Reparameterized and Marginalized Posterior and Predictive Sampling for Complex Bayesian Geostatistical Models

Mary Kathryn Cowles

Jun Yan

Brian Smith

November 19, 2007

Abstract

This paper proposes a four-pronged approach to efficient Bayesian estimation and prediction for complex Bayesian hierarchical Gaussian models for spatial and spatiotemporal data. The method involves reparameterizing the variance/covariance structure of the model, reformulating the means structure, marginalizing the joint posterior distribution, and applying a simplex-based slice sampling algorithm. The approach permits fusion of point-source data and areal data measured at different resolutions and accommodates non-spatial correlation and variance heterogeneity as well as spatial and/or temporal correlation. The method produces Markov chain Monte Carlo samplers with low autocorrelation in the output, so that fewer iterations are needed for Bayesian inference than would be the case with other sampling algorithms.

Keywords: Bayesian inference, data fusion, hierarchical models, Markov chain Monte Carlo, slice sampling

1. INTRODUCTION

Data measured over space and time are essential resources for addressing research questions in environmental science, epidemiology, econometrics, and many other disciplines. Using such data for inference and prediction presents both statistical and computational challenges. As an example, suppose that one wished to predict and map uranium concentrations in the surface soil and rocks over the state of Connecticut, and to quantify the uncertainty in the mapped values. Two kinds of relevant data were produced by the National Uranium Resource Evaluation (NURE), a program initiated by the Atomic Energy Commission (now the Department of Energy) in 1973 to identify uranium resources in the United States. One type is areal data — county averages of surficial uranium produced from aerial radiometric surveys. The other is point-site data — measurements of the concentration of uranium in soil and stream sediment measured at specific locations. The units for both types of measurements are parts per million (ppm). This article proposes a unified framework for analyses combining such data.

The departure point for the unified framework is the simple geostatistical model, which provides a natural and interpretable way to model data (such as the soil and sediment values) measured at irregularly-spaced point sites. Let $\{Y(s_i) : s_i \in D, \quad i = 1, \ldots, n\}$ be the observed point-referenced data over a spatial domain D. Suppose that a $p \times 1$ covariate vector $X(s_i)$ can be used to model the large-scale variation, or spatial trend, of $Y(s_i)$. Let $Y = \{Y(s_1), \ldots, Y(s_n)\}^\top$ and $X = \{X^\top(s_1), \ldots, X^\top(s_n)\}^\top$. A Gaussian geostatistical model for Y consists of spatial trend, spatial correlation, and measurement error:

$$\begin{aligned} &Y = X\beta + Z + \epsilon, \\ &Z \sim N\left(0, \sigma_z^2 \Omega(\phi)\right), \quad \epsilon \sim N(0, \sigma_e^2 I), \end{aligned} \tag{1}$$

where β is a $p \times 1$ vector of covariate coefficients, Z is an $n \times 1$ vector capturing the spatial correlation, and ϵ is a $n \times 1$ vector of independent and identically distributed measurement errors. The distribution of Z is multivariate normal with mean zero and covariance matrix $\sigma_z^2 \Omega(\phi)$, where $\Omega(\phi)$ is the correlation matrix as a function of parameter vector ϕ. In general, the correlation of $Z(s_i)$ and $Z(s_j)$ is modeled as a function of the distance, and possibly orientation, between sites s_i and s_j. Although model (1) is presented in the context of spatial data, the notation remains the same when there is a temporal component, except that the parameter vector ϕ contains additional elements for temporal correlation. Yan, Cowles, Wang, and Armstrong (2007) proposed a reparametrized and marginalized posterior sampling (RAMPS) algorithm to make efficient Bayesian inferences for model (1).

In this paper, we extend the model in (1) to accommodate areal data and data with non-spatial correlation, to permit simultaneous modeling of data with different measurement-error variances and different spatial variances, and to perform prediction. We propose an efficient MCMC sampling algorithm based on reparameterization of the variance/covariance structure, reformulation of the means structure, and a simplex-based slice sampling algorithm. Finally, we apply the approach to producing maps of surficial uranium concentrations in Connecticut.

2. ACCOMODATING DATA FUSION AND NON-SPATIAL HETEROGENEITY

Model (1) is inadequate for the example analysis of combined areal and point-source data on uranium concentrations, where data fusion is needed. Another frequently encountered complexity in real data, although not in the uranium example, is non-spatial heterogeneity, which occurs when variability and correlation other than spatial covariance and independent measurement error are present in the data. An example of both data fusion and non-spatial heterogeneity is Smith and Cowles (2007), who combined areal data (county averages of uranium radiation) with point-source data (long-term radon measurements in specific homes). To capture correlation between multiple radon measurements on the same home, a random offset to the intercept for each home was included in the model. We now extend model (1) to accommodate such analyses.

2.1 A Unified Model Framework

Consider n observed values $Y = \{Y_1, \ldots, Y_n\}^\top$ and S spatial sites in the same spatial domain. The observation vector Y of length n may contain both areal and point-source data. The S spatial sites may include: 1) locations at which observed point-source measurements were made; 2) points on a regular grid, which will be treated as the locations of point source measurements that were averaged to produce the areal averages; and 3) sites at which prediction is desired. Our extended Gaussian geostatistical model for Y consists of fixed effects, non-spatial random effects, spatial random effects, and measurement error:

$$\begin{aligned} &Y = X\beta + W\gamma + KZ + \epsilon, \\ &\gamma \sim N(0, \Omega_\gamma), \quad Z \sim N(0, \Omega_Z), \quad \epsilon \sim N(0, \Omega_\epsilon), \end{aligned} \tag{2}$$

where β is a $p \times 1$ vector of regression coefficients, γ is a $q \times 1$ vector of non-spatial random effects, Z is an $S \times 1$ vector of spatial random effects, ϵ is an $n \times 1$ vector of measurement errors, and the matrices X, W, and K are design matrices for fixed effects, non-spatial random effects, and spatial random effects. The matrix K is defined by

$$K_{ij} = \begin{cases} 1, & Y_i \text{ is a point source datum measured at site } j, \\ \frac{1}{N_i}, & \text{site } j \text{ is one of } N_i \text{ sites contributing to areal average } Y_i, \\ 0, & \text{otherwise.} \end{cases}$$

If Y_i is a point-source measurement, then $N_i = 1$. If Y_i is an areal average, then N_i is roughly proportional to the area of the region over which the measurement Y_i is averaged. The finer the grid of sites used, the closer the proportionality will be.

The variance matrices Ω_γ, Ω_Z, and Ω_ϵ need to be specified in detail. An obvious requirement for Ω_ϵ is to allow different data points in Y to have different measurement-error variances to accommodate fusing data measured using different methods, areal and point source data, etc. In particular, estimation of the measurement-error variance associated with areal data will be affected by the *weights* attributed to the areal averages. If actual areas of the regions are used, then the units (square miles versus square kilometers, etc.)

will determine the estimate of measurement-error variance. If the number of grid points N_i is used as a surrogate for the area, then the resolution of the grid will affect the estimate; the finer the grid, the larger the estimate. Therefore, even if the areal and point source data represent measurements of the same underlying process, separate measurement-error variance parameters should be used for them. For the greatest generality, we also allow for different spatial variances at different sites (possibly due to nonstationarity of the spatial covariance) and different variances for the non-spatial random effects.

Suppose that there are L_γ types of variance of the non-spatial random effects $\sigma^2_{\gamma,i},\ i = 1,\ldots,L_\gamma$; L_Z types of spatial variance $\sigma^2_{Z,i},\ i = 1,\ldots,L_Z$, and L_ϵ types of measurement error variance $\sigma^2_{\epsilon,i},\ i = 1,\ldots,L_\epsilon$. Further, let $r_i,\ i = 1,\ldots,q$, take on an integer value between 1 and L_γ to index which of the L_γ types of random-effect variance goes with random effect γ_i. Similarly, let $v_i, i = 1,\ldots,S$, index which of the L_Z types of spatial variance goes with site i and $m_i,\ i = 1,\ldots,n$ index which of the L_ϵ types of measurement error variance is associated with with observation Y_i. We construct vectors for componentwise variances of γ, Z, and ϵ, respectively, as $V_\gamma = \{\sigma^2_{\gamma,r_1},\ldots,\sigma^2_{\gamma,r_q}\}^\top$, $V_Z = \{\sigma^2_{Z,v_1},\ldots,\sigma^2_{Z,v_S}\}^\top$, and $V_\epsilon = \{\sigma^2_{\epsilon,m_1}/w_1,\ldots,\sigma^2_{\epsilon,m_n}/w_n\}^\top$, where $w_i,\ i = 1,\ldots,n$, is the weight associated with observation i. The weight w_i will be 1 for point-source data, and for areal data, either (a) the number of point-source measurements contributing to the areal average (if known), (b) the area of the region, or (c) the number of grid points N_i. Assuming all these types of random effects are mutually independent, we have $\Omega_\gamma = \mathrm{diag}(V_\gamma)$, $\Omega_Z = \mathrm{diag}(V_Z^{1/2})\Sigma(\phi)\mathrm{diag}(V_Z^{1/2})$, and $\Omega_\epsilon = \mathrm{diag}(V_\epsilon)$, where $\Sigma(\phi)$ is a spatial correlation matrix with parameter vector ϕ. The likelihood is specified by

$$Y \sim N\left(X\beta, \quad W\Omega_\gamma W^\top + K\Omega_Z K^\top + \Omega_\epsilon\right). \tag{3}$$

An equivalent specification of the likelihood that expedites MCMC sampling for prediction is given in Section 3. The reparameterization discussed in Section 2.2 is the same in both settings.

2.2 Reparameterizing the variance parameters

The RAMPS algorithm reparameterizes the variance parameters. Concatenate the vectors of measurement error variances, spatial variances, and random effects variances for a total of $F = L_\gamma + L_Z + L_\epsilon$ variance parameters, $\sigma_1^2, \ldots, \sigma_F^2$. If there are one measurement-error variance, one spatial variance, and no random effects variances, then $\sigma_1^2 \equiv \sigma_e^2$ and $\sigma_2^2 \equiv \sigma_z^2$ as in the special case of Yan et al. (2007). Our reparameterization is in terms of $\kappa = \{\kappa_1, \ldots, \kappa_F\}^\top$ and σ_{tot}^2, where

$$\sigma_{tot}^2 = \sum_{j=1}^{F} \sigma_j^2, \quad \text{and} \quad \kappa_j = \frac{\sigma_j^2}{\sigma_{tot}^2},\ j = 1, 2, \ldots, F. \tag{4}$$

Note that $\kappa_F \equiv 1 - \sum_{j=1}^{F-1} \kappa_j$ and is not a free parameter to be estimated. Let $\kappa_\gamma = V_\gamma / \sigma_{tot}^2$, $\kappa_Z = V_Z / \sigma_{tot}^2$, and $\kappa_\epsilon = V_\epsilon / \sigma_{tot}^2$. Then the likelihood is specified as

$$Y \sim N\left(X\beta,\ \sigma_{tot}^2 \Omega\right) \tag{5}$$

where $\Omega = W \mathrm{diag}(\kappa_\gamma) W^\top + K \mathrm{diag}(\sqrt{\kappa_Z}) \Sigma(\phi) \mathrm{diag}(\sqrt{\kappa_Z}) K^\top + \mathrm{diag}(\kappa_\epsilon)$.

2.3 Prior densities under reparameterization

Prior distributions on $\theta = (\phi, \kappa, \sigma^2, \beta)$ complete the Bayesian model specification. The semiconjugate prior on β given σ_{tot}^2 is

$$\beta | \sigma_{tot}^2 \sim N\left(\mu_\beta, \sigma_{tot}^2 \Sigma_\beta\right) \tag{6}$$

If the modeler prefers to specify the prior for β conditional on one of the individual variances σ_i^2 rather than on σ_{tot}^2, then κ_i becomes an additional multiplicative factor in the prior variance of β. Placing an independent, flat, improper prior on each element of β corresponds to letting the diagonal entries of Σ_β go to infinity.

We place independent priors, each with bounded support, on the spatial correlation parameters ϕ. To date we have used uniform priors over an appropriately chosen region. For example, if ϕ is the range parameter in a spatial correlation function, the boundaries of the uniform distribution represent the largest and smallest reasonable values of the effective range.

To construct priors on σ^2_{tot} and κ, suppose that independent inverse gamma (IG) priors with have been placed on each variance σ^2_j. Then the joint prior density induced on κ and σ^2_{tot} is

$$p(\kappa_1, \kappa_2, \ldots, \kappa_{F-1}, \sigma^2_{tot}) \ = \ \prod_{j=1}^{F} \left[\frac{b_j}{\Gamma(a_j)} \frac{1}{\kappa_j^{a_j+1}} \right] \frac{1}{(\sigma^2_{tot})^{\sum_{j=1}^{F} a_j + 1}} \exp\left(-\frac{1}{\sigma^2_{tot}} \sum_{j=1}^{F} \frac{b_j}{\kappa_j} \right).$$

This joint density may be factored into the product of a marginal density for κ and a conditional density for σ^2_{tot} given κ. The marginal density is

$$p(\kappa_1, \kappa_2, \ldots \ldots, \kappa_{F-1}) \ = \ \prod_{j=1}^{F} \left[\frac{b_j}{\Gamma(a_j)} \frac{1}{\kappa_j^{a_j+1}} \right] \frac{\Gamma\left(\sum_{j=1}^{F} a_j\right)}{\left(\sum_{j=1}^{F} \frac{b_j}{\kappa_j}\right)^{\sum_{j=1}^{F} a_j}}$$

and the conditional density $p(\sigma^2_{tot}|\kappa)$ is IG with parameters $\sum_{j=1}^{F} a_j$ and $\sum_{j=1}^{F} b_j/\kappa_j$. If $F = 2$ and $b_1 = b_2$ then the marginal density of κ_1 simplifies to a Beta density. However, if $F > 2$ then the marginal density of κ does not simplify to a Dirichlet even if all the $b_j, \ j = 1, \ldots, F$ are equal.

3. REFORMULATING THE MEANS STRUCTURE FOR PREDICTION

Commonly, instead of, or in addition to, estimating model parameters, the research goal is to predict the underlying values of the spatial process and/or data values at measured or unmeasured locations. These locations may include some or all of the sites at which data values have been observed, or they may consist entirely of sites for which no observed data values (point-site or areal) are available.

To facilitate the prediction algorithm described in Section 5, we reorder and partition the vector of observed data values and the vector of spatially correlated random effects as follows. Reorder Z as $(Z_p^\top, Z_u^\top)^\top$, where Z_p is the vector of spatial random effects at sites for which prediction is desired, and Z_u is the vector of spatial random effects at sites where prediction is unneeded. For a computational benefit which becomes clear later, we reorder

Y as $(Y_1^\top, Y_2^\top)^\top$ such that the design matrix K becomes block triangular, with the columns of K conformable to the reordered Z. We then reorder the measurement errorvector ϵ, and the design matrices X and W comformably to write the model as

$$\left[\begin{array}{c} Y_1 \\ \hline Y_2 \end{array}\right] = \left[\begin{array}{c|c} X_1 & K_{1,1} \\ \hline X_2 & K_{2,1} \end{array}\right] \left[\begin{array}{c} \beta \\ \hline Z_p \end{array}\right] + \left[\begin{array}{c} 0 \\ \hline K_{2,2} \end{array}\right] Z_u + \left[\begin{array}{c} W_1 \\ \hline W_2 \end{array}\right] \gamma + \left[\begin{array}{c} \epsilon_1 \\ \hline \epsilon_2 \end{array}\right].$$

Note that Y_1 is the vector of data values in Y that are associated exclusively with Z_p while Y_2 is the vector of data values in Y that are associated with both Z_p and Z_u, or Z_u alone.

In the manner of structured MCMC (SMCMC) (Hodges 1998; Sargent, Hodges, and Carlin 2000), all model stages involving the means structure can be reformulated as a single linear model:

$$\left[\begin{array}{c} Y_1 \\ \hline Y_2 \\ \hline 0 \\ \hline \mu_\beta \end{array}\right] = \left[\begin{array}{c|c} X_1 & K_{1,1} \\ \hline X_2 & K_{2,1} \\ \hline 0 & -I \\ \hline I & 0 \end{array}\right] \left[\begin{array}{c} \beta \\ \hline Z_p \end{array}\right] + \left[\begin{array}{c} W_1\gamma + \epsilon_1 \\ \hline K_{2,2}Z_u + W_b\gamma + \epsilon_2 \\ \hline \epsilon_{z_p} \\ \hline \epsilon_\beta \end{array}\right], \qquad (7)$$

In a compact form, denote the model as

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}.$$

Here $\mathbf{Y}$ is a vector of known values, $\mathbf{X}$ is a design matrix of known values, $\mathbf{B}$ contains all the unknown means-related parameters for which estimation or prediction is required, and $\mathbf{E}$ is a vector of multivariate normal variate with mean 0 and covariance matrix $\sigma^2_{tot}\mathbf{\Omega}$ with

$$\mathbf{\Omega} = \left[\begin{array}{c|c|c|c} W_1\Omega_\gamma W_1^\top + \Omega_{\epsilon;1,1} & 0 & 0 & 0 \\ \hline 0 & K_{2,2}\Sigma_{Z;u,u}K_{2,2}^\top + W_2\Omega_\gamma W_2^\top + \Omega_{\epsilon;2,2} & \Sigma_{Z;2,1}K_{2,2}^T & 0 \\ \hline 0 & K_{2,2}^T\Sigma_{Z;p,u} & \Sigma_{Z;p,p} & 0 \\ \hline 0 & 0 & 0 & \Sigma_\beta \end{array}\right]. \qquad (8)$$

The advantage of decomposing of Y into Y_1 and Y_2 is that it leads to a block diagonal $\mathbf{\Omega}$, which facilitates matrix operations.

Note that if prediction is not desired, then Y_1, Z_p, $K_{1,1}$ and $K_{2,1}$ all will be null; thus in this case the expression simplifies to:

$$\left[\begin{array}{c} Y \\ \hline \mu_\beta \end{array}\right] = \left[\begin{array}{c} X \\ \hline I \end{array}\right] \left[\, \beta \,\right] + \left[\begin{array}{c} KZ + W\gamma + \epsilon \\ \hline \epsilon_\beta \end{array}\right] \tag{9}$$

or

$$\mathbb{Y} = \mathbb{X}\beta + \mathbb{E},$$

and the covariance matrix of the error vector $\mathbb{E}$ becomes $\sigma^2_{tot}\boldsymbol{\Xi}$ with

$$\boldsymbol{\Xi} = \left[\begin{array}{c|c} K\Sigma_Z K^\top + W\Omega_\gamma W^\top + \Omega_\epsilon & 0 \\ \hline 0 & \Sigma_\beta \end{array}\right]. \tag{10}$$

Furthermore, if Σ_β in (6) is allowed to go to infinity, then the corresponding blocks of $\boldsymbol{\Omega}^{-1}$ and $\boldsymbol{\Xi}$ will go to zero, and the rows of $\mathbf{Y}$, $\mathbf{X}$, and $\mathbf{E}$ corresponding to μ_β will have no effect on the computations in Section 5. Thus these rows can be omitted from the specification altogether when independent, improper priors are used on the elements of β.

The Bayesian model is completed with specification of prior densities on θ as described in Section 2.3.

4. SLICE SAMPLING ON A SIMPLEX

In developing a Markov chain Monte Carlo algorithm for sampling from the posterior and predictive distributions described in the preceding sections, special attention to the parameters $\{\kappa_1, \kappa_2, \ldots \kappa_F\}$ defined in 4 is required. The remainder of this section describes the "SIMPLICE" algorithm, which is a component of the RAMPS algorithm for fitting the general model described in the preceding sections.

4.1 Simplexes

Consider drawing samples from the posterior distribution of a vector-valued parameter $x = \{\kappa_1, x_2, \ldots, x_F\}$ such that $0 < x_j < 1$ for $j = 1, \ldots, F$ and $\sum_{j=1}^F x_j = 1$. The support of the posterior distribution of such a parameter is the *standard $(F-1)$-simplex* — that is, the regular simplex with vertices $\{e_1, \ldots, e_F\}$, where $e_{ij} = 0$ for $j \neq i$ and $e_{ij} = 1$ for $j = i$. For example, the standard 1-simplex is the line segment connecting (1,0) and (0,1), and the standard 2-simplex is the equilateral triangle with vertices $e_1 = (1, 0, 0)^\top$, $e_2 = (0, 1, 0)^\top$, and $e_3 = (0, 0, 1)^\top$. An *edge* of a simplex is a line segment connecting two adjacent vertices. A *regular* simplex has all edges of equal length. Sampling uniformly from the standard $(F - 1)$-simplex is equivalent to sampling from the Dirichlet distribution with parameters $\alpha = \{\alpha_1 = 1, \alpha_2 = 1, \ldots, \alpha_F = 1\}$. Note that in the case of a 1-simplex, this is the Beta$(1, 1)$, or standard uniform, distribution.

Any point in the interior of a simplex may be expressed as a convex combination of the vertices of the simplex. The coefficients of the vertices in the convex combination are called the *barycentric coordinates* of the point. Let $V = (v_1, \ldots, v_F)$ be the matrix of vertices in Cartesian coordinates of an $(F - 1)$-simplex. Let $T = (t_1, \ldots, t_F)^\top$ be the vector of the barycentric coordinates of a point in the simplex. Then the Cartesian coordinates of the point are given by the map (see, for example, Hörmann, Deydold, and Derflinger 2004, Algorithm 1.10, p.257):

$$T \mapsto \sum_i t_i v_i = VT. \tag{11}$$

Thus, sampling uniformly from an arbitrary $(F - 1)$-simplex may be accomplished by sampling the barycentric coordinates with respect to the vertices $(v_1, \ldots, v_F)$ from a Dirichlet distribution with parameters $\alpha = \{\alpha_1 = 1, \alpha_2 = 1, \ldots, \alpha_F = 1\}$. The Cartesian coordinates of the point may then be obtained from the mapping (11).

4.2 SIMPLICE sampling algorithm

Slice sampling is based on the "fundamental theorem of simulation" (Robert and Casella 2004, Theorem 2.15) which states that simulating a random variable or vector X from the density $f(x)$ is equivalent to sampling uniformly under the graph of $f(x)$ – that is, to simulating (X, Y) from the joint density that is uniform on $\{(x, y) : 0 < y < f(x)\}$. If the draws of y are ignored, then the marginal density of the draws of x is $f(x)$. This theorem also applies if $f(x)$ is known only up to a normalizing constant.

Algorithm 1 Outline of SIMPLICE sampling

```
1: $y \Leftarrow u\ f(x_0)$, where $u$ is a random drawn from $U(0,1)$.
2: $\Delta \Leftarrow$ initial sampling simplex as in Algorithm 2
3: Sample $c \sim \text{Unif}(\Delta)$
4: while $c \notin S_x$ and $f(c) <= y$ do
5:     $\Delta \Leftarrow$ shrunk simplex as in Algorithm 3
6:     Sample $c \sim \text{Unif}(\Delta)$
7: end while
8: return $x_1 = c$
```

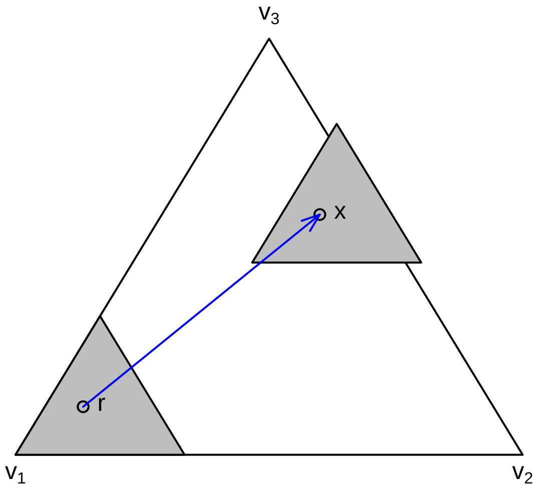
The proposed simplex slice (SIMPLICE) sampling algorithm is adapted from the univariate slice sampling algorithm (Neal 2003). The goal is to sample a new point x_1 from density function f whose support is on a simplex with vertices comprising the columns of a matrix V, using x_0, the value drawn from the previous iteration. Let Unif(A) denote a uniform distribution over region A. Let S_x be the support of $f(x)$. The broad outline is Algorithm 1, followed by specifics.

Line 2 in Algorithm 1 is motivated when f is very concentrated on a small part of its support. In this case, searching the entire standard $(F-1)$ simplex for a point in the "slice" becomes inefficient, particularly if evaluating f is computationally expensive. Starting from a smaller simplex randomly placed over x_0 reduces the average number of evaluations per iteration, possibly at the cost of increasing autocorrelation in the slice sampler output. Steps 3-5 of algorithm 2 create a matrix containing the vertices of a new simplex whose edge length is a specified proportion q of that of the standard $(F-1)$-simplex V_0. Steps 6-7 translate the vertices of this smaller simplex to place it randomly over x_0. The Figure next to Algorithm 2 illustrates how this smaller simplex is placed for a standard 2-simplex with $0 < q < 1$. In the trivial case when $q = 1$, we simply use $V_1 \equiv V_0$.

Pseudo-code is presented in Algorithm 3. The figure next to Algorithm 3 illustrates how the shrinking is done. The idea is that for each vertex that is closer to x_0 than to c, we slide all the other vertices toward this vertex until we hit c. The dark grey and light grey areas

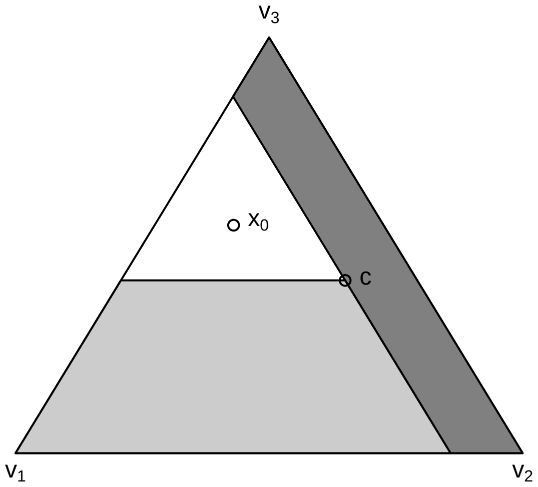
Algorithm 2 Making the initial simplex.

```
$V_1 \Leftarrow V_0$
if $0 < q < 1$ then
    for $i = 2, \ldots, F$ do
        $V_{1,i} \Leftarrow V_{0,i} + q(V_{0,1} - V_{0,i})$
    end for
    Sample $U_0 \sim Dirichlet(1, 1, ..., 1)$
    $V_1 \Leftarrow V_1 + (x_0 - U_0 V_1)J^\top$, where $J^\top = (1, \ldots, 1)$
end if
return $V_1$
```



Algorithm 3 Shrinking simplexes

```
$V_1 \Leftarrow V_0$
$\delta \Leftarrow b_c - b_{x_0}$.
for all $i$ such that $\delta_i < 0$ do
    for all $j \neq i$ do
        $V_{1,j} \Leftarrow V_{0,j} + b_{c,i}(V_{0,i} - V_{0,j})$
        $b_c \Leftarrow V_1^{-1} c_c$
        $V_0 \Leftarrow V_1$
    end for
end for
```



are removed first and second.

Proof of correctness of the algorithm is given in the appendix.

5. MARKOV CHAIN MONTE CARLO SAMPLING ALGORITHM

5.1 Marginalization for posterior and predictive sampling

The RAMPS algorithm is designed to approach independent sampling from the joint posterior distribution of all unknown model parameters $\theta \equiv (\phi, \kappa, \sigma^2_{tot}, \mathbf{B})$. The posterior density of θ given $\mathbf{Y}$ may be factored as

$$p(\theta|Y) = p(\phi, \kappa|Y)p(\sigma^2_{tot}|\phi, \kappa, Y)p(\mathbf{B}|\phi, \kappa, \sigma^2_{tot}, Y). \tag{12}$$

The RAMPS algorithm draws from $p(\theta|Y)$ by sequentially drawing from the three conditional densities on the right hand side of equation (12). We discuss each step in detail next.

5.2 Sampling from $p(\phi, \kappa|Y)$

Step one of an MCMC iteration, say k, is to draw

$$p(\phi^k, \kappa^k) \sim p(\phi, \kappa|Y).$$

Note that this density is the joint posterior marginal density of ϕ and κ, not a conditional density depending on values of other parameters from the previous iteration.

Sampling of this density is most efficient using the formulation in (9). The use of semi-conjugate priors on σ^2_{tot} and β simplifies the process of integrating these parameters out of the resulting joint posterior distribution to obtain the following analytic form of $p(\phi, \kappa|Y)$ up to a normalizing constant:

$$p(\phi, \kappa|Y) \propto |\Xi|^{-1/2} \left|\mathbb{X}^\top \Xi^{-1} \mathbb{X}\right|^{-1/2} \left(\frac{\widehat{RSS}(\phi, \kappa)}{2} + \sum_{j=1}^{F} \frac{b_j}{\kappa_j} \right)^{-\left(\sum_{j=1}^{F} a_j\right) - \frac{n-p^*}{2}} \prod_{j=1}^{F} \left(\kappa_j^{-a_j - 1} \right), \tag{13}$$

where $p^* = 0$ if a proper conjugate multivariate normal prior is used for β and $p^* = p$ if an improper flat prior is used for β, $\widehat{RSS}(\phi, \kappa) = (\mathbb{Y} - \mathbb{X}\hat{\beta})^\top \Xi^{-1} (\mathbb{Y} - \mathbb{X}\hat{\beta})$, and $\hat{\beta}$ is the weighted least squares estimate of β given ϕ and κ (and μ_β and Σ_β if a proper prior on β is used) in the linear model (9).

Algorithm 4 Slice sampling for ϕ and κ

```
$y \Leftarrow u\ f(x_0)$, where $u \sim \text{Unif}(0,1)$ and $x_0 = (\phi_0, \kappa_0)$
$(H \times \Delta) \Leftarrow$ initial sampling hyperrectangle and simplex
Sample $c \sim \text{Unif}(H \times \Delta)$
**while** $c \notin S_\phi \times S_\kappa$ and $f(c) <= y$ **do**
    $H \Leftarrow$ shrunk hypertectangle as in Neal (2003)
    $\Delta \Leftarrow$ shrunk simplex as in Algorithm 3
    Sample $c \sim \text{Unif}(H \times \Delta)$
**end while**
**return** $x_1 = c$
```

If there were a way to draw independent samples from (13), then our algorithm would produce independent draws from the joint posterior distribution of all model parameters. This is, however, not possible, and, within each iteration of the MCMC sampler, we must turn to iterative methods to draw from (13). Yan et al. (2007) found that using slice sampling (Neal 2003) rather than the Metropolis-Hastings algorithm (Hastings 1970) reduced autocorrelation in the MCMC output.

Our slice sampling algorithm for ϕ and κ combines the shrinking-hyperrectangle method described in Neal (2003) for ϕ with the shrinking-simplex method proposed in Section 4 for κ. Details are in Algorithm 4.

Note that evaluating the expression in (13) requires calculating two determinants and a quadratic form. Computing the Cholesky decomposition of $\boldsymbol{\Xi}$ facilitates calculating both the determinant $|\boldsymbol{\Xi}|$ and the quadratic form $\widehat{RSS}$. Unfortunately, Cholesky decomposition is very computationally intensive (of order d^3 on a dense matrix of dimension d). This step of the algorithm, in which (13) may need to be evaluated repeatedly for different values of ϕ and κ before an acceptable candidate is found, relies on the formulation in (9) rather than that in (7). Let n_{z_p} be the length of Z_p. Matrix $\boldsymbol{\Xi}$ is of dimension $(n+p) \times (n+p)$ (or only $n \times n$ with an improper prior on β), whereas $\boldsymbol{\Omega}$ has an additional block of dimension $n_{z_p} \times n_{z_p}$ and therefore would be slower to decompose. The matrix $\mathbb{X}^\top \boldsymbol{\Xi}^{-1} \mathbb{X}$ is only $p \times p$, so computation of its determinant is trivial.

5.3 Sampling from $p(\sigma^2_{tot}|\phi, \kappa, Y)$

With the aforementioned priors on $(\sigma^2_{tot},\ \kappa)$, and β, it can be shown that $p(\sigma^{2,k}_{tot}|\phi^k, \kappa^k, Y)$ is inverse gamma:

$$\mathrm{IG}\left(\sum_{j=1}^{F} a_j + \frac{n-p^*}{2}, \quad \sum_{j=1}^{F} \frac{b_j}{\kappa^k_j} + \frac{\widehat{RSS}(\phi^k, \kappa^k)}{2}\right). \tag{14}$$

Since $\widehat{RSS}(\phi, \kappa)$ has already been calculated in step 1, this step is completely straightforward.

5.4 Sampling from $p(\mathbf{B}|\phi, \kappa, \sigma^2_{tot}, Y)$

It can be shown that $p(\mathbf{B}^k|\phi^k, \kappa^k, \sigma^{2,k}_{tot}, Y)$ is multivariate normal $N\left(\hat{\mathbf{B}}, \quad \sigma^2_{tot}\left[\mathbf{X}^\top\mathbf{\Omega}^{-1}\mathbf{X}\right]^{-1}\right)$, where $\hat{\mathbf{B}}$ is the weighted least squares estimate of $\mathbf{B}$ in the linear model (7).

If only parameter estimation is needed, then $\mathbf{B} = \beta$, and this step involves drawing β from a multivariate normal density, all components of the mean and precision matrix of which have already been calculated in step 1 as described above.

On the other hand, if prediction is also required, then $\mathbf{B}$ includes both β and Z_p. This step may be carried out by plugging $\phi^{(k)}$ and $\kappa^{(k)}$ into the formulation in (7) and calculating $\hat{\mathbf{B}}$ and the precision matrix $\mathbf{X}^\top\mathbf{\Omega}^{-1}\mathbf{X}$. Cholesky decomposition of $\mathbf{\Omega}$ facilitates computation of $\mathbf{X}^\top\mathbf{\Omega}^{-1}\mathbf{X}$. To speed up the Cholesky decomposition, the reformulation in Section 3 has been carefully designed to make $\mathbf{\Omega}$ block diagonal (so that the Cholesky decomposition can be performed on each block individually) and as sparse as possible (so that sparse matrix operations can be used). The matrix $\mathbf{X}^\top\mathbf{\Omega}^{-1}\mathbf{X}$, which must also be Cholesky decomposed, is of dimension $p + n_{Z_p}$, where n_{Z_p} is the number of prediction sites. Keeping this dimension as small as feasible is the rationale for designing the reformulation so that not all sites that contribute to observed data values have to be included in the prediction vector.

6. EVALUATION OF ALGORITHM PERFORMANCE

A small simulation study was conducted to evaluate the frequentist performance of the Bayesian models and computing strategy for combined areal and point-source data. Two hundred datasets were simulated to imitate areal averages for the 99 counties in the state of Iowa as well as point-source data from 250 locations generated randomly over the state. A regularly-spaced grid of 403 points was used to underlie the areal data. The spherical correlation function was used in simulating the spatial process. Distance was calculated as great circle distance in hundred miles. The point-source data had an average of 3 observations at each location, imitating multiple measurements within individual homes. A non-spatial random effect for homes was incorporated into the simulation. In summary, there are 99 areal observations and 750 point-source observations.

For each dataset, three different model fits were performed: (a) using the point-source data only, (b) using the areal averages only, and (c) fusing both kinds of data. In all cases, flat priors were used on β_a and β_p, the intercepts for areal and point-source data respectively. Vague inverse gamma priors with shape and scale parameters set to 0.01 were used for σ_z^2 (variance of the spatial process), $\sigma_{e,p}^2$ and $\sigma_{e,a}^2$ (measurement error variances of point source and areal data), and σ_{re}^2 (variance of non-spatial random effects for homes). A uniform prior on $(0, 3)$ was used for the spatial range parameter ϕ. For each run, a single MCMC sampler was run for 1200 iterations. The first 200 were discarded as burn-in, and posterior means and credible sets were calculated from the remaining 1000 iterations. Results are reported in Table 1.

[Table 1 about here.]

The spatial range parameter ϕ is consistently over-estimated, indicating over-smoothing. Using a finer grid to underlie the areal data might improve estimation of ϕ. The overestimation of the variance parameters is due to using the posterior mean as the point estimate when the marginal posterior density is right skewed; positive bias is expected in this case. With

200 datasets in the simulation, the standard error in estimating the coverage is 0.015. Thus there is no strong evidence of under- or over-coverage except that ϕ has slight under-coverage.

The simulation study also exemplifies the advantages of data fusion relative to analysis of areal data alone or point-source data alone. Bias is smaller and interval widths are narrower while coverage is at least as good when both types of data are used than when either is used alone.

For simpler geostatistical models with the structure in (1), Smith, Yan, and Cowles (2007) report on comparisons between the RAMPS algorithm as implemented in the R package `ramps` and a blocked Metropolis-Hastings algorithm as implemented in the R package `spBayes` (Finley, Banerjee, and Carlin 2007). Although the Metropolis-Hastings algorithm produces more iterations per unit time, it suffers from high autocorrelation in the sampler output. As a result, the RAMPS algorithm is shown to produce from 5.5 to 10.7 times as many effective samples per unit time as the blocked M-H algorithm for parameters in the variance/covariance structure.

7. REAL-DATA EXAMPLE OF PREDICTION

To predict the surface of uranium concentration over the state of Connecticut, both the areal and point-source data were log-transformed prior to analysis as is standard practice with uranium data. A regularly-spaced grid of 1117 points was laid over the state of Connecticut to underlie the county averages for the 8 counties. Preliminary exploratory analysis of the point-source data suggested that the exponential correlation function was the best-fitting one-parameter spatial correlation function for these data. The maximum distance between any two locations in Connecticut is 123 miles. For the analysis, improper flat priors were placed on β_p and β_a, the intercepts for point-source and areal data respectively. Vague inverse gamma priors with both shape and scale parameters set equal to 0.01 were placed on all the variances ($\sigma^2_{e,p}$, $\sigma^2_{e,a}$ and σ^2_z). The uniform prior on the spatial correlation parameter ϕ is set to be $\mathrm{Unif}(0, 100)$, which expresses the belief that the distance at which the spatial correlation decayed to 0.05 was somewhere between 0 and 300 miles. A single MCMC sampler chain was run for 500 iterations. Figure 1 plots the mean and standard deviation of the posterior predictive distribtution of the underlying spatial process at the points on the grid based on the joint analysis of both areal and point-source data. The standard deviation of the posterior predictive distribution is larger at the eastern end of the state, where no point-source data were reported.

[Figure 1 about here.]

8. DISCUSSION

For estimation and prediction using complex spatial and spatiotemporal data, this paper has presented a hierarchical Bayesian geostatistical model and an efficient MCMC-based computational algorithm, called RAMPS, to fit it. Simulation study results indicate that the model performs well with respect to bias and interval coverage for datasets including both areal and point-source data and manifesting non-spatial as well as spatial correlation. The proposed algorithm, based on slice-sampling for a marginalized posterior density with support on the intersection of a hyperrectangle and a simplex, is shown to produce 5 to 10 times as many effective samples per unit time as standard Metropolis-with-Gibbs algorithms for fitting Bayesian geostatistical models. A real data example illustrates the ability of the model and algorithm to estimate the posterior predictive distribution of the underlying spatial process, thereby making it possible to map the estimated surface and to quantify the uncertainty in the map.

He, Hodges, and Carlin (2007) propose a reparameterization and marginalization of the precision parameters (inverses of variances) in simpler gaussian spatial models using intrinsic conditional autoregressive (ICAR) priors on the spatial random effects. Their simplex-based slice sampling algorithm includes a grid-based stepping-out procedure to specify the initial sampling area. This procedure would be feasible only when the dimension of the simplex was very small, whereas the simplex-shrinking approach that we have proposed in our slice-sampling algorithm is much more broadly applicable. They do not have a method of shrinking the sampling area if many candidate points are rejected. Their ICAR-based models are simpler than our geostatistical models in several important respects: ICAR models have no spatial correlation parameters corresponding to ϕ; spatial association is expressed in terms of precision matrices rather than variance/covariance matrices; and their models do not accommodate non-spatial correlation. Nevertheless, their finding of better MCMC sampler performance after reparameterization and marginalization is very consistent with ours.

A. APPENDIX: PROOF OF CORRECTNESS OF THE SIMPLICE ALGORITHM

Our proof of the correctness of our SIMPLICE algorithm for drawing from a density with support on a standard simplex follows the general outline of Neal's proof of the correctness of the single-variable slice sampling algorithm in Section 4.3 of Neal (2003). Two properties must hold in order to guarantee convergence of the Markov chain constructed by the algorithm to the target distribution: (a) the Markov chain must be ergodic and (b) each update must leave the target distribution invariant. We consider here only the case in which $f(x) > 0$ over the entire simplex, and, hence, ergodicity follows by the argument of Neal (2003).

To show invariance, suppose that x_0 is distributed as $f(x)$. What must be shown is that the selection of x_1 in lines 3–8 of Algorithm 1 leaves the joint distribution of x_0 and x_1 invariant. In this setting, this can be accomplished by demonstrating that the updates satisfy *detailed balance*, which means that "the probability density for x_1 to be selected as the next state given that x_0 is the current state is the same as the probability density for x_0 to be selected as the next state given that x_1 is the current state, for any states x_0 and x_1 within S" (Neal 2003, Section 4.2).

Like Neal's single-variable slice sampler algorithm, our simplex slice sampling algorithm requires intermediate steps based on the generation of random variates. As in his proof, we let r denote those random choices and let $\pi(r)$ denote a "one-to-one mapping with Jacobian one (with regard to the real-valued variables), which may depend on x_0 and x_1," and then demonstrate that

$$\begin{aligned} & \Pr(\text{next state} = x_1 \text{ and intermediate choices} = r | \text{current state} = x_0) \\ = \; & \Pr(\text{next state} = x_0 \text{ and intermediate choices} = \pi(r) | \text{current state} = x_1) \end{aligned}$$

The required result is obtained by integrating over all possible values of r.

For specifying the initial sampling simplex, if the entire standard $(F-1)$-simplex is used $(q = 1)$, then there is no randomness at this step and no mapping is required. If $0 < q < 1$, then the placement of the smaller initial sampling simplex around x_0 depends on U_0, a random draw from the Dirichlet density. If x_1 is not inside the initial sampling simplex determined by U_0, then the probability of passing from x_0 to x_1 equals the probability of passing from x_1 to x_0 (both probabilities are 0), so detailed balance holds. We need to show that if the initial sampling simplex contains x_1, then the mapping $\pi(U_0)$ must give the value that would have produced the same initial sampling simplex with x_1 as the starting point. That is, $U_1 = \pi(U_0)$ must satisfy $V_0 + (x_0 - V_0U_0)J^\top = V_0 + (x_1 - V_0U_1)J^\top$ or

$$U_1 = U_0 + V_0^{-1}(x_1 - x_0).$$

U_1 produced by this mapping will be the barycentric coordinates of x_1 with respect to V and, as such, will be a random draw from a Dirichlet with all parameters equal to 1. Thus the probability densities for U_0 and U_1 are the same. Furthermore, the Jacobian of the transformation is one, as required.

The mapping π also maps the sequence of candidate points c generated when starting from x_0 to the same sequence of candidate points if starting from x_1. This sequence of candidate points determines how the sampling simplex is shrunk. The probability density for selecting the first candidate point is obviously the same whether we begin from x_0 or x_1 because the initial sampling simplex is the same. If for any candidate point in the sequence, the simplex-shrinking procedure starting from x_0 (x_1) produces a simplex that does not contain x_1 (x_0), then we again are in the situation where the probability of passing from x_0 to x_1 equals the probability of passing from x_1 to x_0 equals 0, so detailed balance holds.

Otherwise, the candidate points produce the same sequence of shinking sampling simplexes, each containing both x_0 and x_1, so that the probability densities for each rejected candidate, as well as for the final accepted candidate, are the same whether we start from x_0 or x_1. This establishes "detailed balance."

ACKNOWLEDGMENT

The authors thank Vasyl Zhabotynsky for assistance in testing code and evaluating algorithm performance.

REFERENCES

Finley, A. O., Banerjee, S., and Carlin, B. P. (2007), "spBayes: An R Package for Univariate and Multivariate Hierarchical Point-referenced Spatial Models," *Journal of Statistical Software*, 19.

Hastings, W. K. (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, 57, 97–109.

He, Y., Hodges, J. S., and Carlin, B. P. (2007), "Re-considering the variance parameterization in multiple precision models," *Bayesian Analysis*, 2, 529–556.

Hodges, J. S. (1998), "Some Algebra and Geometry for Hierarchical Models, Applied to Diagnostics (Disc: p521-536)," *Journal of the Royal Statistical Society, Series B, Methodological*, 60, 497–521.

Hörmann, W., Deydold, J., and Derflinger, G. (2004), *Automatic Nonuniform Random Variate Generation*, Berlin: Springer.

Neal, R. M. (2003), "Slice Sampling," *The Annals of Statistics*, 31, 705–767.

Robert, C. P. and Casella, G. (2004), *Monte Carlo Statistical Methods*, Springer, second edition ed.

Sargent, D. J., Hodges, J. S., and Carlin, B. P. (2000), "Structured Markov Chain Monte Carlo," *Journal of Computational and Graphical Statistics*, 9, 217–234.

Smith, B. J. and Cowles, M. K. (2007), "Correlating point-referenced radon and areal uranium data arising from a common spatial process," *Journal of the Royal Statistical Society, Series C — Applied Statistics*, 56, 313–326.

Smith, B. J., Yan, J., and Cowles, M. K. (2007), "Unified Geostatistical Modeling for Data Fusion and Spatial Heteroskedasticity with R package ramps," Tech. Rep. 385, The University of Iowa Department of Statistics and Actuarial Science, http://www.stat.uiowa.edu/techrep/.

Yan, J., Cowles, M., Wang, S., and Armstrong, M. (2007), "Parallelizing MCMC for Bayesian Spatiotemporal Geostatistical Models," *Statistics and Computing*, 17, 323–335.

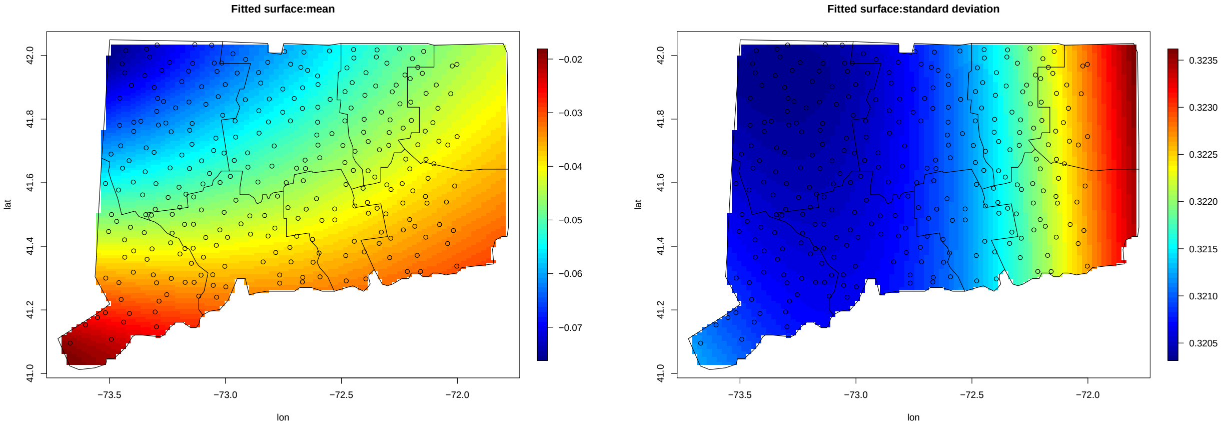


Figure 1: Summary of uranium concentration analysis in Connecticut.

Table 1: Summary of simulation results.

Point-source data only

Parameter	Truth	Mean of Post Means	95% Credible Sets Width	 Coverage
ϕ	100.00	158.660	223.975	0.885
$\sigma^2_{e,p}$	0.25	0.251	0.061	0.940
σ^2_z	0.40	0.631	1.095	0.970
σ^2_{re}	0.20	0.202	0.157	0.935
β_p	0.00	0.001	1.343	0.985

Areal data only

Parameter	Truth	Mean of Post Means	95% Credible Sets Width	 Coverage
ϕ	100.00	159.402	228.098	0.895
$\sigma^2_{e,a}$	0.25	0.242	0.334	0.895
σ^2_z	0.40	0.690	1.317	0.975
β_a	0.50	0.499	1.426	0.985

Fusion of point-source and area data

Parameter	Truth	Mean of Post Means	95% Credible Sets Width	 Coverage
ϕ	100.00	144.362	179.080	0.905
$\sigma^2_{e,p}$	0.25	0.251	0.062	0.940
$\sigma^2_{e,a}$	0.25	0.263	0.250	0.965
σ^2_z	0.40	0.563	0.822	0.980
σ^2_{re}	0.2	0.202	0.138	0.940
β_p	0.00	0.005	1.156	0.985
β_a	0.50	0.501	0.181	0.945