مجله‌ی پژوهش‌های آماری ایران
پژوهشگده‌ی آمار
دوره‌ی ۶، شماره‌ی ۱، بهار و تابستان ۱۳۸۸، صص ۱۴۱–۱۴۷
J. Statist. Res. Iran **6** (2009): 141–147

Evaluation and Application of the Gaussian-Log Gaussian Spatial Model for Robust Bayesian Prediction of Tehran Air Pollution Data

Hamidreza Zareifard and Majid Jafari Khaledi*

Tarbiat Modares University

Extended Abstract. Air pollution is one of the major problems of Tehran metropolis. Regarding the fact that Tehran is surrounded by Alborz Mountains from three sides, the pollution due to the cars traffic and other polluting means causes the pollutants to be trapped in the city and have no exit without appropriate wind guff. Carbon monoxide (CO) is one of the most important sources of pollution in Tehran air. The concentration of carbon monoxide increases remarkably at the city regions with heavy traffic. Due to the negative effects of this gas on breathing metabolism and people brain activities, the modeling and classifying of the CO amounts in order to control and reduce it, is very noteworthy. For this reason Rivaz et al. (2007) using a Gaussian model presented the space-time analysis of the Tehran air pollution based on the observations from 11 stations for measuring the air pollution. Although assuming the Gaussian observations causes the simplicity of the inferences such as prediction, but often this assumption is not true in reality. One of the outrage factors from normality assumption is the outlying observations. For example in Tehran air pollution issue, the Sorkhe Hesar station indicates very low pollution compare to the other stations due to locating in a forest region. Therefore this observation could be considered as an outlying observation. Whereas the presence of such data causes the thickening of distribution tails and increasing the kurtosis coefficient, therefore in this situation normal distribution which has a narrower tails can not be used.

* Corresponding author

Generally identifying and modeling the outlying observations is one of the main issues that statistician have been faced with since long time ago and many different solutions have been presented so far to overcome the problems arising from such observations. Amongst all these solutions, robust methods can be mentioned (Militino et al., 2006, and Cerioli and Riani, 1999). In these methods with normality observations assumption, the aim is to present a robust analysis. But there might be an outlying observation which belongs to the same pattern of other data. In this case applying those distributions with thicker tails compare to the normal distribution could be useful. This matter was evaluated by Jeffreys (1961) for the first time. Maronna (1976) and Lang et al. (1989) evaluated the verifying maximum likelihood estimation for the model in which the errors imitating the student-t distribution. West (1984) also used the scale mixture of normal distribution families for modeling the outlying observations. Fernandez and Steel (2000) also evaluated the existence of posterior distribution and its moments by introducing the improper prior distributions for West model. In the field of geostatistical data, Palacios and Steel (2006) introduced the extended Gaussian model as below by considering the errors distribution from the scale mixture of normal distributions family:

$$Z(x) = f'(x)\beta + \sigma \frac{\varepsilon(x)}{\sqrt{\lambda(x)}} + \tau\rho(x),$$

in which mean surface is assumed to be a linear function of $f'(x) = (f(x_1), \ldots, f(x_n))$ with unknown regression coefficients vector β. Further $\varepsilon(\cdot)$ is a second-order stationary error process with mean 0 and unit variance and a correlation function depending only on the distance between points,

$$\text{corr}\{\varepsilon(x_i), \varepsilon(x_j)\} = C_\theta\left(||x_i - x_j||\right) = C_\theta\left(||h||\right).$$

Also the $\lambda(\cdot)$ random field is considered independent from $\rho(\cdot)$ and $\varepsilon(\cdot)$ fields. In addition to that it is assumed that the $\ln\lambda(\cdot)$ random field is a Gaussian with finite-dimensional distributions:

$$\ln(\lambda) = (\ln\lambda_1, \ldots, \ln\lambda_n)' \sim N_n\left(-\frac{\nu}{2}\underline{1}, \nu C_\theta\right),$$

in which $\underline{1}$ is the units vector and the matrix $C_\theta = (C_\theta(||x_i - x_j||))$. They also considered the correlation function of $C_\theta(\cdot)$ as the Matren flexible class,

$$C_\theta\left(||h||\right) = \frac{1}{2^{\theta_2 - 1}\Gamma(\theta_2)}(\theta_2)^{\theta_2} k_{\theta_2}(\theta_2),$$

where $\theta = (\theta_1, \theta_2)$ with $\theta_1 > 0$ the range parameter and θ_2 the smoothness parameter and where k_{θ_2} is the modified Bessel function of third kind of order θ_2 (Stein, 1999). Also $\rho(\cdot)$ denotes an uncorrelated Gaussian process with mean 0 and unitary variance, which is used for modeling the measurement errors and small-scale variation or the so-called "nugget effect". It is noteworthy that in this model $\rho(\cdot)$ and $\varepsilon(\cdot)$ random fields are considered independent from each other. The σ and τ parameters are positive and the ratio of $\omega^2 = \frac{\tau^2}{\sigma^2}$ indicates the relative importance of the nugget effect.

Based on this model, the likelihood function will be as follows:

$$L(\beta, \sigma^2, \tau^2, \theta; z) = f(z|\beta, \sigma^2, \tau^2, \theta) = \int_{R^+} \dots \int_{R^+} f(z|\beta, \sigma^2, \tau^2, \theta, \Lambda)\, dP_{\lambda_1} \dots dP_{\lambda_n},$$

in which:

$$f(z|\beta, \sigma^2, \tau^2, \theta, \Lambda) = N_n\left(X\beta, \sigma^2\left(\Lambda^{-\frac{1}{2}} C_\theta \Lambda^{-\frac{1}{2}}\right) + \tau^2 I\right),$$

$X = (f(x_1), \dots, f(x_n))'$ and $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$. Therefore due to the frequency method problems in analysis of the model such as maximizing the above likelihood function, the Bayesian approach has been used for inference and prediction.

For Bayesian analysis firstly it is necessary to determine the prior distributions of the model parameters. Since each of the model parameters controls a specific characteristic of the field, therefore it is assumed that all parameters are independent from each other and thus the prior distribution could be written as follows:

$$\pi(\beta, \sigma^2, \omega^2, \nu, \theta) = \pi(\beta)\pi(\sigma^2)\pi(\omega^2)\pi(\nu)\pi(\theta).$$

Berger et al. (2001) indicated that posterior distributions corresponding to the improper priors such as Jefferys' prior might become improper. Therefore to make sure that the posterior is proper, proper priors were considered for each of the model parameters.

In order to predict response variable in the new location, the predictive distribution must be determined. For this objective, if $Z_0 = Z(x_0)$ corresponds to the value of response variable in the location x_0, then the Bayesian predictive distribution is given by

$$f(z_0|z) = \int f(z_0|z, \lambda, \eta, \lambda_0)\pi(\lambda_0|z, \lambda, \eta)\pi(\lambda, \eta|z) d\lambda\ d\eta\ d\lambda_0,$$

in which $\eta = (\beta, \sigma^2, \omega^2, \nu, \theta)$ and λ_0 is mixing variable corresponds to the location x_0. Due to the field is conditionally Gaussian, we have

$$f(z_0, z|\lambda, \eta, \lambda_0) = N_{n+1}\left(\mu^*, \sigma^2 \Lambda^{*\frac{1}{2}} C_\theta \Lambda^{*\frac{1}{2}} + \tau^2 I_{n+1}\right),$$

in which $\boldsymbol{\lambda} = \text{diag}(\boldsymbol{\lambda}, \lambda_0)$ and in addition,

$$\mu^* = \begin{pmatrix} f'(t_0)\beta \\ X\beta \end{pmatrix} \qquad C_\theta^* = \begin{pmatrix} 1 & r'_\theta \\ r_\theta & C_\theta \end{pmatrix},$$

where r_θ is the vector of elements $(C_\theta(||x_0 - x_j||)), j = 1, ..., n$. As a result, $Z_0|z, \lambda, \eta, \lambda_0$ has normal distribution with mean and variance

$$E(Z_0|z, \lambda, \eta, \lambda_0) = f'(t_0)\beta + \lambda_0^{-\frac{1}{2}} r'_\theta \Lambda^{-\frac{1}{2}} \left(\Lambda^{-\frac{1}{2}} C_\theta \Lambda^{-\frac{1}{2}} + \omega^2 I\right)^{-1} (z - X\beta),$$

$$\text{var}(Z_0|z, \lambda, \eta, \lambda_0) = \sigma^2 \left\{\lambda_0^{-1} + \omega^2 - \lambda_0^{-1} r'_\theta \Lambda^{-\frac{1}{2}} \left(\Lambda^{-\frac{1}{2}} C_\theta \Lambda^{-\frac{1}{2}} + \omega^2 I\right)^{-1} \Lambda^{-\frac{1}{2}} r_\theta\right\}.$$

Also, because $p(\ln \lambda_0|\lambda, z, \eta) = p(\ln \lambda_0|\lambda, \nu)$, conditional distribution of $\ln \lambda_0|\lambda, \nu$ is normal with mean and variance

$$E(\ln \lambda_0|\lambda, \nu) = -\frac{\nu}{2} + r'_\theta C_\theta^{-1} \left(\ln \lambda + \frac{\nu}{2}\underline{1}\right),$$

$$\text{var}(\ln \lambda_0|\lambda, \nu) = \nu(1 - r'_\theta C_\theta^{-1} r_\theta).$$

.

Since the analytical calculation of the Bayesian predictive distribution and consequently the Bayesian spatial prediction is very difficult or even impossible, using the Markov chain Monte Carlo methods, a drawing from posterior distribution of $\pi(\lambda, \eta|z)$ was conducted and then by replacing the collected samples in the $\pi(\lambda_0|z, \lambda, \eta)$ and sampling from this distribution and again replacing the collected samples in the $f(z_0|z, \lambda, \eta, \lambda_0)$ and sampling from this distribution, samples of Bayesian predictive distribution of $f(z_0|z)$ in the form of $\{z_0^{(k)}\}_{k=1}^l$ could be generated. Therefore an approximation of Bayesian spatial prediction and predicting variance will be as follows:

$$\hat{Z}_0 = \frac{\sum_{k=1}^l z_0^{(k)}}{l},$$

$$\text{var}(Z_0|z) \approx \frac{\sum_{k=1}^l (z_0^{(k)})^2}{l} - \left\{\frac{\sum_{k=1}^l z_0^{(k)}}{l}\right\}^2.$$

But sampling of the posterior distribution of $\pi(\lambda, \eta|z)$ and consequently conducting the above process is very difficult. Therefore to facilitate the sampling, using the augmentation method, the vector ε was augmented to the joint posterior and thus sampling is performed from posterior distribution $\pi(\lambda, \varepsilon, \eta|z)$ where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$. Therefore, if $\{\lambda^{(i)}, \varepsilon^{(i)}, \eta^{(i)}\}_{i=1}^m$ are samples generated from $\pi(\lambda, \varepsilon, \eta|z)$ then $\{\lambda^{(i)}, \eta^{(i)}\}_{i=1}^m$ are samples generated from posterior distribution $\pi(\lambda, \eta|z)$.

In generalized inverse Gaussian (GIG) model, the observations with small λ_i's tend to be away from the mean surface and to be considered outlier in some ways. In other words, these observations belong to a region with larger observational variance relative to the rest of the space. Therefore Palacios and Steel (2006) tested the below hypothesis using the Bayesian factor to identify the suspicious outlying observations.

$$\begin{cases} H_0 : \lambda_i = 1 \\ H_1 : \lambda_i \neq 1 \end{cases}$$

But since calculating the Bayesian factor is very complex and time consuming, in this paper applying the Highest Posterior Density (HPD) is recommended to determine the outlying observations. Because in this situation the posterior distribution does not have a closed form, and therefore Chen and Shao algorithm (1998) can be used to determine this region. In this algorithm if $\{\lambda_{i_j}\}_{j=1}^n$ denote ergodic MCMC sample from the posterior distribution $\pi(\lambda_i|z)$ and $\lambda_{i(j)}$ is the jth ordered statistic, then an HPD region will be obtained for λ_i as below:

$$R_{k^*}^i(n) = (\lambda_{i_{(k^*)}}, \lambda_{i_{(k^*+[(1-\alpha)n])}}),$$

in which $[(1-\alpha)n]$ denotes the integer part of $(1-\alpha)n$ and k^* is selected in a way that:

$$\lambda_{i_{(k^*+[(1-\alpha)n])}} - \lambda_{i_{(k^*)}} = \min_{1 \leqslant k \leqslant n-[(1-\alpha)n]} (\lambda_{i_{(k+[(1-\alpha)n])}} - \lambda_{i_{(k)}}).$$

We also by using a simulation example evaluated the capability of the Gaussian-Log-Gaussian model to determine the outlying observations. Based on this example, it was observed that under Gaussian-Log-Gaussian model, determining the outlying observations using an HPD region is desirably possible and meanwhile the calculation time of the HPD region is remarkably and noticeably reduced and it is easily determinable.

In the sequel, also using the criteria the mean squared prediction error of the predicting error and cross validation of the simulated data and Tehran air pollution, the GIG model capability to robust Bayesian prediction was evaluated and based on that the appropriate predicting performance of the GIG model was observed compare to the Gaussian model. This could be due to the undesirable effects of the outlying observations on the Gaussian model results and also robustness of the GIG model to the presence of such data.

Keywords. Gaussian-log Gaussian spatial model; robust spatial prediction; Bayesian approach; highest posterior density region; Markov chain Mont Carlo methods; mean square prediction error.

Reference

Berger, J.O., De Oliveira, V. and Sanso, B. (2001). Objective Bayesian analysis of spatially correlated data, *Journal of the American Statistical Association*, **93**, 1361-1374.

Cerioli, A. and Riani, M. (1999). The ordering of spatial data and the detection of multiple outliers, *Journal of Computational and Graphical Statistics*, **8**, 239-258.

Chen, M. and Shao, Q. (1998). Monte Carlo estimation of Bayesian credible and HPD intervals, *Journal of Computational and Graphical Statistics*, **7**, 69-92.

Fernandez, C. and Steel, M.F.J. (2000). Bayesian regression analysis with scale mixtures of normals, *Econometric Theory*, **16**, 80-101.

Jeffreys, H. (1961). *Theory of Probability*, Oxford University Press, London.

Lange, K.L., Little, R.J.A. and Taylor J.M.G. (1989). Robust statistical modeling using the T-distribution, *Journal of the American Statistical Association*, **84**, 881-896.

Militino, A.F., Palacios, M.B. and Ugarte, M.D. (2006). Outliers detection in multivariate spatial linear models, *Journal of Statistical Planning and Inference*, **136**, 125-146.

Maronna, R. (1976). Robust M-estimators of multivariate location and scatter, *Annals of Statistics*, **4**, 51- 67.

Palacios, M.B. and Steel M.F.J. (2006). Non-Gaussian Bayesian geostatistical modeling, *Journal of the American Statistical Association*, **101**, 604-618.

Rivaz, F., Mohammadzadeh, M. and Khaledi, M.J. (2007). Empirical Bayes prediction for space-Time data under separable models, *Journal of statistical Research*, **1**, 45-61

Stein, M.L. (1999). *Interpolation of Spatial Data: Some Theory of Kriging*, Springer-Verlag, New York.

West, M. (1984). Outlier models and prior distributions in Bayesian linear regression, *Journal of the Royal Statistical Society*, Series B, **46**, 431-439.

Hamidreza Zareifard
Department of Statistics,
Faculty of Mathematical Sciences,
Tarbiat Modares University,
Tehran, Iran.
e-mail: *zareifard@modares.ac.ir*

Majid Jafari Khaledi
Department of Statistics,
Faculty of Mathematical Sciences,
Tarbiat Modares University,
Tehran, Iran.
e-mail: *jafari-m@modares.ac.ir*

The full version of the paper, in Persian, appears on pages 1–23.