22S:138
Spatial statistics

Lecture 22
Nov. 28, 2007

Kate Cowles
374 SH, 335-0727
kcowles@stat.uiowa.edu

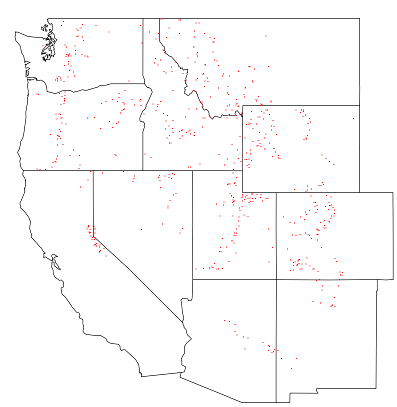
How I became interested in problem

Cowles, M.K., Zimmerman, D.L., Christ, A., and McGinnis, D.L. (2002) Combining Snow Water Equivalent Data from Multiple Sources to Estimate Spatio-Temporal Trends and Compare Measurement Systems. *Journal of Agricultural, Biological, and Environmental Statistics*, **7**, 536-557.

Problem: water supply in western United States

- approximately 75% of annual discharge in western rivers begins as snowpack
- water management decisions depend on estimation of the water contained in the snowpack
- several U.S. government agencies collect data on snow water equivalent (SWE)
  - amount of water in the snow
- we considered annual SWE data from the eleven westernmost of the lower 48 United States ("western U.S.")
  - N =70,745 observations of SWE
  - S = 2027 sites
  - T = 89 years (1910-1998)

SWE Measurement Sites



Goals of our study

- to estimate the temporal trend in SWE over the entire western U.S.
- to characterize how this trend varies spatially
- to investigate whether there are systematic differences in the accuracy and reliability of the measurement systems
- to account appropriately for spatiotemporal correlation structure of data

Motivation for studies of parallelizing MCMC for Bayesian spatial and spatiotemporal models

- preferred model involved a separable correlation structure
  - geostatistical model for spatial component
  - AR 1 model for temporal component
- in 2001, we abandoned the idea of fitting such a model to the dataset as a whole because months of computing time would have been involved
- in 2002-2003, I began collaboration with Marc Armstrong and Shaowen Wang
  - Shaowen heads Grid Research and Education Group at the University of Iowa

Geostatistical models

- natural and interpretable way to model spatial correlation for data measured at irregularly-spaced point sites
- correlation is a function of the distance, and possibly orientation, between sites

Parametric correlation functions

Function	$corr(\phi, d)$
Exponential	$exp(-\phi d)$
Spherical	$\frac{1}{2}(\phi^3 d^3 - 3\phi d + 2), \; d \leq \frac{1}{\phi}$
	$0, \quad d > \frac{1}{\phi}$

- ϕ is a parameter controlling the rate of decay of correlation with increasing distance
- $corr(\phi, d)$ is the correlation between residuals at two sites separated by distance d.

Simple geostatistical model with spatial correlation and additive measurement error

$$\mathbf{Y} \sim N\left(\mathbf{X}^T\boldsymbol{\beta},\ \sigma_s^2\,\Sigma(\phi)\ +\ \sigma_e^2\,I\right)$$

- $\mathbf{X}$ is a matrix of location-specific covariates
- $\boldsymbol{\beta}$ is a vector of coefficients to be estimated
- $\Sigma(\phi)$ is spatial correlation matrix
  - entries are calculated from correlation function
- σ_s^2 is spatial variance
- σ_e^2 is random variance (measurement error variance)
- I is identity matrix
- Bayesian model completed by specification of prior distributions on ϕ, σ_s^2, σ_e^2, and $\boldsymbol{\beta}$

Whiley and Wilson, 2004

- went beyond "embarrassingly parallel" implementation of MCMC in which separate chains are run on different processors
- identified two potential benefits of parallelizing within-iteration MCMC computations for latent spatial Gaussian models
  - reducing time taken to generate required number of samples from (approximation to) target distribution
  - allowing given MCMC algorithm to be applied to target distribution of larger dimension by dividing storage requirements among processors and over distributed memory

- proposed and compared several different parallel MCMC algorithms with respect to run time and mixing
  - all algorithms based on Metropolis updates
- found speedups up to a factor of 5 with 8 processors
- slow mixing

Whiley M and Wilson SP (2004) Parallel algorithms for Markov chain Monte Carlo methods in latent spatial Gaussian models. *Statistics and Computing*, **14**:171-179.

Diggle and Ribeiro, and R package "geoR"

- one practical solution to computational intensiveness
- reparameterized covariance structure

$$\sigma_s^2\Sigma(\phi) + \sigma_e^2 I = \ \sigma_s^2\left(\Sigma(\phi) + taurel I\right)$$

  where $taurel\ =\ \frac{\sigma_e^2}{\sigma_s^2}$
- required treatment of $taurel$
  - fixed, known value, or
  - discrete uniform prior
- semi-conjugate inverse gamma prior on σ_s^2
- normal or flat prior on $\boldsymbol{\beta}$

MCMC algorithm in geoR R package

- based on factoring posterior

$$p(\boldsymbol{\beta}, \phi, \sigma_s^2, ta$$
$$p(\phi, taurel|\mathbf{y}) \times p(\sigma_s^2|\phi, taurel, \mathbf{y}) \times p(\boldsymbol{\beta}|\phi, \sigma_s^2$$

- each MCMC iteration m
  - generate $(\phi^m, \ taurel^m)$ from discrete joint marginal $p(\phi, taurel|\mathbf{y})$
  - generate $\sigma_s^{2\ m}$ from $p(\sigma_s^2|\phi^m, taurel^m, \mathbf{y})$
    - ∗ inverse gamma
  - generate $\boldsymbol{\beta}$ from $p(\boldsymbol{\beta}|\phi^m, \sigma_s^{2\ m}, taurel^m, \mathbf{y})$
    - ∗ multivariate

Diggle, P.J. and Ribeiro Jr, P.J. (2002) Bayesian inference in Gaussian model-based geostatistics. *Geographical and Environmental Modelling*, Vol. 6, No. 2, 129-146.

Our alternative reparameterization

- facilitates prior specification and computing algorithm
- reparameterized covariance matrix

$$\sigma_s^2 \Sigma(\phi) + \sigma_e^2 I = \sigma_{tot}^2 \left[(1-S)\ \Sigma(\phi)\ + S\ I\right]$$

where

$$\sigma_{tot}^2 = \sigma_s^2 + \sigma_e^2$$
$$S = \frac{\sigma_e^2}{\sigma_s^2 + \sigma_e^2}$$

- priors
  - continuous uniform prior on ϕ
    - ∗ endpoints chosen to reflect belief as to largest and smallest possible distances at which spatial correlation could decay to 0
  - $U(0,1)$ prior on S – uniform shrinkage
  - $IG(a,b)$ prior on σ_{tot}^2
  - multivariate normal or flat prior on $\boldsymbol{\beta}$

Spatiotemporal model with separable correlation structure

$$\mathbf{Y} \sim N\Big(\mathbf{X}^T\boldsymbol{\beta},\ \sigma_{tot}^2\ \Big\{(1-S)\ \mathbf{K}\ [\Sigma(\phi) \otimes \Sigma(\rho)]\ \mathbf{K}^T$$

- where $\Sigma(\rho)$ is an AR(1) matrix representing temporal correlation
- K is a matrix of 1's and 0's that matches each observation Y_i with the correct row and column of $\Sigma(\phi) \otimes \Sigma(\rho)$
  - K is not needed if data are "rectangular"
- prior on ρ uniform on (-1,1) or (0,1), slightly bounded away from endpoints

Our sequential MCMC algorithm

- based on factoring posterior

$$p(\ \boldsymbol{\beta}, \phi, \rho, \sigma_{tot}^2, S|$$

$$p(\phi, \rho, S|\mathbf{y})\ \times\ p(\sigma_{tot}^2|\phi, \rho, S, \mathbf{y})\ \times\ p(\ \boldsymbol{\beta}|\phi, \rho, \sigma_{tot}^2$$

- each MCMC iteration m
  - generate $(\phi^m,\ \rho^m,\ S^m,)$ from continuous joint marginal $p(\phi, \rho, S|\mathbf{y})$ using slice sampling
  - generate ${\sigma_{tot}^2}^m$ from $p(\sigma_{tot}^2|\phi^m, \rho^m, S^m, \mathbf{y})$ — inverse gamma
  - generate $\boldsymbol{\beta}$ from $p(\ \boldsymbol{\beta}|\phi^m, \rho^m, {\sigma_{tot}^2}^m, S^m, \mathbf{y})$ — multivariate normal
- all parameters are blocked – would be i.i.d. sampling if there were a way to obtain independent draws from joint posterior marginal of ϕ, ρ, and S

Drawing from joint posterior marginal of ϕ, ρ, and S

$$p(\phi, \rho, S \mid \mathbf{y}) \propto \frac{1}{\left(b + \frac{\left[\mathbf{y} - \mathbf{X}\ \hat{\boldsymbol{\beta}}(\phi,\rho,S)\right]^T \Sigma(\phi,\rho,S)^{-1}\left[\mathbf{y} - \mathbf{X}\right.}{2}\right.}$$

$$\times \frac{1}{|\Sigma(\phi, \rho, S)|^{1/2}}\ \frac{1}{|\mathbf{X}^T \Sigma(\phi, \rho, S)^{-1}\mathbf{X}|}$$

$$\times I_{(0,1)}(S)\ \times\ I_{l,r}(\phi)\ \times\ I_{-1,1}(\rho)$$

- Cholesky decomposition of $\Sigma(\phi, \rho, S)$ enormously reduces computation involved for obtaining required determinants and quadratic form
- trivariate slice sampling (Neal, 2003) attractive due to finite support of all 3 parameters

- in our experience, compared to Metropolis updating, slice sampling
  - results in lower autocorrelation in sampler output since new values are drawn at every iteration
  - at cost of requiring more computationally-expensive evaluations within each iteration
  - results in more "effective samples per second"

Parallelizing the algorithm

- used PLAPACK
  - public-domain parallel linear algebra library for dense matrix operations
  - `www.cs.utexas.edu/~plapack/`
  - van de Geijn, 1997
- correlation matrix is distributed among multiple processors, and Cholesky decomposition is done in parallel
- "master" node "gathers" results from each iteration and does computations involving scalar quantities

Cluster used for timing studies

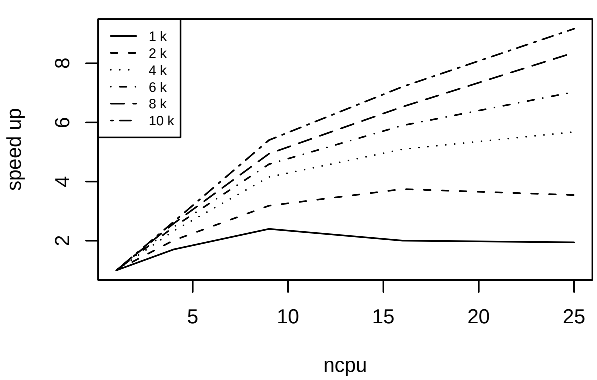
- Beowulf Linux cluster
- 14 nodes
- Dual 1.4GHz CPU
- 1G memory
- No other users

Timing results for spatiotemporal model

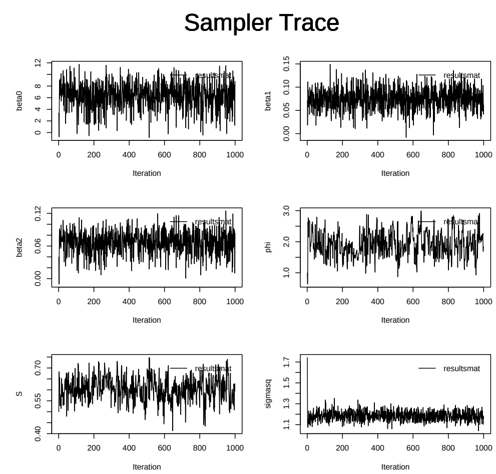
Number of	Sample Size					
CPUs	1000	2000	4000	6000	8000	10000
1	1.00	1.00	1.00	1.00	1.00	1.00
4	1.70	2.01	2.33	2.46	2.58	2.64
9	2.40	3.18	4.16	4.59	4.94	5.40
16	2.00	3.75	5.09	5.90	6.53	7.21
25	1.94	3.54	5.68	7.03	8.36	9.17

- Number of time points: 20
- Number of spatial points: 60, 120, 240, 360, 480, 600
- Results are for 10 MCMC iterations (about 60 likelihood evaluations)

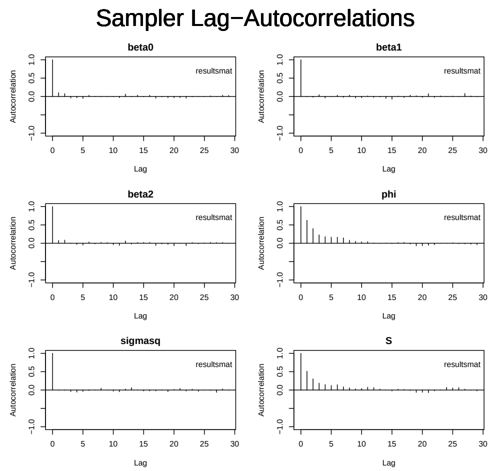
Speedup comparisons



Output for single year of SWE data



Output for single year of SWE data



Work in progress

- tuning to improve performance
  - clever distributed storage
  - multivariate slice-sampling
  - block size for PLAPACK
- porting to TeraGrid
  - NSF-funded Grid composed of extremely high-performance clusters at 8 partner sites
  - decomposing problems so as to minimize *inter*-cluster communication
- application to SWE data and radon data
- extension to nonstationary spatial covariance structures and other more complex models