A Resampling-Based Stochastic Approximation Method for Analysis of Large Geostatistical Data

Article *in* Journal of the American Statistical Association · March 2013

DOI: 10.1080/01621459.2012.746061

5 authors, including:

Faming Liang
University of Florida
109 PUBLICATIONS **1,652** CITATIONS
SEE PROFILE

Yichen Cheng
Georgia State University
10 PUBLICATIONS **72** CITATIONS
SEE PROFILE

Qifan Song
Purdue University
10 PUBLICATIONS **87** CITATIONS
SEE PROFILE

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA

A Resampling-Based Stochastic Approximation Method for Analysis of Large Geostatistical Data

Faming LIANG, Yichen CHENG, Qifan SONG, Jincheol PARK, and Ping YANG

The Gaussian geostatistical model has been widely used in modeling of spatial data. However, it is challenging to computationally implement this method because it requires the inversion of a large covariance matrix, particularly when there is a large number of observations. This article proposes a resampling-based stochastic approximation method to address this challenge. At each iteration of the proposed method, a small subsample is drawn from the full dataset, and then the current estimate of the parameters is updated accordingly under the framework of stochastic approximation. Since the proposed method makes use of only a small proportion of the data at each iteration, it avoids inverting large covariance matrices and thus is scalable to large datasets. The proposed method also leads to a general parameter estimation approach, maximum mean log-likelihood estimation, which includes the popular maximum (log)-likelihood estimation (MLE) approach as a special case and is expected to play an important role in analyzing large datasets. Under mild conditions, it is shown that the estimator resulting from the proposed method converges in probability to a set of parameter values of equivalent Gaussian probability measures, and that the estimator is asymptotically normally distributed. To the best of the authors' knowledge, the present study is the first one on asymptotic normality under infill asymptotics for general covariance functions. The proposed method is illustrated with large datasets, both simulated and real. Supplementary materials for this article are available online.

KEY WORDS: Asymptotic normality; Infill asymptotics; Large spatial data; *U*-statistics.

1. INTRODUCTION

Geostatistics is a branch of spatial statistics that deals with data obtained by sampling a spatially continuous process $\{X(s)\}$, $s \in \mathbb{R}^2$, at a discrete set of locations $\{s_i, i = 1, \ldots, n\}$ in a spatial region of interest $A \subset \mathbb{R}^2$. Consider a Gaussian geostatistical model,

$$Y(s_i) = \mu(s_i) + X(s_i) + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} N(0, \tau^2), \tag{1}$$

where $Y(s_i)$ denotes the observation at location s_i, $\mu(s_i)$ denotes the mean of $Y(s_i)$, $\{X(s_i)\}$ denotes a spatial Gaussian process with $E(X(s_i)) = 0$, $\text{var}(X(s_i)) = \sigma^2$, and $\text{corr}(X(s_i), X(s_j)) = \rho(\|s_i - s_j\|)$ for an appropriate correlation function with Euclidean distance $\|\cdot\|$, and τ^2 denotes the nugget variance. The correlation function is chosen from a certain parametric family, such as the Matérn, exponential, or spherical covariance models (Cressie 1993). Under this model, $\{Y(s)\}$ follows a multivariate Gaussian distribution as follows:

$$[Y(s_1), \ldots, Y(s_n)]^T \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{2}$$

where $\boldsymbol{\mu} = (\mu(s_1), \ldots, \mu(s_n))^T$, $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{R} + \tau^2 \boldsymbol{I}$, $\boldsymbol{I}$ is the $n \times n$ identity matrix, and $\boldsymbol{R}$ is an $n \times n$ correlation matrix with the (i, j)th element $\rho(\|s_i - s_j\|)$. Model (1) is perhaps the most popular model in geostatistics. It can be easily extended to the regression setting with the mean $\mu(s_i)$ being replaced by

$$\mu(s_i) = \beta_0 + \sum_{j=1}^{p} \beta_j \boldsymbol{c}_j(s_i), \tag{3}$$

where $\boldsymbol{c}_j(\cdot)$ denotes the jth explanatory variable, and β_j is the corresponding regression coefficient.

It is generally recognized that the parameter estimation for model (1) suffers from two severe difficulties. First, evaluation of the likelihood function of this model involves inverting the matrix $\boldsymbol{\Sigma}$. This is infeasible when n is large, because the complexity of matrix inversion increases as $O(n^3)$. Second, some parameters of this model may be inconsistently estimable due to the existence of equivalent probability measures for the Gaussian process (Stein 2004; Zhang 2004).

To alleviate the first difficulty, various methods have been proposed in the literature. These methods can be roughly grouped into four categories, namely, covariance tapering, lower dimensional space process approximation, likelihood approximation, and Markov random field approximation. The method of covariance tapering is to set the covariances at large distances to zero but still keep the original covariances for proximate sites, see, for example, Furrer et al. (2006), Kaufman et al. (2008), and Du et al. (2009). The tapered covariance matrix is sparse, and thus can be inverted much more efficiently than inverting a full matrix of the same dimension. However, not all parameters of the covariance function are consistently estimable for this method. The lower dimensional space approximation methods seek to approximate the spatial process $\{X(s)\}$ by a lower dimensional space process $\{\tilde{X}(s)\}$ with the use of smoothing techniques, such as kernel convolutions, moving averages, low-rank splines, or basis functions, see, for example, Wikle and Cressie (1999), Banerjee et al. (2008), and Cressie and Johannesson (2008). Concerns with these methods are adequacy of approximations.

Faming Liang (E-mail: *fliang@stat.tamu.edu*) is Professor, and Yichen Cheng (E-mail: *ycheng@stat.tamu.edu*), Qifan Song (E-mail: *qsong@stat.tamu.edu*), and Jincheol Park (E-mail: *jpark@stat.tamu.edu*) are graduate students, Department of Statistics, Texas A&M University, College Station, TX 77843-3143. Ping Yang (E-mail: *pyang@tamu.edu*) is Professor, Department of Atmospheric Sciences, Texas A&M University, College Station, TX 77843-3143. Liang's research was partially supported by grants from the National Science Foundation (DMS-1007457 and DMS-1106494) and the award (KUS-C1-016-04) made by King Abdullah University of Science and Technology (KAUST). Yang's research was partially supported by the endowment funds related to the David Bullock Harris Chair in Geosciences at the College of Geosciences, Texas A&M University. The authors thank the editor, associate editor, and the referees for their constructive comments that have led to significant improvement of this article.

Journal of the American Statistical Association
March 2013, Vol. 108, No. 501, Theory and Methods
DOI: 10.1080/01621459.2012.746061

For large datasets, the dimension of the approximation process $\{\tilde{X}(s)\}$ can still be very high, degrading the applicability of these methods. The likelihood approximation methods seek to approximate the likelihood function in the spectral domain of the spatial process (Stein 1999; Fuentes 2007; Matsuda and Yajima 2009) or by a product of conditional densities (Vecchia 1988; Jones and Zhang 1997; Stein et al. 2004). Concerns about these methods include adequacy of the likelihood approximation and some implementation issues. Expertise is required for selecting an appropriate spectral density estimate or a sequence of conditional densities. As suggested by its name, the Markov random field approximation method is to approximate the spatial process $\{X(s)\}$ by a Markov random field (Rue and Tjelmeland 2002; Rue and Held 2005), but this method is mainly used for regular lattice data. An exception is Park and Liang (2012), where the authors extend this method to irregular lattice data.

The second difficulty has been addressed by several authors, including Mardia and Marchall (1984), Stein (1999, 2004), Chen et al. (2000), and Zhang (2004), among others. There are two different asymptotics in spatial statistics: expanding-domain asymptotics, where more and more data are collected in increasing domains while the sampling density stays constant, and infill asymptotics, where data are collected by sampling more and more densely in a fixed domain. Asymptotic properties of estimators are quite different under the two asymptotics. For example, the maximum likelihood estimator is consistent and asymptotically normally distributed for many covariance models under expanding-domain asymptotics (Mardia and Marchall 1984), whereas such consistent estimators do not exist under infill asymptotics (Chen et al. 2000). Under infill asymptotics, for exponential and Matérn correlation functions, not all covariance parameters are consistently estimable, and part of them are consistently estimable only after certain reparameterizations (Zhang 2004). Also, it is unclear whether or not the estimates are asymptotically normally distributed.

In this article, we propose a resampling-based stochastic approximation (RSA) method for addressing the two difficulties. At each iteration of RSA, a small subsample is drawn from the full dataset, and then the current estimate of parameters is updated accordingly under the framework of stochastic approximation (Robbins and Monro 1951). Since RSA makes use of only a small proportion of the data at each iteration, it avoids inverting large covariance matrices and thus is scalable to large datasets. Note that RSA is conceptually very different from the existing methods. The existing methods are to approximate the model (1) using a computationally convenient model, whose parameters are usually not directly comparable with the parameters of model (1). Instead, RSA seeks to perform a data dimension reduction, while continuing to work on the model (1). RSA also leads to a general parameter estimation approach, maximum mean log-likelihood estimation (MMLE), which includes the popular maximum (log)-likelihood estimation (MLE) approach, as a special case, and is expected to play a major role for large data analysis. Theoretical properties of the RSA method are explored in this article. Under mild conditions, we show the RSA estimator converges in probability to a set of parameter values of equivalent Gaussian probability measures, and that the estimator is asymptotically normally distributed. To the best of the authors' knowledge, the present study is the first one on asymptotic normality under infill asymptotics for general covariance functions. The numerical results indicate that RSA can work well for large datasets.

The remainder of this article is organized as follows. Section 2 describes the RSA method. Sections 3 and 4 present some theoretical results about RSA. Section 5 illustrates RSA using simulated examples along with comparisons with some existing methods. In Section 6, RSA is applied to a large real data example. Section 7 concludes the article with a brief discussion.

2. A RESAMPLING-BASED STOCHASTIC APPROXIMATION APPROACH

Consider the Gaussian model (1). Although the number of observations can be large, the model contains only a small number of parameters. This motivates us to develop the resampling-based stochastic approximation method, which, at each iteration, makes use of only a small proportion of the data for parameter estimation. The method can be described as follows.

Let $\boldsymbol{Z}(\boldsymbol{s}) = (Y(s_1^*), \ldots, Y(s_m^*))^T$ denote a sample drawn, randomly and without replacement, from the full dataset $\mathcal{Y} = \{Y(s_1), \ldots, Y(s_n)\}$. In statistics, $\boldsymbol{Z}(\boldsymbol{s})$ is also called a subsample of $\mathcal{Y}$. In the following, we will denote $\boldsymbol{Z}(\boldsymbol{s})$ by $(\boldsymbol{Z}, \boldsymbol{S})$, which makes the dependence of $\boldsymbol{Z}$ on the sites $\boldsymbol{S} = (s_1^*, \ldots, s_m^*)$ implicit. Given $\boldsymbol{S}$, we model $\boldsymbol{Z}$ using the same model as for $\mathcal{Y}$; that is,

$$\boldsymbol{Z}|\boldsymbol{S} \sim N_m(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z), \tag{4}$$

where $\boldsymbol{\mu}_z = (\mu(s_1^*), \ldots, \mu(s_m^*))^T$, $\boldsymbol{\Sigma}_z = \sigma^2 \boldsymbol{R}_z + \tau^2 \boldsymbol{I}$, and $\boldsymbol{R}_z$ is an $m \times m$ correlation matrix with the (i, j)th element given by a correlation function $\rho(\|s_i^* - s_j^*\|)$. As in model (1), $\boldsymbol{\mu}_z$ can also be modeled by a regression,

$$\boldsymbol{\mu}_z = \beta_0 \mathbf{1}_m + \sum_{j=1}^{p} \beta_j \boldsymbol{c}_j, \tag{5}$$

where $\mathbf{1}_m$ denotes an m-vector of 1's and $\boldsymbol{c}_j$ denotes the jth explanatory variable of the model.

Let $\boldsymbol{\theta}$ denote the parameter vector of a Gaussian geostatistical model specified by (1) and (3). Therefore, $\boldsymbol{\theta}$ includes τ^2, σ^2, $\beta_0, \ldots, \beta_p$, and the parameters in the correlation function $\rho(\cdot)$. We propose to estimate $\boldsymbol{\theta}$ by minimizing the Kullback–Leibler divergence,

$$\mathrm{KL}(f_{\boldsymbol{\theta}}, g) = -\int\int \log\left(\frac{f_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{s})}{g(\boldsymbol{z}|\boldsymbol{s})}\right) g(\boldsymbol{z}|\boldsymbol{s}) g(\boldsymbol{s}) d\boldsymbol{z} d\boldsymbol{s}, \tag{6}$$

where $f_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{s})$ is a multivariate normal density specified by (4), $g(\boldsymbol{z}|\boldsymbol{s})$ denotes the unknown true density function from which the data were generated although we are using (4) for data analysis, and $g(\boldsymbol{s})$ is the distribution of the sites $\boldsymbol{s}$. Jensen's inequality implies that $\mathrm{KL}(f_{\boldsymbol{\theta}}, g) \geq 0$. As a method of parameter estimation, minimizing the Kullback–Leibler divergence has been widely used in the literature, see, for example, Dowe et al. (1998), Liang and Zhang (2008), and Chen et al. (2009). Using subsamples randomly drawn from $\mathcal{Y}$, the Kullback–Leibler divergence can

be approximated by

$$\widehat{\mathrm{KL}}(f_{\boldsymbol{\theta}}, g \mid \mathcal{Y}) = C - \binom{n}{m}^{-1} \sum_{i=1}^{\binom{n}{m}} \log f_{\boldsymbol{\theta}}(\boldsymbol{z}_i \mid \boldsymbol{s}_i),$$

where C denotes a constant related to the entropy of $g(\boldsymbol{z}, \boldsymbol{s})$, and $\binom{n}{m}$ is the binomial coefficient. Then, the stochastic approximation method can be used to estimate $\boldsymbol{\theta}$ by solving the systems of equations:

$$\begin{aligned} \frac{\partial \widehat{\mathrm{KL}}(f_{\boldsymbol{\theta}}, g|\mathcal{Y})}{\partial \boldsymbol{\theta}} &= -\binom{n}{m}^{-1} \sum_{i=1}^{\binom{n}{m}} \frac{\partial \log f_{\boldsymbol{\theta}}(\boldsymbol{z}_i|\boldsymbol{s}_i)}{\partial \boldsymbol{\theta}} \\ &= -\binom{n}{m}^{-1} \sum_{i=1}^{\binom{n}{m}} H(\boldsymbol{\theta}, \boldsymbol{z}_i, \boldsymbol{s}_i) = 0, \end{aligned} \tag{7}$$

where $H(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{s}) = \partial \log f_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{s})/\partial \boldsymbol{\theta}$ is the first-order derivative of $\log f_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{s})$ with respect to $\boldsymbol{\theta}$, and $(\boldsymbol{z}_i, \boldsymbol{s}_i)$ denotes a random sample drawn from $\mathcal{Y}$. Note that $\partial \widehat{\mathrm{KL}}(f_{\boldsymbol{\theta}}, g|\mathcal{Y})/\partial \boldsymbol{\theta}$ forms a U-statistic with the kernel $H(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{s})$.

For the purpose of illustration, we consider the exponential correlation function given by

$$\rho(h) = \exp(-h/\phi), \tag{8}$$

where h denotes the Euclidean distance between two observations, and ϕ is the correlation parameter. For the exponential model, the respective components of $H(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{s})$ in (7) are given by

$$\begin{cases} H_{\beta_0}(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{s}) = \mathbf{1}_m^T \boldsymbol{\Sigma}_z^{-1}(\boldsymbol{z} - \boldsymbol{\mu}_z), \\ H_{\beta_i}(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{s}) = \boldsymbol{c}_i^T \boldsymbol{\Sigma}_z^{-1}(\boldsymbol{z} - \boldsymbol{\mu}_z), \quad i = 1, \ldots, p, \\ H_{\phi}(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{s}) = -\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}_z^{-1}\sigma^2 \frac{d\boldsymbol{R}_z}{d\phi}\right) \\ \qquad\qquad + \frac{\sigma^2}{2}(\boldsymbol{z} - \boldsymbol{\mu}_z)^T \boldsymbol{\Sigma}_z^{-1} \frac{d\boldsymbol{R}_z}{d\phi} \boldsymbol{\Sigma}_z^{-1}(\boldsymbol{z} - \boldsymbol{\mu}_z), \\ H_{\sigma^2}(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{s}) = -\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}_z^{-1}\boldsymbol{R}_z\right) \\ \qquad\qquad + \frac{1}{2}(\boldsymbol{z} - \boldsymbol{\mu}_z)' \boldsymbol{\Sigma}_z^{-1} \boldsymbol{R}_z \boldsymbol{\Sigma}_z^{-1}(\boldsymbol{z} - \boldsymbol{\mu}_z), \\ H_{\tau^2}(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{s}) = -\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}_z^{-1}\right) + \frac{1}{2}(\boldsymbol{z} - \boldsymbol{\mu}_z)^T \boldsymbol{\Sigma}_z^{-2}(\boldsymbol{z} - \boldsymbol{\mu}_z), \end{cases} \tag{9}$$

where $d\boldsymbol{R}_z/d\phi$ is an $m \times m$-matrix with the (i, j)th element given by $h_{ij}/\phi^2 e^{-h_{ij}/\phi}$, and h_{ij} denotes the Euclidean distance between site i and site j.

The varying truncation stochastic approximation algorithm (Andrieu et al. 2005) was adopted in this article for estimation of $\boldsymbol{\theta}$. Let Θ denote the space of $\boldsymbol{\theta}$, and let $\{\mathcal{K}_s, s \geq 0\}$ be a sequence of compact subsets of Θ such that

$$\bigcup_{s \geq 0} \mathcal{K}_s = \Theta, \quad \text{and} \quad \mathcal{K}_s \subset \mathrm{int}(\mathcal{K}_{s+1}), \quad s \geq 0, \tag{10}$$

where $\mathrm{int}(A)$ denotes the interior of set A. Let $\{a_t\}$ and $\{b_t\}$ be two monotone, nonincreasing, positive sequences. Let $\mathcal{X}_0$ be a subset of $\mathcal{X}$. Let $\mathbb{T} : \Theta \to \mathcal{K}_0$ be a measurable function which maps a point $\boldsymbol{\theta}$ in Θ to a random point in $\mathcal{K}_0$; that is, $\boldsymbol{\theta}$ will be reinitialized in $\mathcal{K}_0$. Let π_t denote the number of truncations performed until iteration t, and $\pi_0 = 0$. The varying truncation stochastic approximation algorithm starts with a random choice of $\boldsymbol{\theta}_0$ in the space $\mathcal{K}_0$, and then iterates between the following steps:

Algorithm 1 Resampling-Based Stochastic Approximation (RSA) Algorithm

(i) Draw $(\boldsymbol{Z}_{t+1}, \boldsymbol{S}_{t+1})$ from the set $\{Y(s_1), \ldots, Y(s_n)\}$ at random and without replacement.

(ii) Update each component of $\boldsymbol{\theta}_t$ in the following equations:

$$\begin{cases} \beta_0^{(t+\frac{1}{2})} = \beta_0^{(t)} + a_{t+1} H_{\beta_0}(\boldsymbol{\theta}_t, \boldsymbol{Z}_{t+1}, \boldsymbol{S}_{t+1}), \\ \beta_i^{(t+\frac{1}{2})} = \beta_i^{(t)} + a_{t+1} H_{\beta_i}(\boldsymbol{\theta}_t, \boldsymbol{Z}_{t+1}, \boldsymbol{S}_{t+1}), \\ \qquad\qquad i = 1, \ldots, p, \\ \phi^{(t+\frac{1}{2})} = \phi^{(t)} + a_{t+1} H_{\phi}(\boldsymbol{\theta}_t, \boldsymbol{Z}_{t+1}, \boldsymbol{S}_{t+1}), \\ (\sigma^2)^{(t+\frac{1}{2})} = (\sigma^2)^{(t)} + a_{t+1} H_{\sigma^2}(\boldsymbol{\theta}_t, \boldsymbol{Z}_{t+1}, \boldsymbol{S}_{t+1}), \\ (\tau^2)^{(t+\frac{1}{2})} = (\tau^2)^{(t)} + a_{t+1} H_{\tau^2}(\boldsymbol{\theta}_t, \boldsymbol{Z}_{t+1}, \boldsymbol{S}_{t+1}). \end{cases}$$

(iii) If $\|\boldsymbol{\theta}_{t+\frac{1}{2}} - \boldsymbol{\theta}_t\| \leq b_t$ and $\boldsymbol{\theta}_{t+\frac{1}{2}} \in \mathcal{K}_{\pi_k}$, where $\|\cdot\|$ denote the Euclidean norm of a vector, then set $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_{t+\frac{1}{2}}$ and $\pi_{t+1} = \pi_t$; otherwise, set $\boldsymbol{\theta}_{t+1} = \mathbb{T}(\boldsymbol{\theta}_t)$ and $\pi_{t+1} = \pi_t + 1$.

To facilitate simulations, one may reparameterize ϕ, σ^2, and τ^2 in logarithms to ensure their positivity during simulations. Let $\tilde{\boldsymbol{\theta}}_n$ denote a solution to Equation (7), where the subscript n indicates its dependence on the sample $\mathcal{Y}$. Let $\hat{\boldsymbol{\theta}}_n^{(t)}$ denote the estimator of $\tilde{\boldsymbol{\theta}}_n$ obtained by the RSA algorithm at iteration t, that is, $\hat{\boldsymbol{\theta}}_n^{(t)} = \boldsymbol{\theta}_t$ as produced in the simulation. The RSA estimator has very nice theoretical properties. In Sections 3 and 4, we consider, respectively, two types of asymptotics, the infill asymptotics of $\tilde{\boldsymbol{\theta}}_n$ and the stochastic approximation asymptotics of $\hat{\boldsymbol{\theta}}_n^{(t)}$. Under mild conditions, we show that both $\tilde{\boldsymbol{\theta}}_n$ and $\hat{\boldsymbol{\theta}}_n^{(t)}$ are asymptotically normally distributed.

Regarding the stopping rule of RSA, we note that the stopping rule for the general stochastic approximation algorithm has been extensively studied in the literature, see for example, Yin (1990) and Glynn and Whitt (1992). A popular stopping rule is sequential stopping; that is, letting the simulation run until the volume of a confidence set of the parameters achieves a prescribed value. The conditions that guarantee the asymptotic validity of this rule were established in Glynn and Whitt (1992). In this article, we adopt a multiple-run variant of this rule: Run the algorithm multiple times for the same dataset, and then determine the number of iterations at which all runs have converged, that is, producing about the same estimates. To diagnose the convergence of RSA, the Gelman–Rubin method (Gelman and Rubin 1992) can be used. The Gelman–Rubin method was designed for diagnosis of convergence of an iterative simulation algorithm which is not necessarily an Markov chain Monte Carlo (MCMC) algorithm.

3. INFILL ASYMPTOTICS OF $\tilde{\boldsymbol{\theta}}_n$

On infill asymptotics of $\tilde{\boldsymbol{\theta}}_n$, we establish two theorems. In Theorem 1, we show that $\tilde{\boldsymbol{\theta}}_n$ converges to a minimizer of (6). In Theorem 2, we show that $\tilde{\boldsymbol{\theta}}_n$ is asymptotically normally

distributed. The major challenge with infill asymptotics is that the correlation between observations is gradually increasing as the number of observations increases; that is, the observation sequence is not a stationary sequence. This nonstationarity disenabled the use of conventional laws of large numbers, and thus, the asymptotic theory, such as consistency and asymptotic normality, is extremely difficult to be established under the framework of infill asymptotics.

3.1 Convergence of $\tilde{\boldsymbol{\theta}}_n$

Let $\mathcal{S}_n = \{X_1, \ldots, X_n\} = \{X(s_1), \ldots, X(s_n)\}$ denote a set of samples drawn from a stationary random field defined on a bounded region. For the infill asymptotic, we note that it generally behaves like an asymptotic of resampling from a finite population: Even infinite samples can be drawn (with replacement) from the finite population, and the accuracy of approximation to the underlying system/process is still limited to the finite population. Motivated by this observation, we introduce an auxiliary finite population for $\mathcal{S}_n$; that is, treating $\mathcal{S}_n$ as a simple random sample from a finite population $\mathcal{S}_N = \{X_1, \ldots, X_n, X_{n+1}, \ldots, X_N\}$, where $X_{n+1}, \ldots, X_N$ are drawn in the same sampling procedure from the same random field as for the samples $\{X_1, \ldots, X_n\}$. Then, the asymptotic of $\tilde{\boldsymbol{\theta}}_n$ can be studied by making use of some known results of finite population U-statistics. As previously mentioned, $\partial \widehat{\mathrm{KL}}(f_{\boldsymbol{\theta}}, g|\mathcal{Y})/\partial \boldsymbol{\theta}$ forms a U-statistic. Introduction of auxiliary populations is crucial to the proof of some results, particularly, the asymptotic normality of $\tilde{\boldsymbol{\theta}}_n$.

Let

$$U_n = \binom{n}{m}^{-1} \sum_{i=1}^{\binom{n}{m}} \psi\left(X_1^{(i)}, \ldots, X_m^{(i)}\right) \tag{11}$$

be a U-statistic defined on the random sample $\{X_1, \ldots, X_n\}$, where $\psi(\cdot)$ is the kernel of the U-statistic. Lemma 1 concerns the convergence of the U-statistic, whose proof can be found in the Appendix.

Lemma 1. Let $\{X_1, \ldots, X_n\}$ be a random sample drawn from a bounded, stationary random field. If the mapping $(x_1, \ldots, x_m) \mapsto \psi(x_1, \ldots, x_m)$ is continuous (a.e.) and $E|\psi(X_1, \ldots, X_m)|^2 < \infty$, then, as $n \to \infty$,

$$U_n \to E(\psi(X_1, \ldots, X_m)) \quad \text{in probability.}$$

Remark 1. A similar convergence result was obtained by Chatterji (1968) (see also Borovskikh 1996) for U-statistics under the assumption that $X_1, \ldots, X_n$ are symmetrically dependent. Lemma 1 relaxes this assumption to the case that $X_1, \ldots, X_n$ are generally dependent but drawn from a bounded stationary random field. We note that under expanding-domain asymptotics, an almost sure convergence of U_n can be obtained.

Suppose that the kernel $\psi(\cdot)$ depends on $\boldsymbol{\theta}$. To indicate this dependence, we rewrite $\psi(\cdot)$ as $\psi_{\boldsymbol{\theta}}(\cdot)$ and define

$$U_n(\boldsymbol{\theta}) = \binom{n}{m}^{-1} \sum_{i=1}^{\binom{n}{m}} \psi_{\boldsymbol{\theta}}\left(X_1^{(i)}, \ldots, X_m^{(i)}\right) \quad \text{and}$$
$$U(\boldsymbol{\theta}) = E(\psi_{\boldsymbol{\theta}}(X_1, \ldots, X_m)). \tag{12}$$

In the rest of this article, $\psi(\cdot)$ and $\psi_{\boldsymbol{\theta}}(\cdot)$ will be used exchangeably. In the context where $\boldsymbol{\theta}$ is not our concern, we will depress $\boldsymbol{\theta}$ from $\psi_{\boldsymbol{\theta}}$ for simplicity of notations.

Suppose that we are interested in maximizing $U(\boldsymbol{\theta})$. Let Θ denote the space of $\boldsymbol{\theta}$, and let $\Theta_0 = \{\boldsymbol{\theta}^* \in \Theta : U(\boldsymbol{\theta}^*) = \sup_{\boldsymbol{\theta}} U(\boldsymbol{\theta})\}$ denote the set of global maximizers of $U(\boldsymbol{\theta})$. To avoid triviality, Θ_0 is assumed not empty. Lemma 2 shows that if Θ is compact, then $\tilde{\boldsymbol{\theta}}_n$ converges in probability to a point in Θ_0 when the sample size becomes large.

Lemma 2. Let $\{X_1, \ldots, X_n\}$ be a random sample drawn from a bounded, stationary random field. Assume the following conditions hold:

(i) The mapping $\boldsymbol{\theta} \mapsto \psi_{\boldsymbol{\theta}}(X_1, \ldots, X_m)$ is continuous for almost all $(X_1, \ldots, X_m)$ and satisfies

$$E|\psi_{\boldsymbol{\theta}}(X_1, \ldots, X_m)|^2 < \infty. \tag{13}$$

(ii) For every sufficiently small ball $O \subset \Theta$, the mapping $(x_1, \ldots, x_m) \mapsto \sup_{\boldsymbol{\theta} \in O} \psi_{\boldsymbol{\theta}}(x_1, \ldots, x_m)$ is measurable and satisfies

$$E|\sup_{\boldsymbol{\theta} \in O} \psi_{\boldsymbol{\theta}}(X_1, \ldots, X_m)|^2 < \infty. \tag{14}$$

Then, for any estimators $\tilde{\boldsymbol{\theta}}_n$ such that $U_n(\tilde{\boldsymbol{\theta}}_n) \geq U_n(\boldsymbol{\theta}^*) + o_p(1)$ for some $\boldsymbol{\theta}^* \in \Theta_0$, for every $\epsilon > 0$ and every compact set $\mathcal{K} \subset \Theta$,

$$P(d(\tilde{\boldsymbol{\theta}}_n, \Theta_0) \geq \epsilon \text{ and } \tilde{\boldsymbol{\theta}}_n \in \mathcal{K}) \to 0,$$

where $d(\cdot, \cdot)$ denotes a distance metric.

Remark 2. The proof of this lemma is given in the online supplementary material, where the lemma is proved in a similar way to that of van der Vaart (1998) in proving the consistency of M-estimators and that of Wald (1949) in proving the consistency of maximum likelihood estimators for a set of iid random variables.

Remark 3. In Lemma 2, Θ is restricted to a compact space. To apply this lemma to a problem whose parameter space is not compact, one needs to show that the estimators are in a compact set eventually or make a suitable compactification for the parameter space.

To study the infill asymptotics of $\tilde{\boldsymbol{\theta}}_n$, we define

$$l_{\boldsymbol{\theta}}(\boldsymbol{z}, \boldsymbol{s}) = \log f_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{s}), \quad M(\boldsymbol{\theta}) = E[l_{\boldsymbol{\theta}}(\boldsymbol{z}, \boldsymbol{s})],$$
$$M_n(\boldsymbol{\theta}) = \binom{n}{m}^{-1} \sum_{i=1}^{\binom{n}{m}} l_{\boldsymbol{\theta}}(\boldsymbol{z}_i, \boldsymbol{s}_i). \tag{15}$$

Thus, $M_n(\boldsymbol{\theta})$ forms a U-statistic estimator of $M(\boldsymbol{\theta})$ with the kernel $l_{\boldsymbol{\theta}}(\boldsymbol{z})$, and minimizing (6) is equivalent to maximizing $M_n(\boldsymbol{\theta})$.

Theorem 1 shows that as $n \to \infty$, $\tilde{\boldsymbol{\theta}}_n$ converges to the set $\Theta_0 = \{\boldsymbol{\theta}^* : El_{\boldsymbol{\theta}^*}(\boldsymbol{Z}, \boldsymbol{S}) = \sup_{\boldsymbol{\theta} \in \Theta} El_{\boldsymbol{\theta}}(\boldsymbol{Z}, \boldsymbol{S})\}$ in probability. Its proof can be found in the Appendix.

Theorem 1. Let $\mathcal{Y} = \{Y(s_1), \ldots, Y(s_n)\}$ denote a random sample drawn from the spatial Gaussian model (1) defined on a bounded region, let $\tilde{\boldsymbol{\theta}}_n$ denote a solution to (7), and let $\Theta_0 = \{\boldsymbol{\theta}^* \in \Theta : El_{\boldsymbol{\theta}^*}(\boldsymbol{Z}, \boldsymbol{S}) = \sup_{\boldsymbol{\theta} \in \Theta} El_{\boldsymbol{\theta}}(\boldsymbol{Z}, \boldsymbol{S})\}$, where $(\boldsymbol{Z}, \boldsymbol{S})$

denotes a random sample of size m drawn from model (1). Assume Θ is compact, then for every $\epsilon > 0$,

$$P(d(\tilde{\boldsymbol{\theta}}_n, \Theta_0) \geq \epsilon) \to 0, \quad \text{as } n \to \infty,$$

where $d(\cdot, \cdot)$ denotes a distance metric.

Remark 4. To accommodate the case that the model is inconsistently estimable, that is, part or all parameters of the model are not consistently estimable, we assume in Theorem 1 that Θ is compact. As implied by Lemma 5, this assumption does not affect the performance of RSA, because RSA can keep $\boldsymbol{\theta}_t$ in a compact set almost surely. In simulations, we may set Θ to a huge set, say, $[-10^{100}, 10^{100}]^{d_\theta}$, which, as a practical matter, is equivalent to set $\Theta = \mathbb{R}^{d_\theta}$. Here, d_θ denotes the dimension of $\boldsymbol{\theta}$.

For the models which are consistently estimable, that is, all parameters of the model are consistently estimable, the compactness assumption of Θ can be removed. In this case, the proof of Theorem 1 can be simply accomplished with a suitable compactification of Θ; that is, setting $l_{\partial\Theta}(\boldsymbol{z}, \boldsymbol{s}) = -\infty$, where $\partial\Theta$ denotes the boundary of Θ. This is permitted in Lemma 2, and it ensures that $\tilde{\boldsymbol{\theta}}_n$ never takes the boundary values when maximizing $M_n(\boldsymbol{\theta})$.

Remark 5. As implied by Theorem 1, RSA leads to a general parameter estimation approach, maximum mean log-likelihood estimation (MMLE), which has included the maximum (log)-likelihood estimation (MLE) approach as a special case. If $m = n$, then RSA is reduced to the MLE. Due to its computational attractiveness, we expect that MMLE will play an important role for large data analysis.

3.2 Asymptotic Normality of $\tilde{\boldsymbol{\theta}}_n$

Assume that the model (1) has been appropriately reparameterized such that all parameters are consistently estimable. Under this assumption, we show that $\tilde{\boldsymbol{\theta}}_n$ is asymptotically normally distributed.

Lemma 3 concerns the asymptotic normality of the U-statistic defined on a set of samples drawn from a bounded stationary random field. It is the key to establishing the asymptotic normality of $\tilde{\boldsymbol{\theta}}_n$. To prove this lemma, we assume that the function $\psi(x_1, \ldots, x_m)$ is continuous (a.e.) and $E|\psi(X_1, \ldots, X_m)|^2 < \infty$. In addition, we impose some constraints on the sampling procedure of $\mathcal{S}_n = \{X_1, \ldots, X_n\}$: $\mathcal{S}_n$ is drawn through a procedure which ensures the following conditions hold: for $1 \leq k \leq m-1$ and any $\alpha > 0$,

$$E|\psi_{k,n}(X_1, \ldots, X_k)|^2 \text{ is uniformly bounded with respect to } n \text{ and } n^\alpha \sigma_{k,n}^2 \to \infty \text{ as } n \to \infty, \tag{16}$$

where $\psi_{k,n}(x_1, \ldots, x_k) = E\{\psi(X_1, \ldots, X_m)|X_1 = x_1, \ldots, X_k = x_k, \mathcal{S}_n\}$ is the conditional expectation of $\psi(X_1, \ldots, X_m)$ based on the finite population $\mathcal{S}_n$, and $\sigma_{k,n}^2 = \text{var}(\psi_{k,n}(X_1, \ldots, X_k))$. Let $\psi_k(x_1, \ldots, x_k) = E\{\psi(X_1, \ldots, X_m)|X_1 = x_1, \ldots, X_k = x_k\}$. Be aware that $E(|\psi_{k,n}(X_1, \ldots, X_k)|^2)$ is actually the second-order sample moments of ψ_k and $\sigma_{k,n}^2$ the sample variance of ψ_k. This assumption essentially requires that the finite sample $\{X_1, \ldots, X_n\}$ resembles the underlying random field such that $\sigma_{k,n}^2$ converges to a constant as $n \to \infty$. This assumption is satisfied except that the sampling procedure is degenerated to drawing samples from a single site or the function $\psi_k(\cdot)$ is degenerated to taking a constant value.

Lemma 3. Let $\mathcal{S}_n = \{X_1, \ldots, X_n\}$ be a random sample drawn from a bounded, stationary random field. Consider the U-statistic defined in (11). Assume the following conditions hold:

(i) The function $\psi(x_1, \ldots, x_m)$ is continuous (a.e.), and $E|\psi(X_1, \ldots, X_m)|^2 < \infty$.
(ii) $\mathcal{S}_n$ satisfies the condition (16).

Then, as $n \to \infty$,

$$(U_n - E(\psi(X_1, \ldots, X_m)))/\sqrt{\text{Var}(U_n)} \Rightarrow N(0, 1),$$

where $\Rightarrow$ denotes the convergence in distribution, and $N(0, 1)$ denotes the standard normal distribution.

Lemma 4 concerns the asymptotic normality of the estimator $\tilde{\boldsymbol{\theta}}_n$, which maximizes $U_n(\boldsymbol{\theta})$ defined in (12). Its proof can be found in the Appendix. In Lemma 4, Θ is assumed to be compact. This assumption stems from Lemma 2, which, as shown below, is the basis of Lemma 4.

Lemma 4. Let $\{X_1, \ldots, X_n\}$ be a random sample drawn from a bounded stationary random field. Assume the following conditions hold:

(i) The parameter space Θ is compact.
(ii) The kernel $\psi_{\boldsymbol{\theta}}(\cdot)$ is twice continuously differentiable on the interior of Θ, and satisfies

$$\begin{aligned} &E|\psi_{\boldsymbol{\theta}}(X_1, \ldots, X_m)|^2 \\ &\quad < \infty, \; E\left\| \frac{\partial}{\partial \boldsymbol{\theta}} \psi_{\boldsymbol{\theta}}(X_1, \ldots, X_m) \right\|^2 \\ &\quad < \infty, \; E\left\| \frac{\partial^2}{\partial \boldsymbol{\theta}^2} \psi_{\boldsymbol{\theta}}(X_1, \ldots, X_m) \right\|^2 < \infty. \end{aligned} \tag{17}$$

(iii) For every sufficiently small ball $O \subset \Theta$, the mapping $(x_1, \ldots, x_m) \mapsto \sup_{\boldsymbol{\theta} \in O} \psi_{\boldsymbol{\theta}}(x_1, \ldots, x_m)$ is measurable and satisfies

$$E|\sup_{\boldsymbol{\theta} \in O} \psi_{\boldsymbol{\theta}}(X_1, \ldots, X_m)|^2 < \infty. \tag{18}$$

(iv) $\mathcal{S}_n$ satisfies the condition (16); that is, there exists a constant C such that for $1 \leq k \leq m-1$,

$$\sup_n E\left(\left\| \frac{\partial}{\partial \boldsymbol{\theta}} \psi_{\boldsymbol{\theta},k}(X_1, \ldots, X_k) \right\|^2 \middle| \mathcal{S}_n \right) < C, \quad \text{a.s.},$$

where $\frac{\partial}{\partial \boldsymbol{\theta}} \psi_{\boldsymbol{\theta},k}(x_1, \ldots, x_k) = E\{\frac{\partial}{\partial \boldsymbol{\theta}} \psi_{\boldsymbol{\theta}}(X_1, \ldots, X_m)|X_1 = x_1, \ldots, X_k = x_k)\}$. In addition, for any $\alpha > 0$ and $1 \leq k \leq m-1$, $n^\alpha \|\boldsymbol{\Sigma}_{k,n}\| \to \infty$ as $n \to \infty$, where $\boldsymbol{\Sigma}_{k,n}$ denotes the sample covariance matrix of $\frac{\partial}{\partial \boldsymbol{\theta}} \psi_{\boldsymbol{\theta},k}(X_1, \ldots, X_k)$.

Then, for any estimators $\tilde{\boldsymbol{\theta}}_n$ such that $U_n(\tilde{\boldsymbol{\theta}}_n) \geq U_n(\boldsymbol{\theta}^*) + o_p(1)$ for some $\boldsymbol{\theta}^* \in \Theta_0$,

$$\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^* \Rightarrow N(0, \boldsymbol{H}_*^{-1} \boldsymbol{\Sigma} \boldsymbol{H}_*^{-1}),$$

where $\boldsymbol{H}_* = E\{\frac{\partial^2 \psi_{\boldsymbol{\theta}}(X_1, \ldots, X_m)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}|_{\boldsymbol{\theta} = \boldsymbol{\theta}^*}\}$ is the expected Hessian of $\psi_{\boldsymbol{\theta}}(X_1, \ldots, X_m)$ at $\boldsymbol{\theta}^*$, and $\boldsymbol{\Sigma}$ is the covariance matrix of the U-statistic defined by the kernel $\frac{\partial \psi_{\boldsymbol{\theta}}(X_1, \ldots, X_m)}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta} = \boldsymbol{\theta}^*}$.

Theorem 2 concerns the asymptotic normality of the minimizer of the Kullback–Leibler divergence, whose proof can be found in the Appendix.

Theorem 2. Let $\{Y(s_1), \ldots, Y(s_n)\}$ be a random sample drawn from the spatial Gaussian model (1) defined on a bounded region, let $\tilde{\boldsymbol{\theta}}_n$ denote a solution to (7), and let $\Theta_0 = \{\boldsymbol{\theta}^* \in \Theta : El_{\boldsymbol{\theta}^*}(\boldsymbol{Z}, \boldsymbol{S}) = \sup_{\boldsymbol{\theta}\in\Theta} El_{\boldsymbol{\theta}}(\boldsymbol{Z}, \boldsymbol{S})\}$, where $(\boldsymbol{Z}, \boldsymbol{S})$ denotes a random sample of size m drawn from model (1). Assume that Θ is compact, the model is consistently estimable, and the sampling procedure of $\{Y(s_1), \ldots, Y(s_n)\}$ satisfies the condition (iv) of Lemma 4 (with $\psi_{\boldsymbol{\theta}}(\cdot) = l_{\boldsymbol{\theta}}(\cdot)$). Then,

$$\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^* \Rightarrow N(0, \boldsymbol{H}_*^{-1}\boldsymbol{\Sigma}\boldsymbol{H}_*^{-1}), \tag{19}$$

where $\boldsymbol{H}_* = E\{\frac{\partial^2 l_{\boldsymbol{\theta}}(\boldsymbol{Z},\boldsymbol{S})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\}$ is the expected Hessian of $l_{\boldsymbol{\theta}}(\boldsymbol{z})$ at $\boldsymbol{\theta}^*$, and $\boldsymbol{\Sigma}$ is the covariance matrix of the U-statistic defined by the kernel $\frac{\partial l_{\boldsymbol{\theta}}(\boldsymbol{Z},\boldsymbol{S})}{\partial\boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$.

Remark 6. A necessary condition for the asymptotic normality of $\tilde{\boldsymbol{\theta}}_n$ is that $\boldsymbol{\theta}$ is consistently estimable. As aforementioned, under this condition, the compactness assumption of Θ can be simply removed via a suitable compactification of Θ. To keep the assumption of Θ consistent with other theorems, Θ is still assumed to be compact here.

Remark 7. As mentioned earlier, RSA is reduced to the MLE if $m = n$. However, Theorem 2 cannot be directly extended to MLE. The reason is as follows: In proving Theorem 2, m is assumed to be fixed while letting $n \to \infty$, but m will increase with n for the case of MLE. It is interesting to point out that Zhu and Stein (2005) conjectured that MLE has the same result as given in (19).

4. STOCHASTIC APPROXIMATION ASYMPTOTICS OF $\hat{\boldsymbol{\theta}}_n^{(t)}$

The stochastic approximation algorithm was first introduced by Robbins and Monro (1951) for solving the mean field function equation

$$h(\boldsymbol{\theta}) = \int_{\mathcal{X}} H(\boldsymbol{\theta}, x) g_{\boldsymbol{\theta}}(x) dx = 0, \tag{20}$$

where $\boldsymbol{\theta} \in \Theta$ is a parameter vector and $g_{\boldsymbol{\theta}}(x)$, $x \in \mathcal{X}$, is a density function depending on $\boldsymbol{\theta}$. The algorithm works by iterating between the following two steps:

Algorithm 2 Stochastic Approximation

- Generate $X_{t+1} \sim g_{\boldsymbol{\theta}_t}(x)$, where t indexes the iteration.
- Set $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + a_t H(\boldsymbol{\theta}_t, X_{t+1})$, where a_t is the gain factor.

Later, this algorithm was applied by Kiefer and Wolfowitz (1952) to solve the optimization problem of the form,

$$\max_{\boldsymbol{\theta}} \int l(\boldsymbol{\theta}, x) g_{\boldsymbol{\theta}}(x) dx, \tag{21}$$

by setting $H(\boldsymbol{\theta}, x) = \partial l(\boldsymbol{\theta}, x)/\partial\boldsymbol{\theta}$ or an estimate of the derivative when it is not explicitly available. For the problem of minimizing the Kullback–Leibler divergence (6) based on a finite set of observations, we have $X = (\boldsymbol{Z}, \boldsymbol{S})$, $H(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{s}) = \partial \log f_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{s})/\partial\boldsymbol{\theta}$ as defined in (7), and $g_{\boldsymbol{\theta}}(\boldsymbol{z}, \boldsymbol{s})$ is a uniform distribution defined on the set of all possible m-subsamples drawn from $\mathcal{Y}$.

To prove the convergence of the stochastic approximation algorithm, one often needs to impose a severe restriction on the growth rate of $h(\boldsymbol{\theta})$. To remove this restriction, Chen and Zhu (1986) proposed a varying truncation stochastic approximation algorithm. The convergence of the modified algorithm can be shown for a wide class of mean field functions, see, for example, Chen (2002). A similar varying truncation stochastic approximation algorithm has also been studied in Andrieu et al. (2005). The main difference between the two varying truncation algorithms is at their reinitialization step. In Chen and Zhu (1986), the simulation is always reinitialized at the same point while in Andrieu et al. (2005), the simulation is reinitialized in a prespecified region. Of course, these two algorithms share similar theoretical properties, such as the sampling path $\{\boldsymbol{\theta}_t\}$ can be kept in a compact set almost surely.

The remainder of this section is organized as follows. In Section 4.1, we present a varying truncation stochastic approximation algorithm in the reinitialization style of Andrieu et al. (2005), and study its convergence. In Section 4.2, we show that the RSA estimator converges to $\tilde{\boldsymbol{\theta}}_n$ almost surely, and it is asymptotically normally distributed.

4.1 Convergence of a Varying Truncation Stochastic Approximation Algorithm

Let Θ be the parameter space, which is not necessarily compact. To make the results more general, we set $\Theta = \mathbb{R}^{d_{\boldsymbol{\theta}}}$ in this section. Let $\{\mathcal{K}_s, s \geq 0\}$ be a sequence of compact subsets of Θ as defined in (10), and let $\{a_t\}$ and $\{b_t\}$ be two monotone, nonincreasing, positive sequences. A general varying truncation stochastic approximation algorithm can be described as follows. It starts with a random choice of $\boldsymbol{\theta}_0$ in the space $\mathcal{K}_0$ and then iterates between the following steps:

Algorithm 3 Varying Truncation Stochastic Approximation

(i) Generate $X_{t+1} \sim g_{\boldsymbol{\theta}_t}(x)$, where t indexes the iteration.
(ii) Set $\boldsymbol{\theta}_{t+\frac{1}{2}} = \boldsymbol{\theta}_t + a_t H(\boldsymbol{\theta}_t, X_{t+1})$, where a_t is the gain factor.
(iii) If $\|\boldsymbol{\theta}_{t+\frac{1}{2}} - \boldsymbol{\theta}_t\| \leq b_t$ and $\boldsymbol{\theta}_{t+\frac{1}{2}} \in \mathcal{K}_{\pi_k}$, then set $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_{t+\frac{1}{2}}$ and $\pi_{t+1} = \pi_t$; otherwise, set $\boldsymbol{\theta}_{t+1} = \mathbb{T}(\boldsymbol{\theta}_t)$ and $\pi_{t+1} = \pi_t + 1$. Here, $\mathbb{T}$ and π_t are defined as in Algorithm 1.

Algorithm 3 is the same as the varying truncation stochastic approximation MCMC algorithm given in Andrieu et al. (2005) except that at each iteration the new sample X_{t+1} is generated through an exact sampler instead of an MCMC sampler. Hence, Algorithm 3 can be viewed as a special varying truncation stochastic approximation MCMC algorithm by viewing the exact sampler as a special MCMC sampler. Let $P_{\boldsymbol{\theta}}$ denote the Markov transition kernel corresponding to the exact sampler. It is easy to see that it is irreducible and aperiodic, admits $g_{\boldsymbol{\theta}_t}(x)$ as the invariant distribution, and satisfies the drift condition given in Andrieu et al. (2005). Therefore, Algorithm 3 has the same convergence as the varying truncation stochastic approximation MCMC algorithm. Lemma 5 is a formal statement for this

convergence, whose proof is given in supplementary material, available online.

Lemma 5. Assume the conditions (A_1), (A_2), and (A_4) (given in the online supplementary material) hold. Let k_π denote the iteration number at which the πth truncation occurs in the simulation. Let $\mathcal{X}_0 \subset \mathcal{X}$ be such that $\sup_{x\in\mathcal{X}_0} V(x) < \infty$ and $\mathcal{K}_0 \subset \mathcal{V}_{C_0}$, where $\mathcal{V}_{C_0}$ is defined in (A_1). Let $\{\boldsymbol{\theta}_t\}$ be given by Algorithm 3. Then, there exists almost surely a number, denoted by π_s, such that $k_{\pi_s} < \infty$ and $k_{\pi_s+1} = \infty$; that is, $\{\boldsymbol{\theta}_t\}$ can be kept in a compact set almost surely. In addition,

$$d(\boldsymbol{\theta}_t, \mathcal{L}) \to 0, \quad \text{a.s.},$$

where $\mathcal{L}$ is defined in (A_1), and $d(\boldsymbol{\theta}, \mathcal{L}) = \inf_{\boldsymbol{\theta}'}\{\|\boldsymbol{\theta} - \boldsymbol{\theta}'\| : \boldsymbol{\theta}' \in \mathcal{L}\}$ denotes a distance measure induced by the Euclidean norm.

Lemma 6 concerns the asymptotic normality of $\boldsymbol{\theta}_t$. Liang and Wu (2010) studied the asymptotic normality of the varying truncation stochastic approximation MCMC estimator. By viewing the exact sampler as a special MCMC sampler, Liang and Wu's result (Theorem 2.2) implies the following lemma. We note that a similar result can also be found in Benveniste et al. (1990) (Theorem 13, chap. 4), where the asymptotic normality is established for conventional stochastic approximation MCMC algorithms (without varying truncation) under slightly different conditions.

Lemma 6. Assume the conditions (A_1), (A_2), (A_3), and (A_4) (given in the online supplementary material) hold. Let the simulation start with a point $(\boldsymbol{\theta}_0, X_0) \in \mathcal{K}_0 \times \mathcal{X}$, where $\mathcal{K}_0 \subset \mathcal{V}_{C_0}$ (defined in (A_1)) and $\sup_{X\in\mathcal{X}} V(X) < \infty$. Let $\{\boldsymbol{\theta}_t\}$ be given by Algorithm 3. Conditioned on $\Lambda(\boldsymbol{\theta}_*) = \{\boldsymbol{\theta}_t \to \boldsymbol{\theta}_*\}$,

$$\frac{\boldsymbol{\theta}_t - \boldsymbol{\theta}_*}{\sqrt{a_t}} \Rightarrow \mathbb{N}(0, \boldsymbol{\Sigma}_{sa}), \tag{22}$$

where $\boldsymbol{\theta}_* \in \mathcal{L}$ as defined in (A_1), $\mathbb{N}(\cdot,\cdot)$ denotes the Gaussian distribution, and

$$\boldsymbol{\Sigma}_{sa} = \int_0^\infty e^{(F'+\zeta I)t}\Gamma e^{(F+\zeta I)t}dt, \tag{23}$$

where F is defined in (A_3), ζ is given in Equation (7) of the online supplementary material, and Γ is defined by

$$\frac{1}{N}\sum_{t=1}^{N} E\left(\epsilon_{t+1}\epsilon_{t+1}^T \middle| \mathcal{F}_t\right) \to \Gamma,$$

with $\epsilon_{t+1} = H(\boldsymbol{\theta}_t, X_{t+1}) - h(\boldsymbol{\theta}_t)$, and $\mathcal{F}_t = \sigma\{\boldsymbol{\theta}_0, X_0, \ldots, \boldsymbol{\theta}_t, X_t\}$ being a σ-algebra formed by $\{\boldsymbol{\theta}_0, X_0, \ldots, \boldsymbol{\theta}_t, X_t\}$.

4.2 Convergence of Algorithm 1

As shown by Stein (2004) and Zhang (2004), the model (1) is inconsistently estimable for some correlation functions, such as exponential and Matérn. To accommodate this case, we restrict Θ to a compact set. This ensures that $\{\tilde{\boldsymbol{\theta}}_n\}$, the solution to (7), lies in a compact set. Under this assumption, the convergence of Algorithm 1 can be established based on Lemma 5. In simulations, Θ can be set to a huge set, for example, $[-10^{100}, 10^{100}]^{d_\theta}$, which, as mentioned before, is equivalent to set $\Theta = \mathbb{R}^{d_\theta}$. For the case that the model is consistently estimable, the compactness constraint of Θ can be simply removed following from Lemmas 5 and 6.

Theorem 3. Let $\{Y(s_1), \ldots, Y(s_n)\}$ be a random sample drawn from a spatial Gaussian model (1), which is defined on a bounded region and has an exponential correlation function. Let $\mathcal{L} = \{\boldsymbol{\theta} : \partial\widehat{\mathrm{KL}}(f_{\boldsymbol{\theta}}, g|\mathcal{Y})/\partial\boldsymbol{\theta} = 0\}$ denote the set of solutions to the system of Equations (7). Assume Θ is compact and let $\{\hat{\boldsymbol{\theta}}_n^{(t)}\}$ be given by Algorithm 1 (i.e., $\hat{\boldsymbol{\theta}}_n^{(t)} = \boldsymbol{\theta}_t$), then $\lim_{t\to\infty} d(\hat{\boldsymbol{\theta}}_n^{(t)}, \mathcal{L}) = 0$ a.s. as $t \to \infty$.

The proof of this theorem is given in the online supplementary material. Although the model (1) is inconsistently estimable for certain correlation functions, for example, exponential or Matérn, it can be reparameterized such that the resulting model is consistently estimable (Zhang 2004). Theorem 4 shows that the RSA estimator $\hat{\boldsymbol{\theta}}_n^{(t)}$ is asymptotically normally distributed, provided that the model or the reparameterized model is consistently estimable. The proof of Theorem 4 is given in the online supplementary material.

Theorem 4. Let $\{Y(s_1), \ldots, Y(s_n)\}$ be a random sample drawn from a spatial Gaussian model (1), which is defined on a bounded region and has an exponential correlation function. Assume the model (1) is consistently estimable and Θ is compact. Let $\{\hat{\boldsymbol{\theta}}_n^{(t)}\}$ be given by Algorithm 1. Then, given $\Lambda(\tilde{\boldsymbol{\theta}}_n) = \{\hat{\boldsymbol{\theta}}_n^{(t)} \to \tilde{\boldsymbol{\theta}}_n\}$,

$$\frac{\hat{\boldsymbol{\theta}}_n^{(t)} - \tilde{\boldsymbol{\theta}}_n}{\sqrt{a_t}} \Rightarrow \mathbb{N}(0, \boldsymbol{\Sigma}_{sa}), \tag{24}$$

where $\tilde{\boldsymbol{\theta}}_n$ denotes a solution to (7) and $\boldsymbol{\Sigma}_{sa}$ is as defined in Lemma 6.

As a summary of Theorems 2 and 4, we note that $\hat{\boldsymbol{\theta}}_n^{(t)}$ is asymptotically normally distributed, and its asymptotic distribution is given by

$$\hat{\boldsymbol{\theta}}_n^{(t)} \Rightarrow N(\boldsymbol{\theta}^*, a_t\boldsymbol{\Sigma}_{sa} + \boldsymbol{H}_*^{-1}\boldsymbol{\Sigma}\boldsymbol{H}_*^{-1}),$$

where $\boldsymbol{H}_*$ and $\boldsymbol{\Sigma}$ are given in Theorem 2 and $\boldsymbol{\Sigma}_{sa}$ is given in Theorem 4. The term $a_t\boldsymbol{\Sigma}_{sa}$ of the covariance matrix represents the part of Monte Carlo error in $\hat{\boldsymbol{\theta}}_n^{(t)}$.

5. SIMULATED EXAMPLES

In this section, we use three simulated examples to illustrate the performance of RSA. These examples address the following issues: (i) How is the RSA estimator related to the MLE?, (ii) How can one choose the value of m?, (iii) Is RSA feasible for very large datasets?, and (iv) Does RSA work under expanding-domain asymptotics?

To apply RSA to our examples, we reparameterize ϕ, σ^2, and τ^2 in their logarithms and set the sequence of compact subsets of Θ as follows:

$$\begin{aligned}\mathcal{K}_{\pi_t} = &[-2-\pi_t, 2+\pi_t] \times [-2-\pi_t, 2+\pi_t] \times [-\pi_t, 4+\pi_t] \\ &\times [-2-\pi_t, 2+\pi_t] \times [-2-\pi_t, 2+\pi_t],\end{aligned} \tag{25}$$

in the order of parameters β_0, β_1, $\log(\phi)$, $\log(\sigma^2)$, and $\log(\tau^2)$, where π_t denotes the number of varying truncations at iteration t. Note that this setting has been used for all examples of this

Table 1. Comparisons of RSA with MLE, the Bayesian method, and the Gaussian predictive process (GPP) method for 50 simulated datasets with nugget effects

Estimator	Size	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\phi}/\hat{\sigma}^2$	$\hat{\tau}^2$	CPU(m)
True	–	1.000	1.000	25.000	1.0	–
	100	1.022 (0.068)	0.998 (0.009)	19.278 (0.768)	0.939 (0.010)	0.3
RSA	300	1.016 (0.065)	1.000 (0.007)	22.046 (0.684)	0.974 (0.009)	6.4
	500	1.013 (0.064)	1.001 (0.007)	23.084 (0.675)	0.977 (0.008)	29.3
	700	0.997 (0.063)	0.999 (0.006)	24.023 (0.659)	0.993 (0.007)	81.5
MLE	–	1.000 (0.061)	1.000 (0.006)	25.269 (0.72)	0.999 (0.007)	19.4
Bayes	–	0.994 (0.058)	1.000 (0.006)	28.560 (0.829)	1.098 (0.009)	93.9
	36	1.031 (0.075)	0.998 (0.008)	35.199 (6.958)	1.809 (0.045)	111.2
GPP	100	1.031 (0.075)	0.998 (0.008)	21.212 (4.683)	1.773 (0.042)	343.0

Size: It refers to the subsample size for RSA and the grid size for GPP. CPU(m): CPU time (in min) cost by a single run on a 3.0 GHz personal computer (all computations of this article were done on the same computer). The numbers in the parentheses denote the standard error of the estimates (this is the same for other tables of this article).

article, including the real data examples studied in Section 6. A different choice of $\mathcal{K}_{\pi_t}$ may result in different numbers of truncations, but should not affect the convergence of RSA.

The sequences $\{a_t\}$ and $\{b_t\}$ are set in the form

$$a_t = \frac{a_0 t_0}{\max(t, t_0)}, \quad b_t = b_0 \left(\frac{t_0}{\max(t, t_0)} \right)^{\eta}, \tag{26}$$

where we set $b_0 = 100, t_0 = 400$, and $\eta = 0.55$ for all examples. For a_0, we tried two different values, 0.01 and 0.001, for different examples. In general, if m is large, we set a_0 to a smaller value; otherwise, we may try a slightly larger value for a_0. The reason is as follows: When m is large, the innovation term defined in (9) tends to have a large magnitude, so a small gain factor sequence can stabilize the convergence of RSA.

5.1 Comparison Studies

In this example, we consider a geostatistical model with measurement errors. The model is specified by (1) and (3) with $\beta_0 = \beta_1 = 1$, $\phi = 25$, $\sigma^2 = 1$, $\tau^2 = 1.0$, and the explanatory variable $\boldsymbol{c}_1$ is generated from a Gaussian distribution with mean 0 and standard deviation 0.5. Using the package *geoR* (Ribeiro Jr. and Diggle 2001), we simulated 50 datasets of size $n = 2000$ with the sampling sites uniformly distributed in a bounded region of $[0, 100] \times [0, 100]$.

For each dataset, RSA was run four times with $m = 100$, 300, 500, and 700, respectively. We set $a_0 = 0.01$ for the runs with $m = 100$, 300, and 500 and $a_0 = 0.001$ for the run with $m = 700$. Each run consisted of 2500 iterations. Our multiple pilot runs indicate that RSA can converge within 2500 iterations for this example. The numerical results are summarized in Table 1.

Stein (1999, 2004) showed that the model (1) is inconsistently estimable for the exponential correlation function. Two probability measures can be equivalent for a sampling path $\{Y(\boldsymbol{s}), \boldsymbol{s} \in A\}$ from any bounded subset A of $\mathbb{R}^d$ if $\phi_1/\sigma_1^2 = \phi_2/\sigma^2$ (see Stein 1999 for $d = 1$ and Stein 2004 for $d > 1$). For this reason, we reported in Table 2 the ratios of the estimates of ϕ and σ^2 instead of their respective estimates. As a contrast, we will show in Section 5.3 that in the scenario of domain expanding, the model is consistently estimable.

5.1.1 Comparison with MLE. For comparison, we calculated the MLE of $\boldsymbol{\theta}$ for each dataset using the package *geoR*. The comparison indicates that RSA works very well for this example. Even when m is as small as 100, the estimates of β_0 and β_1 are extremely accurate. As m increases, the estimates of ϕ/σ^2 and τ^2 are improved and get closer and closer to their true values. Figure 1 compares the MLE and RSA estimates for the 50 datasets. It shows that for each dataset, the estimates from the two methods are close to each other. In the presence of nugget effects, the parameters ϕ, σ^2, and τ^2 are usually difficult to estimate. As shown in Table 1, both τ^2 and ϕ/σ^2 tend to be underestimated when m is small. However, these biases tend to disappear as m increases. This finding is consistent with the results reported in the literature: The nugget effect can reduce the convergence rate of the MLE of the Gaussian process model, see, for example, Chen et al. (2000).

Regarding central processing unit (CPU) time, we note that for this example, 2500 iterations have been excessively long for the convergence of RSA. Figure 2 shows the trajectories of RSA collected in six independent runs for a dataset with nugget effects, and Figure 3 shows the Gelman–Rubin shrink factor for the ϕ/σ^2-trajectories of the six runs (the plots are similar for other trajectories). Both plots indicate that 1000–1500 iterations have been good enough for RSA to converge. Around the 1300th iteration, the upper limit of the 95% confidence level of the Gelman–Rubin shrink factor starts to be lower than 1.1. It is worth noting that the convergence rate of RSA is almost independent of n, the sample size of the full dataset. As shown in the next example, where each dataset consists of 50000 observations, the CPU time cost by RSA is about the same as for this example. However, this is different for MLE, whose CPU time increases as $O(n^3)$. Although it was quite fast for this example, it took an extremely long CPU time for the next example (with

Table 2. A comparison of RSA and MLE for 50 simulated datasets without nugget effects (see Table 1 for notations)

Estimator	Size	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\phi}/\hat{\sigma}^2$
	100	1.027 (0.053)	1.002 (0.002)	24.453 (0.414)
RSA	300	1.008 (0.049)	0.999 (0.002)	24.932 (0.451)
	500	1.025 (0.053)	1.004 (0.002)	24.990 (0.512)
MLE	–	1.031 (0.054)	1.001 (0.001)	25.291 (0.739)
True	–	1.000	1.000	25.000

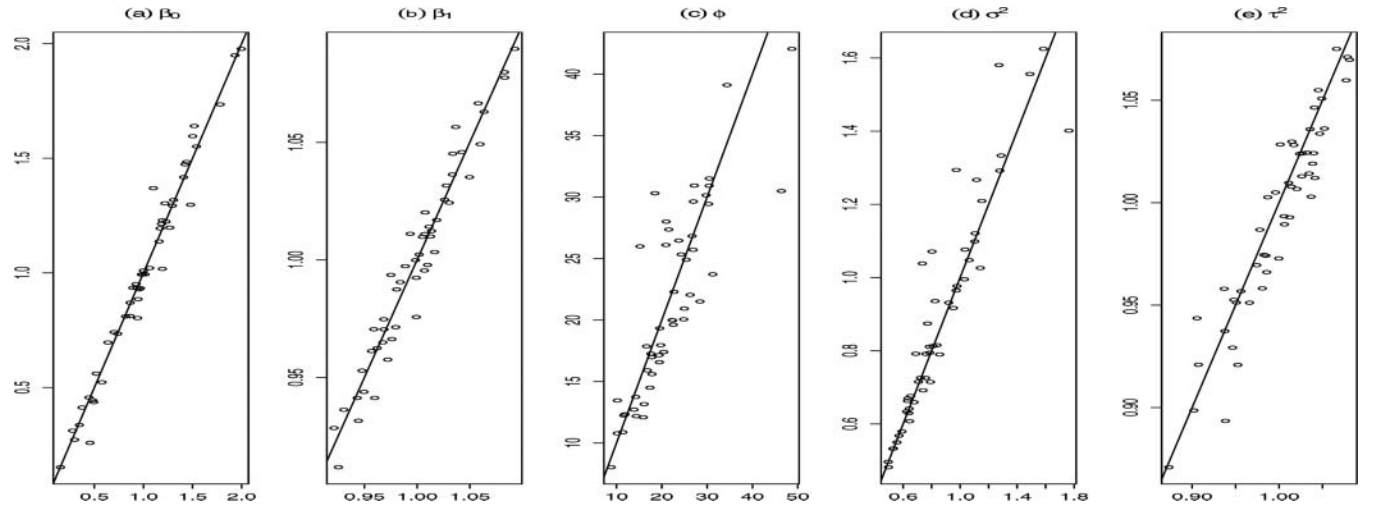

Figure 1. Comparison of the MLE and RSA estimates for the simulated example. The horizontal axis shows the MLE and the vertical axis shows the RSA estimates obtained with $m = 700$: (a) plot for β_0; (b) plot for β_1; (c) plot for ϕ; (d) plot for σ^2; and (e) plot for τ^2.

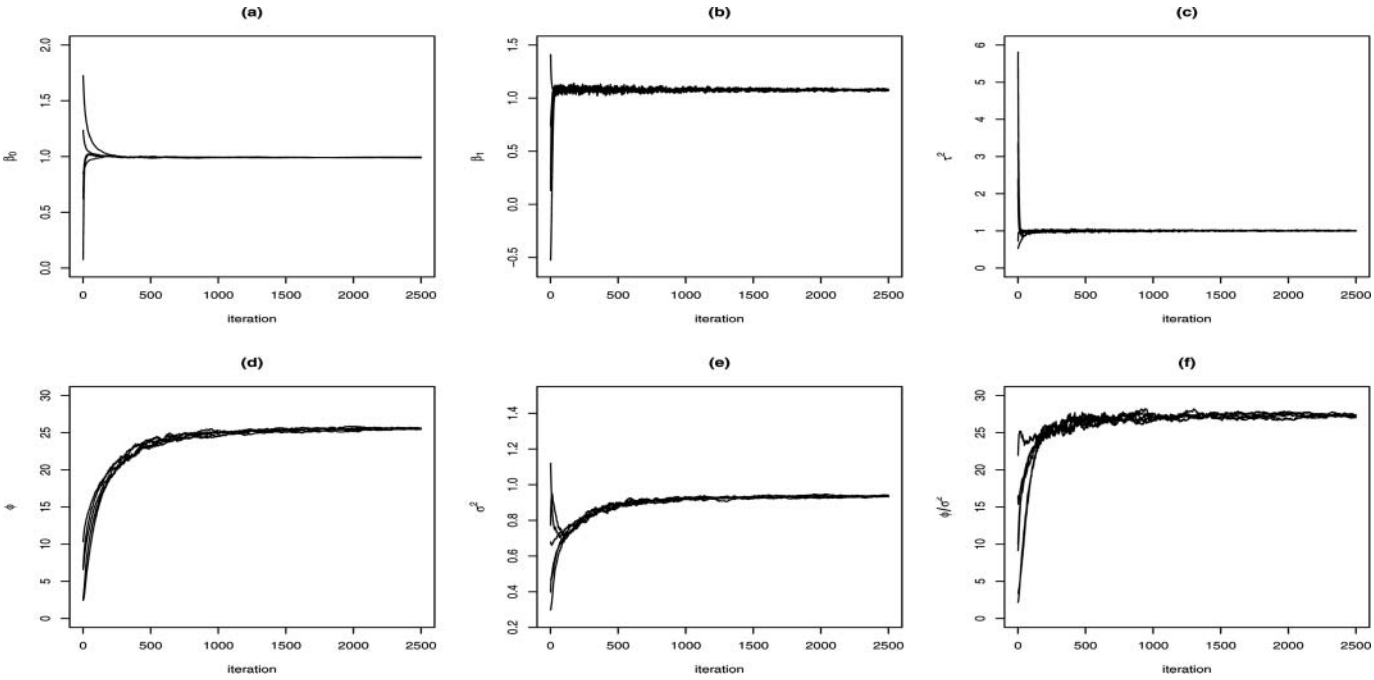

Figure 2. Trajectories of RSA ($m = 700$, 6 runs) for a dataset with nugget effects: (a)–(f) are for β_0, β_1, τ^2, ϕ, σ^2, and ϕ/σ^2, respectively.

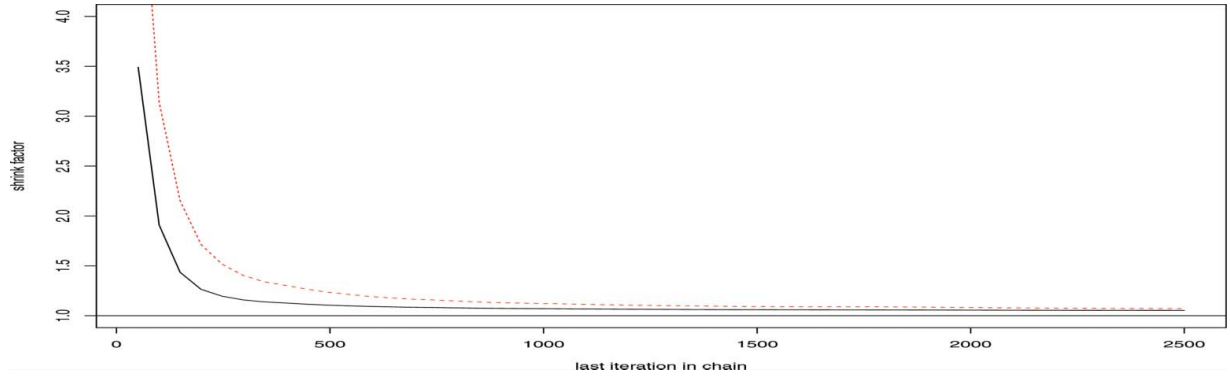

Figure 3. Gelman–Rubin shrink factor for the ϕ/σ^2-trajectories of RSA collected in six runs for one dataset. The solid and dashed lines show the shrink factor and the upper limit of the 95% confidence interval of the shrink factor, respectively. The online version of this figure is in color.

Table 3. RSA for very large datasets: $n = 50{,}000$ and 2500 iterations (see Table 1 for notations)

Estimator	Size	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\phi}/\hat{\sigma}^2$	$\hat{\tau}^2$	CPU(m)
	100	1.029 (0.050)	0.998 (0.005)	19.035 (0.789)	0.941 (0.008)	0.6
	300	1.036 (0.047)	1.004 (0.004)	21.327 (0.633)	0.970 (0.006)	7.3
RSA	500	1.033 (0.045)	1.002 (0.004)	22.204 (0.531)	0.976 (0.006)	30.6
	700	1.006 (0.045)	0.999 (0.002)	23.494 (0.677)	0.989 (0.004)	82.4
True	–	1.000	1.000	25.000	1.0	–

$n = 50000$). Even for a dataset with $n = 11{,}000$ (one of our real data examples), it took 10,340 min on the same computer.

5.1.2 Comparisons with the Bayesian and Predictive Process Methods. For a thorough comparison study, we applied the Bayesian method (Diggle et al. 1998; Diggle and Ribeiro Jr. 2002) and the Gaussian predictive process approximation method (Banerjee et al. 2008) to this example. The former method has been implemented in the R package *geoR* and the latter in *spBayes*. The reason why these two methods were chosen for comparison is twofold: their parameters still match with the parameters of the model (1), and their software is available to the public. Note that some other methods, for example, those based on the likelihood or Markov random field approximation, may contain parameters that do not match with the model (1). In running the Bayesian method, we have followed the instruction of *geoR* to impose uniform priors on β_0, β_1, ϕ, and the nugget-to-sill ratio parameter τ^2/σ^2, impose a reciprocal prior on σ^2 (i.e., $f(\sigma^2) \propto 1/\sigma^2$), and discretize the nugget-to-sill ratio parameter to 21 points which are equally spaced on the interval [0.5, 1.5]. The discretization reduced the simulation time substantially; otherwise, it would be extremely long. Under the default setting of *geoR*, the algorithm was run for 1050 iterations for each dataset. The results are summarized in Table 1.

As previously mentioned, the Gaussian predictive process approximation method belongs to class of lower dimensional space approximation methods, for which the lower dimensional process, that is, the so-called predictive process, was constructed based on the local kriging prediction on a number of prespecified knots. The computational cost of this method is governed by the number of knots. In general, a larger number of knots will cause longer CPU time and lead to a better approximation to the original process. In running this method, we tried two different settings for the knots, a 6×6 grid and a 10×10 grid. We also followed the instruction of *spBayes* to specify a flat prior for β_0 and β_1, and the following priors for the other parameters:

$$\phi \sim \text{Uniform}[1, 50], \quad \sigma^2 \sim \text{IG}(0.1, 0.1), \quad \tau^2 \sim \text{IG}(0.1, 0.1),$$

where $\text{IG}(a, b)$ denotes the inverse-gamma distributions with parameters a and b. For each dataset, the algorithm was run for 2000 iterations, where the first 500 iterations were discarded for burn-in and the samples drawn from the remaining 1500 iterations were used for inference. The results are also summarized in Table 1.

The comparison indicates that RSA outperforms the Bayesian and Gaussian predictive process methods significantly, in both estimation accuracy and CPU time. For the Gaussian predictive process method, even it requires more CPU time than RSA, the estimates of τ^2 are wrong, and the estimates of ϕ/σ^2 are highly varied. The Bayesian method performs better than the Gaussian predictive method, but its estimates for τ^2 and ϕ/σ^2 are both severely upper biased. Compared to these two methods, RSA did a much better job, especially with $m = 700$; it cost less CPU time and produced much more accurate estimates.

Both the Bayesian and Gaussian predictive process methods have been applied to the very large datasets considered in Section 5.2, where each dataset consists of 50000 observations. Due to the $n \times n$-covariance matrix storage problem in their software, both methods failed to produce any results for those datasets. Note that both the Bayesian and Gaussian predictive process methods involve inversions of $n \times n$-matrices, although in the latter the computation can be reduced based on the Sherman–Woodbury–Morrison matrix identity (e.g., Harville 1997). In contrast, as shown in Table 3, RSA works well for those large datasets with almost the same CPU time as for this example. Note that RSA does not involve any computations of $n \times n$ matrices. This feature makes RSA uniquely appealing in computer memory in addition to its attractiveness in computational time and estimation accuracy.

5.1.3 On Nugget Effect. We have also considered the data without measurement errors, that is, the nugget effect $\tau^2 = 0$. We simulated 50 datasets with $\beta_0 = \beta_1 = 1$, $\phi = 25$, $\sigma^2 = 1$ and the explanatory variable $\boldsymbol{c}_1$ being generated from Normal $N(0, 0.5^2)$. Each dataset consisted of $n = 2000$ observations with the sampling sites being uniformly distributed in the region $[0, 100] \times [0, 100]$. RSA was run for these datasets with $a_0 = 0.001$. The results are summarized in Table 2. A comparison with Table 1 indicates that RSA can work very well for the data without nugget effects. With only $m = 300$, the parameters could have been estimated rather accurately. This result is very encouraging, which indicates that MMLE can work as a general parameter estimation approach for the models of big data.

5.2 RSA for Very Large Data

In this section, we explored the performance of RSA for very large datasets. We simulated 50 datasets from models (1) and (3) with $\beta_0 = \beta_1 = 1$, $\phi = 25$, $\sigma^2 = 1$, $\tau^2 = 1.0$, and the explanatory variable $\boldsymbol{c}_1$ being generated from the Normal $N(0, 0.5^2)$. Each dataset consisted of 50,000 observations uniformly positioned in the region $[0, 100] \times [0, 100]$. RSA was run for these datasets with $a_0 = 0.01$ for $m = 100$, 300, and 500 and $a_0 = 0.001$ for $m = 700$. The numerical results are summarized in Table 3. Even for such large datasets, RSA still work well. More importantly, its CPU time is almost independent of n. Therefore, RSA can be applied to very large datasets. As mentioned earlier, MLE failed for these datasets.

Taking a closer look at Tables 1 and 3, it is easy to see that for a fixed value of m, the bias of the estimate of ϕ/σ^2 tends to

Table 4. RSA for the datasets simulated in the scenario of domain expanding (see Table 1 for notations)

Estimator	Size	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\phi}$	$\hat{\sigma}^2$	$\hat{\tau}^2$	$\hat{\phi}/\hat{\sigma}^2$
	100	0.996	1.008	2.591	0.857	1.138	7.409
		(0.008)	(0.009)	(0.236)	(0.044)	(0.041)	(3.407)
RSA	300	0.988	1.022	1.932	1.052	0.949	1.893
		(0.009)	(0.009)	(0.040)	(0.022)	(0.018)	(0.067)
	500	0.995	1.015	1.950	1.041	0.953	1.932
		(0.008)	(0.008)	(0.041)	(0.021)	(0.017)	(0.070)
MLE	–	0.988	1.016	2.021	1.033	0.967	2.002
	–	(0.007)	(0.008)	(0.045)	(0.019)	(0.016)	(0.067)
True	–	1.0	1.0	2.0	1.0	1.0	2.0

increase with n, although not significantly. This suggests that a slightly larger value of m should be used for a large dataset. This is reasonable: A large dataset needs a large subset to represent itself.

5.3 RSA under the Scenario of Domain Expanding

In this section, we explored the performance of RSA for the scenario of domain expanding. We simulated 50 datasets from models (1) and (3) with $\beta_0 = \beta_1 = 1$, $\phi = 2$, $\sigma^2 = 1$, $\tau^2 = 1$, and the explanatory variable $\boldsymbol{c}_1$ being generated from the Normal $N(0, 0.5^2)$. Each dataset consisted of 2000 observations uniformly positioned in the region $[0, 100] \times [0, 100]$. Since the region is so large relative to the true value of ϕ, this example mimics the scenario of domain expanding and it is known that the model is consistently estimable in this scenario.

RSA was run for these datasets with $a_0 = 0.01$, and each run consisted of 2500 iterations. The numerical results were summarized in Table 4. The CPU times are similar to those reported in Table 1. Certain patterns can be observed in Table 4: (i) For all values of m, $\hat{\sigma}^2 + \hat{\tau}^2 = 2$ approximately holds; and (ii) when m is large, both $\hat{\sigma}^2$ and $\hat{\tau}^2$ converge to their true values. This example shows that RSA also works under domain expanding asymptotics.

6. REAL DATA ANALYSIS

In this section, we illustrate the application of RSA to large, irregularly spaced spatial data. Two examples are examined. The first is for the observations from weather stations in the United States, and the second is for the observations from a gold mine. Due to space limitations, only the first example is presented next, and the second example is presented in the supplementary material, available online.

We consider the precipitation data from the National Climatic Data Center for the years 1895 to 1997, which are available at */www.image.ucar.edu/GSP/Data/US.monthly.met/*. This dataset has been examined by several authors, for example, Johns et al. (2003), Furrer et al. (2006), and Kaufman et al. (2008). Johns et al. (2003) focused on missing observations imputation, and Furrer et al. (2006) and Kaufman et al. (2008) used the datasets to illustrate the covariance tapering method. In this analysis, we analyze the monthly total precipitation anomalies, which are defined as the monthly totals standardized by the long-run mean and standard deviation for each station. The dataset we considered is the precipitation anomalies of April 1948, which has been used as a demonstration dataset in the software *KriSp* (Furrer 2006). The reason why we chose to work on this dataset is twofold. First, the dataset is large, consisting of 11,918 stations. Note that part of the data was imputed by Johns et al. (2003), but for the purpose of illustration, we follow Furrer (2006) to treat all data as real observations. Second, the data show no obvious nonstationarity or anisotropy. Otherwise, it would require a more complicated model, such as a mixture spatial model, than is considered here.

In our analysis, we first divide the data into two parts, a random subset of 11,000 observations as the training set and the remaining 918 observations as the test set. RSA was applied to the training data with $m = 500$ and $m = 700$. For each setting of m, RSA was run for five times with $a_0 = 0.001$ and each run consisted of 2500 iterations.

The results are summarized in Table 5. It indicates that RSA works very stable for this example. The standard deviations of all parameters are quite small.

To assess the quality of RSA estimates, we measure their prediction performance on the test set. For a given estimate (totally, 10 estimates have been produced by RSA in 10 runs), the conditional mean $E(Y(s_0)|Y(s_1), \ldots Y(s_n))$ is calculated, where s_0 denotes a prediction site and $\{Y(s_1), \ldots, Y(s_n)\}$ denotes the training data. Since the error process in (1) is Gaussian, the conditional mean $E(Y(s_0)|Y(s_1), \ldots, Y(s_n))$ coincides with the kriging predictor. However, this predictor involves inverting a large matrix, $11{,}000 \times 11{,}000$, for this example. To reduce computational time, we predict $Y(s_0)$ based on only the observations that lie in a neighborhood around s_0. As discussed in Cressie (1993), the choice of neighborhood size, denoted by δ, may depend on the range (ϕ), the nugget-to-sill ratio (τ^2/σ^2), and the spatial configuration of data locations. However, there is no simple relationship between δ and those parameters. For this reason, we tried different values of δ, including 25, 40, 50, and 100 miles. On average, each point has approximately 10, 24, 37, and 132 neighboring points for the four neighborhood sizes, respectively. In what follows, we will call this prediction method local kriging. The mean squared prediction errors (MSPEs) were calculated for each value of δ and each model estimate

Table 5. Numerical results of RSA for anomalies of monthly precipitation for April 1948

Method	Size	$\hat{\beta}_0$	$\hat{\phi}$	$\hat{\sigma}^2$	$\hat{\tau}^2$	CPU(m)
RSA	500	0.163 (0.000)	183.71 (0.45)	0.825 (0.003)	0.059 (0.000)	29.6
	700	0.161 (0.001)	179.38 (1.15)	0.829 (0.001)	0.057 (0.000)	84.1
MLE		0.138	164.20	0.807	0.057	10,340.4

The estimates were calculated by averaging over five independent runs with standard deviations given in the parentheses (see Table 1 for notations).

Table 6. MSPEs for anomalies of monthly precipitation in April 1948

Methods	m	Neighborhood size (δ) 25	40	50	100
Local Kriging	500	0.118 (9.1×10^{-6})	0.116 (8.1×10^{-5})	0.125 (1.2×10^{-4})	0.147 (6.3×10^{-5})
	700	0.118 (1.7×10^{-5})	0.117 (8.6×10^{-5})	0.126 (1.1×10^{-4})	0.147 (7.3×10^{-5})
	MLE	0.118	0.119	0.129	0.148
Tapering	500	0.297 (9.7×10^{-5})	0.160 (2.9×10^{-5})	0.136 (1.3×10^{-5})	0.115 (1.8×10^{-5})
	700	0.297 (1.4×10^{-4})	0.160 (3.7×10^{-5})	0.136 (1.6×10^{-5})	0.115 (1.8×10^{-5})
	MLE	0.279	0.150	0.130	0.111

The values reported in the table are calculated by averaging over five runs, and the numbers in parentheses denote the standard deviations of the averaged MSPEs.

obtained previously. Total, there were 40 MSPEs calculated and they are summarized in Table 6. The results indicate that the local kriging method with about 20 neighboring points can provide a sufficiently precise prediction. This is consistent with the results of Furrer et al. (2006), where it is stated that a tapering radius with 16–24 points is sufficient for this example. A natural question is why the predictions produced with $\delta = 50$ and 100 are worse than those produced with $\delta = 25$ and 40. Possible reasons include (i) the mistake in model specification; that is, the data may (slightly) violate the assumptions of stationarity and isotropy; and (ii) the error in parameter estimation.

Given the estimate of parameters, the prediction can also be done using the covariance tapering method. Furrer et al. (2006) showed that tapering the covariance matrix with an appropriate compactly supported correlation function reduces the computational burden significantly and still leads to an asymptotically optimal mean squared prediction error. Note that Furrer et al. (2006) assumed that the covariance function is known. For this example, we have the estimated covariance matrices tapered by a spherical correlation function with different range parameter values, $\phi = 25$, 40, 50, and 100 miles. The resulting MSPEs are summarized in Table 6. Due to its covariance adjustment, it is not a surprise that the tapering method produced better predictions than the local kriging method when the neighborhood is large. This suggests that the covariance tapering method can be used in conjunction with RSA for prediction of large datasets.

For comparison, the covariance tapering method was also rerun with the covariance function given in Furrer et al. (2006), which is a mixture of two exponential covariance functions with respective parameters $(\phi, \sigma^2) = (40.73, 0.277)$ and $(\phi, \sigma^2) = (523.73, 0.722)$, and the range parameter of the spherical correlation function being set to 50 miles. The resulting MSPE is 0.132, which is slightly worse than the best values reported in Table 6. For a thorough comparison, MLE was also applied to this example with the estimation results reported in Table 5. It cost extremely long CPU time, 10,340.4 min, on the same computer. The resulting MSPEs are shown in Table 6.

Figure 4(b)–(d) show the prediction surfaces of the monthly precipitation anomaly, which are evaluated on a regular 0.065×0.12 latitude/longitude grid within the conterminous U.S., roughly at the solution of the NOAA data product. Comparing to Figure 4(a), it is easy to see that our predictions are rather accurate, which match the observed images very well.

In summary, RSA and MLE perform similarly for this example in both parameter estimation and prediction, but MLE takes much more CPU time than RSA. This example shows that RSA is advantageous for large spatial data.

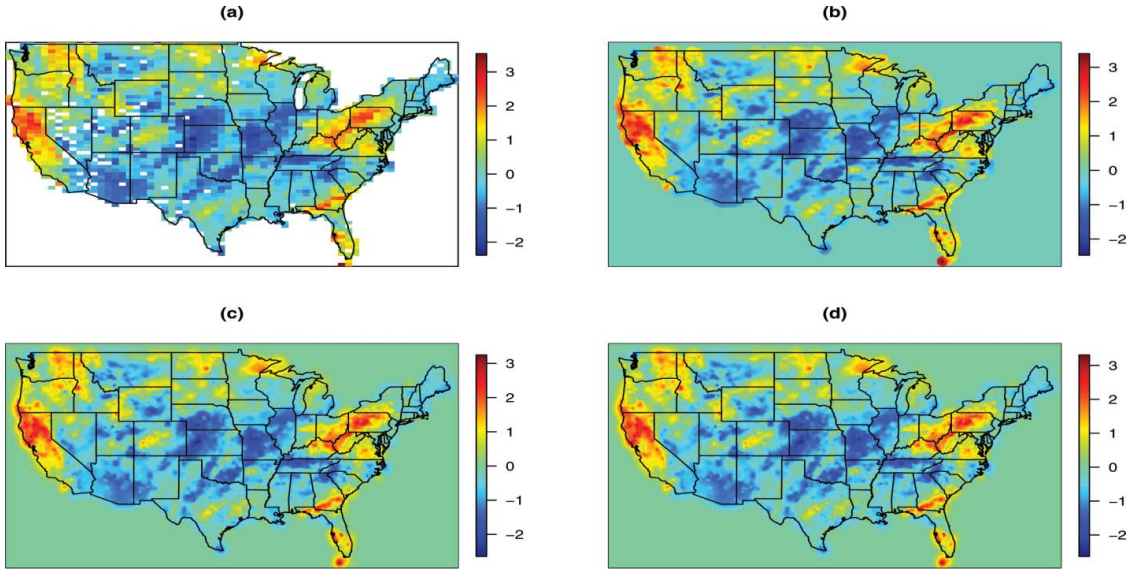

Figure 4. Observed and predicted precipitation anomaly for April 1948 (500×400 grid). (a) Observed image. (b) Predicted image by local kriging with $\delta = 40$ miles for an RSA estimate obtained with $m = 500$. (c) Predicted image by covariance tapering with $\delta = 100$ miles for an RSA estimate obtained with $m = 500$. (d) Predicted image by covariance tapering with $\delta = 100$ miles for the MLE. The online version of this figure is in color.

7. DISCUSSION

In this article, we have proposed the RSA method for analysis of large spatial data. At each iteration of RSA, a small subsample is drawn from the full dataset, and then the current estimates of the parameters are updated accordingly under the framework of stochastic approximation. Since RSA makes use of only a small proportion of the data at each iteration, it avoids inverting large covariance matrices and thus, is scalable to very large datasets. Under mild conditions, we show that the RSA estimator converges in probability to a set of parameter values of equivalent Gaussian probability measures, and that the estimator is asymptotically normally distributed. The numerical examples indicate that RSA can work well for large datasets. RSA also leads to a general parameter estimation approach, MMLE, for big data, which has included MLE as a special case.

For implementation of RSA, two issues need to be addressed. The first issue is about the subsample size m, which plays the role of sample size for the mean log-likelihood function. To indicate the dependence of $\tilde{\boldsymbol{\theta}}_n$ on m, we write $\tilde{\boldsymbol{\theta}}_{m,n}$ for $\tilde{\boldsymbol{\theta}}_n$ in what follows. Obviously, if the model is consistently estimable, then $\tilde{\boldsymbol{\theta}}_{m,n}$ is consistent for $\boldsymbol{\theta}$ as both m and n go to infinity. A question of interest is for a given value of n, how to choose m such that $\tilde{\boldsymbol{\theta}}_{m,n}$ is close to the MLE of $\boldsymbol{\theta}$. Intuitively, as also suggested by our numerical results, m should increase with n such that the subsample can resemble characteristics of the whole dataset and thus $\tilde{\boldsymbol{\theta}}_{m,n}$ is close to the MLE of $\boldsymbol{\theta}$. To find such a value of m, practically we can gradually increase the value of m until $\tilde{\boldsymbol{\theta}}_{m,n}$ becomes stable as a function of m. In theory, it is still unclear how m should grow with n in the scenario of fixed domain. Further study of this issue is of great interest.

The second issue is about the subsampling scheme. For mathematical simplicity, this article advocates the simple random sampling scheme, that is, each sample is drawn from the full dataset at random and without replacement. An alternative scheme is stratified sampling, which may be more attractive than the simple random sampling scheme from the perspective of parameter estimation. When the sites $s_1, \ldots, s_n$ are not uniformly distributed in the given region, stratified sampling is potentially more efficient than simple random sampling in terms of subsample sizes. In other words, to achieve the same estimation accuracy, stratified sampling may need a smaller subsample size than simple random sampling. To use stratified sampling in RSA, a little extra theoretical work need to be done to ensure that the resulting estimator still converges to the right place.

Regarding the relationship between RSA and bootstrap, we note that they are closely related: Equation (7) is the m out of n without replacement bootstrap estimator (e.g., Politis et al. 2001; Bickel and Sakov 2008) of the partial derivative of the Kullback–Leibler divergence defined in (6). For dependent data, the bootstrap method has been extensively studied under the expanding-domain asymptotics (e.g., Hall 1985; Davison and Hinkley 1997; Lahiri 2003). These studies are typically conducted for a block resampling procedure, which attempts to retain the dependence structure of the data, under some mixing conditions for the process. This ensures that the resampling scheme produces replicates that are asymptotically independent, identically distributed. However, under the infill asymptotics, there is no general analog to mixing conditions. Except for Loh and Stein (2008), where the authors considered bootstrapping one-dimensional Gaussian random fields defined on a regular grid, we are not aware of any previous results demonstrating the validity of bootstrapping under infill asymptotics. RSA provides a way for parameter estimation based on the bootstrapped log-likelihood function under the infill asymptotics. Bootstrapping may provide us useful tools for a further study of the property of the RSA method.

In this article, RSA is only applied to geostatistical data. Extending it to spatiotemporal data is straightforward. RSA can also be easily modified to accommodate the missing data problem by adding a missing data imputation step in its procedure. This is very important for real data analysis and will be explored elsewhere. In addition to the spatial data, RSA can also be applied to large independent data in which the observations are mutually independent. For large independent data, a popular method is divide-and-conquer (e.g., Xi et al. 2009; Lin and Xi 2011), which works in three steps: partition the dataset into a number of small subsets, analyze each subset separately, and then aggregate the results from each subset to get the final result. Compared to the divide-and-conquer method, RSA is more general: RSA can work for both dependent and independent data, while the divide-and-conquer method can only work for the latter. Given its generality and computational attractiveness, we expect that RSA will play an important role in big data analysis.

APPENDIX

In this appendix, the lemmas and theorems are only proved for the case $\psi(\cdot)$ taking values in one-dimensional space $\mathbb{R} = (-\infty, \infty)$. Extending the results to the case of multiple-dimensional space is straightforward. To facilitate the proofs, we introduce the following notations:

$$\mathcal{S}_n = \{X_1, \ldots, X_n\}, \quad \mathcal{S}_N = \{X_1, \ldots, X_n, X_{n+1}, \ldots, X_N\},$$
$$q = 1 - \frac{n}{N},$$
$$\psi_k(x_1, \ldots, x_k) = E\{\psi(X_1, \ldots, X_m) | X_1 = x_1, \ldots, X_k = x_k\}, \quad 1 \le k \le m-1,$$

$$\varphi_{k,N}(x_1, \ldots, x_k) = E\{\psi(X_1, \ldots, X_m) | X_1 = x_1, \ldots, X_k = x_k, \mathcal{S}_N\}, \quad 1 \le k \le m-1,$$
$$\sigma_k^2 = \mathrm{var}(\psi_k(X_1, \ldots, X_k)), \quad \sigma_{k,N}^2 = \mathrm{var}(\psi_{k,N}(X_1, \ldots, X_k)), \quad 1 \le k \le m-1,$$

Proof of Lemma 1

To prove this lemma, it suffices to show that $\mathrm{var}(U_n) \to 0$ as $n \to \infty$. To calculate $\mathrm{var}(U_n)$, we construct an auxiliary set $\mathcal{S}_N = \{X_1, \ldots, X_n, X_{n+1}, \ldots, X_N\}$. Thus, $\mathcal{S}_n$ can be viewed as a simple random sample of $\mathcal{S}_N$, and U_n can be viewed as a U-statistic defined on the finite population $\mathcal{S}_N$. By Zhao and Chen (1990), we have

$$\mathrm{var}(U_n \mid \mathcal{S}_N) = \frac{qm^2}{n}\sigma_{1,N}^2 + O\left(\frac{q}{nN}\right) + O\left(\frac{q^2}{n^2}\right). \tag{A.1}$$

Since $\psi(\cdot)$ is continuous (a.e.) and $E|\psi(X_1, \ldots, X_m)|^2 < \infty$, it follows from Lahiri (1996) that

$$\sigma_{1,N}^2 \xrightarrow{p} \sigma_1^2, \quad \text{as } N \to \infty, \tag{A.2}$$

where $\xrightarrow{p}$ denotes convergence in probability. Lahiri (1996) studied the convergence of an empirical measure of the spatial process under infill asymptotics, and his result (Corollary of Theorem 4) implies that for a bounded stationary random field, the sample mean of any continuous

function will converge in probability to its true mean (provided existence) as the sample size becomes large.

Therefore, by (A.1) and (A.2),

$$\operatorname{var}(U_n|\mathcal{S}_\infty) = \lim_{N\to\infty} \operatorname{var}(U_n|\mathcal{S}_N) = O_p\left(\frac{1}{n}\right), \tag{A.3}$$

where $O_p(c_n) = c_n O_p(1)$, and $O_p(1)$ denotes a sequence of random variables converging in probability to a constant.

On the other hand, U_n, as a linear function $\psi(X_1^{(i)}, \ldots, X_m^{(i)})$, is also a.e. continuous and possesses the second-order moment. It follows from Lahiri (1996) that $\operatorname{var}(U_n|\mathcal{S}_N) \xrightarrow{p} \operatorname{var}(U_n)$ as $N \to \infty$; that is,

$$\operatorname{var}(U_n|\mathcal{S}_\infty) = \operatorname{var}(U_n) + o_p(1), \tag{A.4}$$

where $o_p(1)$ is short for a sequence of random variables that converge to zero in probability. By (A.3) and (A.4), we have

$$\operatorname{var}(U_n) = O_p\left(\frac{1}{n}\right) - o_p(1) = o_p(1). \tag{A.5}$$

Since $\operatorname{var}(U_n)$ is a constant sequence, (A.5) further implies that $\operatorname{var}(U_n) \to 0$. This concludes the proof of Lemma 1.

To prove Lemma 2, we introduce another lemma, whose proof is left to the reader.

Lemma 7. Let $X_i^{(k)}$, $k = 1, \ldots, q$, denote q sequences of random variables. If q is finite and for all $1 \le k \le q$,

$$X_i^{(k)} \xrightarrow{p} a_k, \quad \text{as } i \to \infty, \tag{A.6}$$

then

$$\max_k X_i^{(k)} \xrightarrow{p} \max_k a_k, \quad \text{as } i \to \infty.$$

Proof of Theorem 1

Because $\boldsymbol{Z}$ follows a multivariate Gaussian distribution, $l_{\boldsymbol{\theta}}(\boldsymbol{z}, \boldsymbol{s})$ is continuous and nonpositive, and

$$l_{\boldsymbol{\theta}}(\boldsymbol{z}, \boldsymbol{s}) = -\frac{m}{2}\log(2\pi) - \frac{1}{2}\log(|\boldsymbol{\Sigma}_z|) - \frac{1}{2}(\boldsymbol{z} - \boldsymbol{\mu}_z)^T \boldsymbol{\Sigma}_z^{-1}(\boldsymbol{z} - \boldsymbol{\mu}_z).$$

Since the normal distribution possesses finite moments of any order, (13) and (14) are satisfied. By Lemma 2, we have $d(\tilde{\boldsymbol{\theta}}_n, \Theta_0) \to 0$ in probability.

Proof of Lemma 3

The proof of this lemma is partly based on the proof of Zhao and Chen (1990) for the asymptotic normality of finite population U-statistics, and the proof of Lahiri (1996) for the convergence of empirical measures under infill asymptotics. In the following, we will omit the details for the parts that are from these references.

As in the proof of Lemma 1, we construct an auxiliary finite population $\mathcal{S}_N = \{X_1, \ldots, X_n, X_{n+1}, \ldots, X_N\}$ with N being fixed to $N = n^2$. Since the auxiliary samples $\{X_{n+1}, \ldots, X_N\}$ are drawn in the same sampling procedure from the same random field as for $\{X_1, \ldots, X_n\}$, $\mathcal{S}_N$ also satisfies the condition (16). Thus,

$$\frac{E(|\psi_{k,N}|^2)}{n\sigma_{1,N}^2} = \frac{E(|\psi_{k,N}|^2)}{\sqrt{N}\sigma_{1,N}^2} \to 0, \quad \text{a.s.}, \quad \text{as } N \to \infty. \tag{A.7}$$

Therefore, by Corollary 2.1 of Zhao and Chen (1996) (the setting $N = n^2$ and (A.7) imply that the conditions required by Corollary 2.1 of Zhao and Chen (1996) are satisfied), we have

$$(U_n - E(U_n|\mathcal{S}_N))/\sqrt{\operatorname{var}(U_n|\mathcal{S}_N)} \Rightarrow N(0, 1). \tag{A.8}$$

Since $\psi(\cdot)$ is continuous (a.e.) and its second-order moment exists, it follows from Lahiri (1996) that

$$E(U_n|\mathcal{S}_N) \xrightarrow{p} E(U_n), \quad \operatorname{var}(U_n|\mathcal{S}_N) \xrightarrow{p} \operatorname{var}(U_n), \quad \text{as } N \to \infty.$$

By the setting $N = n^2$, which implies $n \to \infty \Leftrightarrow N \to \infty$, and Slutsky's theorem, we have

$$(U_n - E(U_n))/\sqrt{\operatorname{var}(U_n)} \Rightarrow N(0, 1). \tag{A.9}$$

This complete the proof of the lemma.

Proof of Lemma 4

In this proof, we assume that $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$ is a d-dimensional vector. Since $\tilde{\boldsymbol{\theta}}_n$ maximizes $U_n(\boldsymbol{\theta})$, it solves the first-order equation given by

$$\binom{n}{m}^{-1} \sum_{i=1}^{\binom{n}{m}} \omega_{\boldsymbol{\theta}}\left(x_1^{(i)}, \ldots, x_m^{(i)}\right) = 0,$$

where $\omega_{\boldsymbol{\theta}}(x_1, \ldots, x_m) = (\partial\psi_{\boldsymbol{\theta}}(x_1, \ldots, x_m)/\partial\theta_1, \ldots, \partial\psi_{\boldsymbol{\theta}}(x_1, \ldots, x_m)/\partial\theta_d)^T$. Let

$$\boldsymbol{H}_{\boldsymbol{\theta}}\left(x_1^{(i)}, \ldots, x_m^{(i)}\right) = \frac{\partial^2\psi_{\boldsymbol{\theta}}\left(x_1^{(i)}, \ldots, x_m^{(i)}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}$$

be the Hessian of $\psi_{\boldsymbol{\theta}}(\cdot)$. Then, by a mean-value expansion,we can write

$$0 = \binom{n}{m}^{-1} \sum_{i=1}^{\binom{n}{m}} \omega_{\boldsymbol{\theta}}\left(X_1^{(i)}, \ldots, X_m^{(i)}\right) = \binom{n}{m}^{-1} \sum_{i=1}^{\binom{n}{m}} \omega_{\boldsymbol{\theta}^*}\left(X_1^{(i)}, \ldots, X_m^{(i)}\right) + \left\{\binom{n}{m}^{-1} \sum_{i=1}^{\binom{n}{m}} \tilde{\boldsymbol{H}}_i\right\}(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*),$$

where $\tilde{\boldsymbol{H}}_i$ is equal to the Hessian $\boldsymbol{H}_{\boldsymbol{\theta}}(x_1^{(i)}, \ldots, x_m^{(i)})$ except that whose each row is evaluated at a different mean value of $\boldsymbol{\theta}$ between $\tilde{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}^*$. By Lemma 2, whose conditions are satisfied by conditions (i)–(iii), $\tilde{\boldsymbol{\theta}}_n$ converges to $\boldsymbol{\theta}^*$ in probability. Furthermore, $\tilde{\boldsymbol{H}}_i$ is continuous in $\boldsymbol{\theta}$ and its second-order moment exists, so we have

$$\binom{n}{m}^{-1} \sum_{i=1}^{\binom{n}{m}} \tilde{\boldsymbol{H}}_i \xrightarrow{p} \boldsymbol{H}_*,$$

by Lemma 1. Since the model has been assumed to be consistently estimable, $\boldsymbol{H}_*$ is nonsingular. Therefore,

$$\left\{\binom{n}{m}^{-1} \sum_{i=1}^{\binom{n}{m}} \tilde{\boldsymbol{H}}_i\right\}^{-1} \xrightarrow{p} \boldsymbol{H}_*^{-1}, \tag{A.10}$$

and thus,

$$(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) = -\left\{\binom{n}{m}^{-1} \sum_{i=1}^{\binom{n}{m}} \tilde{\boldsymbol{H}}_i\right\}^{-1} \left\{\binom{n}{m}^{-1} \sum_{i=1}^{\binom{n}{m}} \omega_{\boldsymbol{\theta}^*}\left(X_1^{(i)}, \ldots, X_m^{(i)}\right)\right\}. \tag{A.11}$$

Note that $E(\omega_{\boldsymbol{\theta}^*}(X_1, \ldots, X_m)) = 0$ under regularity conditions. By Lemma 3, whose conditions are satisfied by conditions (ii) and (iv),

$$\binom{n}{m}^{-1} \sum_{i=1}^{\binom{n}{m}} \omega_{\boldsymbol{\theta}^*}\left(X_1^{(i)}, \ldots, X_m^{(i)}\right) \Rightarrow N(0, \boldsymbol{\Sigma}), \tag{A.12}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of the U-statistic defined by the kernel $\omega_{\boldsymbol{\theta}^*} = \partial\psi_{\boldsymbol{\theta}}/\partial\boldsymbol{\theta}|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$. By (A.10), (A.11), and (A.12), we conclude the proof of the lemma.

Proof of Theorem 2

Since $\boldsymbol{Z}$ follows a multivariate normal distribution, it is easy to see that the kernel $l_{\boldsymbol{\theta}}(\boldsymbol{Z}, \boldsymbol{S})$ of the U-statistic $M_n(\boldsymbol{\theta})$ satisfies conditions (i)–(iii) of Lemma 4. Assuming that the sampling process satisfies the condition (iv), by Lemma 4, $\tilde{\boldsymbol{\theta}}_n$ is asymptotically normally distributed as described in (19).

SUPPLEMENTARY MATERIALS

Section 1: A real data example, goldmine samples.
Section 2: The proof of Lemma 3.2.
Section 3: The proofs of Lemma 4.1, Theorem 4.1, and Theorem 4.2.

[Received March 2011. Revised October 2012.]

REFERENCES

Andrieu, C., Moulines, É., and Priouret, P. (2005), "Stability of Stochastic Approximation Under Verifiable Conditions," *SIAM Journal of Control and Optimization*, 44, 283–312. [327,330]

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008), "Gaussian Predictive Process Models for Large Spatial Data Sets," *Journal of the Royal Statistical Society,* Series B, 70, 825–848. [325,334]

Benveniste, A., Métivier, M., and Priouret, P. (1990), *Adaptive Algorithms and Stochastic Approximations*, Springer-Verlag: Berlin. [331]

Bickel, P. J., and Sakov, A. (2008), "On the Choice of m in the m out of n Bootstrap and Confidence Bounds for Extrema," *Statistica Sinica*, 18, 967–985. [337]

Borovskikh, Y. V. (1996), *U-Statistics in Banach Spaces*, Utrecht: VSP. [328]

Chatterji, S. D. (1968), "Martingale Convergence and the Radon–Nikodym Theorem in Banach Spaces," *Mathematica Scandinavica*, 22, 21–41. [328]

Chen, B., Hu, J., Zhu, Y., and Sun, Z. (2009), "Parameter Identifiability With Kullback–Leibler Information Divergence Criterion," *International Journal of Adaptive Control and Signal Processing*, 23, 940–960. [326]

Chen, H. F. (2002), *Stochastic Approximation and Its Applications*, Dordrecht: Kluwer Academic Publishers. [330]

Chen, H. F., and Zhu, Y. M. (1986), "Stochastic Approximation Procedures With Randomly Varying Truncations," *Scientia Sinica,* Series A, 29, 914–926. [330]

Chen, H. S., Simpson, D. G., and Ying, Z. (2000), "Infill Asymptotics for a Stochastic Process Model With Measurement Error," *Statistica Sinica*, 10, 141–156. [326,332]

Cressie, N. A. C. (1993), *Statistics for Spatial Data* (2nd ed.), New York: Wiley. [325,335]

Cressie, N., and Johannesson, G. (2008), "Fixed Rank Kriging for Very Large Spatial Data Sets," *Journal of the Royal Statistical Society,* Series B, 70, 209–226. [325]

Davison, A. C., and Hinkley, D. V. (1997), *Bootstrap Methods and Their Applications*, Cambridge: Cambridge University Press. [337]

Diggle, P. J., and Ribeiro Jr, P. J. (2002), "Bayesian Inference in Gaussian Model-Based Geostatistics," *Geograhical & Environmental Modelling*, 6, 129–146. [334]

Diggle, P. J., Tawn, J., and Moyeed, R. (1988), "Model Based Geostatistics" (with discussion), *Applied Statistics*, 47, 299–350. [334]

Dowe, D. L., Baxter, R. A., Oliver, J. J., and Wallace, C. S. (1998), "Point Estimation Using the Kullback–Leibler Loss Function and MML," in *Research and Development in Knowledge Discovery and Data Mining*, Lecture Notes in Computer Sciences, (Vol. 1394), pp. 87–95. [326]

Du, J., Zhang, H., and Mandrekar, V. S. (2009), "Fixed-Domain Asymptotic Properties of Tapered Maximum Likelihood Estimators," *The Annals of Statistics*, 37, 3330–3361. [325]

Fuentes, M. (2007), "Approximate Likelihood for Large Irregularly Spaced Spatial Data," *Journal of the American Statistical Association*, 102, 321–331. [326]

Furrer, R. (2006), "*KriSp*: An R Package for Covariance Tapered Kriging of Large Datasets Using Sparse Matrix Techniques," *inside.mines.edu/∼rfurrer/software/KriSp*. [325,335,336]

Furrer, R., Genton M. G., and Nychka, D. (2006), "Covariance Tapering for Interpolation of Large Spatial Datasets," *Journal of Computational and Graphical Statistics*, 15, 502–523. [335]

Gelman, A., and Rubin, D. B. (1992), "Inference From Iterative Simulations Using Multiple Sequences" (with discussion), *Statistical Science*, 7, 457–511. [327]

Glynn, P. W., and Whitt, W. (1992), "The Asymptotic Validity of Sequential Stopping Rules for Stochastic Simulations," *Annals of Applied Probability*, 2, 180–198. [327]

Hall, P. (1985), "Resampling a Coverage Pattern," *Stochastic Processes and their Applications*, 20, 231–246. [337]

Harville, D. A. (1997), *Matrix Algebra from a Statistician's Perspective*, New York: Springer. [334]

Johns, C. J., Nychka, D., Kittel, T. G. F., and Daly, C. (2003), "Infilling Sparse Records of Spatial Fields," *Journal of the American Statistical Association*, 98, 796–806. [335]

Jones, R. H., and Zhang, Y. (1997), "Models for Continuous Stationary Space–Time Processes," in *Modeling Longitudinal and Spatially Correlated Data: Methods, Applications and Future Directions*, eds. P. J. Diggle, W. G. Warren, and R. D. Wolfinger, New York: Springer. [326]

Kaufman, C. G., Schervish, M. J., and Nychka, D. W. (2008), "Covariance Tapering for Likelihood-Based Estimation in Large Spatial Data Sets," *Journal of the American Statistical Association*, 103, 1545–1555. [325,335]

Kiefer, J., and Wolfowitz, J. (1952), "Stochastic Estimation of the Modulus of a Regression Function," *Annals of Mathematical Statistics*, 23, 462–466. [330]

Lahiri, S. N. (1996), "On Inconsistency of Estimators Based on Spatial Data Under Infill Asymptotics," *Sankhyā: The Indian Journal of Statistics*, 58, 403–417. [337,338]

——— (2003), *Resampling Methods for Dependent Data*, New York: Springer. [337]

Liang, F., and Wu, M. (2010), "Population Stochastic Approximation MCMC Algorithm and Its Weak Convergence," Technical Report, Department of Statistics, Texas A&M University. [331]

Liang, F., and Zhang, J. (2008), "Estimating the False Discovery Rate Using the Stochastic Approximation Algorithm," *Biometrika*, 95, 961–977. [326]

Lin, N., and Xi, R. (2011), "Aggregated Estimating Equation Estimation," *Statistics and Its Interface*, 4, 73–83. [337]

Loh, J. M., and Stein, M. L. (2008), "Spatial Bootstrap With Increasing Observations in a Fixed Domain," *Statistica Sinica*, 18, 667–688. [337]

Mardia, K. V., and Marshall, R. J. (1984), "Maximum Likelihood Estimation of Models for Residual Covariance in Spatial Regression," *Biometrika*, 71, 135–146. [326]

Matsuda, Y., and Yajima, Y. (2009), "Fourier Analysis of Irregularly Spaced Data on $\mathbb{R}^d$," *Journal of the Royal Statistical Society,* Series B, 71, 191–217. [326]

Park, J., and Liang, F. (2012), "Bayesian Analysis of Geostatistical Models With an Auxiliary Lattice," *Journal of Computational and Graphical Statistics*, 21, 453–475. [326]

Politis, D. N., Romano, J. P., and Wolf, M. (2001), "On the Asymptotic Theory of Subsampling," *Statistica Sinica*, 11, 1105–1124. [337]

Ribeiro Jr., P. J., and Diggle, P. J. (2001), "GeoR: A Package for Geostatistical Analysis," *R-NEWS*, 1, 15–18. [332]

Robbins, H., and Monro, S. (1951), "A Stochastic Approximation Method," *The Annals of Mathematical Statistics*, 22, 400–407. [326,330]

Rue, H., and Held, L. (2005), *Gaussian Markov Random Fields: Theory and Applications*, Boca Raton, FL: Chapman & Hall/CRC. [326]

Rue, H., and Tjelmeland, H. (2002), "Fitting Gaussian Markov Random Fields to Gaussian Field," *Scandinavian Journal of Statistics*, 29, 31–49. [326]

Stein, M. L. (1999), *Interpolation of Spatial Data: Some Theory for Kriging*, New York: Springer. [326,332]

——— (2004), "Equivalence of Gaussian Measures for Some Nonstationary Random Fields," *Journal of Statistical Planning and Inference*, 123, 1–11. [325,326,331,332]

Stein, M. L., Chi, Z., and Welty, L. J. (2004), "Approximating Likelihoods for Large Spatial Data Sets," *Journal of the Royal Statistical Society,* Series B, 66, 275–296. [326]

Vecchia, A. V. (1988), "Estimation and Model Identification for Continuous Spatial Processes," *Journal of the Royal Statistical Society,* Series B, 50, 297–312. [326]

van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge: Cambridge University Press. [328]

Wald, A. (1949), "Note on the Consistency of the Maximum Likelihood Estimate," *The Annals of Mathematical Statistics*, 20, 595–601. [328]

Wikle, C., and Cressie, N. (1999), "A Dimension-Reduced Approach to Space–Time Kalman Filtering," *Biometrika*, 86, 815–829. [325]

Xi, R., Lin, N., and Chen, Y. (2009), "Compression and Aggregation for Logistic Regression Analysis in Data Cubes," *IEEE Transactions on Knowledge and Data Engineering*, 21, 479–492. [337]

Yin, G. (1990), "A Stopping Rule for the Robbins–Monro Method," *Journal of Optimization Theory and Applications*, 67, 151–173. [327]

Zhang, H. (2004), "Inconsistent Estimation and Asymptotically Equal Interpolations in Model-Based Geostatistics," *Journal of the American Statistical Association*, 99, 250–261. [325,326,331]

Zhao, L., and Chen, X. (1990), "Normal Approximation for Finite Population U-Statistics," *Acta Mathematicae Applicatae Sinica*, 6, 263–272. [337,338]

Zhu, Z., and Stein, M. L. (2005), "Spatial Sampling Design for Parameter Estimation of the Covariance Function," *Journal of Statistical Planning and Inference*, 134, 583–603. [330]