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.

March 3, 2010

The Big n issue

Univariate spatial regression

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{w} + \boldsymbol{\epsilon},$$

- 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 $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{w}^* + \boldsymbol{\epsilon}$. But what $\mathbf{Z}$?

- Consider "knots" $\mathscr{S}^* = \{\mathbf{s}_1^*, \ldots, \mathbf{s}_{n^*}^*\}$ with $n^* << n$.
- Let $\mathbf{w}^* = \{w(\mathbf{s}_i^*)\}_{i=1}^{n^*}$
- $\mathbf{Z}(\boldsymbol{\theta}) = \{\text{cov}(w(\mathbf{s}_i), w(\mathbf{s}_j^*))\}'\{\text{var}(\mathbf{w}^*)\}^{-1}$ is $n \times n^*$.

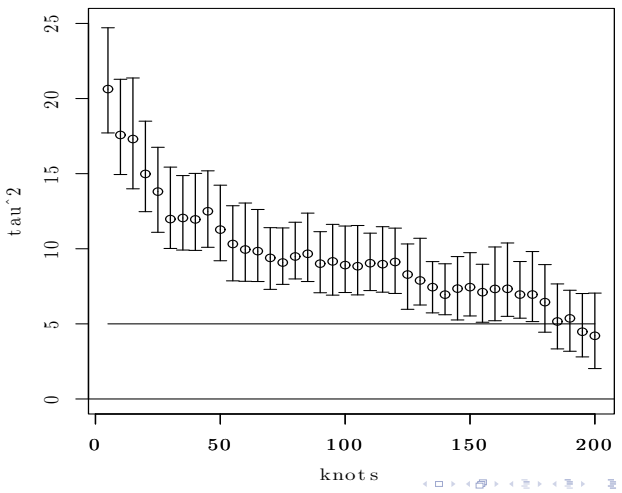
Predictive process regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}(\boldsymbol{\theta})\mathbf{w}^* + \boldsymbol{\epsilon},$$

- Fitting requires only $n^* \times n^*$ matrix computations ($n^* << n$).
- Key attraction: The above arises as a process model: $\tilde{w}(\mathbf{s}) \sim GP(0, \sigma_w^2 \tilde{\rho}(\cdot; \boldsymbol{\phi}))$ instead of $w(\mathbf{s})$.
- $\tilde{\rho}(\mathbf{s}_1, \mathbf{s}_2; \boldsymbol{\phi}) = \text{cov}(w(\mathbf{s}_1), \mathbf{w}^*)\text{var}(\mathbf{w}^*)^{-1}\text{cov}(\mathbf{w}^*, w(\mathbf{s}_2))$

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

$\tilde{w}_{\tilde{\epsilon}}(\mathbf{s}) = \tilde{w}(\mathbf{s}) + \tilde{\epsilon}(\mathbf{s})$, where

$\tilde{\epsilon}(\mathbf{s}) \overset{indep}{\sim} N(0, \sigma_w^2(1 - \mathbf{r}(\mathbf{s}, \boldsymbol{\phi})' R^{*-1}(\boldsymbol{\phi})\mathbf{r}(\mathbf{s}, \boldsymbol{\phi})))$.

Maximum likelihood estimates of τ^2:

# of Knots	Predictive Process	Rectified Predictive Process
25	1.56941	1.00786
36	1.65688	1.15386
64	1.45169	1.08358
100	1.37916	1.09657
225	1.27391	1.08985
400	1.22429	1.09489
625	1.21127	1.09998
exact	1.14414	1.14414

Illustration from:

Finley, A.O., S. Banerjee, P. Waldmann, and T. Ericsson. (2008) Hierarchical spatial modeling of additive and dominance genetic variance for large spatial trial datasets. Biometrics. DOI:10.1111/j.1541-0420.2008.01115.x

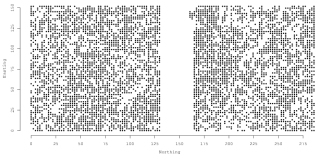
Univariate random effects models

Modeling genetic variation in Scots pine (*Pinus sylvestris* L.), long-term progeny study in northern Sweden.

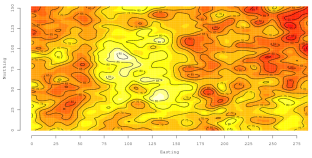
Quantitative genetics: studies the inheritance of polygenic traits, focusing upon estimation of additive genetic variance, σ^2_a, and the heritability $h^2 = \sigma^2_a/\sigma^2_{Tot}$, where the σ^2_{Tot} represents the total genetic and unexplained variation.

A high heritability, h^2, should result in a larger selection response (i.e., a higher probability for genetic gain in future generations).

Observed trees



Observed height



Data overview:

- established in 1971 (by Skogforsk)
- partial diallel design of 52 parent trees
- 8,160 planted randomly on 2.2m squares
- 1997 reinventory of 4,970 surviving trees, height, DBH, branch angle, etc.

Genetic effects model:

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

- $\mathbf{a} = [a_i]_{i=1}^n \sim MVN(\mathbf{0}, \sigma^2_a\mathbf{A})$
- $\mathbf{d} = [d_i]_{i=1}^n \sim MVN(\mathbf{0}, \sigma^2_d\mathbf{D})$
- $\boldsymbol{\epsilon} = [\epsilon_i]_{i=1}^n \sim N(\mathbf{0}, \tau^2 I_n)$

A and **D** are fixed relationship matrices (See e.g., Henderson, 1985; Lynch and Walsh,1998)

Note, genetic variance is further partitioned into additive and the non-additive *dominance* component σ^2_d

Genetic effects model:

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

- Common feature is systematic heterogeneity among observational units (i.e., violation of $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \tau^2 I_n)$)
- 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)

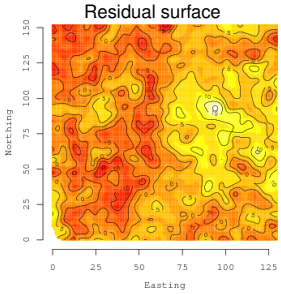
Genetic model results

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

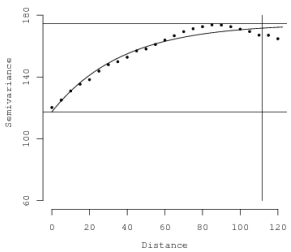
	Non-spatial	
Parameter	Add.	Add. Dom.
β	72.53 (69.66, 75.08)	72.27 (70.04, 74.57)
σ^2_a	31.94 (18.30, 49.85)	25.23 (14.12, 43.96)
σ^2_d	–	22.37 (11.24, 40.11)
τ^2	133.60 (121.18, 144.70)	116.14 (100.51, 127.76)
h^2	0.19 (0.12, 0.28)	0.15 (0.09, 0.26)

Genetic model results, cont'd.

Residual surface



Residual semivariogram



So, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \tau^2 I_n)$. Consider a spatial model.

Previous approaches to accommodating residual spatial dependence:

- Manipulate the mean function
  - constructing covariates using residuals from neighboring units (see e.g., Wilkinson et al., 1983; Besag and Kempton, 1986; Williams, 1986)
- Geostatitical
  - spatial process formed $AR(1)_{col} \otimes AR(1)_{row}$ (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.

Spatial model for genetic trials:

$$Y(\mathbf{s}_i) = \mathbf{x}^T(\mathbf{s}_i)\boldsymbol{\beta} + a_i + d_i + w(\mathbf{s}_i) + \epsilon_i,$$

- $\mathbf{a} = [a_i]_{i=1}^n \sim MVN(\mathbf{0}, \sigma_a^2 \mathbf{A})$
- $\mathbf{d} = [d_i]_{i=1}^n \sim MVN(\mathbf{0}, \sigma_d^2 \mathbf{D})$
- $\mathbf{w} = [w(\mathbf{s}_i)]_{i=1}^n \sim MVN(\mathbf{0}, \sigma_w^2 C(\boldsymbol{\theta}))$
- $\boldsymbol{\epsilon} = [\epsilon_i]_{i=1}^n \sim N(\mathbf{0}, \tau^2 I_n)$

Tools used to estimate parameters:

- Markov chain Monte Carlo (MCMC) - iterative
  - Gibbs sampler (β, **a**, **d**, **w**)
  - Metropolis-Hastings and Slice samplers ($\boldsymbol{\theta}$)

Here MCMC is computationally infeasible because of Big-N!

Trick to sample genetic effects:

Gibbs draw for random effects, e.g., $\mathbf{a}|\cdot \sim MVN(\boldsymbol{\mu}_{a|\cdot}, \Sigma_{a|\cdot})$, where calculating $\Sigma_{a|\cdot} = \left[\frac{1}{\sigma_a^2}\mathbf{A}^{-1} + \frac{I_n}{\tau^2}\right]^{-1}$ is computationally expensive!

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

Thus, $\Sigma_{a|\cdot} = P^T \left(\frac{1}{\sigma_a^2}\Lambda^{-1} + \frac{1}{\tau^2} I\right)^{-1} P$ to achieve computational benefits.

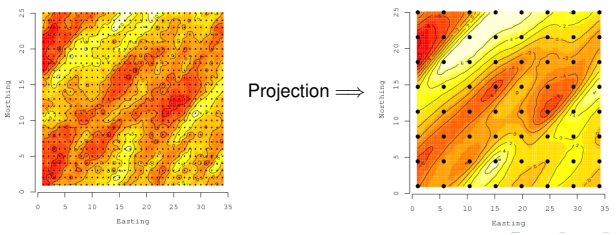
Unfortunately, this *trick* does not work for **w**. Rather, we proposed the knot-based *predictive process*.

Corresponding *predictive process* model:

$$Y(\mathbf{s}_i) = \mathbf{x}^T(\mathbf{s}_i)\boldsymbol{\beta} + a_i + d_i + \tilde{w}(\mathbf{s}_i) + \epsilon_i,$$

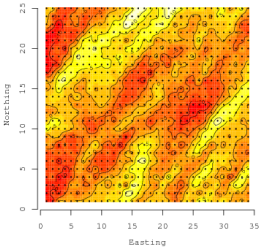
- $\tilde{w}(\mathbf{s}_i) = \mathbf{c}(\mathbf{s}_i; \boldsymbol{\theta})^T C(\boldsymbol{\theta})^{*-1}(\boldsymbol{\theta}) \mathbf{w}^*$

where, $\mathbf{w}^* = [w(\mathbf{s}_i^*)]_{i=1}^m \sim MVN(\mathbf{0}, C^*(\boldsymbol{\theta}))$ and $C^*(\boldsymbol{\theta}) = [C(\mathbf{s}_i^*, \mathbf{s}_j^*; \boldsymbol{\theta})]_{i,j=1}^m$

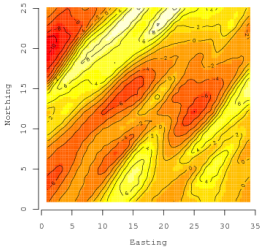


$\tilde{\mathbf{w}}$ can accomodate complex spatial dependence structures, E.g., anisotropic Matérn correlation function: $\rho(\mathbf{s}_i, \mathbf{s}_j; \boldsymbol{\theta}) = \left(1/\Gamma(\nu)2^{\nu-1}\right)\left(2\sqrt{\nu d_{ij}}\right)^{\nu} \kappa_{\nu}(2\sqrt{\nu d_{ij}})$, where $d_{ij} = (\mathbf{s}_i - \mathbf{s}_j)^T \Sigma^{-1} (\mathbf{s}_i - \mathbf{s}_j)$, $\Sigma = G(\psi)\Lambda^2 G^T(\psi)$. Thus, $\boldsymbol{\theta} = (\nu, \psi, \Lambda)$.

Simulated



Predicted



Genetic + spatial effects models

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

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

Model	p_D	DIC
Non-spatial		
Add.	306.40	15,618.09
Add. Dom.	555.92	15,547.85
Spatial Isotropic		
64 Knots	639.77	14,877.51
144 Knots	739.61	14,814.89
256 Knots	802.29	14,771.64
Spatial Anisotropic		
64 Knots	678.82	14,884.13
144 Knots	748.89	14,823.90
256 Knots	806.46	14,781.53

Genetic + spatial effects models results

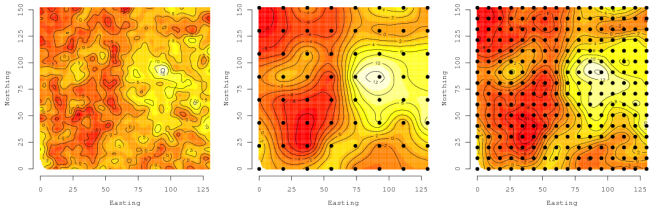
Parameter credible intervals, 50% (2.5%, 97.5%) for the isotropic Matérn and 64 and 256 knots Scots pine trial.

	Spatial	
Parameter	64 Knots	256 Knots
β	72.53 (69.00, 76.05)	74.21 (69.66, 79.66)
σ^2_a	26.87 (17.14, 41.82)	33.03 (18.19, 53.69)
σ^2_d	11.69 (6.00, 34.27)	13.96 (7.65, 27.05)
σ^2_w	41.84 (23.71, 73.34)	50.36 (30.24, 88.10)
τ^2	89.55 (72.11, 99.65)	80.75 (67.90, 96.16)
ν	0.83 (0.31, 1.46)	0.47 (0.26, 1.28)
ϕ	0.05 (0.02, 0.09)	0.04 (0.02, 0.09)
Eff. Range	71.00 (44.66, 127.93)	74.59 (45.22, 129.83)
h^2	0.21 (0.13, 0.31)	0.25 (0.15, 0.39)

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

Genetic + spatial effects models results, cont'd.

Genetic model residuals $\tilde{w}(\mathbf{s})$, 64 knots $\tilde{w}(\mathbf{s})$, 256 knots



Predictive process – balance model richness with computational feasibility (e.g., 4,970×4,970 vs. 64×64).

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 $C(\boldsymbol{\theta})$)