Estimation of Spatial Dependence with GEE

YOON DONG LEE[1] , HYEMI CHOI[2]

ABSTRACT

We consider an efficient parametric estimation method of spatial dependence in weak stationary processes. Spatial dependence is modeled through variogram and correlogram. Most of parametric estimation methods of correlogram use two step method; nonparametric estimation and parametric integration. We bind these two steps into one step by using GEE method instead of least squares type optimization. Our one step method is more efficient statistically and gives a clear interpretation of related concepts used in traditional two step methods.

Keywords. Spatial dependence, Variogram, GEE, Quasi-likelihood

1. INTRODUCTION

In spatial models, usually the dependence of stationary spatial process $\{Z(\boldsymbol{s}) : \boldsymbol{s} \in I\!R^d\}$ is described by $2\gamma(\boldsymbol{h}) = \mathrm{var}(\mathrm{Z}(\boldsymbol{s}+\boldsymbol{h})-\mathrm{Z}(\boldsymbol{s}))$ or $\rho(\boldsymbol{h}) = E(Z(\boldsymbol{s}+\boldsymbol{h})Z(\boldsymbol{s}))$, which are called as *variogram* and *correlogram* respectively. Note that $2\gamma(\boldsymbol{h}) = 2\sigma^2(1-\rho(\boldsymbol{h}))$ when $\sigma^2 = \mathrm{var}(\mathrm{Z}(\boldsymbol{s}))$. For variogram estimation, various easy-to-use methods have been proposed.

The methods proposed by Journel and Huijbregts (1978) and Cressie (1985) are divided into two steps, the nonparametric estimation step and the parametric integration step. In the first step, they estimated variogram for a preselected set $H = \{\boldsymbol{h}_1, \ldots, \boldsymbol{h}_K\}$ of lags by the Matheron's estimator (1962) which is the average of the squared differences $(Z(\boldsymbol{s}_1) - Z(\boldsymbol{s}_2))^2$ for all possible pairs of the observed values $Z(\boldsymbol{s}_1)$ and $Z(\boldsymbol{s}_2)$ satisfying $\boldsymbol{s}_1 = \boldsymbol{s}_2+\boldsymbol{h}$ for $\boldsymbol{s}_1, \boldsymbol{s}_2 \in \mathbf{D}$, the sampling region. In case of irregularly spaced data, Matheron's method is modified as follows: let $T(\boldsymbol{h})$ be a predetermined set of lag $\boldsymbol{h}'$ regarded to be $\boldsymbol{h}$ approximately and average the squared differences at the pairs satisfying $\boldsymbol{s}_1 = \boldsymbol{s}_2 + \boldsymbol{h}'$, $\boldsymbol{h}' \in T(\boldsymbol{h})$. In the

[1]Department of Applied Statistics,University of Suwon
[2]Department of Statistics, Seoul National University

second step, the model variogram $2\gamma(\boldsymbol{h};\boldsymbol{\beta})$ defined through the parameter $\boldsymbol{\beta}$ is estimated by plugging $\widehat{\boldsymbol{\beta}}$, which is obtained by versions of least squares estimator (LSE), based on $2\hat{\gamma}(\boldsymbol{h})$, $\boldsymbol{h} \in H$ at the first step .

In this paper, we propose one step estimation method of correlogram $\rho(\boldsymbol{h};\boldsymbol{\beta})$, directly using the squared differences $Y(\boldsymbol{s}_1, \boldsymbol{s}_2)$, $(\boldsymbol{s}_1, \boldsymbol{s}_2) \in \mathbf{D}^2$. More efficient estimator $\hat{\boldsymbol{\beta}}$ than traditionally used LSEs is proposed, which is based on generalized estimating equation method and its asymptotic properties are shown in this paper.

The generalized estimating equations (GEE) suggested by Liang and Zeger (1986) provides a general tool to handle non-normal dependent data. Albert and McShane (1995) applied GEE to the estimation problem for binary spatial data. They considered the estimation of the mean structure and dependence structure simultaneously by iteratively solving the two estimating equations: one for mean and the other for dependence. They suggested two step method for the dependence estimation.

This paper is only concerned with dependence estimation. The proposed method is expected to be more statistically efficient than the two step estimation methods, since it reduces the loss of the captured information in the first step of the two step method. Moreover, this one step method reduces ambiguity and arbitrariness in selecting the set H of lags of interest and the tolerance region $T(\boldsymbol{h})$, $\boldsymbol{h} \in H$. The major difficulties in handling $Y(\boldsymbol{s}_1, \boldsymbol{s}_2)$, $(\boldsymbol{s}_1, \boldsymbol{s}_2) \in \mathbf{D}^2$ directly, instead of the estimated variogram $2\hat{\gamma}(\boldsymbol{h})$, $\boldsymbol{h} \in H$, reside in the non-normality of the quantity $Y(\boldsymbol{s}_1, \boldsymbol{s}_2)$ and dependence among the quantities.

Throughout the paper, we assume that process $Z(\cdot)$ is normally or nearly normally distributed with the mean 0 and the finite variance σ^2. We describe the proposed method in section 2. The statistical and computational methods to reduce the computational burden are also outlined.

2. GEE FOR CORRELOGRAM ESTIMATION

When the process $Z(\boldsymbol{s})$ is Gaussian with mean 0 and variance σ^2, the squared term $Y(\boldsymbol{s}_1, \boldsymbol{s}_2) = (Z(\boldsymbol{s}_1) - Z(\boldsymbol{s}_2))^2$ has a chi-squared distribution with one degree of freedom and a scale parameter $2\gamma(\boldsymbol{s}_1 - \boldsymbol{s}_2;\boldsymbol{\beta})$, $\boldsymbol{\beta} \in I\!R^q$, *i.e.* $Y(\boldsymbol{s}_1, \boldsymbol{s}_2) \sim 2\gamma(\boldsymbol{s}_1 - \boldsymbol{s}_2;\boldsymbol{\beta}) \cdot \chi^2(1)$. The mean and the variance of $Y(\boldsymbol{s}_1, \boldsymbol{s}_2)$ are $2\gamma(\boldsymbol{s}_1 - \boldsymbol{s}_2;\boldsymbol{\beta})$

and $2\{2\gamma(s_1 - s_2;\beta)\}^2$, respectively. The correlation $r((s_1, s_2), (s_3, s_4))$ between $Y(s_1, s_2)$ and $Y(s_3, s_4)$ is expressed in terms of the correlogram $\rho(\boldsymbol{h})$ as follows:

$$r((s_1, s_2), (s_3, s_4)) = \frac{\{\rho(s_1 - s_3) + \rho(s_2 - s_4) - \rho(s_1 - s_4) - \rho(s_2 - s_3)\}^2}{4\{1 - \rho(s_1 - s_2)\}\{1 - \rho(s_3 - s_4)\}}. \tag{2.1}$$

By modeling $\rho(\boldsymbol{h})$ with respect to a parameter β, $r((s_1, s_2), (s_3, s_4))$ depends on β, say $r((s_1, s_2), (s_3, s_4); \beta)$. For notational convenience, let $[s_1, s_2]$ be (s_1, s_2) or (s_2, s_1), and $Y[s_1, s_2]$ denote $Y(s_1, s_2)$ or $Y(s_2, s_1)$, noting that $Y(s_1, s_2) = Y(s_2, s_1)$. Now, consider $Y[s_1, s_2]$ only for the paired sampling sites between which the distance is less than a fixed constant λ,

$$\mathcal{D}_n^\lambda = \{[s_i, s_j] : 0 < d(s_i, s_j) < \lambda, \quad s_i, s_j \in \mathbf{D}_n\}$$

where $[s_1, s_2]$ denotes the pair of two points s_1, s_2 satisfying $s_1 \prec s_2$ with a complete ordering $\prec$ on $\mathbf{D}$, and $d(\cdot, \cdot)$ is a metric defined on $\mathbb{R}^d \times \mathbb{R}^d$. A set $\mathbf{D}_n$ of sampling sites denotes $\mathbf{D}$ of which number of elements is n. Let n_λ be the number of elements contained in $\mathcal{D}_n^\lambda$. When we define $\boldsymbol{Y}_n^\lambda$ to be the vector of Ys corresponding to all paired sampling sites in $\mathcal{D}_n^\lambda$ in an order defined on $\mathcal{D}_n^\lambda$. Note that $\boldsymbol{Y}_n^\lambda$ is a n_λ-dimensional vector. The fixed constant λ plays a role of the maximum lag. As in time series analysis, the maximum lag can be an appropriately defined bounded value regardless of sample size n (*cf.* Box and Jenkins, 1976, p33). Thus we assume that $n_\lambda = O(n)$.

The expectation $\boldsymbol{\mu}_n(\beta)$ and the (scaled) covariance $V_n(\beta)$ of $\boldsymbol{Y}_n^\lambda$ are given by

$$\begin{aligned}
\boldsymbol{\mu}_n(\beta) &= 2\sigma^2\{1 - \rho(s_i - s_j;\beta)\}, \quad \text{an } n_\lambda \times 1 \text{ vector}, \\
V_n(\beta) &= A_n(\beta) R_n(\beta) A_n(\beta), \quad \text{an } n_\lambda \times n_\lambda \text{ matrix},
\end{aligned}$$

where β is a q-dimensional vector of unknown parameters, $R_n(\beta)$ is an $n_\lambda \times n_\lambda$ matrix of which elements are $r([s_i, s_j], [s_k, s_l]; \beta)$, $[s_i, s_j], [s_k, s_l] \in \mathcal{D}_n^\lambda$, and $A_n(\beta)$ a diagonal matrix with diagonal elements $(1 - \rho(s_i - s_j;\beta))$ and all off-diagonal elements zero, defined along the same order of index used in defining $\boldsymbol{Y}_n^\lambda$.

Now, the generalized estimating equation for our model is defined as

$$U_n(\beta) \equiv \Gamma_n^T(\beta) V_n^{-1}(\beta)(\boldsymbol{Y}_n^\lambda - \boldsymbol{\mu}_n(\beta)) = \mathbf{0} \tag{2.2}$$

where $\Gamma_n(\beta)$ is a $n_\lambda \times q$ matrix of which rows are defined to be $(\partial/\partial\beta)\{\rho(s_1 - s_2;\beta)\}$ for all $[s_1, s_2] \in \mathcal{D}_n^\lambda$.

The solution $\hat{\beta}_n$ of (2.2) is obtained by the Gauss-Newton type iteration, obtained from Fisher's scoring approximation,

$$\beta^{new} = \beta^{old} - (1/2\sigma^2) \cdot M_n^{-1}(\beta^{old}) U_n(\beta^{old}) \quad ,$$

where $M_n(\beta) = \Gamma_n^T(\beta) V_n^{-1}(\beta) \Gamma_n(\beta)$, as given in McCullagh and Nelder (1989). The asymptotic covariance matrix of $\hat{\beta}_n$ is $2M_n^{-1}(\beta)$. The consistency and asymptotic normality of GEE estimator $\hat{\beta}_n$ of the true value β_* are obtained as follows.

PROPOSITION 2.1. *Under mild regularity conditions, $\hat{\beta}_n \to \beta_*$ in probability as n goes to ∞, and $\sqrt{n}(\hat{\beta}_n - \beta_*)$ is asymptotically multivariate Gaussian with zero mean and covariance matrix*

$$\Sigma(\beta_*) = \lim_{n \to \infty} 2\, n \cdot M^{-1}(\beta_*) \quad .$$

The proof is based on the relations

$$\begin{aligned}
&\mathrm{cov}(\boldsymbol{Y}_{\mathrm{n}}^{\lambda}) = 2(2\sigma^2)^2 \mathrm{V}(\beta_*), \mathrm{E}(\mathrm{U_n U_n^T}) = 2(2\sigma^2)^2 \mathrm{M}(\beta_*) \\
&E((\partial/\partial\beta) U_n) = -(2\sigma^2) M(\beta_*) \\
&\mathrm{cov}(\sqrt{\mathrm{n}}\hat{\beta}_{\mathrm{n}}) \simeq \{\mathrm{E}((\partial/\partial\beta)\mathrm{U_n})\}^{-1} \mathrm{E}(\mathrm{U_n U_n^T}) \{\mathrm{E}((\partial/\partial\beta)\mathrm{U_n})\}^{-1},
\end{aligned}$$

which follows from the view point of the standardized estimating function in Heyde (1997) and covariance of quasi-likelihood estimators for dependent observations in McCullagh and Nelder (1989, p332). The result needs the mixing property defined on the process $Z(\cdot)$ and so does on $Y(\cdot)$. Otherwise, it is shown analogously to Liang and Zeger (1986) under m-dependence condition. The proposition was stated in the view point of *increasing-domain asymptotics*, in which the region where the process is observed is increased as the number of observation increases, as usual in asymptotics of time series. In spatial setting, other asymptotics such as *infill-domain asymptotics* and *mixed-domain asymptotics* are possible to consider, of which details are found in Lahiri, Lee and Cressie (2002) and Lahiri (200x).

The main difficulty in estimation of $\hat{\beta}_n$ is the computation of a large matrix $V_n(\beta)$ and its inversion when n is quite large. To reduce the dimension n_λ of $\boldsymbol{Y}_n^{\lambda}$, taking smaller λ, or more generally performing the proposed method on a smaller subset $\widetilde{\mathcal{D}}_n^{\lambda}$ of $\mathcal{D}_n^{\lambda}$ seems possible without any difference in related results.

The elements of the set $\widetilde{\mathcal{D}}_n^\lambda$ are selected by one's preference as for the selection of the set H of lags and the tolerance region $T(\cdot)$ in LSEs. However, for the purpose of formalization we use the set $\mathcal{D}_n^\lambda$ and $\boldsymbol{Y}_n^\lambda$ themselves.

REFERENCES

Albert, P. and McShane, M. (1995). A Generalized Estimating Equations Approach for Spatially Correlated Binary Data: Applications to the Analysis of Neuroimaging Data. *Biometrics*, **51**, 627-638.

Box, G.E.P. and Jenkins, G.M. (1976). *Time Series Analysis: Forecasting and Control*, 2nd edition, Holden-Day, San Francisco, CA.

Cressie, N. (1985). Fitting variogram models by weighted least squares. *Journal of the International Association for Mathematical Geology*, **17**, 563-586.

Heyde, C.C. (1997). *Quasi-Likelihood and Its Application: A Genral Approach to Optimal Parameter Estimation*, Springer-Verlag, New York.

Journel, A.G. and Huijbregts, C.J. (1978). *Mining Geostatistics.* Academic Press, London.

Lahiri, S.N. (200x). Central Limit Theorems for Weighted Sums of a Spatial Process Under a Class of Stochastic and Fixed Design. (???)

Lahiri, S., Lee, Y.D. and Cressie, N. (2002) On asymptotic distribution and asymptotic efficiency of least squares estimators of spatial variogram parameters. *Journal of Statistical Planning and Inference*, **103**, 65-85.

Liang, K. and Zeger, S (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13-22.

Matheron, G. (1962). Traite de Geostatistique Appliquee, Tom **I**. *Memoires du Bureau de Recherches Geologiques et Minieres*, No. **14**. Editions Technip, Paris.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Model*, 2nd edition, Chapman and Hall.