Introduction to Spatial Data and Models

Sudipto Banerjee¹ and Andrew O. Finley²

¹ Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, U.S.A.
² Department of Forestry & Department of Geography, Michigan State University, Lansing Michigan, U.S.A.

October 15, 2012

Researchers in diverse areas such as climatology, ecology, environmental health, and real estate marketing are increasingly faced with the task of analyzing data that are:

- highly multivariate, with many important predictors and response variables,
- geographically referenced, and often presented as maps,
- temporally correlated, as in longitudinal or other time series structures.

⇒ motivates hierarchical modeling and data analysis for complex spatial (and spatiotemporal) data sets.
Type of spatial data

- **point-referenced data**, where $Y(s)$ is a random vector at a location $s \in \mathbb{R}^d$, where $s$ varies continuously over $D$, a fixed subset of $\mathbb{R}^d$ that contains an $r$-dimensional rectangle of positive volume;

- **areal data**, where $D$ is again a fixed subset (of regular or irregular shape), but now partitioned into a finite number of areal units with well-defined boundaries;

- **point pattern data**, where now $D$ is itself random; its index set gives the locations of random events that are the spatial point pattern. $Y(s)$ itself can simply equal 1 for all $s \in D$ (indicating occurrence of the event), or possibly give some additional covariate information (producing a marked point pattern process).

First step in analyzing data

- First Law of Geography: Mean + Error
- Mean: first-order behavior
First step in analyzing data
First Law of Geography: Mean + Error
Mean: first-order behavior
Error: second-order behavior (covariance function)

EDA tools examine both first and second order behavior
Preliminary displays: Simple locations to surface displays

First Law of Geography
\[ \text{data} = \text{mean} + \text{error} \]

Spatial surface observed at finite set of locations
\[ \mathcal{S} = \{ s_1, s_2, \ldots, s_n \} \]
Tessellate the spatial domain (usually with data locations as vertices)
Fit an interpolating polynomial:
\[ f(s) = \sum_i w_i(\mathcal{S}; s)f(s_i) \]

"Interpolate" by reading off \( f(s_0) \).
Issues:
- Sensitivity to tessellations
- Choices of multivariate interpolators
- Numerical error analysis
Scallops data: image and contour plots

Introduction to spatial data and models

Elements of point-level modelling

- Point-level modelling refers to modelling of spatial data collected at locations referenced by coordinates (e.g., lat-long, Easting-Northing).

**Fundamental concept:** Data from a spatial process \( \{Y(s) : s \in D\} \), where \( D \) is a fixed subset in Euclidean space.

**Example:** \( Y(s) \) is a pollutant level at site \( s \)

**Conceptually:** Pollutant level exists at all possible sites
Introduction to spatial data and models  Stationary Gaussian processes

Suppose our spatial process has a mean, $\mu(s) = E(Y(s))$, and that the variance of $Y(s)$ exists for all $s \in D$.

- **Strong stationarity**: If for any given set of sites, and any displacement $h$, the distribution of $(Y(s_1), \ldots, Y(s_n))$ is the same as $(Y(s_1 + h), \ldots, Y(s_n + h))$.
- **Weak stationarity**: Constant mean $\mu(s) = \mu$, and $\text{Cov}(Y(s), Y(s+h)) = C(h)$: the covariance depends only upon the displacement (or separation) vector.
- **Strong stationarity implies weak stationarity**
- **The process is Gaussian if $Y = (Y(s_1), \ldots, Y(s_n))$ has a multivariate normal distribution.**
Variograms

- Suppose we assume $E[Y(s + h) - Y(s)] = 0$ and define 
  $E[Y(s + h) - Y(s)]^2 = Var(Y(s + h) - Y(s)) = 2\gamma(h)$. 
This is sensible if the left hand side depends only upon $h$. 
Then we say the process is intrinsically stationary.

Note that intrinsic stationarity defines only the first and second 
moments of the differences $Y(s + h) - Y(s)$. It says nothing 
about the joint distribution of a collection of variables 
$Y(s_1), \ldots, Y(s_n)$, and thus provides no likelihood.

Intrinsic Stationarity and Ergodicity

- Relationship between $\gamma(h)$ and $C(h)$:
  
  $2\gamma(h) = Var(Y(s + h)) + Var(Y(s)) - 2Cov(Y(s + h), Y(s))$
  
  $= C(0) + C(0) - 2C(h)$
  
  $= 2[C(0) - C(h)]$.

- Easy to recover $\gamma$ from $C$. The converse needs the 
  additional assumption of ergodicity: $\lim_{|u| \to \infty} C(u) = 0$.
Isotropy

When $γ(h)$ or $C(h)$ depends upon the separation vector only through the distance $∥h∥$, we say that the process is isotropic. In that case, we write $γ(∥h∥)$ or $C(∥h∥)$. Otherwise we say that the process is anisotropic.

Isotropic processes are popular because of their simplicity, interpretability, and because a number of relatively simple parametric forms are available as candidates for $C$ (and $γ$). Denoting $∥h∥$ by $t$ for notational simplicity, the next two tables provide a few examples...

Some common isotropic variograms

<table>
<thead>
<tr>
<th>Model</th>
<th>Variogram, $γ(t)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>$γ(t) = \begin{cases} \tau^2 + \sigma^2 &amp; \text{if } t &gt; 0 \ 0 &amp; \text{otherwise} \end{cases}$</td>
</tr>
<tr>
<td>Spherical</td>
<td>$γ(t) = \begin{cases} \tau^2 + \sigma^2 + \tau^2 \left[2\phi t - \frac{1}{2}(\phi t)^3\right] &amp; \text{if } 0 &lt; t \leq 1/\phi \ \tau^2 + \sigma^2 (1 - \exp(-\phi t)) &amp; \text{if } t &gt; 0 \end{cases}$</td>
</tr>
<tr>
<td>Exponential</td>
<td>$γ(t) = \begin{cases} \tau^2 + \sigma^2(1 - \exp(-\phi t^2)) &amp; \text{if } t &gt; 0 \ 0 &amp; \text{otherwise} \end{cases}$</td>
</tr>
<tr>
<td>Exponential</td>
<td>$γ(t) = \begin{cases} \tau^2 + \sigma^2(1 - \exp(-</td>
</tr>
<tr>
<td>Matérn</td>
<td>$γ(t) = \begin{cases} \tau^2 + \sigma^2 [1 - (1 + \phi t)e^{-\phi t}] &amp; \text{if } t &gt; 0 \ 0 &amp; \text{o/w} \end{cases}$</td>
</tr>
</tbody>
</table>

Examples: Spherical Variogram

$γ(t) = \begin{cases} \tau^2 + \sigma^2 [\frac{2}{\phi t} - \frac{1}{2}(\phi t)^3] & \text{if } 0 < t \leq 1/\phi \\ 0 & \text{if } t = 0. \end{cases}$

Examples: Spherical Variogram

$γ(t) = \begin{cases} \tau^2 + \sigma^2 & \text{if } t > 1/\phi \\ 0 & \text{if } t < 1/\phi \\ \tau^2 & \text{if } t = 0. \end{cases}$

While $γ(0) = 0$ by definition, $γ(0^+) \equiv \lim_{t \to 0^+} γ(t) = \tau^2$; this quantity is the nugget.
**Introduction to spatial data and models**

**Isotropy**

- While \( \gamma(0) = 0 \) by definition, \( \gamma(0^+) \equiv \lim_{t \to 0^+} \gamma(t) = \tau^2 \); this quantity is the **nugget**.
- \( \lim_{t \to \infty} \gamma(t) = \tau^2 + \sigma^2 \); this asymptotic value of the semivariogram is called the **sill**. (The sill minus the nugget, \( \sigma^2 \) in this case, is called the **partial sill**.)

Some common isotropic covariograms

<table>
<thead>
<tr>
<th>Model</th>
<th>Covariance function, ( C(t) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>( C(t) ) does not exist</td>
</tr>
<tr>
<td>Spherical</td>
<td>( C(t) = \begin{cases} \sigma^2 \left( 1 - \frac{1}{2} \phi t + \frac{1}{2} \phi^2 t^2 \right) &amp; \text{if } t \geq 1/\phi \ \tau^2 + \sigma^2 &amp; \text{if } 0 &lt; t \leq 1/\phi \ \tau^2 + \sigma^2 &amp; \text{otherwise} \end{cases} )</td>
</tr>
<tr>
<td>Exponential</td>
<td>( C(t) = \sigma^2 \exp(-\phi t) )</td>
</tr>
<tr>
<td>Exponential, powered</td>
<td>( C(t) = \begin{cases} \sigma^2 \exp(-</td>
</tr>
<tr>
<td>Matérn at ( \nu = 3/2 )</td>
<td>( C(t) = \begin{cases} \sigma^2 \left( 1 + \phi t \right) \exp(-\phi t) &amp; \text{if } t &gt; 0 \ \tau^2 + \sigma^2 &amp; \text{otherwise} \end{cases} )</td>
</tr>
</tbody>
</table>

Notes on exponential model

\[
C(t) = \begin{cases} 
\tau^2 + \sigma^2 & \text{if } t = 0 \\
\sigma^2 \exp(-\phi t) & \text{if } t > 0 
\end{cases}
\]

- We define the **effective range**, \( t_0 \), as the distance at which this correlation has dropped to only 0.05. Setting \( \exp(-\phi t_0) \) equal to this value we obtain \( t_0 \approx 3/\phi \), since \( \log(0.95) \approx -3 \).
Notes on exponential model

\[
C(t) = \begin{cases} 
\sigma^2 + \sigma^2 & \text{if } t = 0 \\
\sigma^2 \exp(-\phi t) & \text{if } t > 0 
\end{cases}
\]

- We define the effective range, \( t_0 \), as the distance at which this correlation has dropped to only 0.05. Setting \( \exp(-\phi t_0) = 0.05 \), since \( \log(0.05) \approx -3 \).

- Finally, the form of \( C(t) \) shows why the nugget \( \sigma^2 \) is often viewed as a "nonspatial effect variance," and the partial sill \( (\sigma^2) \) is viewed as a "spatial effect variance."

The Matérn Correlation Function

- Much of statistical modelling is carried out through correlation functions rather than variograms

\[
C(t) = \begin{cases} 
\sigma^2 + \sigma^2 & \text{if } t = 0 \\
\frac{2^{\nu-1} (2\sqrt{\nu t\phi})^\nu K_\nu(2\sqrt{\nu t\phi})}{\pi^\nu \Gamma(\nu)} & \text{if } t > 0 
\end{cases}
\]

\( K_\nu \) is the modified Bessel function of order \( \nu \) (computationally tractable)

The Matérn Correlation Function

- Much of statistical modelling is carried out through correlation functions rather than variograms

\[
C(t) = \begin{cases} 
\sigma^2 + \sigma^2 & \text{if } t = 0 \\
\frac{2^{\nu-1} (2\sqrt{\nu t\phi})^\nu K_\nu(2\sqrt{\nu t\phi})}{\pi^\nu \Gamma(\nu)} & \text{if } t > 0 
\end{cases}
\]

\( K_\nu \) is the modified Bessel function of order \( \nu \)

- \( \nu \) is a smoothness parameter (a fractal) controlling process smoothness

How do we select a variogram? Can the data really distinguish between variograms?

Empirical Variogram:

\[
\gamma(t) = \frac{1}{2|N(t)|} \sum_{s_i, s_j \in N(t)} (Y(s_i) - Y(s_j))^2
\]

where \( N(t) \) is the number of points such that \( ||s_i - s_j|| = t \) and \( |N(t)| \) is the number of points in \( N(t) \).
How do we select a variogram? Can the data really distinguish between variograms?

Empirical Variogram:

\[ \gamma(t) = \frac{1}{2|N(t)|} \sum_{s_i, s_j \in N(t)} (Y(s_i) - Y(s_j))^2 \]

where \( N(t) \) is the number of points such that \( |s_i - s_j| = t \)
and \( |N(t)| \) is the number of points in \( N(t) \).

Grid up the \( t \) space into intervals \( I_1 = (0, t_1), I_2 = (t_1, t_2), \)
and so forth, up to \( I_K = (t_{K-1}, t_K) \). Representing \( t \) values in each interval by its midpoint, we define:

\[ N(t_k) = \{(s_i, s_j) : |s_i - s_j| \in I_k\}, k = 1, \ldots, K. \]
# Basics of Bayesian Inference

**Sudipto Banerjee** and Andrew O. Finley

1. Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, U.S.A.
2. Department of Forestry & Department of Geography, Michigan State University, Lansing Michigan, U.S.A.

October 15, 2012

## Bayesian principles

- **Classical statistics:** model parameters are fixed and unknown.
- A Bayesian thinks of parameters as random, and thus having distributions (just like the data). We can thus think about unknowns for which no reliable frequentist experiment exists, e.g. $\theta =$ proportion of US men with untreated prostate cancer.
- A Bayesian writes down a prior guess for parameter(s) $\theta$, say $p(\theta)$. He then combines this with the information provided by the observed data $y$ to obtain the posterior distribution of $\theta$, which we denote by $p(\theta | y)$.
- All statistical inferences (point and interval estimates, hypothesis tests) then follow from posterior summaries. For example, the posterior means/medians/modes offer point estimates of $\theta$, while the quantiles yield credible intervals.

## Principles of Bayesian Inference

- The key to Bayesian inference is “learning” or “updating” of prior beliefs. Thus, posterior information $\geq$ prior information.
- Is the classical approach wrong? That may be a controversial statement, but it certainly is fair to say that the classical approach is limited in scope.
- The Bayesian approach expands the class of models and easily handles:
  - repeated measures
  - unbalanced or missing data
  - nonhomogenous variances
  - multivariate data
  – and many other settings that are precluded (or much more complicated) in classical settings.

## Basics of Bayesian inference

- Calculations (numerical and algebraic) are usually required only up to a proportionaly constant. We, therefore, write the posterior as:
  \[
  p(\theta | y, \lambda) \propto p(\theta | \lambda) \times f(y | \theta).
  \]
- If $\lambda$ are known/fixed, then the above represents the desired posterior. If, however, $\lambda$ are unknown, we assign a prior, $p(\lambda)$, and seek:
  \[
  p(\theta, \lambda | y) \propto p(\lambda)p(\theta | \lambda)f(y | \theta).
  \]
- The proportionality constant does not depend upon $\theta$ or $\lambda$:
  \[
  \frac{1}{p(y)} = \int p(\lambda)p(\theta | \lambda)f(y | \theta)d\lambda d\theta
  \]
- The above represents a joint posterior from a hierarchical model. The marginal posterior distribution for $\theta$ is:
  \[
  p(\theta | y) = \int p(\lambda)p(\theta | \lambda)f(y | \theta)d\lambda.
  \]

### Bayesian inference: point estimation

- **Point estimation** is easy: simply choose an appropriate distribution summary: posterior mean, median or mode.
- **Mode** sometimes easy to compute (no integration, simply optimization), but often misrepresents the “middle” of the distribution – especially for one-tailed distributions.
- **Median**: probably the best compromise in being robust to tail behaviour although it may be awkward to compute as it needs to solve:
  \[
  \int_{-\infty}^{\theta_{med\text{ian}}} p(\theta | y)d\theta = \frac{1}{2}.
  \]
A simple example: Normal data and normal priors

Interpret: Posterior mean is a weighted mean of prior mean and data ... $N\left(\bar{y}, \sigma^2\tau^2\right)$.

Prior mean: $\mu, \tau^2$ are known.

Posterior distribution of $\theta$:

$$p(\theta | y) \propto N(\mu, \tau^2) \times N(\bar{y}, \sigma^2\tau^2)$$

$$= N\left(\theta \mid \frac{n\mu}{n\sigma^2 + n\tau^2}, \frac{n\sigma^2 + n\tau^2}{\sigma^2 + n\tau^2}\right)$$

Integration (quadrature) by “Monte Carlo integration”.

Then clearly $P(\theta \in (q_L, q_U) | y) = 1 - \alpha$.

Example: Consider a single data point $y$. A

Another simple example: The Beta-Binomial model

Example: Let $Y$ be the number of successes in $n$ independent trials.

$$P(Y = y | \theta) = f(y | \theta) = \frac{n!}{y!(n-y)!} \theta^y (1 - \theta)^{n-y}$$

Prior: $p(\theta) = Beta(\theta | a, b)$.

$$p(\theta) \propto \theta^{a-1} (1 - \theta)^{b-1}$$

Prior mean: $\mu = a/(a+b)$; Variance $ab/(a+b)^2(a+b+1)$

Posterior distribution of $\theta$:

$$p(\theta | y) = Beta(\theta | a + y, b + n - y)$$

We do this in two stages using composition sampling:

- First draw $\theta_2^j \sim p(\theta_2 | y)$, $j = 1, \ldots, M$.
- Next draw $\theta_1^j \sim p\left(\theta_1 | \theta_2^j, y\right)$.

This sampling scheme produces exact samples, $\{\theta_1^j, \theta_2^j\}_{j=1}^M$ from the posterior distribution $P(\theta_1, \theta_2 | y)$.

Gelfand and Smith (JASA, 1990) demonstrated automatic marginalization: $\{\theta_2^j\}_{j=1}^M$ are samples from $p(\theta_2 | y)$ and (of course) $\{\theta_1^j\}_{j=1}^M$ are samples from $p(\theta_1 | \theta_2, y)$.

In effect, composition sampling has performed the following “integation”:

$$p(\theta_1 | y) = \int p(\theta_1 | \theta_2, y) p(\theta_2 | y) d\theta_2.$$
Bayesian predictions

- Suppose we want to predict new observations, say \( \tilde{y} \), based upon the observed data \( y \). We will specify a joint probability model \( p(\tilde{y}, y | \theta) \), which defines the conditional predictive distribution:

\[
p(\tilde{y} | y, \theta) = \frac{p(\tilde{y}, y | \theta)}{p(y | \theta)}
\]

- Bayesian predictions follow from the posterior predictive distribution that averages out the \( \theta \) from the conditional predictive distribution with respect to the posterior:

\[
p(\tilde{y} | y) = \int p(\tilde{y} | y, \theta)p(\theta | y)d\theta.
\]

- This can be evaluated using composition sampling:
  - First obtain: \( \theta^{(j)} \sim p(\theta | y) \), \( j = 1, \ldots, M \)
  - For \( j = 1, \ldots, M \) sample \( \tilde{y}^{(j)} \sim p(\tilde{y} | y, \theta^{(j)}) \)
- The \( \{\tilde{y}^{(j)}\}_{j=1}^{M} \) are samples from the posterior predictive distribution \( p(\tilde{y} | y) \).

Some remarks on sampling-based inference

- Direct Monte Carlo: Some algorithms (e.g., composition sampling) can generate independent samples exactly from the posterior distribution. In these situations there are NO convergence problems or issues. Sampling is called exact.

- Markov Chain Monte Carlo (MCMC): In general, exact sampling may not be possible/feasible. MCMC is a far more versatile set of algorithms that can be invoked to fit more general models. Note: anywhere where direct Monte Carlo applies, MCMC will provide excellent results too.

- Convergence issues: There is no free lunch! The power of MCMC comes at a cost. The initial samples do not necessarily come from the desired posterior distribution. Rather, they need to converge to the true posterior distribution. Therefore, one needs to assess convergence, discard output before the convergence and retain only post-convergence samples. The time of convergence is called burn-in.

- Diagnosing convergence: Usually a few parallel chains are run from rather different starting points. The sample values are plotted (called trace-plots) for each of the chains. The time for the chains to “mix” together is taken as the time for convergence.

- Good news! All this is automated in WinBUGS. So, as users, we need to only configure how to specify good Bayesian models and implement them in WinBUGS.
Marginal and conditional distributions

- With a flat prior on $\beta$ we obtain, after some algebra, the conditional posterior distribution:
  \[ p(\beta | \sigma^2, y) \sim N(\beta | (X^TX)^{-1}X^Ty, \sigma^2(X^TX)^{-1}). \]

- The conditional posterior distribution of $\beta$ would have been the desired posterior distribution had $\sigma^2$ been known.

- Since that is not the case, we need to obtain the marginal posterior distribution by integrating out $\sigma^2$ as:
  \[ p(\beta | y) = \int p(\beta | \sigma^2, y)p(\sigma^2 | y)d\sigma^2 \]

- Can we solve this integration using composition sampling? YES: if we can generate samples from $p(\sigma^2 | y)$!

Ingredients of a linear model include an $n \times 1$ response vector $y = (y_1, \ldots, y_n)^T$ and an $n \times p$ design matrix (e.g. including regressors) $X = [x_1, \ldots, x_p]$, assumed to have been observed without error. The linear model:
\[ y = X\beta + \epsilon; \quad \epsilon \sim N(0, \sigma^2I) \]

- The linear model is the most fundamental of all serious statistical models encompassing:
  - ANOVA: $y$ is continuous, $x_i$'s are categorical
  - REGRESSION: $y$ is continuous, $x_i$'s are continuous
  - ANCOVA: $y$ is continuous, some $x_i$'s are continuous, some categorical.

- Unknown parameters include the regression parameters $\beta$ and the variance $\sigma^2$. We assume $X$ is observed without error and all inference is conditional on $X$.

For Bayesian analysis, we will need to specify priors for the unknown regression parameters $\beta$ and the variance $\sigma^2$.

- Consider independent flat priors on $\beta$ and $\log \sigma^2$:
  \[ p(\beta) \propto 1; \quad p(\log(\sigma^2)) \propto 1 \text{ or equivalently } p(\beta, \sigma^2) \propto \frac{1}{\sigma^2} \]

- None of the above two “distributions” are valid probabilities (they do not integrate to any finite number). So why is it that we are even discussing them?

- It turns out that even if the priors are improper (that’s what we call them), as long as the resulting posterior distributions are valid we can still conduct legitimate statistical inference on them.

So, we need to find the marginal posterior distribution of $\sigma^2$. With the choice of the flat prior we obtain:
\[
p(\sigma^2 | y) \propto \frac{1}{(\sigma^2)^{n-p+1}} \exp \left( -\frac{(n-p)s^2}{2\sigma^2} \right) = IG(\sigma^2 | n-p/2, (n-p)s^2/2),
\]

where $s^2 = \hat{\sigma}^2 = \frac{1}{n-p}y^T(I-P_X)y$.

- This is known as an inverted Gamma distribution (also called a scaled chi-square distribution)
  \[ IG(\sigma^2 | (n-p)/2, (n-p)s^2/2). \]

- In other words: $((n-p)s^2/\sigma^2 | y) \sim \chi^2_{n-p}$ (with $n-p$ degrees of freedom). A striking similarity with the classical result: The distribution of $\sigma^2$ is also characterized as $(n-p)s^2/\sigma^2$ following a chi-square distribution.
The Gibbs sampler

- Now we are ready to carry out composition sampling from $p(\beta, \sigma^2 \mid y)$ as follows:
  - Draw $M$ samples from $p(\sigma^2 \mid y)$:
    \[
    \sigma^2(j) \sim IG\left(\frac{n-p}{2}, \frac{(n-p)\sigma^2}{2}\right), \quad j = 1, \ldots, M
    \]
  - For $j = 1, \ldots, M$, draw from $p(\beta \mid \sigma^2(j), y)$:
    \[
    \beta(j) \sim N\left((X^T X)^{-1}X^T y, \sigma^2(j)(X^T X)^{-1}\right)
    \]
  - The resulting samples $\{\beta(j), \sigma^2(j)\}_{j=1}^M$ represent $M$ samples from $p(\beta, \sigma^2 \mid y)$.
  - $\{\beta(j)\}_{j=1}^M$ are samples from the marginal posterior distribution $p(\beta \mid y)$. This is a multivariate $t$ density:
    \[
    p(\beta \mid y) \propto \left(1 + \frac{(\beta - \hat{\beta})^T (X^T X)(\beta - \hat{\beta})}{(n-p)s^2}\right)^{-n/2}.
    \]

Bayesian predictions from the linear model

- Suppose we have observed the new predictors $\tilde{X}$, and we wish to predict the outcome $\tilde{y}$. We specify $p(\tilde{y}, y \mid \theta)$ to be a normal distribution:
  \[
  \begin{pmatrix}
  \tilde{y} \\
  y
  \end{pmatrix} \sim N\left(
  \begin{pmatrix}
  X \\
  \tilde{X}
  \end{pmatrix} \beta, \sigma^2 I \right)
  \]
- Note $p(\tilde{y} \mid y, \beta, \sigma^2) = p(\tilde{y} \mid \beta, \sigma^2) = N(\tilde{y} \mid \tilde{X}\beta, \sigma^2 I)$.
- The posterior predictive distribution:
  \[
  p(\tilde{y} \mid y) = \int p(\tilde{y} \mid y, \beta, \sigma^2)p(\beta, \sigma^2 \mid y)d\beta d\sigma^2
  \]
  \[
  = \int p(\tilde{y} \mid \beta, \sigma^2)p(\beta, \sigma^2 \mid y)d\beta d\sigma^2.
  \]
- By now we are comfortable evaluating such integrals:
  - First obtain: $\{\beta(j), \sigma^2(j)\} \sim p(\beta, \sigma^2 \mid y), \quad j = 1, \ldots, M$
  - Next draw: $\tilde{y}^{(j)} \sim N(X\beta(j), \sigma^2(j))$.

Composition sampling for linear regression

- The marginal distribution of each individual regression parameter $\beta_j$ is a non-central univariate $t_{n-p}$ distribution. In fact,
  \[
  \beta(j) - \hat{\beta}_j \sim \sqrt{\frac{s^2}{(X^T X)_{jj}}} \cdot t_{n-p}.
  \]
  The 95% credible intervals for each $\beta_j$ are constructed from the quantiles of the $t$-distribution. The credible intervals exactly coincide with the 95% classical confidence intervals, but the interpretation is direct: the probability of $\beta_j$ falling in that interval, given the observed data, is 0.95.
- Note: an intercept only linear model reduces to the simple univariate $N(y_j \mid \mu, \sigma^2/n)$ likelihood, for which the marginal posterior of $\mu$ is
  \[
  \frac{\mu - \bar{y}}{s/\sqrt{n}} \sim t_{n-1}.
  \]
The Gibbs sampler

In principle, the Gibbs sampler will work for extremely complex hierarchical models. The only issue is sampling from the full conditionals. They may not be amenable to easy sampling – when these are not in closed form. A more general and extremely powerful - and often easier to code - algorithm is the Metropolis-Hastings (MH) algorithm.

This algorithm also constructs a Markov Chain, but does not necessarily care about full conditionals.

The Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm: Start with an initial value for \( \theta = \theta^{(0)} \). Select a candidate or proposal distribution from which to propose a value of \( \theta \) at the \( j \)-th iteration: \( q(\theta^{(j-1)}, \nu) \). For example, \( q(\theta^{(j-1)}, \nu) = N(\theta^{(j-1)}, \nu) \) with \( \nu \) fixed.

Compute

\[
R = \frac{p(\theta^{*} | Y)p(q(\theta^{(j-1)} | \theta^{*}, \nu))}{p(q(\theta^{(j-1)} | Y)p(\theta^{(j-1)}, \nu))}
\]

If \( R \geq 1 \) then set \( \theta^{(j)} = \theta^{*} \). If \( R \leq 1 \) then draw \( U \sim (0, 1) \). If \( U \leq R \) then \( \theta^{(j)} = \theta^{*} \). Otherwise, \( \theta^{(j)} = \theta^{(j-1)} \).

Repeat for \( j = 1, \ldots, M \). This yields \( \theta^{(1)}, \ldots, \theta^{(M)} \), which, after a burn-in period, will be samples from the true posterior distribution. It is important to monitor the acceptance ratio \( R \) of the sampler through the iterations. Rough recommendations: for vector updates \( R \approx 20\% \), for scalar updates \( R \approx 40\% \). This can be controlled by “tuning” \( \nu \).

Popular approach: Embed Metropolis steps within Gibbs to draw from full conditionals that are not accessible to directly generate from.

Example

For the linear model, our parameters are \((\beta, \Sigma)\). We write \( \theta \equiv (\beta, \log(\sigma^2)) \) and, at the \( j \)-th iteration, propose \( \theta^{*} \sim N(\theta^{(j-1)}, \Sigma) \). The log transformation on \( \sigma^2 \) ensures that all components of \( \theta \) have support on the entire real line and can have meaningful proposed values from the multivariate normal.

But we need to transform our prior to \( p(\beta, \log(\sigma^2)) \).

Let \( z = \log(\sigma^2) \) and assume \( p(\theta, z) = p(\beta)p(z) \). Let us derive \( p(z) \). REMEMBER, we need to adjust the position. Then \( p(z) = p(\beta)p(z|\theta) = \bar{p}(z|\theta) \). The position here is \( z^\ast = \theta^\ast \).

Let \( p(\theta) = 1 \) and \( p(z^2) = \Gamma(\nu^2|\nu, 1) \). Then log posterior is:

\[
-\frac{\nu}{2} \log(\sigma^2) + 1 + \frac{1}{2} \sum_{i=1}^{n} (Y_i - X_i \beta)^2 - \frac{1}{2\nu} \log(\sigma^2).
\]

A symmetric proposal distribution, say \( q(z|\theta^{(j-1)}, \Sigma) = N(\theta^{(j-1)}, \Sigma) \), cancels out in \( R \). In practice it is better to compute \( \log(q(z|\theta^{(j-1)}, \Sigma)) \) and, at the \( j \)-th iteration, propose \( \theta^{(j)} \sim N(\theta^{(j-1)}, \Sigma) \). For the proposed, \( p(q(z|\theta^{(j-1)}, \Sigma)) \) is a 4d variance-covariance matrix, and \( d = \dim(\theta) = \nu + 1 \).

If \( \log U \geq 0 \) then set \( \theta^{(j)} = \theta^{(j-1)} \). If \( \log U \leq 0 \) then the distribution \( U \sim (0, 1) \). If \( U \leq r \) then set \( \theta^{(j)} = \theta^{(j)} \). Otherwise, \( \theta^{(j)} = \theta^{(j-1)} \).

Repeat the above procedure for \( j = 1, \ldots, M \) to obtain samples \( \theta^{(1)}, \ldots, \theta^{(M)} \).
Spatial Autoregressive Models

Sudipto Banerjee¹ and Andrew O. Finley²

¹ Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, U.S.A.
² Department of Forestry & Department of Geography, Michigan State University, Lansing Michigan, U.S.A.

October 15, 2012

---

Areal unit data

Maps of raw standard mortality ratios (SMR) of lung and esophagus cancer between 1991 and 1998 in Minnesota counties

---

Key Issues

- Is there spatial pattern? Spatial pattern implies that observations from units closer to each other are more similar than those recorded in units farther away.
- Do we want to smooth the data? Perhaps to adjust for low population sizes (or sample sizes) in certain units? How much do we want to smooth?
- Inference for new areal units? Is prediction meaningful here? If we modify the areal units to new units (e.g. from zip codes to county values), what can we say about the new counts we expect for the latter given those for the former? This is the Modifiable Areal Unit Problem (MAUP) or Misalignment.

---

Proximity matrices

$W$, entries $w_{ij}$, ($w_{ii} = 0$); choices for $w_{ij}$:

- $w_{ij} = 1$ if $i, j$ share a common boundary (possibly a common vertex)
- $w_{ij}$ is an inverse distance between units
- $w_{ij} = 1$ if distance between units is $\leq K$
- $w_{ij} = 1$ for m nearest neighbors.

$\tilde{W}$ need not be symmetric.

$\tilde{W}$: standardize row $i$ by $w_{i+} = \sum_j w_{ij}$ (row stochastic but need not be symmetric).

$W$ elements often called “weights”; nicer interpretation?
Proximity matrices

- Note that proximity matrices are user-defined.
- We can define distance intervals, \((0, d_1], (d_1, d_2], \) and so on.
  - First order neighbours: all units within distance \(d_1\).
  - First order proximity matrix \(W^{(1)}\): Analogous to \(W\), \(w^{(1)}_{ij} = 1\) if \(i\) and \(j\) are first order neighbors; 0 otherwise.
  - Second order neighbours: all units within distance \(d_2\), but separated by more than \(d_1\).
  - Second order proximity matrix \(W^{(2)}\); \(w^{(2)}_{ij} = 1\) if \(i\) and \(j\) are second order neighbors; 0 otherwise.
  - And so on...

The areal correlogram is a useful tool to study spatial association with areal data.

Working with \(I\), we can replace \(w_{ij}\) with \(w^{(1)}_{ij}\) taken from \(W^{(1)}\) and compute \(I^{(1)}\).

Next replace \(w_{ij}\) with \(w^{(2)}_{ij}\) taken from \(W^{(2)}\) and compute \(I^{(2)}\), etc.

Plot \(I^{(r)}\) vs. \(r\)

- If there is a spatial pattern, we expect \(I^{(r)}\) to decline in \(r\) initially and then vary about 0.

The areal correlogram is a useful tool to study spatial association with areal data.

To smooth \(Y_i\), replace with \(\bar{Y}_i = \sum w_{ij}Y_j\). Note: \(K\)-nearest neighbours (KNN) regression falls within this framework.

More generally,

\[(1 - \alpha)Y_i + \alpha\bar{Y}_i\]

Linear (convex) combination, shrinkage

Model-based smoothing, e.g.,

\[E(Y_i | \{Y_j, j = 1, 2, ..., n\})\]

- There are analogues for areal data of the empirical correlation function and the variogram.
- Moran’s \(I\): analogue of lagged autocorrelation

\[I = \frac{n}{\sum w_{ij}} \sum (Y_i - \bar{Y})(Y_j - \bar{Y}) \]

\(I\) is not supported on \([-1, 1]\).

- Geary’s \(C\): analogue of Durbin-Watson statistic

\[C = \frac{(n - 1) \sum w_{ij}(Y_i - Y_j)^2}{\sum w_{ij} \sum (Y_i - Y_j)^2}\]

Both are asymptotically normal if \(Y_i\) are i.i.d., the first with mean \(-1/(n-1)\) and the second with mean 1.

Significance testing using a Monte Carlo test, permutation invariance

Spatial smoothers

- To smooth \(Y_i\), replace with \(\bar{Y}_i = \sum w_{ij}Y_j\). Note: \(K\)-nearest neighbours (KNN) regression falls within this framework.

More generally,

\[(1 - \alpha)Y_i + \alpha\bar{Y}_i\]

Linear (convex) combination, shrinkage

Model-based smoothing, e.g.,

\[E(Y_i | \{Y_j, j = 1, 2, ..., n\})\]

Markov Random Fields

- First, consider \(Y = (y_1, y_2, ..., y_n)\) and consider the set \(\{p(y_i | y_j, j \neq i)\}\)
- We know \(p(y_1, y_2, ..., y_n)\) determines \(\{p(y_i | y_j, j \neq i)\}\) (full conditional distributions)
- \(??\) Does \(\{p(y_i | y_j, j \neq i)\}\) determine \(p(y_1, y_2, ..., y_n)\)? If so, we call the joint distribution a Markov Random Field.

In general we cannot write down an arbitrary set of conditionals and assert that they determine the joint distribution. Example:

\[Y_i | Y_j \sim N(\alpha_0 + \alpha_1Y_j, \sigma_1^2) \]

\[Y_2 | Y_1 \sim N(\beta_0 + \beta_1Y_1, \sigma_2^2)\]

The first equation implies that \(E[Y_i] = \alpha_0 + \alpha_1E[Y_j]\), i.e., \(E[Y_i]\) is linear in \(E[Y_j]\). The second equation implies that \(E[Y_2] = \beta_0 + \beta_1E[Y_1]\), i.e., \(E[Y_2]\) is linear in \(E[Y_1]\). Clearly this isn’t true in general. Hence no joint distribution.
Also \( p(y_1, \ldots, y_n) \) may be improper even if all the full conditionals are proper.
\[
p(y_1, y_2) \propto \exp\left( (y_1 - y_2)^2 \right)
\]
But \( p(Y_2 | Y_1) \propto N(Y_2 | Y_1, 2) \) and \( p(Y_1 | Y_2) \propto N(Y_2, 1) \), yet the joint distribution is improper.

- Compatibility: Brook’s Lemma. Let \( y_0 = (y_{10}, \ldots, y_{n0}) \) be any fixed point in the support of \( p(\cdot) \).

\[
p(y_1, \ldots, y_n) = \frac{p(y_1 | y_2, \ldots, y_n)}{p(y_1 | y_2, \ldots, y_{n0})} \frac{p(y_2 | y_3, \ldots, y_n)}{p(y_2 | y_3, \ldots, y_{n0})} \cdots \frac{p(y_n | y_1, \ldots, y_{n-1})}{p(y_n | y_1, \ldots, y_{n-1}, y_{n0})}.
\]

If LHS is proper, the fact that it integrates to 1 determines the normalizing constant!

Suppose we want:
\[
p(y_i | y_j, j \neq i) = p(y_i | y_j \in \partial i)
\]

When does the set \{ \( p(y_i | y_j \in \partial i) \) \} uniquely determine \( p(y_1, y_2, \ldots, y_n) \)?

To answer this question, we need the following important concepts:

- Clique: A clique is a set of cells such that each element is a neighbor of every other element. We use notation \( i \sim j \) if \( i \) is a neighbor of \( j \) and \( j \) is a neighbor of \( i \).
- Potential: A potential of order \( k \) is a function of \( k \) arguments that is exchangeable in these arguments. The arguments of the potential would be the values taken by variables associated with the cells for a clique of size \( k \).

Gibbs distribution: \( p(y_1, \ldots, y_n) \) is a Gibbs distribution if it is a function of the \( y_i \)'s only through potentials on cliques:
\[
p(y_1, \ldots, y_n) \propto \exp\left( -\gamma \sum \phi(k)(y_{a1}, \ldots, y_{ak}) \right),
\]
where \( \phi(k) \) is a potential of order \( k \), \( M_k \) is the set of all cliques of size \( k \) and is indexed by \( a \), and \( \gamma > 0 \) is a scale parameter.

Hammersley-Clifford Theorem: If we have a Markov Random Field (i.e., \{ \( p(y_i | y_j \in \partial i) \) \}) uniquely determine \( p(y_1, y_2, \ldots, y_n) \), then the latter is a Gibbs distribution.

Geman and Geman (1984) result: If we have a joint Gibbs distribution, then we have a Markov Random Field.

Returning to \( \mathcal{W} \), let \( b_{ij} = w_{ij}/w_{i+} \) and \( \tau_i^2 = \tau^2/w_{i+} \), so
\[
p(y_1, y_2, \ldots, y_n) \propto \exp\left\{ -\frac{1}{2\tau^2} \mathbf{y}^T (D_w - W) \mathbf{y} \right\}
\]
where \( D_w \) is diagonal with \( D_{wi} = w_{i+} \) and thus
\[
p(y_1, y_2, \ldots, y_n) \propto \exp\left\{ -\frac{1}{2\tau^2} \sum_{\mathbf{r} \neq \mathbf{y}} w_{ij}(y_i - y_j)^2 \right\}
\]
Caution: \( (D_w - W)\mathbf{1} = 0 \). Intrinsic autoregressive (IAR) model; improper, so requires a constraint (e.g., \( \sum_i y_i = 0 \))

Not a valid data model, but only as a random effects model!
SAR models
Simultaneous Auto-Regressive (SAR) models
We may write the CAR model as:
\[ y = By + \epsilon \Rightarrow (I-B)y = \epsilon; \]
where \( G \) is the number of “islands” in the map. In fact, \( n-G \) is the rank of \( D_w - W \).

The impropriety can be remedied in an obvious way. Redefine the CAR as:
\[ p(y_1, y_2, \ldots, y_n) \propto |D_w - \rho W|^{1/2} \exp\{-\frac{1}{2\tau^2} y'(D_w - \rho W)y\}, \]
where \( \rho \) is chosen to make \( D_w - \rho W \) non-singular. This is guaranteed if \( \rho \in (1/\lambda_{(1)}, 1) \), where \( \lambda_{(1)} \) is the minimum eigenvalue of \( D^{-1/2}W D^{-1/2} \). In practice, the bound \( \rho \in (0, 1) \) is often preferred.

Example of a hierarchical model with CAR effects.
Consider the areal data disease mapping model:
\[ Y_i \mid \mu_i \sim \text{Po}(E_i \mu^i); \quad \text{where} \]
\[ Y_i = \text{observed disease count}, \]
\[ E_i = \text{expected count (known)}, \quad \text{and} \]
\[ \mu_i = x_i' \beta + \phi_i; \quad \text{the } x_i \text{ are explanatory variables} \]

The \( \phi_i \) capture regional clustering via a conditionally autoregressive (CAR) prior,
\[ \phi_i \mid \phi_{j \neq i} \sim N\left(\tilde{\phi}_i, \frac{\tau^2}{m_i}\right); \quad \text{where} \]
\[ \tilde{\phi}_i = \frac{1}{m_i} \sum_{j \neq i} \phi_j; \]
\( \tilde{\phi}_i \) is the set of “neighbours” of region \( i \), and \( m_i \) is the number of these neighbours.

To \( \rho \) or not to \( \rho \)?
- Advantages:
  - makes distribution proper
  - adds parametric flexibility
  - \( \rho = 0 \) interpretable as independence

- Disadvantages:
  - why should we expect \( y_i \) to be a proportion of average of neighbors - sensible spatial interpretation?
  - calibration of \( \rho \) as a correlation, e.g.,
    \[ \rho = 0.80 \text{ yields } 0.1 \leq I \leq 0.15, \]
    \[ \rho = 0.90 \text{ yields } 0.2 \leq I \leq 0.25, \]
    \[ \rho = 0.99 \text{ yields } I \leq 0.5 \]
  - So, used with random effects, scope of spatial pattern may be limited

Simultaneous Auto-Regressive (SAR) models
- We may write the CAR model as:
  \[ y = By + \epsilon \Rightarrow (I-B)y = \epsilon; \]
Since \( y \sim N(0, (I-B)^{-1}D) \), we have
\[ \epsilon \sim N(0, D(I-B)^{-1}). \]
- Instead of letting \( y \) induce the distribution of \( \epsilon \), let \( \epsilon \) induce a distribution for \( y \). Letting \( \epsilon \sim N(0, D) \), where \( D \) is diagonal, \( D_{ii} = \sigma_i^2 \) and let:
  \[ y_i = \sum_{j=1}^{n} b_{ij} y_j + \epsilon_i. \]
Assuming \((I-B)^{-1}\) exists, we obtain:
\[ y \sim N\left(0, (I-B)^{-1}D(I-B)^{-1}\right). \]

Comments on SAR models
- Often we take \( B = \rho W \). If \( \rho \in (1/\lambda_{(1)}, 1/\lambda_{(n)}) \), where \( \lambda_{(1)} \) and \( \lambda_{(n)} \) are the minimum and maximum eigenvalues of \( W \). This ensures \((I-\rho W)^{-1}\) exists.
- Alternatively, we can replace \( W \) with \( \tilde{W} = \{w_{ij}/w_{ii}\} \) where \( w_{ii} \) is the sum of the elements in the \( i \)-th row of \( W \). Then \(|\rho| < 1 \) ensures existence of \((I-\rho \tilde{W})^{-1}\).
- Often SAR models are also applied to point-referenced data where \( \tilde{W} \) is taken to be the inter-point distance.
SAR models

Two variants:

- The SAR “lag model”:

  \[ y = By + X\beta + \epsilon. \]

- The SAR “residual” or “error model”:

  \[ (I - B)(y - X\beta) = \epsilon; \Rightarrow y = By + (I - B)X\beta + \epsilon. \]

SAR models are well suited to maximum likelihood estimation but not at all for MCMC fitting of Bayesian models. Because it is difficult to introduce SAR random effects (in the CAR framework this is easy because of the hierarchical conditional representation).
Hierarchical Modelling for Univariate Spatial Data

Sudipto Banerjee\(^1\) and Andrew O. Finley\(^2\)

\(^1\) Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, U.S.A.
\(^2\) Department of Forestry & Department of Geography, Michigan State University, Lansing Michigan, U.S.A.

October 15, 2012

Algorithmic Modelling

- Spatial surface observed at finite set of locations \( \mathcal{S} = \{s_1, s_2, \ldots, s_n\} \)
- Tessellate the spatial domain (usually with data locations as vertices)
- Fit an interpolating polynomial:
  \[
  f(s) = \sum_i w_i(\mathcal{S}; s) f(s_i)
  \]
- “Interpolate” by reading off \( f(s_0) \).
- Issues:
  - Sensitivity to tessellations
  - Choices of multivariate interpolators
  - Numerical error analysis

Simple linear model

\[
Y(s) = \mu(s) + \epsilon(s),
\]

- Response: \( Y(s) \) at location \( s \)
- Mean: \( \mu = X(s)\beta \)
- Error: \( \epsilon(s) \overset{\text{iid}}{\sim} N(0, \tau^2) \)
Simple linear model

\[ Y(s) = \mu(s) + \epsilon(s). \]

Assumptions regarding \( \epsilon(s) \):
- \( \epsilon(s)^{\text{iid}} \sim N(0, \tau^2) \)
- \( \epsilon(s_i) \) and \( \epsilon(s_j) \) are uncorrelated for all \( i \neq j \)

Spatial Gaussian processes (GP):
- Say \( w(s) \sim GP(0, \sigma^2\rho(\cdot)) \) and
  \[ \text{Cov}(w(s_1), w(s_2)) = \sigma^2 \rho(\phi; \|s_1 - s_2\|) \]
- Let \( w = [w(s_i)]_{i=1}^D \), then
  \[ w \sim N(0, \sigma^2 R(\phi)), \text{ where } R(\phi) = [\rho(\phi; \|s_i - s_j\|)]_{i,j=1}^D \]

Realization of a Gaussian process:
- Changing \( \phi \) and holding \( \sigma^2 = 1 \):
  \[ w \sim N(0, \sigma^2 R(\phi)), \text{ where } R(\phi) = [\rho(\phi; \|s_i - s_j\|)]_{i,j=1}^D \]
- Correlation model for \( R(\phi) \):
  e.g., exponential decay
  \[ \rho(\phi; t) = \exp(-\phi t) \text{ if } t > 0. \]
- Effective range,
  \[ t_0 = \ln(0.05)/\phi \approx 3/\phi \]
**Univariate spatial models**

Sources of variation

- **Univariate spatial regression**

\[ w \sim N(0, \sigma_w^2 R(\phi)) \] defines complex spatial dependence structures.

E.g., anisotropic Matérn correlation function:

\[ r(s_i, s_j) = \left( \frac{d_i}{\nu} \right)^\phi \nu - 1 \]

\[ d_i = \sqrt{(s_i - s_j)^T \Sigma^{-1} (s_i - s_j)} \]

where \( \Sigma = \sigma^2 R(\phi) + \tau^2 I \). Thus, \( \phi = (\nu, \phi, \Lambda) \).

**Hierarchical modelling**

- **First stage:**

\[ y|\beta, w, \tau^2 \sim \prod_{i=1}^n N(Y(s_i) | X^T(s_i)\beta + w(s_i), \tau^2) \]

- **Second stage:**

\[ w|\sigma^2, \phi \sim N(0, \sigma^2 R(\phi)) \]

- **Third stage:** Priors on \( \Omega = (\beta, \tau^2, \sigma^2, \phi) \)

**Simple linear model + random spatial effects**

\[ Y(s) = \mu(s) + w(s) + \epsilon(s), \]

- **Response:** \( Y(s) \) at some site
- **Mean:** \( \mu = X^T(s)\beta \)
- **Spatial random effects:** \( w(s) \sim GP(0, \sigma^2 \rho(\phi; ||s_1 - s_2||)) \)
- **Non-spatial variance:** \( \epsilon(s) \iid N(0, \tau^2) \)
Univariate spatial models

Hierarchical modelling

- First stage:
  \[ y|\beta, w, \tau^2 \sim \prod_{i=1}^{n} N(Y(s_i)|X^T(s_i)\beta + w(s_i), \tau^2) \]

- Second stage:
  \[ w|\sigma^2, \phi \sim N(0, \sigma^2 R(\phi)) \]

- Third stage: Priors on \( \Omega = (\beta, \tau^2, \sigma^2, \phi) \)
  - Marginalized likelihood:
    \[ y|\Omega \sim N(X\beta, \sigma^2 R(\phi) + \tau^2 I) \]

- Note: Spatial process parametrizes \( \Sigma \):
  \[ y = X\beta + \epsilon, \epsilon \sim N(0, \Sigma), \Sigma = \sigma^2 R(\phi) + \tau^2 I \]

Bayesian Computations

- Choice: Fit \( y|\Omega \times [\Omega] \) or \( |y|\beta, w, \tau^2 \times w|\sigma^2, \phi \times [\Omega] \).

- Conditional model:
  - conjugate full conditionals for \( \sigma^2, \tau^2 \) and \( w \) – easier to program.

- Marginalized model:
  - need Metropolis or Slice sampling for \( \sigma^2, \tau^2 \) and \( \phi \). Harder to program.
  - But, reduced parameter space \( \Rightarrow \) faster convergence
  - \( \sigma^2 R(\phi) + \tau^2 I \) is more stable than \( \sigma^2 R(\phi) \).

- But what about \( R^{-1}(\phi) ?? \) EXPENSIVE!

Where are the \( w \)'s?

- Interest often lies in the spatial surface \( w|y \).
Where are the \( w \)'s?

- Interest often lies in the spatial surface \( w \)\( y \).
- They are recovered from
  \[
  [w|y, X] = \int [w|\Omega, y, X] \times [\Omega|y, X] d\Omega
  \]
  using posterior samples:
  - Obtain \( \Omega^{(1)}, \ldots, \Omega^{(n)} \sim [\Omega|y, X] \)
  - For each \( \Omega^{(i)} \), draw \( w^{(i)} \sim [w|\Omega^{(i)}, y, X] \)

**NOTE:** With Gaussian likelihoods \( [w|\Omega, y, X] \) is also Gaussian. With other likelihoods this may not be easy and often the conditional updating scheme is preferred.

Residual plot: \( [w(s)|y] \)
Univariate spatial models

- Often we need to predict \( Y(s) \) at a new set of locations \( \{\tilde{s}_0, \ldots, \tilde{s}_m\} \) with associated predictor matrix \( \tilde{X} \).

- Sample from predictive distribution:
  \[
  \tilde{y} | y, X, \tilde{X} = \int \tilde{y} \Omega | y, X, \tilde{X} \, d\Omega = \int \tilde{y} | y, \Omega, X, \tilde{X} \times \Omega | y, X \, d\Omega,
  \]
  \( \tilde{y} \Omega, X, \tilde{X} \) is multivariate normal. Sampling scheme:
  - Obtain \( \Omega^{(1)}, \ldots, \Omega^{(G)} \sim \{\Omega | y, X\} \)
  - For each \( \Omega^{(g)} \), draw \( \tilde{y}^{(g)} | y, \Omega^{(g)}, X, \tilde{X} \).

Colorado data illustration

- Modelling temperature: 507 locations in Colorado.
- Simple spatial regression model:
  \[
  Y(s) = X^T(s)\beta + w(s) + \epsilon(s)
  \]

- \( w(s) \sim GP(0, \sigma^2 \rho(\cdot; \phi, \nu)) \)
- \( \epsilon(s) \sim iid N(0, \tau^2) \)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean (2.5%, 97.5%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2.827 (2.131, 3.866)</td>
</tr>
<tr>
<td>[Elevation]</td>
<td>-0.426 (-0.527, -0.333)</td>
</tr>
<tr>
<td>Precipitation</td>
<td>0.927 (0.863, 0.972)</td>
</tr>
<tr>
<td>( \sigma^2 )</td>
<td>0.134 (0.051, 1.245)</td>
</tr>
<tr>
<td>( \phi )</td>
<td>7.39E-3 (4.71E-3, 51.21E-3)</td>
</tr>
<tr>
<td>Range</td>
<td>278.2 (38.8, 476.3)</td>
</tr>
<tr>
<td>( \tau^2 )</td>
<td>0.051 (0.022, 0.092)</td>
</tr>
</tbody>
</table>

Temperature residual map

Elevation map

Residual map with elev. as covariate
Hierarchical Modelling for non-Gaussian Spatial Data

Sudipto Banerjee\(^1\) and Andrew O. Finley\(^2\)

\(^1\) Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, U.S.A.
\(^2\) Department of Forestry & Department of Geography, Michigan State University, Lansing, Michigan, U.S.A.

October 15, 2012

Spatial Generalized Linear Models

- Often data sets preclude Gaussian modelling: \(Y(s)\) may not even be continuous
- Example: \(Y(s)\) is a binary or count variable
  - species presence or absence at location \(s\)
  - species abundance from count at location \(s\)
  - continuous forest variable is high or low at location \(s\)
- Replace Gaussian likelihood by exponential family member
  - Diggle Tawn and Moyeed (1998)

Hierarchical Modelling for non-Gaussian Spatial Data

Sudipto Banerjee\(^1\) and Andrew O. Finley\(^2\)

\(^1\) Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, U.S.A.
\(^2\) Department of Forestry & Department of Geography, Michigan State University, Lansing, Michigan, U.S.A.

October 15, 2012

Spatial Generalized Linear Models

- First stage: \(Y(s)\) are conditionally independent given \(\beta\) and \(w(s)\), so \(f(y(s)|\beta, w(s), \gamma)\) equals
  \[
  h(y(s), \gamma) \exp \left( \gamma y(s) - \psi(y(s)) \right)
  \]
  where \(\gamma(E(Y(s))) = \eta(s) = x^T(s)\beta + w(s)\) (canonical link function) and \(\gamma\) is a dispersion parameter.
- Second stage: Model \(w(s)\) as a Gaussian process,
  \[
  w \sim N(0, \sigma^2 R(\phi))
  \]
- Third stage: Priors and hyperpriors.
- No process for \(Y(s)\), only a valid joint distribution
- Not sensible to add a pure error term \(\epsilon(s)\)

Hierarchical Modelling for non-Gaussian Spatial Data

Sudipto Banerjee\(^1\) and Andrew O. Finley\(^2\)

\(^1\) Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, U.S.A.
\(^2\) Department of Forestry & Department of Geography, Michigan State University, Lansing, Michigan, U.S.A.

October 15, 2012

Spatial Generalized Linear Models

- We are modelling with spatial random effects
- Introducing these in the transformed mean encourages means of spatial variables at proximate locations to be close to each other
- Marginal spatial dependence is induced between, say, \(Y(s)\) and \(Y(s')\), but observed \(Y(s)\) and \(Y(s')\) need not be close to each other
- Second stage spatial modelling is attractive for spatial explanation in the mean
- First stage spatial modelling more appropriate to encourage proximate observations to be close.

Binary spatial regression: forest/non-forest

We illustrate a non-Gaussian model for point-referenced spatial data:
- Objective is to make pixel-level prediction of forest/non-forest across the domain.

Illustration from:

Binary spatial regression: forest/non-forest

We illustrate a non-Gaussian model for point-referenced spatial data:

- Objective is to make pixel-level prediction of forest/non-forest across the domain.
- Data: Observations are from 500 georeferenced USDA Forest Service Forest Inventory and Analysis (FIA) inventory plots within a 32 km radius circle in MN, USA.
- The response $Y(s)$ is a binary variable, with

$$Y(s) = \begin{cases} 1 & \text{if inventory plot is forested} \\ 0 & \text{if inventory plot is not forested} \end{cases}$$

Parameter estimates (posterior medians and upper and lower data:

- Assume vague flat prior for $\beta$, a Uniform(3/32, 3/0.5) prior for $\phi$, and an inverse-Gamma(2,.) prior for $\sigma^2$.
- Parameters updated with Metropolis algorithm using target log density:

$$\ln p(Y | \Omega) \propto -\left(\phi + 1 + n \right) \ln (\sigma^2) - \frac{\phi}{2} \ln (|R(\phi)|) - \frac{1}{2\sigma^2} w^T R(\phi)^{-1} w$$

We fit a generalized linear model where

$$Y(s) \sim \text{Bernoulli}(p(s)), \quad \text{logit}(p(s)) = x^T(s)\beta + w(s).$$

Assume vague flat for $\beta$, a Uniform(3/32, 3/0.5) prior for $\phi$, and an inverse-Gamma(2,.) prior for $\sigma^2$.

The model can be written:

$$Y(s) = \begin{cases} 1 & \text{if inventory plot is forested} \\ 0 & \text{if inventory plot is not forested} \end{cases}$$

- Observed covariates include the coexisting pixel values for 3 dates of 30 x 30 resolution Landsat imagery.

Posterior parameter estimates

Parameter estimates (posterior medians and upper and lower 2.5 percentiles):

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Estimates: 50% (2.5%, 97.5%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept ((\beta_0))</td>
<td>82.39 (74.36, 120.46)</td>
</tr>
<tr>
<td>AprilTC1 ((\beta_1))</td>
<td>-0.27 (-0.45, -0.11)</td>
</tr>
<tr>
<td>AprilTC2 ((\beta_2))</td>
<td>0.17 (0.07, 0.29)</td>
</tr>
<tr>
<td>AprilTC3 ((\beta_3))</td>
<td>-0.24 (-0.43, -0.08)</td>
</tr>
<tr>
<td>JulyTC1 ((\beta_4))</td>
<td>-0.04 (-0.25, 0.17)</td>
</tr>
<tr>
<td>JulyTC2 ((\beta_5))</td>
<td>0.09 (-0.01, 0.19)</td>
</tr>
<tr>
<td>JulyTC3 ((\beta_6))</td>
<td>0.01 (-0.15, 0.16)</td>
</tr>
<tr>
<td>OctTC1 ((\beta_7))</td>
<td>-0.43 (-0.68, -0.22)</td>
</tr>
<tr>
<td>OctTC2 ((\beta_8))</td>
<td>-0.03 (-0.19, 0.14)</td>
</tr>
<tr>
<td>OctTC3 ((\beta_9))</td>
<td>-0.26 (-0.46, -0.07)</td>
</tr>
<tr>
<td>0^2</td>
<td>1.358 (0.39, 2.44)</td>
</tr>
<tr>
<td>log(0.05)/(\phi)</td>
<td>0.00182 (0.00065, 0.00322)</td>
</tr>
<tr>
<td></td>
<td>1644.19 (932.33, 4606.7)</td>
</tr>
</tbody>
</table>

Covariates and proximity to observed FIA plot will contribute to increase precision of prediction.
Median of posterior predictive distributions

97.5%-2.5% range of posterior predictive distributions

CDF of holdout area’s posterior predictive distributions

Classification of 15 20 × 20 pixel areas (based on visual inspection of imagery) into non-forest (●), moderately forest (○), and forest (no marker).
Point-referenced spatial data often come as multivariate measurements at each location.

Examples:
- Environmental monitoring: stations yield measurements on ozone, NO, CO, and PM$_{2.5}$.
- Community ecology: assemblages of plant species due to water availability, temperature, and light requirements.
- Forestry: measurements of stand characteristics age, total biomass, and average tree diameter.
- Atmospheric modeling: at a given site we observe surface temperature, precipitation and wind speed.
Multivariate spatial modelling

Point-referenced spatial data often come as multivariate measurements at each location.

Examples:
- Environmental monitoring: stations yield measurements on ozone, NO, CO, and PM$_2.5$.
- Community ecology: assemblages of plant species due to water availability, temperature, and light requirements.
- Forestry: measurements of stand characteristics, age, total biomass, and average tree diameter.
- Atmospheric modeling: at a given site we observe surface temperature, precipitation, and wind speed

We anticipate dependence between measurements

Point-referenced spatial data often come as multivariate measurements at each location.

Examples:
- Environmental monitoring: stations yield measurements on ozone, NO, CO, and PM$_2.5$.
- Community ecology: assemblages of plant species due to water availability, temperature, and light requirements.
- Forestry: measurements of stand characteristics, age, total biomass, and average tree diameter.
- Atmospheric modeling: at a given site we observe surface temperature, precipitation, and wind speed

We anticipate dependence between measurements at a particular location across locations

Each location contains $m$ spatial regressions

$$Y_k(s) = \mu_k(s) + w_k(s) + \epsilon_k(s), \quad k = 1, \ldots, m.$$  

Mean: $\mu(s) = [\mu_k(s)]_{k=1}^m = [x_k^T(s)]_{k=1}^m \beta_k$  

Cov: $w(s) = [w_k(s)]_{k=1}^m \sim MVGP(0, \Gamma_w(\cdot, \cdot))$  

$$\Gamma_w(s, s') = [\text{Cov}(w_k(s), w_k(s'))]_{k,k'=1}^m$$

Error: $\epsilon(s) = [\epsilon_k(s)]_{k=1}^m \sim MVN(0, \Psi)$

$\Psi$ is an $m \times m$ p.d. matrix, e.g. usually Diag($\tau_k^2$)$_{k=1}^m$.

Properties:
- $\Gamma_w(s', s) = \Gamma_w(s, s')$
- $\lim_{s \to s'} \Gamma_w(s, s')$ is p.d. and $\Gamma_w(s, s) = \text{Var}(w(s))$.
- For sites in any finite collection $\mathcal{S} = \{s_1, \ldots, s_n\}$:
  $$\sum_{i=1}^n \sum_{j=1}^n u^i_i u^j_j \Gamma_w(s_i, s_j) \geq 0 \text{ for all } u_i, u_j \in \mathbb{R}^m.$$  

Any valid $\Gamma_w$ must satisfy the above conditions.

The last property implies that $\Sigma_w$ is p.d.

In complete generality:
- $\Gamma_w(s', s)$ need not be symmetric.
- $\Gamma_w(s, s')$ need not be p.d. for $s \neq s'$.
Moving average or kernel convolution of a process:
- Let \( Z(s) \sim GP(0, \rho(s, s')) \). Use kernels to form:
  \[\begin{align*}
  w_j(s) &= \int \kappa_j(u) Z(s + u) du = \int \kappa_j(s - s') Z(s') ds' \\
  \Gamma_w(s, s') &= (i, j)\text{-th element:}
  
  \left[ \Gamma_w(s - s') \right]_{i,j} = \int \kappa_i(s - s' + u) \kappa_j(u') \rho(u, u') du / u' 
  \end{align*}\]

Convolution of Covariance Functions:
- \( \rho_1, \rho_2, \ldots, \rho_m \) are valid covariance functions. Form:
  \[\left[ \Gamma_w(s - s') \right]_{i,j} = \int \rho_i(s - s' - t) \rho_j(t) dt \]

Constructive approach, contd.
- When \( s = s' \), \( \Gamma_v(s, s) = I_m \), so:
  \[\Gamma_w(s, s) = A(s) A^T (s)\]
  - \( A(s) \) identifies with any square-root of \( \Gamma_w(s, s) \). Can be taken as lower-triangular (Cholesky).
  - \( A(s) \) is unknown!
    - Should we first model \( A(s) \) to obtain \( \Gamma_w(s, s) \)?
    - Or should we model \( \Gamma_w(s, s') \) first and derive \( A(s) \)?
    - \( A(s) \) is completely determined from within-site associations.

Constructive approach, contd.
- If \( A(s) = A \):
  - \( w(s) \) is stationary when \( v(s) \) is.
    - \( \Gamma_w(s, s') \) is symmetric.
    - \( \Gamma_v(s, s') = \rho(s, s') I_m \Rightarrow \Gamma_w = \rho(s, s') A A^T \)
  - Last specification is called **intrinsic** and leads to **separable** models:
    \[\Sigma_w = H(\phi) \otimes \Lambda; \Lambda = A A^T\]

Let \( y = [Y(s_i)]_{i=1}^n \) and \( w = [W(s_i)]_{i=1}^n \).
- First stage:
  \[y \mid \beta, w, \Psi \sim \prod_{i=1}^n MVN (Y(s_i) | X(s_i)^T \beta + w(s_i), \Psi)\]
Multivariate spatial modeling

- Let \( y = \{Y(s_i)\}_{i=1}^n \) and \( w = \{W(s_i)\}_{i=1}^n \).
- First stage:
  \[
  y | \beta, \Psi \sim \text{MVN}(X\beta, \Sigma_w(\Phi))
  \]
  where \( \Sigma_w(\Phi) = \{\Gamma_w(s_i, s_j; \Phi)\}_{i,j=1}^n \).
- Second stage:
  \[
  w | \theta \sim \text{MVN}(0, \Sigma_w(\Phi))
  \]
  Marginalized likelihood:
  \[
  y|\beta, \Psi \sim \text{MVN}(X\beta, \Sigma_w(\Phi) + I \otimes \Psi)
  \]
- Third stage: Priors on \( \Omega = (\beta, \Psi, \Phi) \).

Bayesian computations

- Choice: Fit as \( y | \Omega \times [\Omega] \) or as \( y|\beta, w, \Psi \times [\omega] \times [\Omega] \).
- Conjugate distributions are available for \( \Psi \) and other variance parameters. Easy to program.
- Marginalized model:
  - Conjugate distributions are available for \( \Psi \) and other variance parameters. Easy to program.
  - Marginalized model:
    - Need Metropolis or Slice sampling for most variance-covariance parameters. Harder to program.
    - But reduced parameter space (no \( w \)'s) results in faster convergence
    - \( \Sigma_w(\Phi) + I \otimes \Psi \) is more stable than \( \Sigma_w(\Phi) \).

Bayesian computations

- Choice: Fit as \( y | \Omega \times [\Omega] \) or as \( y|\beta, w, \Psi \times [\omega] \times [\Omega] \).
- Conjugate distributions are available for \( \Psi \) and other variance parameters. Easy to program.
- Marginalized model:
  - Need Metropolis or Slice sampling for most variance-covariance parameters. Harder to program.
  - But reduced parameter space (no \( w \)'s) results in faster convergence
  - \( \Sigma_w(\Phi) + I \otimes \Psi \) is more stable than \( \Sigma_w(\Phi) \).
  - But what about \( \Sigma_w^{-1}(\Phi) \)? Matrix inversion is EXPENSIVE \( O(n^3) \).
Recovering the \( w \)'s?

- Interest often lies in the spatial surface \( w'y \).
- They are recovered from
  \[
  \mathbf{w} | \mathbf{y}, X = \int \mathbf{w} | \Omega, y, X \times | \Omega, y, X | d\Omega
  \]
  using posterior samples:
  - Obtain \( \Omega^{(1)}, \ldots, \Omega^{(S)} \sim | \Omega, y, X | \)

- NOTE: With Gaussian likelihoods \( | \Omega, y, X \) is also Gaussian. With other likelihoods this may not be easy and often the conditional updating scheme is preferred.

Recovering the \( w \)'s?

- Interest often lies in the spatial surface \( w'y \).
- They are recovered from
  \[
  \mathbf{w} | \mathbf{y}, X = \int \mathbf{w} | \Omega, y, X \times | \Omega, y, X | d\Omega
  \]
  using posterior samples:
  - Obtain \( \Omega^{(1)}, \ldots, \Omega^{(S)} \sim | \Omega, y, X | \)
  - For each \( \Omega^{(s)} \), draw \( \mathbf{w}^{(s)} \sim | \mathbf{w}^{(s)}, y, X | \)

Often we need to predict \( Y(s) \) at a new set of locations \( \{s_0, \ldots, s_m\} \) with associated predictor matrix \( X \).

- Sample from predictive distribution:
  \[
  \mathbb{E} \{ \mathbf{y} | \mathbf{y}, X, X \} = \int \mathbb{E} \{ \mathbf{y} | \mathbf{y}, X, X \} | d\Omega
  = \int \mathbb{E} \{ \mathbf{y} | \mathbf{y}, X, X \} | d\Omega
  \]

\( \mathbb{E} \{ \mathbf{y} | \mathbf{y}, X, X \} \) is multivariate normal. Sampling scheme:

- Obtain \( \Omega^{(1)}, \ldots, \Omega^{(S)} \sim | \Omega, y, X | \)
- For each \( \Omega^{(s)} \), draw \( \mathbf{y}^{(s)} \sim \mathbb{E} \{ \mathbf{y} | \mathbf{y}, \Omega^{(s)}, X, X \} \).
Study objectives:
- Evaluate methods for multi-source forest attribute mapping
- Find the “best” model, given the data
- Produce maps of biomass and uncertainty, by tree species

Study area:
- USDA FS Bartlett Experimental Forest (BEF), NH
- 1,053 ha heavily forested
- Major tree species: American beech (BE), eastern hemlock (EH), red maple (RM), sugar maple (SM), and yellow birch (YB)

Response variables:
- Metric tons of total tree biomass per ha
- Measured on 437 \( \frac{1}{2} \) ha plots
- Models fit using random subset of 218 plots
- Prediction at remaining 219 plots

Illustration from:
Covariates

- DEM derived elevation and slope
- Spring, Summer, Fall Landsat ETM+ Tasseled Cap features (brightness, greenness, wetness)

Candidate models

Each model includes 55 covariates and 5 intercepts, therefore, \(X^T\) is 1090 \times 60.

Different specifications of variance structures:
- Non-spatial multivariate \(\text{Diag}(\Psi) = \tau^2\)
- \(\text{Diag}(K), \text{ same } \phi, \text{ Diag}(\Psi)\)
- \(K, \text{ same } \phi, \text{ Diag}(\Psi)\)
Candidate models

Each model includes 55 covariates and 5 intercepts, therefore, $X^T$ is $1090 \times 60$.

Different specifications of variance structures:

1. Non-spatial multivariate $Diag(\Psi) = \tau^2$
2. $Diag(K)$, same $\phi$, $Diag(\Psi)$
3. $K$, same $\phi$, $Diag(\Psi)$
4. $Diag(K)$, different $\phi$, $Diag(\Psi)$
5. $K$, different $\phi$, $Diag(\Psi)$

Model comparison

Deviance Information Criterion (DIC):

$$D(\Omega) = -2 \log L(Data | \Omega)$$
$$D(\Omega) = E_{\Omega}D(\Omega)$$

$$pD = D(\Omega) - D(\bar{\Omega}) \quad \bar{\Omega} = E_{\Omega}[\Omega]$$

$$DIC = D(\Omega) + pD.$$ Lower DIC is better.

Selected model

- Model 5: $K$, different $\phi$, $Diag(\Psi)$
- Parameters: $K = 15$, $\phi = 5$, $Diag(\Psi) = 5$

Focus on spatial cross-covariance matrix $K$ (for brevity).

Posterior inference of $\text{cor}(K)$, e.g., 50 (2.5, 97.5) percentiles:

<table>
<thead>
<tr>
<th></th>
<th>BE</th>
<th>EH</th>
</tr>
</thead>
<tbody>
<tr>
<td>BE</td>
<td>0.16 (0.13, 0.21)</td>
<td>1</td>
</tr>
<tr>
<td>RM</td>
<td>-0.20 (-0.23, -0.15)</td>
<td>0.45 (0.26, 0.66)</td>
</tr>
<tr>
<td>SM</td>
<td>-0.20 (-0.22, -0.17)</td>
<td>-0.12 (-0.16, -0.09)</td>
</tr>
<tr>
<td>YB</td>
<td>0.07 (0.04, 0.08)</td>
<td>0.22 (0.20, 0.25)</td>
</tr>
</tbody>
</table>

These relationships expressed in mapped random spatial effects, $w$. 

Bartlett Experimental Forest

Selected model

- Model 5: $K$, different $\phi$, $Diag(\Psi)$
- Parameters: $K = 15$, $\phi = 5$, $Diag(\Psi) = 5$
Selected model

- Model 5: $K$, different $\phi$, $\text{Diag}(\Psi)$
- Parameters: $K = 15$, $\phi = 5$, $\text{Diag}(\Psi) = 5$

Focus on spatial cross-covariance matrix $K$ (for brevity).

Posterior inference of $\text{cor}(K)$, e.g., 50 (2.5, 97.5) percentiles:

<table>
<thead>
<tr>
<th></th>
<th>BE</th>
<th>EH</th>
<th>..</th>
</tr>
</thead>
<tbody>
<tr>
<td>BE</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EH</td>
<td>0.16 (0.13, 0.21)</td>
<td>1</td>
<td>..</td>
</tr>
<tr>
<td>RM</td>
<td>-0.20 (-0.23, -0.15)</td>
<td>0.45 (0.26, 0.66)</td>
<td>..</td>
</tr>
<tr>
<td>SM</td>
<td>-0.20 (-0.22, -0.17)</td>
<td>-0.12 (-0.16, -0.09)</td>
<td>..</td>
</tr>
<tr>
<td>YB</td>
<td>0.07 (0.04, 0.08)</td>
<td>0.22 (0.20, 0.25)</td>
<td>..</td>
</tr>
</tbody>
</table>

These relationships expressed in mapped random spatial effects, $w$.

Summary

Proposed Bayesian hierarchical spatial methodology:
- Partition sources of uncertainty
- Provides hypothesis testing
- Reveal spatial patterns and missing covariates
Summary

Proposed Bayesian hierarchical spatial methodology:
- Partition sources of uncertainty
  - Provides hypothesis testing
  - Reveal spatial patterns and missing covariates
- Allow flexible inference
  - Access parameters' posterior distribution
  - Access posterior predictive distribution
- Provide consistent prediction of multiple variables
  - Maintains spatial and non-spatial association

Extendable model template:
- Cluster plot sample design – multiresolution models
- Non-continuous response – general linear models
- Obs. over time and space – spatiotemporal models
Hierarchical Modelling for Spatio-temporal Data

Sudipto Banerjee\(^1\) and Andrew O. Finley\(^2\)

\(^1\) Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, U.S.A.
\(^2\) Department of Forestry & Department of Geography, Michigan State University, Lansing, Michigan, U.S.A.

October 15, 2012

Spatio-temporal Models

- Separable form:
  \[ C(s - s', t - t') = \sigma^2 \rho_1(s - s'; \phi_1) \rho_2(t - t'; \phi_2) \]

- Nonseparable form:
  \[ C(s - s', t - t') = \sigma^2 \rho_1(s - s'; \phi_1) \rho_2(t - t'; \phi_2) \]

Specification:
- Again point-referenced vs. areal unit data
- Continuous time vs. discretized time

\[ \Rightarrow \text{association in space, association in time} \]

For point-referenced data, \( t \) continuous, Gaussian
\[ Y(s, t) = \mu(s, t) + w(s, t) + \epsilon(s, t) \]

non-Gaussian data, \( q(EY(s, t) = \mu(s, t) + w(s, t) \]

Don't treat time as a third coordinate \( (s, t) \)
\[ \text{Cov}(Y(s, t), Y(s', t')) = C(s - s', t - t') \]

- Time discretized, \( Y_t(s), t = 1, 2, \ldots T \)
- Type of data: time series or cross-sectional

Mixing of separable covariance functions

Spectral domain approaches
### Spatio-temporal Models

- Time discretized, $Y_t(s), t = 1, 2, ...T$
- Type of data: time series or cross-sectional
- For time series data, exploratory analysis:
  - Arrange into an $n \times T$ matrix $Y$ with entries $Y_t(s)$
  - Center by row averages of $Y$ yields $Y_{row}$
  - Center by column averages of $Y$ yields $Y_{col}$
  - Sample spatial covariance matrix: $\frac{1}{n} Y_{row} Y_{row}^T$
  - Sample autocorrelation matrix: $\frac{1}{n} Y_{col} Y_{col}^T$
Spatio-temporal Models

Modeling: \( Y_t(s) = \mu_t(s) + w_t(s) + \epsilon_t(s) \), or perhaps
\( g(E(Y_t(s))) = \mu_t(s) + w_t(s) \)

For \( \epsilon_t(s) \), ... = \( w_{t-1}(s) + \eta_t(s) \), independent spatial process innovations

Time discretized, \( Y_t(s), t = 1, 2, \ldots, T \)

Type of data: time series or cross-sectional

For time series data, exploratory analysis:
- Arrange into an \( n \times T \) matrix \( Y \) with entries \( Y_{t,s} \)
- Center by row averages of \( Y \) yields \( Y_{rows} \)
- Center by column averages of \( Y \) yields \( Y_{cols} \)
- Sample spatial covariance matrix: \( \frac{1}{T} Y_{rows} Y_{rows}' \)
- Sample autocorrelation matrix: \( \frac{1}{n} Y_{cols} Y_{cols}' \)
- \( E, \) residuals matrix after a regression fitting, Empirical orthogonal functions (EOF)

Dynamic spatiotemporal models

**Measurement Equation**

\[
Y(s,t) = \mu(s,t) + \epsilon(s,t); \quad \epsilon(s,t) \overset{iid}{\sim} N(0, \sigma^2_{\epsilon}).
\]

\[
\mu(s,t) = X(s,t)' \hat{\beta}(s,t).
\]

\[
\hat{\beta}(s,t) = \beta_t + \beta(s,t)
\]

**Transition Equation**

\[
\beta_t = \beta_{t-1} + \eta_t, \quad \eta_t \overset{iid}{\sim} N(0, \Sigma_{\eta}).
\]

\[
\beta(s,t) = \beta(s, t-1) + w(s,t).
\]
Big N problem

Introduction

The Big \( n \) issue

Univariate spatial regression
\[ Y = X\beta + w + \epsilon, \]

- Estimation involves \((\sigma^2 R(\phi) + \tau^2 I)^{-1}\), which is \(n \times n\).
- Matrix computations occur in each MCMC iteration.
- Known as the “Big-N problem” in geostatistics.
- Approach: Use a model \( Y = X\beta + Zw^* + \epsilon \). But what \( Z \)?

Hierarchical Modelling for Large Spatial Datasets

Sudipto Banerjee\(^1\) and Andrew O. Finley\(^2\)

\(^1\) Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, U.S.A.
\(^2\) Department of Forestry & Department of Geography, Michigan State University, Lansing Michigan, U.S.A.

October 15, 2012

Big N problem
Predictive Process Models

- Consider “knots” \( \mathcal{S}^* = \{s^*_1, \ldots, s^*_n^*\} \) with \( n^* < < n \).
- Let \( \mathbf{w}^* = \{w(s^*_1)\}_{i=1}^{n^*} \).
- \( \mathbf{Z}(\theta) = \{\text{cov}(w(s_i), w^*(s_j))\} \{\text{var}(w^*)\}^{-1} \) is \( n \times n^* \).

Predictive process regression model
\[ Y = X\beta + Z(\theta)\mathbf{w}^* + \epsilon, \]

- Fitting requires only \( n^* \times n^* \) matrix computations (\( n^* < < n \)).
- Key attraction: The above arises as a process model:
  \[ \tilde{w}(\mathbf{s}) \sim GP(0, \sigma_w^2 \tilde{\rho}(\cdot; \phi)) \] instead of \( w(\mathbf{s}) \).
  \[ \tilde{\rho}(s_1, s_2; \phi) = \text{cov}(w(s_1), \mathbf{w}^*) \text{var}(\mathbf{w}^*)^{-1} \text{cov}(\mathbf{w}^*, w(s_2)) \]

Knots: A “Knotty” problem??

- Knot selection: Regular grid? More knots near locations we have sampled more?
- Formal spatial design paradigm: maximize information metrics (Zhu and Stein, 2006; Diggle & Lophaven, 2006)
- Geometric considerations: space-filling designs (Royle & Nychka, 1998); various clustering algorithms
- Compare performance of estimation of range and smoothness by varying knot size.
- Stein (2007, 2008): method may not work for fine-scale spatial data
- Still a popular choice – seamlessly adapts to multivariate and spatiotemporal settings.

A rectified predictive process is defined as
\[ \tilde{w}_R(\mathbf{s}) = \tilde{w}(\mathbf{s}) + \tilde{\epsilon}(\mathbf{s}), \]

where
\[ \tilde{\epsilon}(\mathbf{s}) \overset{\text{iid}}{\sim} N(0, \sigma^2_w (1 - R(\phi)/R^{-1}(\phi) R(\phi))). \]

Maximum likelihood estimates of \( \tau^2 \):
Illustration from:


Genetic effects model:

\[ Y_i = \mathbf{x}_i^T \beta + a_i + d_i + \epsilon_i, \]

- Common feature is systematic heterogeneity among observational units (i.e., violation of \( \epsilon \sim N(0, \tau^2 I_n) \))
- Spatial heterogeneity arises from:
  - soil characteristics
  - micro-climates
  - light availability
- Residual correlation among units as a function of distance and/or direction = erroneous parameter estimates (e.g., biased \( h^2 \))

Parameter credible intervals, 50% (2.5%, 97.5%) for the non-spatial models Scots pine trial.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta )</td>
<td>72.53 (69.66, 75.08)</td>
<td>72.27 (70.04, 74.57)</td>
</tr>
<tr>
<td>( \sigma_a^2 )</td>
<td>31.94 (18.30, 49.85)</td>
<td>25.23 (14.12, 43.96)</td>
</tr>
<tr>
<td>( \tau^2 )</td>
<td>~</td>
<td>~</td>
</tr>
<tr>
<td>( \epsilon^2 )</td>
<td>133.60 (121.18, 144.70)</td>
<td>116.14 (100.51, 127.76)</td>
</tr>
<tr>
<td>( h^2 )</td>
<td>0.19 (0.12, 0.28)</td>
<td>0.15 (0.09, 0.26)</td>
</tr>
</tbody>
</table>
Previous approaches to accommodating residual spatial dependence:
- Manipulate the mean function
- Geostatistical
  - spatial process formed $AR(1)_c \otimes AR(1)_r$ (Martin, 1990; Cullis et al., 1998)
  - classical geostatistical method (Zimmerman and Harville, 1991)

All are computationally feasible, but ad hoc and/or restrictive from a modeling perspective.

Trick to sample genetic effects:

Gibbs draw for random effects, e.g., $a_i \sim MVN \left( \mu_{ai}, \Sigma_{ai} \right)$, where calculating $\Sigma_{ai} = \left( \frac{1}{\nu_i} \mathbf{A}^{-1} + \frac{\nu_i}{\sigma_i^2} \right)^{-1}$ is computationally expensive!

However $\mathbf{A}$ and $\mathbf{D}$ are known, so use initial spectral decomposition i.e., $\mathbf{A}^{-1} = \mathbf{P}^{-1} \Lambda^{-1} \mathbf{P}^T$.

Thus, $\Sigma_{ai} = \mathbf{P}^T \left( \frac{1}{\nu_i} \Lambda^{-1} + \frac{1}{\sigma_i^2} I \right)^{-1} \mathbf{P}$ to achieve computational benefits.

Corresponding predictive process model:

$w_{i,j} = c(s_i; \theta)^T C(\theta)^{-1} (\theta) w^*$

where, $w^* = [w(s_i)]_{i=1}^{m} \sim MVN \left( \mathbf{0}, C(\theta) \right)$ and $C(\theta) = [C(s_i; s_j; \theta)]_{i,j=1}^{m}$

$w$ can accommodate complex spatial dependence structures, e.g., anisotropic Matérn correlation function:

$\rho(s_i, s_j; \theta) = \left( \frac{1}{\nu_i} \right)^{2 \nu^{-1}} 2^{\nu-1} \left( 2 \sqrt{\nu_i d_{ij}} \right)^{\nu} \kappa_{\nu} \left( 2 \sqrt{\nu_i d_{ij}} \right)$, where $d_{ij} = (s_i - s_j)^T \Sigma^{-1} (s_i - s_j)$, $\Sigma = G(\psi) \Lambda^2 G^T(\psi)$. Thus, $\theta = (\nu, \psi, \Lambda)$.
### Illustration

#### Univariate random effects models

Genetic + spatial effects models

- Candidate spatial models (i.e., specifications of $C^*(\theta)$):
  - $AR(1)_{col} \otimes AR(1)_{row}$
  - isotropic Matérn
  - anisotropic Matérn
- Each model evaluated using 64, 144, and 256 knot grids.
- Model choice using Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002)

```
<table>
<thead>
<tr>
<th>Model</th>
<th>$P_D$</th>
<th>DIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-spatial Add.</td>
<td>306.40</td>
<td>15,618.1</td>
</tr>
<tr>
<td>Add. Dom.</td>
<td>555.92</td>
<td>15,547.85</td>
</tr>
<tr>
<td>Spatial Isotropic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>64 Knots</td>
<td>639.77</td>
<td>14,877.51</td>
</tr>
<tr>
<td>144 Knots</td>
<td>739.61</td>
<td>14,814.89</td>
</tr>
<tr>
<td>256 Knots</td>
<td>802.29</td>
<td>14,771.64</td>
</tr>
<tr>
<td>Spatial Anisotropic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>64 Knots</td>
<td>678.82</td>
<td>14,884.13</td>
</tr>
<tr>
<td>144 Knots</td>
<td>748.89</td>
<td>14,823.90</td>
</tr>
<tr>
<td>256 Knots</td>
<td>806.46</td>
<td>14,781.53</td>
</tr>
</tbody>
</table>
```

#### Table: Model comparisons using the DIC criterion for the Scots pine dataset.

Genetic + spatial effects models results

Parameter credible intervals, 50% (2.5%, 97.5%) for the isotropic Matérn and 64 and 256 knots Scots pine trial.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>64 Knots</th>
<th>256 Knots</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta$</td>
<td>72.53 (69.00, 76.05)</td>
<td>74.21 (69.66, 77.66)</td>
</tr>
<tr>
<td>$\sigma_w^2$</td>
<td>26.87 (17.14, 41.82)</td>
<td>33.03 (18.19, 53.69)</td>
</tr>
<tr>
<td>$\sigma_d^2$</td>
<td>11.69 (6.00, 34.27)</td>
<td>13.96 (7.65, 27.05)</td>
</tr>
<tr>
<td>$\tau^2$</td>
<td>41.84 (23.71, 73.34)</td>
<td>50.36 (30.24, 88.10)</td>
</tr>
<tr>
<td>$\nu$</td>
<td>89.55 (72.11, 99.65)</td>
<td>80.75 (67.90, 96.16)</td>
</tr>
<tr>
<td>$h^2$</td>
<td>0.83 (0.31, 1.46)</td>
<td>0.47 (0.26, 1.28)</td>
</tr>
<tr>
<td>$\phi$</td>
<td>0.05 (0.02, 0.09)</td>
<td>0.04 (0.02, 0.09)</td>
</tr>
<tr>
<td>Eff. Range</td>
<td>71.00 (44.66, 127.93)</td>
<td>74.59 (45.22, 129.83)</td>
</tr>
</tbody>
</table>

- Decrease in $\tau^2$ due to removal of spatial variation, results in increase in $h^2$ (i.e., ~0.25 vs. ~0.15 with confounding).

#### Summary

Challenge - to meet modeling needs:

- ensure computationally feasible
  - reduce algorithmic complexity = cheap tricks (e.g., spectral decomp. of $A$ prior to MCMC)
  - reduce dimensionality = predictive process
- maintain richness and flexibility
  - focus on the model not how to estimate the parameters = embrace new tools (MCMC) for estimating highly flexible hierarchical models
- truly acknowledge sources of uncertainty
  - propagate uncertainty through hierarchical structures (e.g., recognize uncertainty in $C(\theta)$)

Predictive process – balance model richness with computational feasibility (e.g., 4,970×4,970 vs. 64×64).
Brief notes on setting up semi-high performance computing environments

October 15, 2012

Computing environments

We have two different computing environments for fitting demanding models to large space and/or time data sets.

1. A **distributed system** consists of multiple autonomous computers (nodes) that communicate through a computer network. A computer program that runs in a distributed system is called a distributed program. Message Passing Interface (MPI) is a specification for an Application Programming Interface (API) that allows many computers to communicate with one another (implementations in C, C++, and Fortran.)

2. A **shared memory multiprocessing system** consists of a single computer with memory that may be simultaneously accessed by one or more programs running on multiple central processing units (CPUs). The OpenMP (Open Multi-Processing) is an API that supports shared memory multiprocessing programming (implementations in C, C++, and Fortran).

My lab currently favors **shared memory multiprocessing system**.

- We buy rack mounted units (e.g., Sun Fire X4170 Server with 2 quad-core Intel Xeon Processor 5500 Series and 48 GB of RAM ~10-15k) running the Linux operating systems.

So what kind of speed up to expect from threaded BLAS and LAPACK libraries. Mean computing times of dpotrf:
Computing environments

See [http://blue.for.msu.edu/comp-notes](http://blue.for.msu.edu/comp-notes) for some simple examples of C++ with MKL and Rmath libraries along with associated Makefile files (I'll add more examples shortly and upon request).

- Many core and contributed packages (including spBayes) call Basic Linear Algebra Subprograms (BLAS) and LAPACK (Linear Algebra PACKage) Fortran libraries.
- Substantial computing gains:
  - processor specific threaded BLAS/LAPACK implementation (e.g., MKL or AMD's Core Math Library (ACML))
  - processor specific compilers (e.g., Intel's icc/ifort)

Compiling R to call MKL's BLAS and LAPACK libraries (rather than stock serial versions).

```
MKL_LIB_PATH="/opt/intel/composer_xe_2011_sp1.10.319/mkl/lib/intel64"
export LD_LIBRARY_PATH=$MKL_LIB_PATH
MKL="-L$MKL_LIB_PATH -lmkl_intel_lp64 -lmkl_intel_thread -lmkl_core -liomp5 -lpthread -lm"
./configure --with-blas="$MKL" --with-lapack
```

Time needed to collect 100 MCMC samples using spLM and threaded vs. non-threaded BLAS/LAPACK on a Intel Core 2 Quad processor and Ubuntu 8.10 Linux OS. R compiled with GNU gcc and gfortran.

![Graph showing time needed to collect 100 MCMC samples](image-url)