Hierarchical Modeling for Multivariate Spatial Data

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.

March 4, 2013

Each location contains m spatial regressions
\[ Y_k(s) = \mu_k(s) + w_k(s) + \epsilon_k(s), \quad k = 1, \ldots, m. \]

- **Mean:** \( \mu(s) = [\mu_1(s), \ldots, \mu_m(s)]^{\top} \)
- **Cov:** \( w(s) \sim MVGP(0, \Gamma_w(\cdot)) \) with
  \[ \Gamma_w(s, s') = [\text{Cov}(w_k(s), w_{k'}(s'))]_{k,k'=1}^m \]

Error:
\[ \epsilon(s) = [\epsilon_1(s), \ldots, \epsilon_m(s)]^{\top} \sim MN(0, \Psi) \]

- **\( \Psi \) is an \( m \times m \) p.d. matrix, e.g. usually \( \text{Diag}(\tau_k^2) \).**

Properties:
- \( \Gamma_w(s, s') = \Gamma_w^T(s', s) \)
- \( \lim_{s \to 0} \Gamma_w(s, s') \) is p.d. and \( \Gamma_w(s, s) = V_{\text{ar}}(w(s)) \).
- For sites in any finite collection \( \mathcal{S} = \{s_1, \ldots, s_n\} \):
  \[ \sum_{i=1}^n \sum_{j=1}^n u_i^T \Gamma_w(s_i, s_j) u_j \geq 0 \text{ for all } u_i, u_j \in \mathbb{R}^m. \]

Any valid \( \Gamma_w \) must satisfy the above conditions.
- The last property implies that \( \Sigma_w \) is p.d.
- In general complexity:
  - \( \Gamma_w(s, s') \) need not be symmetric.
  - \( \Gamma_w(s, s') \) need not be p.d. for \( s \neq s' \).

- 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

Moving average or kernel convolution of a process:
- Let \( Z(s) \sim GP(0, \rho(s)) \). Use kernels to form:
  \[ w_j(s) = \int \kappa_j(u) Z(s + u) du = \int \kappa_j(s - s') Z(s') ds' \]
- \( \Gamma_w(s - s') \) has \( (i, j) \)-th element:
  \[ \Gamma_w(s - s')_{i,j} = \int \kappa_i(s - s' - u) \kappa_j'(u) du = \int \kappa_i(s - s' - t) \kappa_j(t) dt. \]

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

- Let \( v_k(s) \sim GP(0, \rho_k(s, s')) \), for \( k = 1, \ldots, m \) be \( m \) independent GP’s with unit variance.
- Form the simple multivariate process \( \mathbf{v}(s) = [v_k(s)]_{k=1}^m \):
  \[
  \mathbf{v}(s) \sim MVGP(0, \Gamma_v(\cdot, \cdot))
  \]
  with \( \Gamma_v(s, s') = \text{Diag}(\rho_k(s, s'))_{k=1}^m \).
- Assume \( \mathbf{w}(s) = \Lambda(s) \mathbf{v}(s) \) arises as a space-varying linear transformation of \( \mathbf{v}(s) \). Then:
  \[
  \Gamma_w(s, s') = \Lambda(s) \Gamma_v(s, s') \Lambda^T(s')
  \]
  is a valid cross-covariance function.

Constructive approach, contd.

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

Bayesian computations

- Choice: Fit as \([y|\Omega] \times [\Omega]\) or as \([y|\beta, \Psi] \times [\mathbf{w}|\Psi] \times [\Omega]\).
- Conditional 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 \( \mathbf{w} \)'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 \( O(n^3) \).

Hierarchical modelling

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

Recovering the \( \mathbf{w} \)'s?

- Interest often lies in the spatial surface \( \mathbf{w} \).
- They are recovered from
  \[
  [\mathbf{w} \mid \Omega, \mathbf{y}, X] = \int [\mathbf{w} \mid \Omega, \mathbf{y}, X] d\Omega
  \]
  using posterior samples:
  - Obtain \( \Omega(1), \ldots, \Omega(\tau) \sim [\Omega \mid \mathbf{y}, X] \)
  - For each \( \Omega(i) \), draw \( \mathbf{w}(i) \sim [\mathbf{w} \mid \Omega(i), \mathbf{y}, X] \)
- **NOTE:** With Gaussian likelihoods \( [\mathbf{w} \mid \Omega, \mathbf{y}, X] \) is also Gaussian. With other likelihoods this may not be easy and often the conditional updating scheme is preferred.
Often we need to predict \( Y(\mathbf{s}) \) at a new set of locations \( \{ \mathbf{s}_0, \ldots, \mathbf{s}_m \} \) with associated predictor matrix \( X \).

Sample from predictive distribution:

\[
\begin{align*}
\mathbb{E} \left[ \mathbf{y} | \mathbf{y}, X, \hat{X} \right] &= \int \mathbf{y} \Omega \mathbf{y}, X, \hat{X} \right] d\Omega \\
&= \int \mathbb{E} \left[ \mathbf{y} | \mathbf{y}, \Omega \mathbf{y}, X, \hat{X} \right] d\Omega,
\end{align*}
\]

\( \mathbf{y}, \Omega, X, \hat{X} \) is multivariate normal. Sampling scheme:

- Obtain \( \Omega^{(1)}, \ldots, \Omega^{(G)} \sim \mathbb{E} \left[ \mathbf{y} | \mathbf{y}, X \right] \)
- For each \( \Omega^{(g)} \), draw \( \mathbf{y}^{(g)} \sim \mathbb{E} \left[ \mathbf{y} | \mathbf{y}, \Omega^{(g)}, X, \hat{X} \right] \).

Slight digression – why we fit a model:

- Association between response and covariates, \( \beta \), (e.g., ecological interpretation)
- Residual spatial and/or non-spatial associations and patterns (i.e., given covariates)
- Subsequent prediction

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 \( \frac{1}{10} \) ha plots
- Models fit using random subset of 218 plots
- Prediction at remaining 219 plots

Illustration from:

Illustration Bartlett Experimental Forest

Covariates

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

Model comparison

Deviance Information Criterion (DIC):

\[
D(\Omega) = -2 \log L(Data | \Omega) \\
D(\hat{\Omega}) = E_{\hat{\Omega}}[D(\Omega)]
\]

\[
pD = D(\Omega) - D(\hat{\Omega}); \hat{\Omega} = E_{\hat{\Omega}}[\Omega]
\]

\[
 DIC = \frac{D(\Omega) + pD}{N}
\]

Lower DIC is better.

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. \(Diag(K), \text{ same } \phi, Diag(\Psi)\)
3. \(K, \text{ same } \phi, Diag(\Psi)\)
4. \(Diag(K), \text{ different } \phi, Diag(\Psi)\)
5. \(K, \text{ different } \phi, Diag(\Psi)\)
6. \(K, \text{ different } \phi, \Psi\)

Selected model

- Model 5: \(K, \text{ different } \phi, Diag(\Psi)\)
- 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>RM</td>
<td>-0.20 (-0.23, -0.15)</td>
<td>-0.20 (-0.23, -0.15)</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, \(w\).
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