Challenges in Spatial Statistics: Large Data

Douglas Nychka, Colorado School of Mines

Instructor: Douglas Nychka

- Professor at Mines, Director of the Data Science Program
- Ph.D., University of Wisconsin, Statistics, 1983
- Focus: Curve and surface fitting, statistical computing
- Fellow ASA, IMS
- Senior Scientist Emeritus, National Center for Atmospheric Research
- Developer of the `fields` and `LatticeKrig` R Packages

Outline of the workshop

- Module 1: Large spatial data
- Module 2: Multivariate Spatial Data
- Module 3: Dynamical models over space and time

Outline

Sections:

- Large spatial data and linear algebra
- Representing curves with basis functions
- Fixed rank Kriging
- Spatial Autoregressions (SAR)
- LatticeKrig

Part 1 Large spatial data and linear algebra

A. Cholesky

D. Krige

Recap of Kriging

- **Recall:**

$$\begin{bmatrix} \mathbf{X_1} \\ \cdots \\ \mathbf{X_2} \end{bmatrix} \sim MN\left(\overbrace{\begin{bmatrix} \mu_1 \\ \cdots \\ \mu_2 \end{bmatrix}}^{\mu}, \overbrace{\begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix}}^{\mathbf{\Sigma}} \right)$$

- Suppose we observe $\mathbf{X_1}$ and want to predict $\mathbf{X_2}$.

D. Krige

-

$$\mathbf{X}_2 | \mathbf{X_1} \sim MN(\mu_2 + \mathbf{\Sigma}_{21} \mathbf{\Sigma}_{11}{}^{-1} (\mathbf{X_1} - \mu_1), \mathbf{\Sigma}_{22} - \mathbf{\Sigma}_{21} \mathbf{\Sigma}_{11}{}^{-1} \mathbf{\Sigma}_{12})$$

Computing these expressions

- $\boldsymbol{\Sigma_{11}}$: the covariance matrix for the observations.

If there are 1000 observations, this matrix is 1000×1000.

- To find MLEs also need the determinant of Σ_{11}.

Computing expressions with Σ_{11}^{-1} and $|\Sigma_{11}|$ grow as the cube of the number of observations.

Twice as many observations will take $8 = 2^3$ times longer.

Its all about the Cholesky

For the linear algebra fans ...

Spatial statistics computations make heavy use of the Cholesky decomposition.

- A a positive definite, symmetric matrix

Cholesky decomposition is $A = LL^{\mathrm{T}}$ where L is a *lower triangular* matrix.

- Compute $\mathbf{y}^{\mathrm{T}} A^{-1} \mathbf{y}$ by

$$\mathbf{y}^{\mathrm{T}} A^{-1} \mathbf{y} = \mathbf{y}^{\mathrm{T}} (LL^{\mathrm{T}})^{-1} \mathbf{y} = (L^{-1}\mathbf{y})^{\mathrm{T}} (L^{-1}\mathbf{y}) = \mathbf{w}^{\mathrm{T}} \mathbf{w}$$

$\mathbf{w}$ *solves* the linear system $L\mathbf{w} = \mathbf{y}$.

Solving a triangular system is very efficient.

- Compute determinant A.

$$|A| = |LL^{\mathrm{T}}| = |L||L^{\mathrm{T}}| = |L|^2$$

The determinant of a triangular matrix is the product of the diagonal elements.

Sparse matrices

- A is sparse if it has many zeros
(Typically we want the number of non-zero elements to grow linearly with the number of dimensions.)

- If A is sparse to find Ax skip over the zero elements to speedup multiplication

- If A is sparse Cholesky decomposition can also be sparse this will speed up solving linear systems.

More on Sparse matrices

A banded matrix with its Cholesky decomposition $A = LL^T$

$$A = \begin{bmatrix} 9 & -3 & 0 & 0 & 0 \\ -3 & 10 & -3 & 0 & 0 \\ 0 & -3 & 10 & -3 & 0 \\ 0 & 0 & -3 & 10 & -3 \\ 0 & 0 & 0 & -3 & 10 \end{bmatrix} \text{ and } L = \begin{bmatrix} 3 & 0 & 0 & 0 & 0 \\ -1 & 3 & 0 & 0 & 0 \\ 0 & -1 & 3 & 0 & 0 \\ 0 & 0 & -1 & 3 & 0 \\ 0 & 0 & 0 & -1 & 3 \end{bmatrix}$$

- With L triangular and sparse very fast to evaluate/ solve for $L^{-1}x$.

This means it is fast to evaluate.

$$x^T A^{-1} x = (L^{-1}x)^T L^{-1} x \text{ and } |A| = |L||L^T| = |L|^2$$

More on Sparse matrices

Order matters:

$$A = \begin{bmatrix} x & 0 & 0 & 0 & x \\ 0 & x & 0 & 0 & x \\ 0 & 0 & x & 0 & x \\ 0 & 0 & 0 & x & x \\ x & x & x & x & x \end{bmatrix} \text{ factors as } L = \begin{bmatrix} x & 0 & 0 & 0 & 0 \\ 0 & x & 0 & 0 & 0 \\ 0 & 0 & x & 0 & 0 \\ 0 & 0 & 0 & x & 0 \\ x & x & x & x & x \end{bmatrix}$$

But

$$A = \begin{bmatrix} x & x & x & x & x \\ x & x & 0 & 0 & 0 \\ x & 0 & x & 0 & 0 \\ x & 0 & 0 & x & 0 \\ x & 0 & 0 & 0 & x \end{bmatrix} \text{ factors as } L = \begin{bmatrix} x & 0 & 0 & 0 & 0 \\ x & x & 0 & 0 & 0 \\ x & x & x & 0 & 0 \\ x & x & x & x & 0 \\ x & x & x & x & x \end{bmatrix}$$

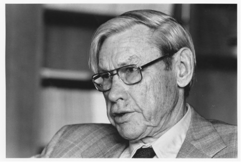
- Permute rows and columns of A to increase sparsity.

E.g. AMD is an ordering algorithm to find approximate minimum degree of a sparse matrix

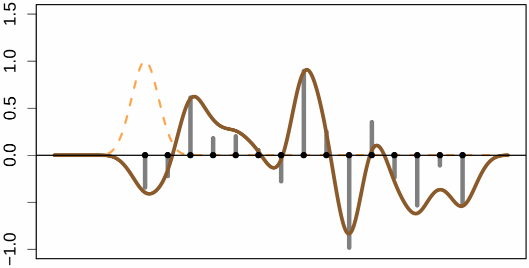
Our strategy

Formulate statistical models for spatial data that lead to sparse linear algebra.

Part 2 Basis functions for curve fitting



M. Stone K. Weierstrauss



Representing a curve

Start with your favorite m basis functions $\{b_1(s), b_2(s), \ldots, b_m(s)\}$
The curve has the form

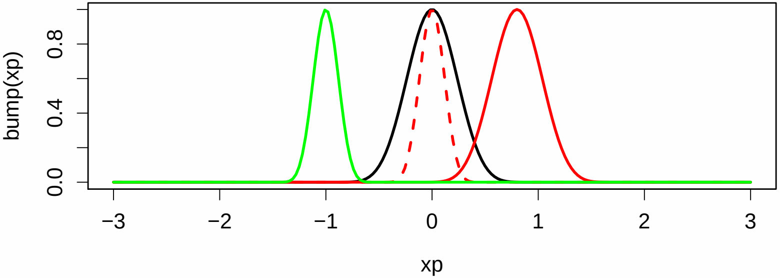
$$g(s) = \sum_{k=1}^{m} b_k(s) c_k$$

where $\mathbf{c} = (c_1, \ldots, c_m)$ are the coefficients.

- The basis functions are fixed
- Based on data find the coefficients.
- m does not have to be the same as the number of observations.

Many spatial statistics problems have this general form or can be approximated by it.

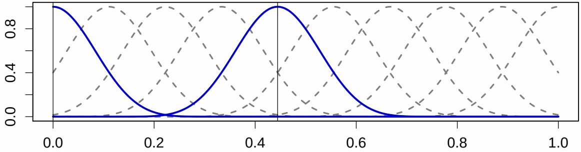
Example of basis functions



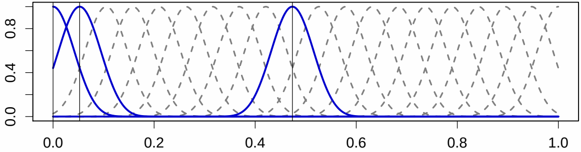
- Build a basis by translating and scaling a bump shaped curve
- Not your usual sine/cosine or polynomials!
- Bsplines not required!

Two Bases

10 Functions:

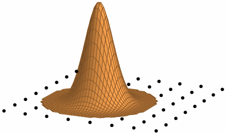


20 Functions:

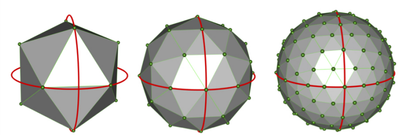


Use both together ($10 + 20 = 30$ functions) to represent two different scales of detail.

In two dimensions



Example of a 2-d bump



Lattice on a sphere

Defining the bump

$$b(\mathbf{s}) = \Phi(\|\mathbf{s} - \mathbf{u}\|/\alpha)$$

Φ a fixed bump shaped function, **u** the knot, and α is a scale factor.

- Gaussian, $\Phi(d) = e^{-d^2}$
- Wendland (2,2),

$$\Phi(d) = (1-d)^6(35d^2 + 18d + 3)/3 \quad (d \le 1) \text{ zero otherwise}$$

Basis function matrix

The *basis matrix*:

$$X_{i,k} = b_k(\mathbf{s}_i)$$

rows index locations, columns index the basis functions.

and so

$$\mathbf{g} = X\mathbf{c}$$

- If the basis functions have compact range then X is sparse.
- If X is sparse then so will $X^T X$.

Part 3 Fixed Rank Kriging

See: N. Cressie and G. Johannesson. (2008)

A model for the coefficients

$$g(s) = \sum_k b_k(s) c_k \text{ and } \mathbf{c} \sim N(0, \Omega)$$

$g(s)$ is a now a spatial process because $\mathbf{c}$ is a random vector.

More about this random effects model

Suppose:

- Basis functions are bumps centered at the knots $u_1, u_2, \ldots u_m$

- Use a spatial covariance to model dependence among coefficients using knot locations.

An Example of Ω

$$Cov(c_k, c_k) = \Omega_{k,k} = e^{-|u_k - u_k|/\alpha}$$

$$g(s) = \sum_k b_k(s) c_k$$

is now a *random* curve.

The covariance function

Using linear statistics:

$$Cov(g(s), g(s')) = \sum_{j,k} b_j(s) b_k(s') \Omega_{j,k}$$

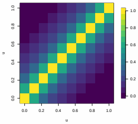
The covariance matrix for g at the observations has the simple formula

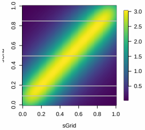
$$X \Omega X^{\mathrm{T}}$$

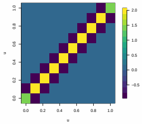
An example

Ten Wendland basis function scale of .4, exponential covariance with range .2.

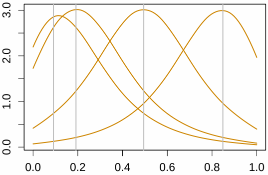
Covariance of $\mathbf{c}$ Covariance of $g(s)$ Precision of $\mathbf{c}$







Four slices of the $g(s)$ covariance matrix



Hard to the see the 10 RBFs! Looks a lot like a Matern, smoothness =1.

Estimating the coefficients

Basic idea: find **c** *given* **y**

Based on the multivariate normal or BLUE,

$$\hat{c} = (X^T X + \Omega^{-1})^{-1} X^T \mathbf{y}$$

or

$$(X^T X + \Omega^{-1})\hat{c} = X^T \mathbf{y}$$

- This is fixed rank Kriging.
- Also known as ridge regression estimate.
- Better to work directly with the inverse of Ω, $Q = \Omega^{-1}$
- Compute using the Cholesky!

If X is sparse and Ω^{-1} is sparse then this is now a sparse linear algebra problem.

Part 4
Spatial Autoregessions (SARs)

SAR models for $\mathbf{c}$

A 1-D case

Some coefficients:

$$. \quad c_{k-2} \quad c_{k-1} \quad c_k \quad c_{k+1} \quad c_{k+2} \quad .$$

Some weights:

$$0 \quad 0 \quad -1 \quad \texttt{a} \quad -1 \quad 0 \quad 0$$

A spatial autoregression:

$$\texttt{a}c_k - (c_{k-1} + c_{k+1}) = \texttt{a}c_k - c_{k-1} - c_{k+1} = e_k$$

$\{e_k\}$ are iid $N(0,1)$

Combining coefficients

$$B\mathbf{c} = \mathbf{e}$$

where $\mathbf{e} \sim N(0, I)$
B a matrix where each row has 3 nonzero weights:
a diagonal element, a and two first order neighbors (-1).

- a parameter needs to be greater than 2
- Precision matrix $Q = B^T B$, this is Ω^{-1}!
- Covariance matrix for $\mathbf{c}$ is $\Omega = Q^{-1} = B^{-1}$
- B and Q are sparse matrices.

NOTE: For practical use the variance and correlation range of this process is related to a and it is useful to normalize to a fixed variance.

SAR in two dimensions

Some coefficients:

$$
\begin{array}{ccccc}
\cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & c_{j-1,k} & \cdot & \cdot \\
\cdot & c_{j,k-1} & c_{j,k} & c_{j,k+1} & \cdot \\
\cdot & \cdot & c_{j+1,k} & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot
\end{array}
$$

Some weights:

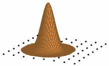
$$
\begin{array}{ccccc}
. & . & . & . & . \\
. & . & -1 & . & . \\
. & -1 & \mathtt{a} & -1 & . \\
. & . & -1 & . & . \\
. & . & . & . & .
\end{array}
$$

- Same concept although indexing is more difficult
- B is a sparse matrix with 5 nonzero elements on each row.
- a must be greater than 4.

Part 5 *LatticeKrig*

LatticeKrig model

A specific, Fixed Rank Kriging model

1. Basis functions at regular knots and compact support (zero beyond fixed range). Use the Wendland function. 

2. Coefficients follow a SAR model
   – for first or second order neighbors.

.	.	.	.	.
.	.	-1	.	.
.	-1	a	-1	.
.	.	-1	.	.
.	.	.	.	.

3. $\hat{\mathbf{c}}$ found by Kriging $(X^T X + \Omega^{-1})\hat{c} = X^T \mathbf{y}$

Why all this trouble?
Basis functions and SAR model give sparse matrices

Some practical additions

The Lattice Krig model should give reasonable covariance functions and follow standard Kriging results.

- Add a linear function to the basis.
- Add several different scales of basis functions together to approximate standard covariance functions.
- Normalize the SAR/basis functions to give a process with a unit variance

Parameters in the model

$$\mathbf{y}_i = g(\mathbf{s}_i) + \epsilon_i$$

- $Var(g(\mathbf{s}_i)) = \sigma^2$
- $Var(\epsilon_i) = \tau^2$

- a parameter in the SAR
- `NC` Number of basis functions in each dimension
- `nlevel` Number of multiresolution levels.
- `nu` Smoothness

`NC` chosen based on resolution of g
`nlevel` as large as possible (~ 3).
`nu` tracks the Matern interpretation, is hard to estimate from data and is also specified.

Summary

- Standard Kriging model breaks down with large data.
- An approximate model can be used based on basis functions and random coefficients
- Choosing compact basis functions and a SAR lead to sparse matrices and fast computation.
- The LatticeKrig model can be tuned to approximate standard Kriging results but for large data sets.

Thanks!