Spatial Parametric Bootstrap Method for Analysis of Finite Strain Data in Geology

Nasrollah Iranpanah*
Department of Statistics, University of Isfahan, Isfahan, Iran

Abstract:
The bootstrap is a computer-based method for assigning measures of accuracy to statistical estimators. Efron considers the bootstrap in the case where observations are independent and identically distributed (iid bootstrap). Sometimes, the iid bootstrap is applied for analysis of dependent data (e.g. time series and spatial data) incorrectly. For example, in geology used iid bootstrap for analysis of finite strain data from a Sheeprock thrust sheet that are spatially dependent. In this paper, we consider the iid bootstrap and parametric bootstrap for the analysis of spatial data and compare these two methods in two Monte Carlo simulation studies. We then use the parametric bootstrap to analyze Mukul's data.

Keywords: Finite strain data; Variogram parameters; iid bootstrap; Parametric bootstrap; Monte Carlo simulation.
Mathematics Subject Classification (2010): 62M30, 62F40, 62M40.

1 Introduction

Spatial models Cressie (1993) have been extensively used in many disciplines, such as structural geology to analyze dependent data collected from different spatial locations. Determination of the spatial correlation structure of the data and prediction are two important problems in statistical analysis of spatial data. To do so, a parametric variogram

*Speaker: n.iranpanah@gmail.com

model is often fitted to the empirical variogram of the data. Since there is no closed form for the variogram parameters estimators, they are usually computed numerically. The bootstrap method can be used to estimate the bias, variance and distribution of variogram parameter estimator. The bootstrap technique Efron (1979) is a very general method to measure the accuracy of a statistical estimator, in particular for parameter estimation from independent identically distributed (iid) variables. This independence assumption is important, and mistakenly assuming independence when spatial dependence is present can lead to misleading conclusions.

For spatially dependent data (e.g. time series and spatial data), the block bootstrap and parametric bootstrap methods can be used. Like the iid bootstrap for independent data, the block bootstrap provides tools for statistical analysis of dependent data such as spatial data Hall (1985); Lahiri (2003) without requiring stringent structural assumptions. Iranpanah *et al* (2009) introduced spatial semi-parametric bootstrap method for spatial data analysis. Fernandez-Casal *et al* (2020) introduced nonparametric bootstrap approach for unconditional risk mapping under heteroscedasticity.

Finite strain data is a source of information when studying the geometry and kinematics of thrust sheets in geology and in the construction of retrodeformable balanced cross-sections. Mukul *et al* (2004) used the iid bootstrap for the analysis of some finite strain data collected from 56 locations in a deformed volume of rock from the Sheeprock thrust sheet in the Provo salient of the Sevier fold-and-thrust belt in western United States. Iranpanah *et al* (2009) showed that independence is not a valid assumption for this data set, and they determined the best variogram model for spatial prediction of these data.

In this paper, we propose the use of the spatial parametric bootstrap (SPB) method for such data in order to correctly accommodate the presence of spatial correlation. The iid bootstrap and SPB methods are compared in two Monte Carlo simulation studies, where it is shown that SPB is substantially more accurate iid bootstrap when spatial correlation is present. Further, the SPB method is used to estimate the bias, variance and distribution of the sample mean and variogram parameter estimators in an analysis of the finite strain data.

2 Spatial Statistics

Usually a random field $\{Z(s) : s \in D\}$ is used for modeling spatial data, where the index set D is a subset of Euclidean space R^d, $d \geq 1$. Suppose $\mathbf{Z} = (Z(s_1), \ldots, Z(s_N))^T$ is a vector of observations of a second order stationary random field $Z(\cdot)$ with constant unknown mean $\mu = E[Z(s)]$ and covariogram $\sigma(h) = \mathrm{Cov}[Z(s), Z(s+h)]; s, s+h \in D$.

At a given location $s_0 \in D$, the best linear unbiased predictor for $Z(s_0)$ is the ordinary kriging predictor, $\hat{Z}(s_0) = \lambda^T \mathbf{Z}$, with kriging variance $\sigma_k^2(s_0) = E[\hat{Z}(s_0) - Z(s_0)]^2 = \sigma(0) - \lambda^T \sigma + m$, where $\lambda^T = (\sigma + \mathbf{1}m)^T \Sigma^{-1}$ and $m = (1 - \mathbf{1}^T \Sigma^{-1} \sigma)(\mathbf{1}^T \Sigma^{-1} \mathbf{1})^{-1}$, $\mathbf{1} = (1, \ldots, 1)^T$, $\sigma = (\sigma(s_0 - s_1), \ldots, \sigma(s_0 - s_N))^T$ and Σ is an $N \times N$ matrix with $(i, j)^{th}$ element given by $\sigma(s_i - s_j)$ Cressie (1993).

In reality, the covariogram is unknown and needs to be estimated based on the observations. An empirical estimator of covariogram is defined as $\hat{\sigma}(h) = N_h{}^{-1} \sum_{N(h)} [(Z(s) - \bar{Z})(Z(s+h) - \bar{Z})]$, where $\bar{Z} = N^{-1} \sum_{i=1}^{N} Z(s_i)$ is the sample mean, $N(h) = \{(s_i, s_j) : s_i - s_j \simeq h; i, j = 1, \cdots, N\}$ and N_h is the number of elements of $N(h)$. To fit a valid parametric covariogram model $\sigma(h; \theta)$ such as the exponential, spherical, or Gaussian models, several methods such as maximum likelihood (ML), restricted maximum likelihood (REML), ordinary least squares (OLS) and generalized least squares (GLS) can be applied to estimate θ. Kent and Mohammadzadeh (1999) obtained a spectral approximation to the likelihood for an intrinsic random field. This approximation can be used to compute $\hat{\theta}$ numerically.

3 IID Bootstrap

Let $X_1, \ldots, X_n$ be iid random variables with common unknown distribution F. Suppose, $\mathbf{X} = \{X_1, \ldots, X_n\}$ denotes the random data and let $T = t(\mathbf{X}; F)$, be a random variable of interest. The goal is to find an accurate approximation to the unknown distribution of T or to some population characteristic, e.g., the variance of T. The bootstrap method of Efron (1979) provides an effective way of addressing these problems without any model assumptions on F. The iid bootstrap algorithm is done with the following steps:

Step 1. *Bootstrap sample.*
Given $\mathbf{X}$, we draw a simple random sample $\mathbf{X}^* = \{X_1^*, \ldots, X_n^*\}$ with replacement from $\mathbf{X}$. Thus, conditional on $\mathbf{X}$, $\{X_1^*, \ldots, X_n^*\}$ are iid random variables with $P_*(X_1^* = X_i) = \frac{1}{n}$, $i = 1, \ldots, n$, where P_* denotes the conditional probability given $\mathbf{X}$. Hence, the common distribution of X_i^*'s is given by the empirical distribution $F_n(x) = n^{-1} \sum_{i=1}^{n} I(X_i \leq x)$, where $I(A)$ is the indicator function equal to 1 when A is true and equal to 0 otherwise.

Step 2. *Bootstrap version T^*.*
We define the bootstrap version T^* of T by replacing $\mathbf{X}$ with $\mathbf{X}^*$ and F with F_n as $T^* = t(\mathbf{X}^*; F_n)$.

Step 3. *Bootstrap estimators.*
We can estimate the population characteristics, e,g. the bias, variance of T or the unknown

distribution of T as following bootstrap estimators

$$\text{Bias}_*(T^*) = E_*(T^*) - \hat{T}, \text{Var}_*(T^*) = E_*[(T^*) - E_*(T^*)]^2, G_*(t) = P_*(T^* \leq t).$$

Step 4. *Monte Carlo approximation.*
When closed-form expressions for bootstrap estimators are not available, determine their precision measures by Monte Carlo simulation as follows. We repeat steps 1-2, B times (e.g., B=1000) to obtain bootstrap replicates $T_1^*, \ldots, T_B^*$ of T^*. The Monte Carlo approximations to the above bootstrap estimators are given by

$$\widehat{\text{Bias}}_*(T^*) = \frac{1}{B}\sum_{b=1}^{B} T_b^* - \hat{T}, \widehat{\text{Var}}_*(T^*) = \frac{1}{B}\sum_{b=1}^{B}(T_b^* - \frac{1}{B}\sum_{b=1}^{B} T_b^*)^2, \widehat{G}_*(t) = \frac{1}{B}\sum_{b=1}^{B} I(T_b^* \leq t).$$

4 Spatial Parametric Bootstrap

Suppose $\mathbf{Z} = (Z(s_1), \cdots, Z(s_N))^T$ are observations of a stationary random field $\{Z(s) : s \in D\}$ with unknown constant mean $\mu = E(Z(s))$ and unknown covariance matrix Σ. The Cholesky decomposition allows Σ to be decomposed as the matrix product $\Sigma = LL^T$, where L is a lower triangular $N \times N$ matrix. Let $\mathbf{Z} = \mu\mathbf{1} + L\epsilon$, where $\epsilon = (\epsilon(s_1), \ldots, \epsilon(s_N))^T$ is a vector of iid random variables with zero mean and unit variance from a normal distribution. The SPB algorithm is described by following steps:

Step 1. *Bootstrap Sample.*
Let the spatial dependence structure be estimated by the covariance matrix $\hat{\Sigma}$ and let $\hat{L}$ be the lower triangular $N \times N$ matrix from Cholesky decomposition $\hat{\Sigma} = \hat{L}\hat{L}^T$. The bootstrap sample $\mathbf{Z}^* = (Z^*(s_1), \cdots, Z^*(s_N))^T$ can be determined as $\mathbf{Z}^* = \hat{\mu}\mathbf{1} + \hat{L}\epsilon^*$, where $\epsilon^* = (\epsilon^*(s_1), \ldots, \epsilon^*(s_N))^T$ is a vector of iid random variables from $N(0, 1)$.

Steps 2-4.
We continue steps 2-4 in iid bootstrap algorithm.

5 Simulation Study

In this section, we conducted two simulation studies to compare iid bootstrap and SPB variance estimators of $\sigma_{\text{O}}^2 = \text{Var}[\sqrt{N}(\bar{Z} - \mu)]$ and $\sigma_{\text{G}}^2 = \text{Var}[\sqrt{N}(\hat{\hat{\mu}} - \mu)]$, where $\bar{Z} = N^{-1}\sum_{i=1}^{N} Z(s_i)$ and $\hat{\hat{\mu}}$ are the OLS and GLS plug-in estimators of μ.

Let $\{Z(s) : s \in D \subset N^2\}$ be a zero mean stationary Gaussian process with the exponential covariogram

$$\sigma(h;\theta) = \begin{cases} c_0 + c_1 & ||h|| = 0 \\ c_1 \exp\{\frac{-||h||}{a}\} & ||h|| \neq 0, \end{cases}$$

Table 1: True values of $\sigma^2_{\rm O} = N\text{Var}(\bar{Z})$, estimates of the normalized bias and MSE for iid bootstrap and SPB variance estimators $\hat{\sigma}^2_{\rm O}$ based on 10000 simulations.

		$\theta_1 = (1,1,1)^T$			$\theta_2 = (0,2,2)^T$		
Method	n	$\sigma^2_{\rm O}$	Bias	MSE	$\sigma^2_{\rm O}$	Bias	MSE
IIDB	6	5.279	-0.649	0.429	19.994	-0.928	0.862
SPB			-0.261	0.356		-0.344	0.455
IIDB	12	6.311	-0.691	0.479	32.074	-0.945	0.893
SPB			-0.064	0.277		-0.076	0.334
IIDB	24	6.890	-0.712	0.507	40.598	-0.952	0.907
SPB			0.018	0.163		0.049	0.222

where $\theta = (c_0, c_1, a)^T$ are nugget effect, partial sill and range, respectively. For the parameter values $\theta_1 = (1,1,1)^T$ (weak dependence) and $\theta_2 = (0,2,2)^T$ (strong dependence), we generate $\mathbf{Z} = \{Z(s_1), \cdots, Z(s_N)\}$ as realization of the Gaussian random field $Z(\cdot)$ over three rectangular regions $D = n \times n;\ n = 6, 12, 24$ on the planar integer grid.

Example 1. Comparison of iid bootstrap and SPB: estimating the variance of the sample mean

In this example, we compare the iid bootstrap and SPB estimators of $\sigma^2_{\rm O} = N\text{Var}(\bar{Z}) = N^{-1}\mathbf{1}^T\Sigma\mathbf{1}$. To apply the iid bootstrap method, we identify the bootstrap sample $\mathbf{Z}^* = \{Z^*(s_1), \cdots, Z^*(s_N)\}$ as simple random sampling with replacement of $\mathbf{Z}$. The iid bootstrap (IIDB) estimator $\hat{\sigma}^2_{\rm O,IIDB} = N\text{Var}_*(\bar{Z}^*)$ of $\sigma^2_{\rm O}$, where $\bar{Z}^* = N^{-1}\sum_{i=1}^N Z^*(s_i)$ has a closed form given by

$$\hat{\sigma}^2_{\rm O,IIDB} = S_N^2 = N^{-1}\sum_{i=1}^N [Z(s_i) - \bar{Z}]^2.$$

To define the SPB version of the $Z(\cdot)$-process over D, we apply steps of 1-4 in section 4. First, covariance matrix estimate $\hat{\Sigma}$ is computed by the plug-in estimator $\hat{\sigma}(h;\theta) = \sigma(h;\hat{\theta})$, where $\hat{\theta} = (\hat{c}_0, \hat{c}_1, \hat{a})^T$ is the ML estimator of θ. Then the bootstrap vector $\epsilon^* = (\epsilon^*(s_1), \ldots, \epsilon^*(s_N))^T$ is generated as iid from $N(0,1)$. Finally, the SPB sample $\mathbf{Z}^* = (Z^*(s_1), \ldots, Z^*(s_N)))^T$ is given by $\mathbf{Z}^* = \hat{\mu}\mathbf{1} + \hat{L}\epsilon^*$. The SPB estimator $\hat{\sigma}^2_{\rm O,SPB} = N\text{Var}_*(\bar{Z}^*)$ of $\sigma^2_{\rm O}$ has a closed form given by $\hat{\sigma}^2_{\rm O,SPB} = N^{-1}\mathbf{1}^T\hat{\Sigma}\mathbf{1}$.

Table 1 shows the true values of $\sigma^2_{\rm O}$ and estimates of the normalized bias $E(\hat{\sigma}^2_{\rm O})/\sigma^2_{\rm O} - 1$ and mean square error MSE $= E(\hat{\sigma}^2_{\rm O}/\sigma^2_{\rm O} - 1)^2$ by the iid bootstrap and SPB methods based on 10000 simulations for each region D and covariance parameters θ_1 and θ_2. The simulation result indicates that the SPB estimators are dramatically preferable to the iid bootstrap versions, especially for stronger dependence structure and larger sample sizes. The reason is that spatial dependence in the data makes the iid bootstrap inappropriate. The simulation result also indicates for the SPB method the normalized bias and MSE decrease as n increases whereas for the iid bootstrap method the normalized bias and MSE increase as n increases.

6 Analysis of the Finite Strain Data

Strain is an important component of the total displacement field in the emplacement of a thrust sheet. The finite strain tensor in a penetratively deformed thrust sheet is a spatial variable. In this section, we apply SPB method for the finite strain data collected by Mukul *et al* (2004). These data are collected from a part of the Sheeprock thrust sheet in the southern Sheeprock Mountain and the West Tintic Mountains, north-central Utah. The strain was measured in the quartzite of the Sheeprock thrust sheet and the geostatistics analysis was illustrated using the X/Z strain axial ratios. A sample from 56 locations was collected systematically from the quartzite on a square grid with spacing of approximately 625 m on the map along and across the strike of the Sheeprock thrust. The sample grids covered a 5 km $\times$ 3 km rectangular area in the Sheeprock thrust sheet. Iranpanah *et al* (2009) choose the best variogram model from 64 different variogram models and compared these variogram models with Mukul's model Mukul *et al* (2004) for the analysis of the spatial kriging predictor. Mukul *et al* (2004) used the iid bootstrap method for analysis of these data, in spite of the fact that they are spatially dependent. Our goal is find estimates of the bias, variance and distribution for the sample mean and variogram parameter estimators with SPB method for these data.

The first step in the SPB algorithm is the estimation of the correlation structure. To estimate of correlation structure, a valid parametric variogram $2\gamma(\cdot,\theta)$ is fitted to the empirical variogram $2\hat{\gamma}(\cdot)$. The spherical variogram

$$2\gamma(h;\theta) = \begin{cases} 0 & ||h|| = 0 \\ c_0 + c_1(\frac{3}{2}\frac{||h||}{a} - \frac{1}{2}(\frac{||h||}{a})^3) & 0 < ||h|| \leq a \\ c_0 + c_1 & ||h|| \geq a \end{cases}$$

is fitted to the finite strain data by ML method with $\hat{\theta} = (\hat{c}_0, \hat{c}_1, \hat{a}) = (0.0045, 0.0030, 2818)$. The correlation structure can be estimated from fitted semi-variogram function as $\hat{\Sigma} = \sigma(h;\hat{\theta}) = \sigma(0;\hat{\theta}) - \gamma(h;\hat{\theta})$. Then, we determine $\hat{L}$ from Cholesky decomposition $\hat{\Sigma} = \hat{L}\hat{L}^T$. Finally, the bootstrap sample is determined as $\mathbf{Z}^* = \hat{\mu}\mathbf{1} + \hat{L}\epsilon^*$, where bootstrap vector ϵ^* is generated using iid random variables $N(0,1)$.

Next suppose that the sample mean $T_1 = \bar{Z}$ and the variogram parameter estimates $\hat{\theta} = (T_2, T_3, T_4) = (\hat{c}_0, \hat{c}_1, \hat{a})$ are of interest, where $T_i = t_i(\mathbf{Z})$; $i = 1, 2, 3, 4$. For these data $\bar{Z} = 1.28$ and also $\hat{\theta} = (\hat{c}_0, \hat{c}_1, \hat{a}) = (0.0045, 0.0030, 2818)$. The SPB version T_i^* of T_i is $T_i^* = t_i(\mathbf{Z}^*)$, where $\mathbf{Z}^*$ is the SPB sample. We estimate the precision measures $\text{Bias}(T_i)$ and

$\text{Var}(T_i)$ and distribution $G_{T_i}(t)$ by the SPB method. Table 2 shows the estimates of SPB bias and SPB variance for the sample mean and the estimates of variogram parameters based on B=1000 times bootstrap replicates. Figure 1 shows a histogram of the sample mean and estimates of the variogram parameters.

Table 2: Estimates of SPB bias and SPB variance for sample mean and parameter variogram estimators based on B=1000 times bootstrap replicates for the finite strain data.

T_i^*	Bias_*	Var_*
$\bar{Z}^*$	0.0133	0.0020
c_0^*	0.0014	2.5×10^{-6}
c_1^*	0.0022	6.7×10^{-5}
a^*	-487	$1.4 \times 10^{+6}$

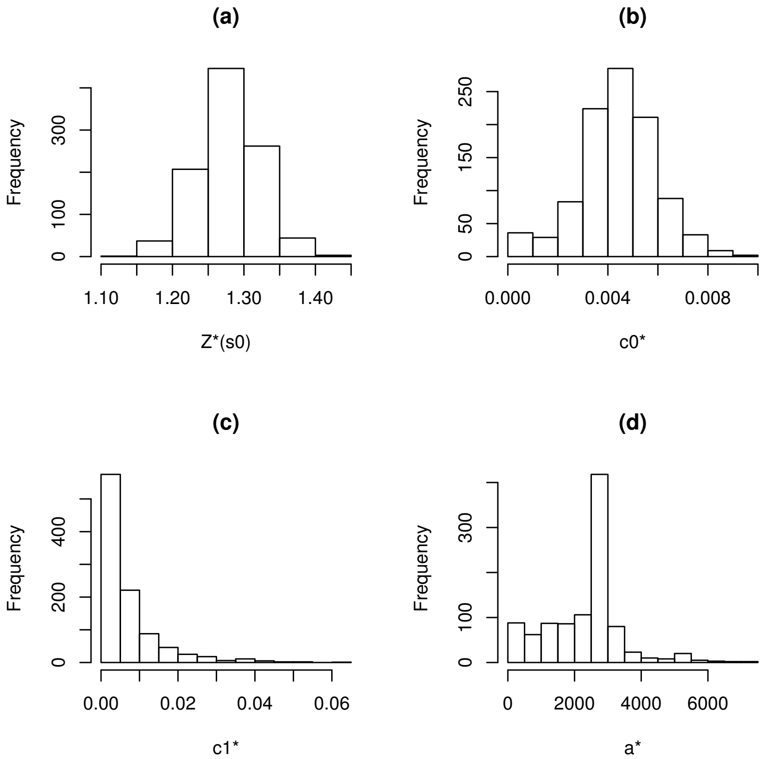

Figure 1: Histogram of (a) sample mean and parameter variogram estimators (b) nugget effect, (c) partial sill, (d) range based on B=1000 bootstrap replicates for the finite strain data.

7 Conclusion

Sometimes, the iid bootstrap method has been used incorrectly for the analysis of spatial data. The SPB method is a more reliable method which can be applied to assign measures of accuracy for estimators in spatial statistics. In this paper, we applied SPB to estimate the bias, variance and distribution of the sample mean and the variogram parameter estimates of a set of finite strain data. We compared iid bootstrap and SPB methods for sample mean variance in simulation studies. In the simulation studies, the SPB estimators are more accurate than the iid bootstrap estimators, especially for stronger dependence structure and larger sample size.

References

Cressie, N. (1993), *Statistics for Spatial Data*. John Wiley, New York.

Efron, B. (1979), Bootstrap Method; Another Look at the Jackknife. *The Annals of Statistics* **7**, 1-26.

Fernandez-Casal, R., Castillo-Paez, S., Garcia-Soidan, P. (2020), nonparametric bootstrap approach for unconditional risk mapping under heteroscedasticity. *Spatial Statistics*. **40**, 1-14.

Hall, P. (1985), Resampling a Coverage Pattern. *Stochastic Processes and their Applications*. **20**, 231-246.

Iranpanah, N., Mohammadzadeh, M., Vahidi Asl, M.Q. Yassaghi, A. (2009), Spatial Data Analysis of Finite Strain Data Across a Thrust Sheet using R. *Computer and Geosciences*. **35**, 626-634.

Iranpanah, N., Mohammadzadeh, M., Taylor, C.C., (2009), Comparison between Block and Semi-Parametric Bootstrap Spatial Data Analysis. *Computation Statistics And Data Analysis*. **55**, 578-587.

Kent, J.T., and Mohammadzadeh, M. (1999), Spectral Approximation to the Likelihood for an Intrinsic Random Field. *Journal of Multivariate Analisis* **70**, 136-155.

Lahiri, S.N., (2003), *Resampling Methods for Dependent Data*. Springer-Verlag, New York.

Mukul, M., Roy, D., Satpathy, S., and Kumar, V.A., (2004), Bootstrapped Statistics: a more Robust Approach to the Analysis of Finite Strain Data. *Journal of Structural Geology* **26**, 595-600.