ is no longer linear.

2.2 Model selection methods

In this paper, we focus on spatial prediction. Our objective is to select $\gamma \in \Gamma$ that minimizes $E\|\hat{\boldsymbol{S}}_\gamma(\hat{\boldsymbol{\theta}}_\gamma) - \boldsymbol{S}\|^2$. Some commonly used model selection criteria, such as AIC, BIC, and the corrected AIC (AICc) criterion (Hurvich and Tsai 1989), are special cases of the generalized information criterion (Nishii 1984) given by

$$\mathrm{GIC}_\lambda(\gamma) = -2\ell_\gamma\left(\boldsymbol{Z}, \hat{\boldsymbol{\theta}}_\gamma^{(ml)}, \hat{\boldsymbol{\beta}}_\gamma^{(ml)}\right) + \lambda\, p_\gamma, \tag{5}$$

with $\lambda = 2$, $\log(n)$, and $2n/(n - p_\gamma - 2)$, respectively, where $\ell_\gamma\left(\boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\beta}_\gamma\right)$ is the log-likelihood function of $\boldsymbol{Z}$ based on model γ, and $\hat{\boldsymbol{\theta}}_\gamma^{(ml)}$ and $\hat{\boldsymbol{\beta}}_\gamma^{(ml)}$ are the corresponding ML estimates. However, these criteria are not aim to minimize $E\|\hat{\boldsymbol{S}}_\gamma(\hat{\boldsymbol{\theta}}_\gamma) - \boldsymbol{S}\|^2$, and hence may not be ideal for spatial prediction purpose, as demonstrated in some simulation examples in Sect. 4.

With the goal of spatial prediction in mind, we now introduce a class of conditional information criteria by conditioning on $\boldsymbol{\eta}$. For notational simplicity, we first assume that the covariance parameter vector $\boldsymbol{\theta}$ is known. Then the CAIC criterion proposed by Vaida and Blanchard (2005) is

$$\mathrm{CAIC}(\gamma; \boldsymbol{\theta}) = \frac{1}{n}\left\{\left\|\boldsymbol{Z} - \hat{\boldsymbol{S}}_\gamma(\boldsymbol{\theta})\right\|^2 + 2\mathrm{tr}\left(\boldsymbol{H}_\gamma(\boldsymbol{\theta})\right)\sigma_\varepsilon^2\right\}, \tag{6}$$

where $\mathrm{tr}\left(\boldsymbol{H}_\gamma(\boldsymbol{\theta})\right)$, the trace of $\boldsymbol{H}_\gamma(\boldsymbol{\theta})$, is the effective degrees of freedom, which tends to be larger for a larger (more complex) model. As shown by Vaida and Blanchard (2005),

$$E\left(\mathrm{CAIC}(\gamma; \boldsymbol{\theta})\right) = \frac{1}{n} E\left\|\hat{\boldsymbol{S}}_\gamma(\boldsymbol{\theta}) - \boldsymbol{S}\right\|^2 + \sigma_\varepsilon^2 .$$

Consequently, CAIC appears more appropriate than AIC for variable selection when prediction is the main goal of the analysis. When $\boldsymbol{\theta}$ is unknown, we can apply CAIC by plugging-in an estimate of $\boldsymbol{\theta}$ in (6).

In general, CAIC tends to select an over-complex model particularly when the underlying true model is small. A natural remedy is to consider applying a larger penalty in (6), leading to the following conditional generalized information criterion (CGIC) indexed by a penalty parameter $\lambda > 0$:

$$\mathrm{CGIC}_\lambda(\gamma) = \frac{1}{n}\left\{\left\|\boldsymbol{Z} - \hat{\boldsymbol{S}}_\gamma(\hat{\boldsymbol{\theta}}_\gamma)\right\|^2 + \lambda\,\mathrm{tr}\big(\boldsymbol{H}_\gamma(\hat{\boldsymbol{\theta}}_\gamma)\big)\hat{\sigma}_\varepsilon^2\right\}; \quad \gamma \in \Gamma, \tag{7}$$

where $\hat{\boldsymbol{S}}_\gamma(\hat{\boldsymbol{\theta}}_\gamma) = \boldsymbol{H}_\gamma(\hat{\boldsymbol{\theta}}_\gamma)\boldsymbol{Z}$, $\hat{\boldsymbol{\theta}}_\gamma$ is some estimator of $\boldsymbol{\theta}$ based on model γ, and $\hat{\sigma}_\varepsilon^2$ is an estimate of σ_ε^2 in (7) independent of $\gamma \in \Gamma$. For example, it can be estimated by REML based on the full model. For a given penalty λ, CGIC in (7) selects the model, $\hat{\gamma}(\lambda) \equiv \arg\min_{\gamma\in\Gamma} \mathrm{CGIC}_\lambda(\gamma)$. The corresponding spatial predictor of $S(\boldsymbol{s})$ at $\boldsymbol{s} \in D$ is given by $\hat{S}_{\hat{\gamma}(\lambda)}(\boldsymbol{s}; \hat{\boldsymbol{\theta}}_{\hat{\gamma}(\lambda)})$, where $\hat{\boldsymbol{\theta}}_{\hat{\gamma}(\lambda)}$ is the estimator of $\boldsymbol{\theta}$ based on the selected model $\hat{\gamma}(\lambda)$. The criterion (7) includes CAIC (corresponding to $\lambda = 2$) and conditional BIC (CBIC) (corresponding to $\lambda = \log(n)$) as special cases. Clearly, if we know the underlying true model depends only on a small number of explanatory variables in advance, it would be preferable to choose a criterion with a larger penalty that penalizes more for a larger model, and vice versa. But in practice, the underlying true model is unknown, and hence a CGIC criterion may perform well only in some situations, which is not adaptive. Without knowing the underlying true model, Shen and Ye (2002) proposed to select the penalty λ of (5) from data in linear regression model selection and termed the method as adaptive model selection. We shall adopt this approach and further generalize it for spatial model averaging.

3 The proposed method

As in Shen and Ye (2002), we could consider $\{\hat{S}_{\hat{\gamma}(\lambda)}(\boldsymbol{s}; \hat{\boldsymbol{\theta}}_{\hat{\gamma}(\lambda)}) : \lambda \in \Lambda\}$ as the class of candidate spatial predictors for some $\Lambda \subset (0, \infty)$. However, $\hat{S}_{\hat{\gamma}(\lambda)}(\boldsymbol{s}; \hat{\boldsymbol{\theta}}_{\hat{\gamma}(\lambda)})$ is discontinuous in $\boldsymbol{Z}$ and hence unstable (Breiman 1996). Motivated from Breiman (1996), we consider a stabilized version of $\hat{S}_{\hat{\gamma}(\lambda)}(\boldsymbol{s}; \hat{\boldsymbol{\theta}}_{\hat{\gamma}(\lambda)})$ obtained by using perturbed data, $\boldsymbol{Z}^* = (Z_1^*, \ldots, Z_n^*)' \equiv \boldsymbol{Z} + \tau\boldsymbol{\xi}$, where $\boldsymbol{\xi} \sim N(\boldsymbol{0}, \sigma_\varepsilon^2\boldsymbol{I})$ is independent of $\boldsymbol{Z}$, and $\tau > 0$ is the perturbation size. When σ_ε^2 is unknown, the estimator $\hat{\sigma}_\varepsilon^2$ in (7) is used for computing $\boldsymbol{Z}^*$. Our proposed stabilized predictor of $S(\boldsymbol{s})$ corresponding to $\hat{S}_{\hat{\gamma}(\lambda)}(\boldsymbol{s}; \hat{\boldsymbol{\theta}}_{\hat{\gamma}(\lambda)})$ is given by

$$E\big(\hat{S}^*_{\hat{\gamma}^*(\lambda)}(\boldsymbol{s}; \hat{\boldsymbol{\theta}}^*_{\hat{\gamma}^*(\lambda)})\big|\boldsymbol{Z}\big); \quad \lambda \in \Lambda, \tag{8}$$

where $\hat{\gamma}^*(\lambda)$ is the model selected by CGIC_λ based on $\boldsymbol{Z}^*$:

$$\hat{\gamma}^*(\lambda) = \arg\min_{\gamma\in\Gamma}\left\{\left\|\boldsymbol{Z}^* - \hat{\boldsymbol{S}}^*_\gamma(\hat{\boldsymbol{\theta}}^*_\gamma)\right\|^2 + \lambda\,\mathrm{tr}\big(\boldsymbol{H}_\gamma(\hat{\boldsymbol{\theta}}^*_\gamma)\big)\big(1+\tau^2\big)\sigma_\varepsilon^2\right\},$$

$\hat{\boldsymbol{S}}^*_\gamma(\hat{\boldsymbol{\theta}}^*_\gamma) = \big(\hat{S}^*_\gamma(\boldsymbol{s}_1; \hat{\boldsymbol{\theta}}^*_\gamma), \ldots, \hat{S}^*_\gamma(\boldsymbol{s}_n; \hat{\boldsymbol{\theta}}^*_\gamma)\big)' = \boldsymbol{H}_\gamma(\hat{\boldsymbol{\theta}}^*_\gamma)\boldsymbol{Z}^*$ is the UK predictor of $\boldsymbol{S}$ based on $\boldsymbol{Z}^*$ and model γ, and $\hat{\boldsymbol{\theta}}^*_\gamma$ is the estimator of $\boldsymbol{\theta}$ obtained in exact the same way as

$\hat{\boldsymbol{\theta}}_\gamma$ but based on $\boldsymbol{Z}^*$. Since $\hat{\gamma}^*(\lambda)$ may be different from $\hat{\gamma}(\lambda)$ for different $\boldsymbol{Z}^*$, the proposed stabilized predictor $E\big(\hat{S}^*_{\hat{\gamma}^*(\lambda)}(\boldsymbol{s};\hat{\boldsymbol{\theta}}^*_{\hat{\gamma}^*(\lambda)})\big|\boldsymbol{Z}\big)$ in (8) can be regarded as a type of model averaging predictor centered around $\hat{S}_{\hat{\gamma}(\lambda)}(\boldsymbol{s};\hat{\boldsymbol{\theta}}_{\hat{\gamma}(\lambda)})$. Clearly, a larger τ tends to produce a smoother but more stable predictor. We found in some simulation experiments (e.g., Table 5) that spatial prediction is typically not sensitive to the choose of τ when τ is between 0.5 and 0.9. Therefore, we fix $\tau = 0.5$ throughout the simulation and the data analysis unless stated otherwise.

As shown by Huang and Chen (2007), $E\big(\hat{S}^*_{\hat{\gamma}^*(\lambda)}(\boldsymbol{s};\hat{\boldsymbol{\theta}}^*_{\hat{\gamma}^*(\lambda)})\big|\boldsymbol{Z}\big)$ is infinitely differentiable. Applying Stein's lemma (1981), the corresponding effective degrees of freedom is given by

$$df_\lambda \equiv \sum_{i=1}^{n} \frac{\partial}{\partial Z_i} E\big(\hat{S}^*_{\hat{\gamma}^*(\lambda)}(\boldsymbol{s}_i;\hat{\boldsymbol{\theta}}^*_{\hat{\gamma}^*(\lambda)})\big|\boldsymbol{Z}\big) = \frac{1}{\tau^2\sigma_\varepsilon^2}\sum_{i=1}^{n} cov\big(\hat{S}^*_{\hat{\gamma}^*(\lambda)}(\boldsymbol{s}_i;\hat{\boldsymbol{\theta}}^*_{\hat{\gamma}^*(\lambda)}), Z_i^*\big|\boldsymbol{Z}\big), \tag{9}$$

and an unbiased estimator of $E\big\|E\big(\hat{\boldsymbol{S}}^*_{\hat{\gamma}^*(\lambda)}(\hat{\boldsymbol{\theta}}^*_{\hat{\gamma}^*(\lambda)})\big|\boldsymbol{Z}\big) - \boldsymbol{S}\big\|^2$ satisfies

$$\big\|\boldsymbol{Z} - E\big(\hat{\boldsymbol{S}}^*_{\hat{\gamma}^*(\lambda)}(\hat{\boldsymbol{\theta}}^*_{\hat{\gamma}^*(\lambda)})\big|\boldsymbol{Z}\big)\big\|^2 + 2\sigma_\varepsilon^2 df_\lambda - n\sigma_\varepsilon^2. \tag{10}$$

Note that (10) is usually referred to as Stein's unbiased risk estimate (SURE). Thus we can simply select λ by minimizing SURE in (10) and obtain

$$\hat{\lambda} \equiv \arg\min_{\lambda\in\Lambda}\left\{\big\|\boldsymbol{Z} - E\big(\hat{\boldsymbol{S}}^*_{\hat{\gamma}^*(\lambda)}(\hat{\boldsymbol{\theta}}^*_{\hat{\gamma}^*(\lambda)})\big|\boldsymbol{Z}\big)\big\|^2 + 2\sigma_\varepsilon^2 df_\lambda\right\}. \tag{11}$$

Consequently, the model selected by $\mathrm{CGIC}_{\hat{\lambda}}$ criterion is

$$\hat{\gamma}(\hat{\lambda}) \equiv \arg\min_{\gamma\in\Gamma}\left\{\big\|\boldsymbol{Z} - \hat{\boldsymbol{S}}_\gamma(\hat{\boldsymbol{\theta}}_\gamma)\big\|^2 + \hat{\lambda}\, tr\big(\boldsymbol{H}_\gamma(\hat{\boldsymbol{\theta}}_\gamma)\big)\sigma_\varepsilon^2\right\},$$

and the resulting stabilized predicted surface is $\big\{E\big(\hat{S}^*_{\hat{\gamma}^*(\hat{\lambda})}(\boldsymbol{s};\hat{\boldsymbol{\theta}}^*_{\hat{\gamma}^*(\hat{\lambda})})\big|\boldsymbol{Z}\big) : \boldsymbol{s}\in D\big\}$.

In practice, it suffices to select among a few discrete penalty values due to the continuity of $E\big(\hat{\boldsymbol{S}}^*_{\hat{\gamma}^*(\lambda)}(\hat{\boldsymbol{\theta}}^*_{\hat{\gamma}^*(\lambda)})\big|\boldsymbol{Z}\big)$ in λ. For example, we consider $\Lambda = \{1, 2, \log(n), 2\log(n)\}$ in the simulation study and the data analysis, which involves two typical criteria: CAIC with $\lambda = 2$ and CBIC with $\lambda = \log(n)$. Note that as demonstrated in Sect. 4, the selected penalty in (11) tends to be small when the size of true model is large, and vice versa. This adaptive feature is attractive, because it automatically adjusts to select an appropriate model $\hat{\gamma}(\hat{\lambda})$, around which a model averaging predictor $E\big(\hat{S}^*_{\hat{\gamma}^*(\hat{\lambda})}(\boldsymbol{s};\hat{\boldsymbol{\theta}}^*_{\hat{\gamma}^*(\hat{\lambda})})\big|\boldsymbol{Z}\big)$ of $S(\boldsymbol{s})$ is obtained, regardless of whether the underlying true model is small or large.

In practice, df_λ in (9) can be computed using a simple Monte Carlo (MC) method by simulating some replicates of perturbed data, and then approximate the covariances in the righthand side of the equality in (9) based on the corresponding sample covariances.

4 Simulations

We conducted a simulation study to examine the performance of the proposed model averaging method. We consider the model given by (1) and (2) with $D = [0, 1]^2$. In the simulation, we consider two different examples (Examples I and II) corresponding to two different sets of explanatory variables. In Example I, $p = 9$ explanatory variables were independently generated from a standard Gaussian white noise process. In addition, the spatial process $\eta(\cdot)$ was generated from a zero-mean Gaussian stationary process with the exponential covariance function having two combinations of parameters $(\nu, a, \sigma_\eta^2) = (0.5, 0.5, 1)$ and $(0.5, 0.1, 1)$ in (3) corresponding to weak and strong spatial dependence, respectively (see Fig. 1). The noise variance in (2) is set to be $\sigma_\varepsilon^2 = 1$, and is assumed known throughout the simulation. We consider five cases corresponding to five different choices of $\boldsymbol{\beta} = (0, \beta_1, \ldots, \beta_9)'$ indexed by $k = 1, 3, 5, 7, 9$ as the underlying true models. For each k, the regression coefficients are given by $\beta_1 = \cdots = \beta_k = \sqrt{7/k}$, and $\beta_{k+1} = \cdots = \beta_p = 0$ so that the signal-to-noise ratio (SNR) is controlled at 8, where SNR is defined as the ratio of the variance of the signal $S(\boldsymbol{s}_i)$ to the noise variance σ_ε^2. We generated data, $\{(x_1(\boldsymbol{s}_i), \ldots, x_9(\boldsymbol{s}_i), Z_i) : i = 1, \ldots, n\}$ with $n = 32$, by sampling at 64×64 regular grid points in D using simple random sampling. For each case, we consider all possible combinations of explanatory variables by selecting among $\Gamma = 2^{\{1,\ldots,9\}}$ with $\gamma = \emptyset$ representing the intercept-only model.

The setup for Example II is basically the same as that in Example I except we select a polynomial up to the second order, $\beta_0 + \beta_1 x + \beta_2 y + \beta_3 x^2 + \beta_4 xy + \beta_5 y^2$, where x and y are the x and y coordinates, and consider only the weak dependence case with $a = 0.5$ in (3). Hence there are $p = 5$ explanatory variables, x, y, xy, x^2, y^2, in this

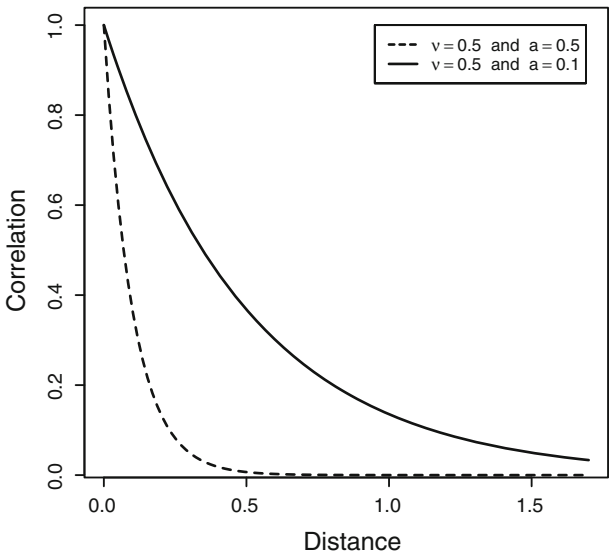

Fig. 1 Matérn correlation functions

example. We consider three different cases with different β values for the underlying model: βx, $\beta(x + y + xy)$, and $\beta(x + y + x^2 + xy + y^2)$, where β is chosen so that SNR $= 8$. Similar to Example I, we consider selecting among $\Gamma = 2^{\{1,\ldots,5\}}$ for each case.

We applied the proposed geostatistical model averaging method, referred to as GMA hereafter, with the covariance function of $\eta(\cdot)$ given by the Matérn class of (3). We selected among four CGIC criteria with $\Lambda = \{1, 2, \log(n), 2\log(n)\}$ in (11) and estimated the Matérn parameters using REML, where df_λ in (11) is computed using the MC method based on 100 replicates as mentioned in the last paragraph of Sect. 3 for each $\lambda \in \Lambda$. The proposed GMA method is compared with five different model selection criteria, including AIC, BIC, CAIC, CBIC, and AICc, and their corresponding model averaging methods (see Burnham and Anderson 2002), denoted as AIC-MA, BIC-MA, CAIC-MA, CBIC-MA, and AICc-MA, where a model averaging predictor of $\boldsymbol{S}$ is given by $\sum_{\gamma\in\Gamma} w_\gamma \hat{\boldsymbol{S}}_\gamma(\hat{\boldsymbol{\theta}}_\gamma)$ with weights

$$w_\gamma = \frac{\exp\left(-\frac{1}{2}\mathrm{GIC}_\lambda(\gamma)\right)}{\sum_{\gamma^*\in\Gamma}\exp\left(-\frac{1}{2}\mathrm{GIC}_\lambda(\gamma^*)\right)} \quad \text{or} \quad \frac{\exp\left(-\frac{1}{2}\mathrm{CGIC}_\lambda(\gamma)\right)}{\sum_{\gamma^*\in\Gamma}\exp\left(-\frac{1}{2}\mathrm{CGIC}_\lambda(\gamma^*)\right)}; \quad \gamma \in \Gamma.$$

The results of various methods are assessed in terms of the mean squared prediction error (MSPE): $\frac{1}{n}\|\hat{\boldsymbol{S}} - \boldsymbol{S}\|^2$, where $\hat{\boldsymbol{S}}$ is a generic predictor of $\boldsymbol{S}$.

The results based on 200 simulation replicates for Examples I and II are shown in Tables 1 and 2, respectively. In general, conditional information criteria perform better than the unconditional counterparts, and model averaging predictors perform better than the corresponding model selection predictors in terms of MSPE. As expected, both CAIC and CBIC perform well only in some situations with CBIC performing better than CAIC when the size of underlying true model is small, and vice versa. The proposed GMA method performs better than CAIC and CBIC for all cases. It also performs better than both CAIC-MA and CBIC-MA for almost all cases.

Tables 3 and 4 show the distribution of the penalties selected from GMA for each case in Examples I and II. Evidently, GMA tends to select a large penalty when the size of true model is small, and vice versa, as if the size of true model is known in advance. This adaptive feature can not be achieved by either CAIC or CBIC.

Finally, we performed a sensitivity study regarding the choice of perturbation size τ for the proposed GMA method. We also investigated whether it is advantageous to optimize the search of λ over $(0, \infty)$. The results for Example II are displayed in Table 5, which show that MSPEs are not sensitive to the perturbation size when τ is between 0.5 and 0.9. In addition, there seems no clear advantage to search λ over $(0, \infty)$. Similar results can be seen for all cases in Example I as well. Although we can further apply SURE to optimize the search of (λ, τ) over $(0, \infty) \times (0, \infty)$, it appears effective and computationally more efficient to fix τ at 0.5 and search over a small number of λ values.

Table 1 MSPE performance of various methods for Example I with weak ($a = 0.5$) and strong ($a = 0.1$) spatial dependence based on 200 simulation replicates, where the values in parentheses are the corresponding standard errors

a	Criterion	$\lvert\{j : \beta_j \neq 0\}\rvert$									
		1		3		5		7		9	
0.5	AIC	.557	(.021)	.538	(.021)	.547	(.021)	.541	(.021)	.556	(.021)
	AICc	.542	(.025)	.498	(.023)	.540	(.024)	.559	(.023)	.657	(.024)
	BIC	.530	(.024)	.515	(.023)	.566	(.024)	.554	(.022)	.569	(.022)
	CAIC	.305	(.010)	.345	(.014)	.469	(.022)	.499	(.020)	.535	(.018)
	CBIC	.255	(.010)	.313	(.016)	.467	(.025)	.506	(.021)	.589	(.021)
	AIC-MA	.467	(.020)	.465	(.019)	.506	(.020)	.522	(.019)	.560	(.019)
	AICc-MA	.425	(.021)	.426	(.020)	.487	(.021)	.533	(.018)	.625	(.017)
	BIC-MA	.434	(.021)	.439	(.020)	.494	(.020)	.523	(.019)	.575	(.017)
	CAIC-MA	.256	(.009)	.309	(.013)	.423	(.019)	.470	(.017)	.533	(.015)
	CBIC-MA	.224	(.009)	.289	(.013)	.440	(.022)	.491	(.019)	.588	(.018)
	GMA	.214	(.009)	.296	(.012)	.386	(.014)	.449	(.014)	.490	(.013)
0.1	AIC	.448	(.021)	.452	(.022)	.452	(.022)	.464	(.022)	.467	(.022)
	AICc	.416	(.026)	.383	(.024)	.393	(.024)	.425	(.024)	.557	(.024)
	BIC	.405	(.026)	.421	(.025)	.423	(.023)	.446	(.023)	.469	(.022)
	CAIC	.219	(.011)	.272	(.016)	.322	(.017)	.392	(.019)	.441	(.018)
	CBIC	.154	(.010)	.253	(.023)	.314	(.021)	.378	(.019)	.471	(.020)
	AIC-MA	.339	(.019)	.367	(.020)	.392	(.020)	.435	(.021)	.475	(.019)
	AICc-MA	.284	(.019)	.318	(.020)	.356	(.020)	.430	(.020)	.559	(.017)
	BIC-MA	.295	(.019)	.333	(.020)	.370	(.020)	.430	(.020)	.496	(.018)
	CAIC-MA	.171	(.008)	.234	(.013)	.295	(.015)	.374	(.016)	.454	(.016)
	CBIC-MA	.136	(.007)	.225	(.019)	.288	(.016)	.381	(.017)	.489	(.016)
	GMA	.116	(.009)	.191	(.011)	.278	(.013)	.347	(.013)	.408	(.013)

Table 2 MSPE performance of various methods for Example II based on 200 simulation replicates, where the values in parentheses are the corresponding standard errors

Criterion	$\lvert\{j : \beta_j \neq 0\}\rvert$					
	1		3		5	
AIC	.372	(.025)	.373	(.025)	.361	(.024)
AICc	.203	(.009)	.250	(.011)	.238	(.009)
BIC	.203	(.009)	.253	(.012)	.236	(.009)
CAIC	.198	(.009)	.246	(.009)	.240	(.008)
CBIC	.180	(.009)	.260	(.010)	.242	(.008)
GMA	.168	(.007)	.234	(.008)	.212	(.008)

5 Application

We applied the proposed GMA method to the mercury data set previously analyzed by Hoeting and Olsen (see Chap. 1 of Peck et al. 1998) using multiple linear regression for assessing mercury levels of fish in Maine lakes. It is known that mercury is a toxic metal, which may damage the human nervous system if the mercury level in the human body is above the safety limit. For example, the state government of Maine suggests that the mercury level in parts per million (ppm) should be less than 0.43. To assess

Table 3 Distributions of the selected penalties for Example I based on 200 simulation replicates

a	$\lvert\{j : \beta_j \neq 0\}\rvert$	$\hat{\lambda}$				Average $\hat{\lambda}$
		1	2	$\log(n)$	$2\log(n)$	
0.5	1	14	23	24	139	5.53
	3	16	43	31	110	4.86
	5	25	53	70	52	3.67
	7	48	89	54	9	2.38
	9	89	98	13	0	1.65
0.1	1	9	7	16	168	6.21
	3	7	18	34	141	5.69
	5	16	29	60	95	4.70
	7	36	76	78	10	2.64
	9	89	99	12	0	1.64

Table 4 Distributions of the selected penalties for Example II based on 200 simulation replicates

$\lvert\{j : \beta_j \neq 0\}\rvert$	$\hat{\lambda}$				Average $\hat{\lambda}$
	1	2	$\log(n)$	$2\log(n)$	
1	0	42	46	112	5.10
3	1	84	54	61	3.89
5	0	83	65	52	3.76

Table 5 MSPE performance of GMA for various perturbation sizes τ in Example II based on 200 simulation replicates, where the penalty parameter λ is chosen from either $\Lambda = \{1, 2, \log(n), 2\log(n)\}$ or a continuous interval $\Lambda^* = (0, \infty)$, and the standard errors of MSPEs are all less than or equal to 0.011

τ	$\lvert\{j : \beta_j \neq 0\}\rvert = 1$		$\lvert\{j : \beta_j \neq 0\}\rvert = 3$		$\lvert\{j : \beta_j \neq 0\}\rvert = 5$	
	Λ	Λ^*	Λ	Λ^*	Λ	Λ^*
0.1	.189	.204	.254	.262	.253	.264
0.5	.168	.174	.234	.234	.212	.220
0.9	.159	.166	.221	.227	.198	.206

whether the fish in Maine lakes are safe to eat, it is important to estimate mercury levels in fish particularly for lakes where no observations are taken and identify the important explanatory variables responsible for elevated levels of mercury.

The data set consists of mercury level (in ppm) as the response and 10 explanatory variables sampled by US Environmental Protection Agency at 110 lakes in Maine (see Fig. 2). The 10 explanatory variables include number of fish, elevation (in feet), surface area (in acres), maximum depth (in feet), lake type (in three categories: oligotrophic, eutrophic, and mesotrophic), lake stratification (in two categories: yes and no), drainage area (in square miles), runoff factor (in percentage of rainwater or melted snow flow into rivers and streams), flushing rate (in number flushes per year), and DAM (in two categories: all natural flowage and some man-made flowage in the drainage

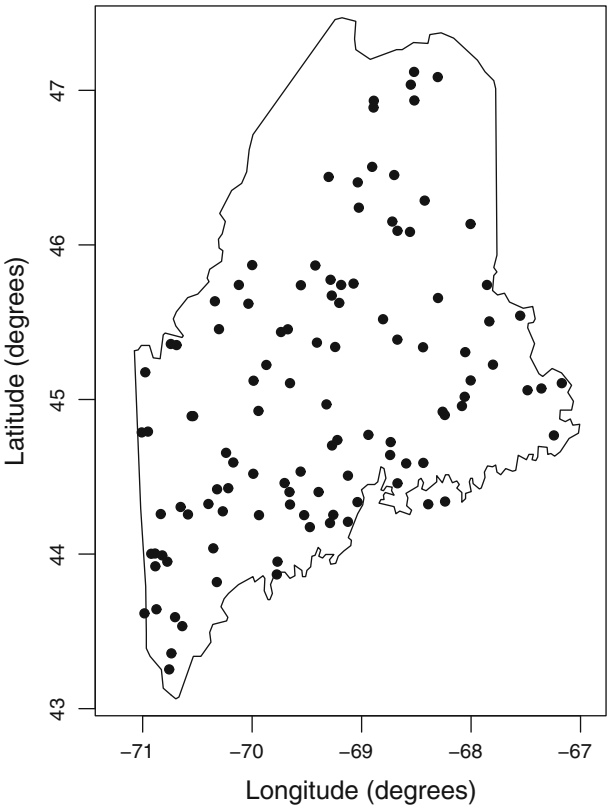

Fig. 2 Locations of 110 lakes in Maine

area). We consider all possible combinations of explanatory variables, leading to 2^{10} candidate models to be selected over.

We applied the model of (2) with Z_i and $S(\boldsymbol{s}_i)$ being the observed and the true mercury levels in fish at lake $\boldsymbol{s}_i$, and $x_1(\boldsymbol{s}_i), \ldots, x_{10}(\boldsymbol{s}_i)$ being the corresponding 10 explanatory variables, where the covariance function of $\eta(\cdot)$ is modeled by the exponential covariance model with $\nu = 0.5$ in (3). We applied CAIC, CBIC, and the proposed GMA method with $\Lambda = \{1, 2, \log(n), 2\log(n)\}$ in (11), where the covariance parameters $\boldsymbol{\theta}$ are estimated using REML and df_λ in (11) is computed using the MC method based on 100 replicates. In addition, σ_ε^2 in (11) is estimated by $\hat{\sigma}_\varepsilon^2 = 0.0876$ using REML based on the full model.

The results for CAIC and CBIC are shown in Table 6. Comparing between the two criteria, CBIC selects a smaller model having only one variable (i.e., elevation). In contrast, CAIC selects a more complex model having two additional variables (surface area and lake stratification). Our method selects $\hat{\lambda} = \log(n)$ corresponding to CBIC. The predicted mercury level surface based on our GMA method is shown in Fig. 3. Among the 110 lakes, 75 of them (shown as solid triangles in Fig. 3) particularly in southeast of Maine have mercury levels higher than the safety limit (i.e., 0.43 ppm).

Table 6 Estimated regression parameters and covariance parameters for CAIC and CBIC, where the values in parentheses are the corresponding standard errors under the selected models

Criterion	Regression parameters				Covariance parameters		
	Intercept	Elevation	Surface area	Lake stratification	a	σ_η^2	σ_ε^2
CAIC	.64	-2.9×10^{-4}	-2×10^{-5}	.083	.546	.034	.075
	(.10)	(1×10^{-4})	(1.5×10^{-5})	(.059)			
CBIC	.66	-2.7×10^{-4}	–	–	.534	.026	.081
	(.08)	(1×10^{-4})	–	–			

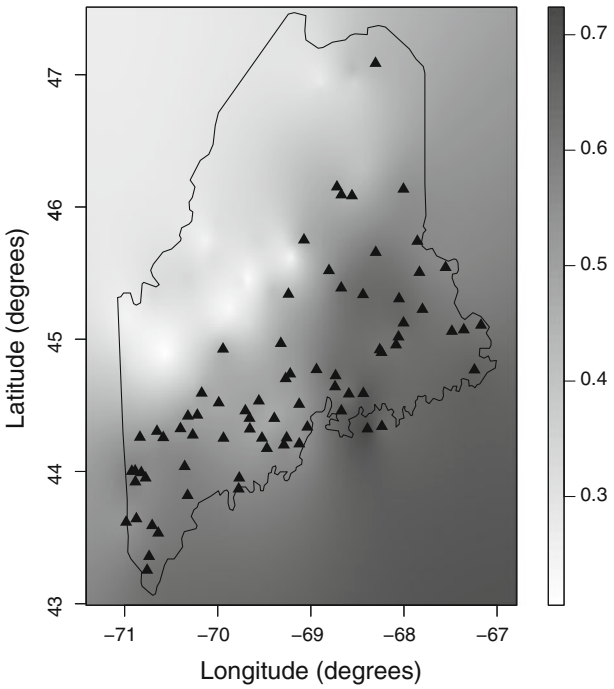

Fig. 3 Locations of 75 lakes in Maine with mercury levels higher than 0.43 ppm

6 Discussion

Motivated from the adaptive model selection idea of Shen and Ye (2002) and the stabilization method of Breiman (1996), in this paper, we developed a model averaging method for geostatistical regression, which performs well for spatial prediction in some simulation experiments. Further theoretical justification is of interest but appears to be very difficult particularly under the fixed domain asymptotic framework, and hence is beyond the scope of this paper.

Acknowledgments This work was supported by the National Science Council of Taiwan under Grants NSC 98-2118-M-018-003-MY2 and NSC 97-2118-M-001-001-MY3. The authors thank the editor, the associate editor, and three anonymous referees for helpful comments and suggestions.

References

Abramowitz M, Stegun IA (1965) Handbook of mathematical functions. Dover, New York

Akaike H (1973) Information theory and the maximum likelihood principle. In International symposium on information theory (V. Petrov and F. Csáki eds.), Akademiai Kiádo, Budapest, pp. 267–281

Breiman L (1996) Heuristics of instability and stabilization in model selection. Ann Stat 24: 2350–2383

Burnham KP, Anderson DR (2002) Model selection and multimodel inference: a practical information-theoretic approach, 2nd edn. Springer-Verlag, New York

Hoeting JA, Madigan D, Raftery AE, Volinsky CT (1999) Bayesian model averaging: a tutorial (with Discussion). Stat Sci 14: 382–401

Hoeting JA, Davis RA, Merton AA, Thompson SE (2006) Model selection for geostatistical models. Ecol Appl 16: 87–98

Huang HC, Chen CS (2007) Optimal geostatistical model selection. J Am Stat Assoc 102: 1009–1024

Hurvich CM, Tsai CL (1989) Regression and time series model selection in small samples. Biometrika 76: 297–307

Matérn B (1986) Spatial variation, 2nd edn. Springer, New York, (Lecture Notes in Statistics)

Nishii R (1984) Asymptotic properties of criteria for selection of variables in multiple regression. Ann Stat 12: 758–765

Patterson HD, Thompson R (1971) Recovery of inter-block information when block sizes are unequal. Biometrika 58: 545–554

Peck R, Haugh LD, Goodman A (eds) (1998) Statistical case studies: a collaboration between academe and industry. ASA-SIAM Series on statistics and applied probability 3 and 4

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6: 461–464

Shao J (1997) An asymptotic theory for model selection (with discussion). Stat Sin 7: 221–264

Shen X, Ye J (2002) Adaptive model selection. J Am Stat Assoc 97: 210–221

Stein C (1981) Estimation of the mean of a multivariate normal distribution. Ann Stat 9: 1135–1151

Vaida F, Blanchard S (2005) Conditional Akaike information for mixed-effects models. Biometrika 92: 351–370

Author Biographies

Chun-Shu Chen graduated from the National Central University, Taiwan, in 2007 with a Ph.D. degree in Statistics. In 2008, he joined the Institute of Statistics and Information Science at the National Changhua University of Education, Taiwan. He is currently an Assistant Professor in the same University and his research activities are focused on spatial statistics and model selection.

Hsin-Cheng Huang is research fellow in the Institute of Statistical Science at Academia Sinica, Taiwan. His research interests include spatial and space-time models, model selection, and wavelet methods.