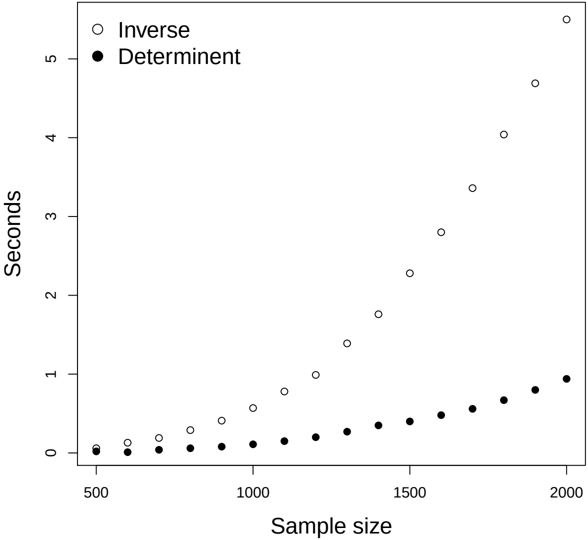
Methods to deal with large datasets

Applied Spatial Statistics

Problems caused by large spatial datasets

- Spatial models are problematic for large sample sizes
- Covariance matrix operations increase cubicly in the sample size (see next slide)
- Fortunately, there have been major computational advances on this problem in the last ten years
- This lecture will provide a high-level overview of this work
- Your midterm exam will be a deeper dive

Covariance matrix operation times



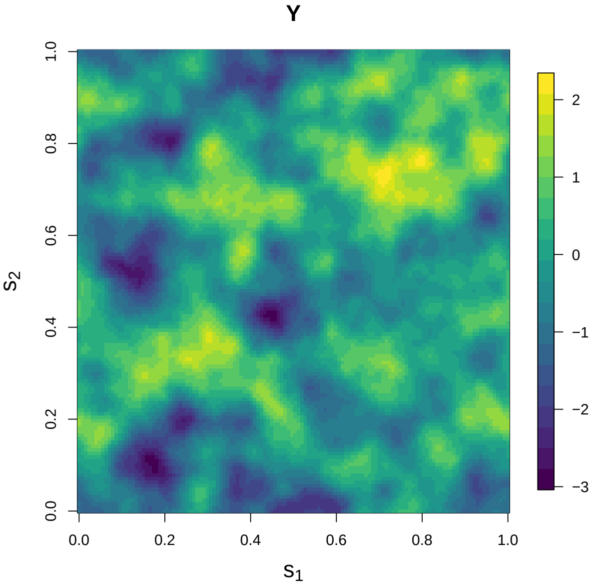
Outline of this lecture

- Low-rank methods
- Spectral methods
- Sparse-matrix methods
- Divide-and-conquer methods

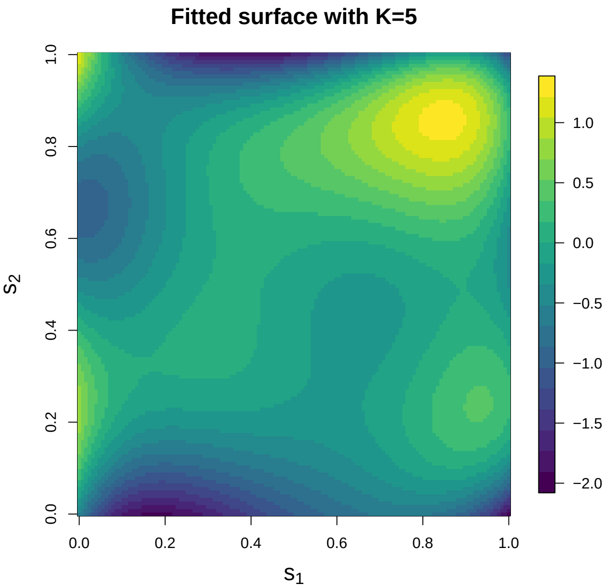
Low-rank methods

- Low-rank methods essentially turn the problem into a linear mixed model
- The covariates are constructed as functions of $\mathbf{s} = (s_1, s_2)$
- The model is $Y_i = \sum_{j=1}^{p} X_j(\mathbf{s}_i)\beta_j + \varepsilon_i$
- There are many choices for the covariates, $X_j(\mathbf{s})$
- First-order polynomial is $X_1(\mathbf{s}) = s_1$ and $X_2(\mathbf{s}) = s_2$
- Second-order polynomial adds $X_3(\mathbf{s}) = s_1^2$, $X_4(\mathbf{s}) = s_2^2$ and $X_5(\mathbf{s}) = s_1 s_2$

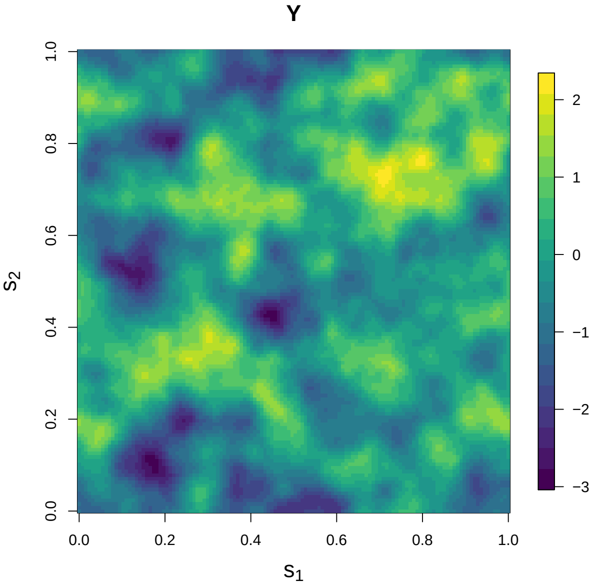
Simulated dataset



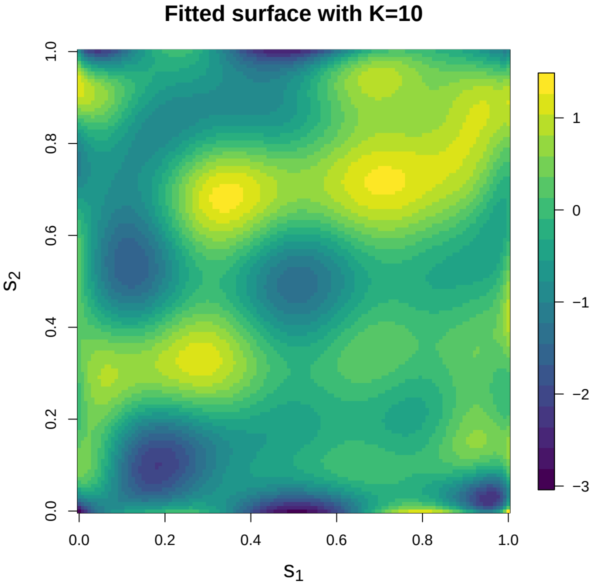
Fitted $K = 5$ order polynomial trend ($p = 20$)



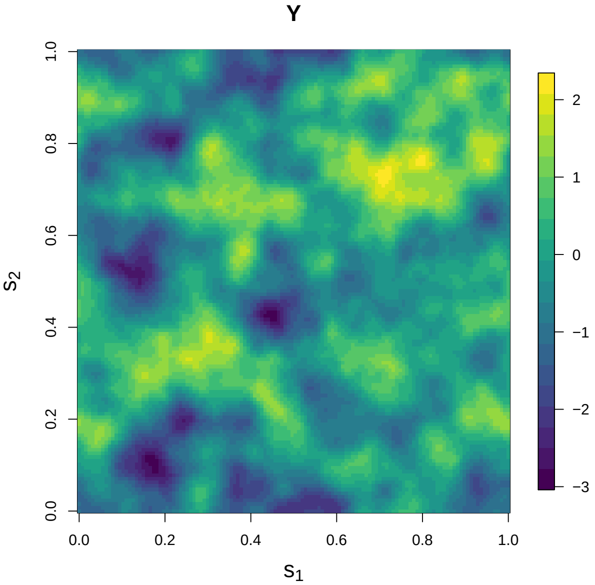
Simulated dataset



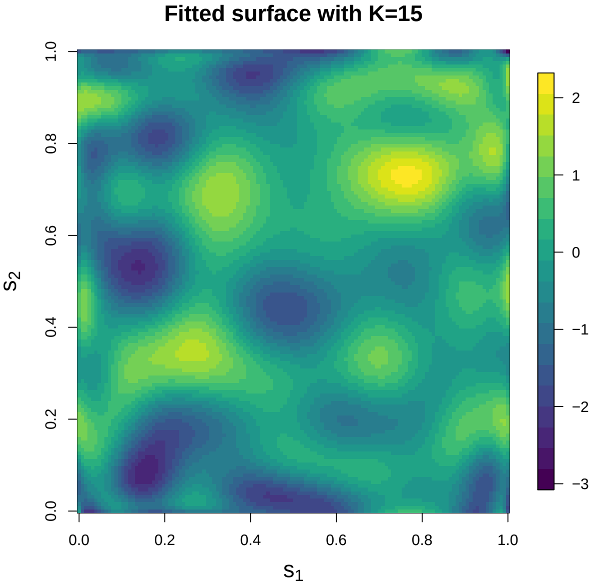
Fitted $K = 10$ order polynomial trend ($p = 65$)



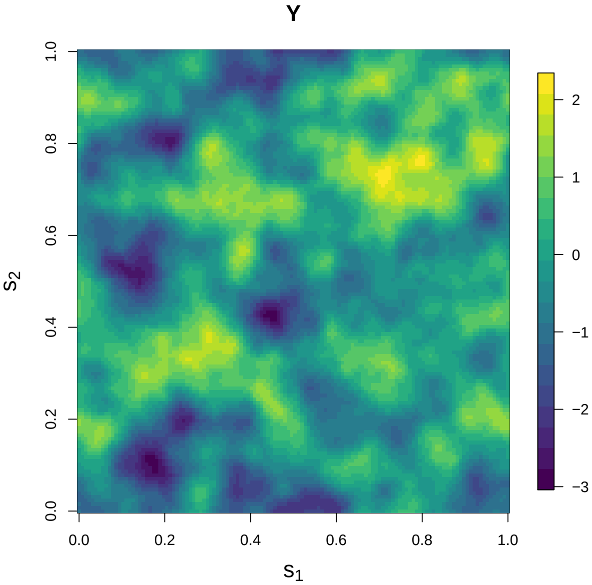
Simulated dataset



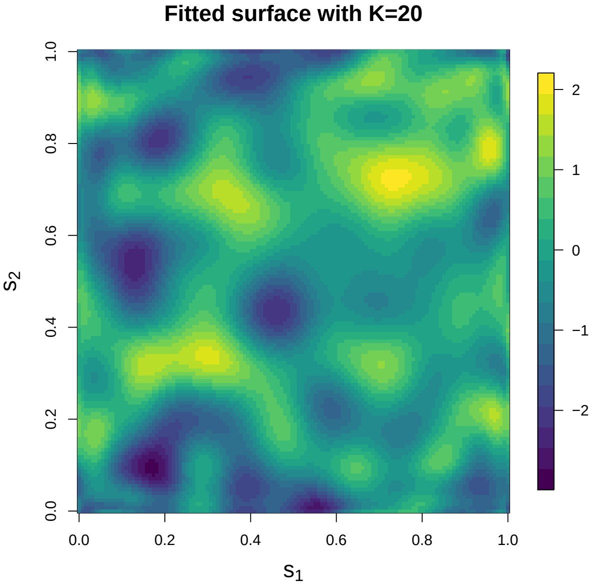
Fitted $K = 15$ order polynomial trend ($p = 135$)



Simulated dataset



Fitted $K = 20$ order polynomial trend ($p = 230$)



Low-rank methods

- There are many choices for basis functions: predictive process, fixed-rank Kriging, splines, wavelets, lattice Kriging, etc

- Generally p must be near n for good prediction

- To avoid over-fitting, usually the β_j are given a prior distribution/complexity penalty

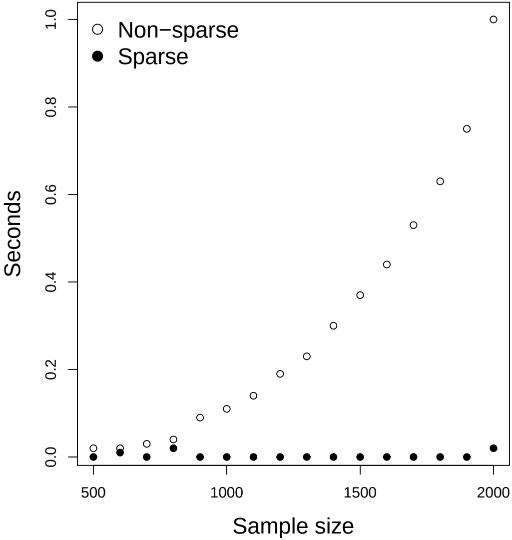
Spectral methods

- Spectral methods are super fast for data on a regular grid (i.e., columns and rows)
- Many large datasets are on a grid, e.g., the satellite data
- For stationary data on a 2D regular grid, the fast Fourier transform decorrelates the data
- The allows the observations to be treated as independent, which eliminates all matrix operations

Sparse matrix methods

- A sparse matrix is one with many entries equal zero
- For example, setting all correlations less than 0.01 to zero gives a sparse covariance matrix
- Sparsity can dramatically improve computation times
- The next slide shows the time to compute the determinant of sparse and non-sparse matrices

Sparse matrix methods



Sparse matrix methods

- Covariance tapering sets small correlations to zero
- There also methods that force the inverse covariance (precision) matrix to be sparse
- Veccia approximation
- Nearest neighbor Gaussian process (NNGP)
- Stochastic partial differential equation (SPDE) model

Divide-and-conquer (DnC) methods

- DnC methods split the data into smaller batches and compile the batch results
- Simple method:
  1. Divide the spatial domain into quadrants
  2. Compute the MLE for each quadrant
  3. Take the average of the MLEs as the final estimate
- It is tricky to decide how to group the observations and deal with correlation between groups
- The next slide shows times to compute the determinant of an $n \times n$ matrix and ten $(n/10) \times (n/10)$ matrices

Divide-and-conquer methods