Spatio-Temporal Smoothing for Very Large Datasets

Satellite Based Vegetation Measurements

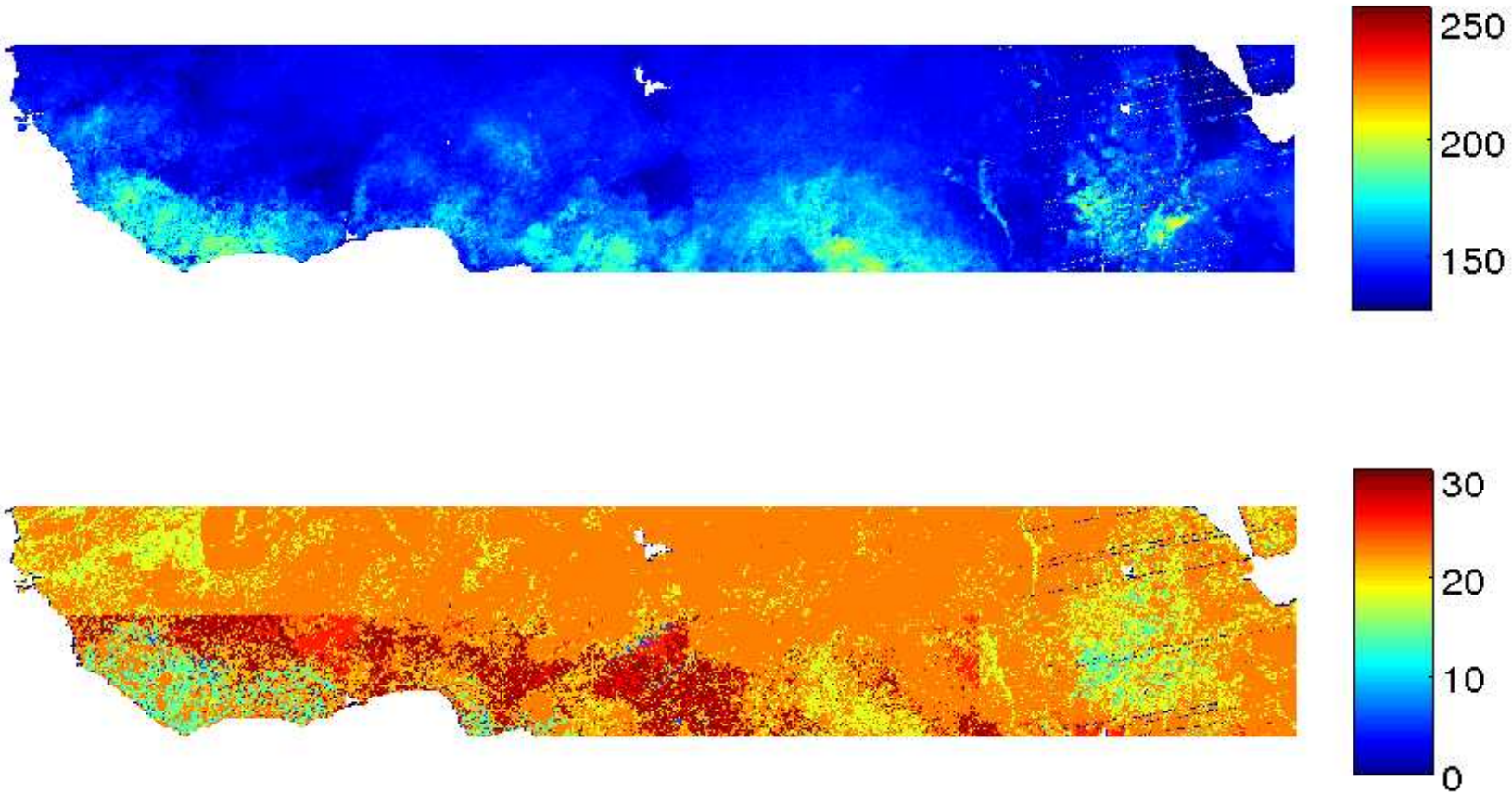
Johan Lindström[1]

[1]Centre for Mathematical Sciences
Lund University

Pan-American Advanced Study Institute
Búzios, June 25, 2014

LUND
UNIVERSITY

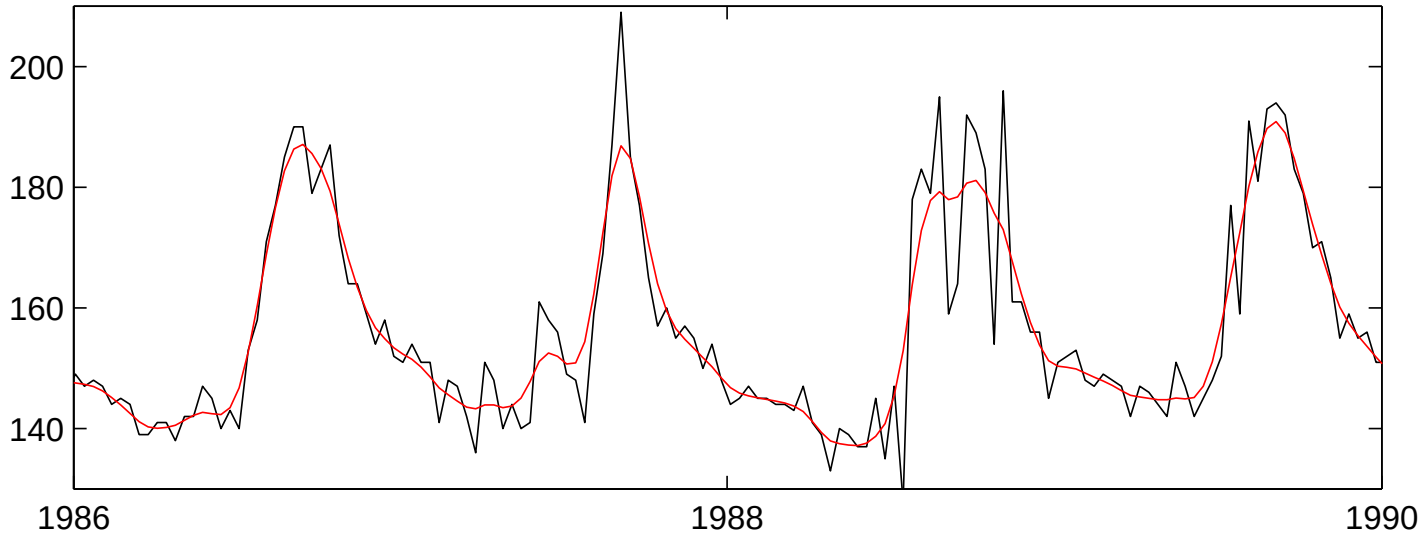
Satellite measured vegetation — The African Sahel



NDVI $[0, 255]$ and cloud $[0, 31]$ data from February 1985.

Smoothed version of the NDVI Data

Smooth the data to fill in missing values and remove noise due to cloud cover, etc.



Important ecological questions:

- Plant phenology (start and end of season)
- Plant productivity (integral)

What can the methods handle?

Assuming a standard desktop computer.

	N
Standard Kriging	$\sim 10^3$
Big N-methods	$\sim 10^5$
Satellite data	$\sim 10^8$

Big N-methods: process convolution (Higdon, 2001), tapering (Furrer et al., 2006), predictive processes (Banerjee et al., 2008), block composite (Eidsvik et al., 2014), SPDE/GMRF (Lindgren et al., 2011).

Space-time is often **worse**. For the SPDE-methods:

Dimension	Time for $R = \text{chol}(Q)$	$\text{nnz}(Q)$	$\text{nnz}(R)$
$\mathbb{R}^1$	$\mathcal{O}\left(N^1\right)$	$c^1 N$	$\sim N^1$
$\mathbb{R}^2$	$\mathcal{O}\left(N^{1.5}\right)$	$c^2 N$	$\sim N^{1+1/2}$
$\mathbb{R}^3$	$\mathcal{O}\left(N^2\right)$	$c^3 N$	$\sim N^{1+2/3}$
$\mathbb{R}^d$	???	$c^d N$	$\sim N^{1+(d-1)/d}$

Can't we just wait for better computers?

Current Satellite data (1982–)

$10\ \text{day}$ temporal resolution

$\sim 8\ \text{km}$ spatial resolution

New EU earth observation platform (launch 2015–)

$3 - 5\ \text{day}$ temporal resolution

$\leq 1\ \text{km}$ spatial resolution

Available data grows faster than processing power!

The standard likelihood

Log-likelihood

$$l(\theta|\boldsymbol{y}) = -\frac{1}{2}\log|\Sigma(\theta)| - \frac{1}{2}\Big(\boldsymbol{y} - \mu(\theta)\Big)^\top \Sigma(\theta)^{-1}\Big(\boldsymbol{y} - \mu(\theta)\Big),$$

with derivatives

$$\begin{aligned}\frac{\partial l(\theta|\boldsymbol{y})}{\partial \theta} = &-\frac{1}{2}\operatorname{tr}\left(\Sigma^{-1}\frac{\partial \Sigma}{\partial \boldsymbol{\theta}}\right) + (\boldsymbol{y} - \mu)^\top \Sigma^{-1}\frac{\partial \mu}{\partial \boldsymbol{\theta}} \\ &+ \frac{1}{2}(\boldsymbol{y} - \mu)^\top \Sigma^{-1}\frac{\partial \Sigma}{\partial \boldsymbol{\theta}}\Sigma^{-1}(\boldsymbol{y} - \mu).\end{aligned}$$

Spline smoothing (Wahba, 1981)

$$\mathrm{GCV}(\theta) = \frac{\|\boldsymbol{y} - \boldsymbol{S}(\theta)\boldsymbol{y}\|_2^2}{\Big(n - \operatorname{tr}(\boldsymbol{S}(\theta))\Big)^2}, \qquad \boldsymbol{S}(\theta) \propto \Big(\boldsymbol{A}^\top \boldsymbol{A} + \boldsymbol{Q}(\theta)\Big)^{-1}$$

Parameter estimates

Parameters are estimated by one of:

1. $\text{argmax}_\theta \, l(\theta|\boldsymbol{y})$.
2. Solving: $\nabla l(\theta|\boldsymbol{y}) = 0$.
3. $\text{argmin}_\theta \, \text{GCV}(\theta)$.

and we need to compute:

$\log \lvert\Sigma\rvert$	For case 1.
$\Sigma^{-1}\boldsymbol{y}$	For all cases.
$\text{tr}\left(\Sigma^{-1}\right)$	For cases 2, 3.

These can be approximated to high accuracy (Aune et al., 2014; Anitescu et al., 2012; Stein et al., 2013)

The Hutchinson (1990) trace estimator

The trace of a matrix Σ^{-1} is given by

$$\operatorname{tr}\left(\Sigma^{-1}\right) \approx \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{e}_k^\top \Sigma^{-1} \boldsymbol{e}_k$$

where the elements of $\boldsymbol{e}_k$ are iid $\{-1, 1\}$ with probability $1/2$.

$$\mathsf{E}\left(\boldsymbol{e}^\top \Sigma^{-1} \boldsymbol{e}\right) = \operatorname{tr}\left(\Sigma^{-1}\mathsf{E}(\boldsymbol{e}\boldsymbol{e}^\top)\right) = \operatorname{tr}\left(\Sigma^{-1}\boldsymbol{I}\right)$$

$$\mathsf{V}\left(\boldsymbol{e}^\top \Sigma^{-1} \boldsymbol{e}\right) = 2\left(\left\|\Sigma^{-1}\right\|_F^2 - \sum_i \left(\Sigma_{ii}^{-1}\right)^2\right) = 2\sum_{i\neq j}\left(\Sigma_{ij}^{-1}\right)^2$$

Usefull matrix algebra Harville (1997); Petersen and Pedersen (2012)

Probing vectors

A time series example ($K = 5$):

$$\mathcal{V} = \begin{bmatrix} \pm 1 & 0 & 0 & 0 & 0 \\ 0 & \pm 1 & 0 & 0 & 0 \\ 0 & 0 & \pm 1 & 0 & 0 \\ 0 & 0 & 0 & \pm 1 & 0 \\ 0 & 0 & 0 & 0 & \pm 1 \\ \pm 1 & 0 & 0 & 0 & 0 \\ 0 & \pm 1 & 0 & 0 & 0 \\ & & \vdots & & \end{bmatrix} \quad \mathsf{V}(\mathcal{V}\mathcal{V}^\top) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 0 & \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & \\ & & \vdots & & & & & \ddots \end{bmatrix}$$

$\boldsymbol{v}_k$ columns of $\mathcal{V}$.

Similar argument in $\mathbb{R}^2$.

In general a **NP-complete graph-colouring** problem, but **near optimal gready algorithms** exist.

Probing vectors II

Variance for the original estimator:

$$\mathrm{V}\left(\frac{1}{K}\sum_{k=1}^{K} \boldsymbol{e}_k^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{e}_k\right) = \frac{2}{K}\sum_{i\neq j}\left(\boldsymbol{\Sigma}^{-1}_{ij}\right)^2$$

Variance for the new estimator:

$$\mathrm{V}\left(\sum_{k=1}^{K} \boldsymbol{v}_k{}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{v}_k\right) = 2\left(\left\|\boldsymbol{\Sigma}^{-1}\circ \mathrm{V}(\boldsymbol{\nu}\boldsymbol{\nu}^\top)\right\|_F^2 - \sum_{i=1}^{n}\left(\boldsymbol{\Sigma}^{-1}_{ii}\right)^2\right)$$

$$= 2\sum_{\substack{i=j+mK \\ m\in\mathbb{Z}^+}}\left(\boldsymbol{\Sigma}^{-1}_{ij}\right)^2$$

This is better if elements in $\boldsymbol{\Sigma}^{-1}$ **decrease away from the diagonal**

Parameter estimates (again)

Parameters are estimated by one of:

1. $\text{argmax}_\theta \, l(\theta|\boldsymbol{y})$.
2. Solving: $\nabla l(\theta|\boldsymbol{y}) = 0$.
3. $\text{argmin}_\theta \, \text{GCV}(\theta)$.

and we need to compute:

$\log \lvert\Sigma\rvert$	For case 1.
$\Sigma^{-1}\boldsymbol{y}$	For all cases.
$\text{tr}\left(\Sigma^{-1}\right)$	For cases 2, 3.

We can reduce $\text{tr}\left(\Sigma^{-1}\right)$ to computations of $\Sigma^{-1}\boldsymbol{b}$.

Aune et al. (2014) uses $\log \lvert\Sigma\rvert = \text{tr} \log \Sigma$, and a series expansion of $\log \Sigma$.

Using the trace estimator the problem reduces to computing

$$x = A^{-1}b \qquad \text{for different } b$$

Preconditioned Conjugate Gradient

Iterative method for solving

$$Ax = b$$

using a **preconditioner**, M, and computing only Ax and $M^{-1}x$.

A suitable preconditioner M has

- $M^{-1}x$ easy to compute
- $M^{-1}A \approx I$.

Classical solutions

- Use $\text{diag}(A)^{-1}$.
- For sparse matrices use the (modified) incomplete Cholesky.

Comments

The Good

- No need to store the full matrices.
- $\mathcal{O}\left(N^1\right)$ for sparse matrices.
- Anitescu et al. can be seen as GEE $\rightarrow$ parameter uncertainties

The Bad

- Still $\mathcal{O}\left(N^2\right)$ for covariances matrices.
- Only Aune et al. evaluates the likelihood.

The Ugly

- Numerical linear solvers in MATLAB and R are reasonable, but not outstanding.
- Sparse matrix support is "lacking" in MATLAB and bad in R.
- Optimal preconditioner depends on the problem.

Smoothing splines (Wahba, 1981)

Find $\hat{\mu}(\boldsymbol{s})$ such that

$$\hat{\mu} = \underset{\mu}{\operatorname{argmin}} \frac{1}{n} \sum_i \|y_i - \mu_i\|_2^2 + \lambda \int \left(\frac{\partial^\alpha \mu(\boldsymbol{s})}{\partial \boldsymbol{s}^\alpha} \right)^2 \mathrm{d}\boldsymbol{s}$$

$\hat{\mu}$ is given as the solutions to an SPDE (Wahba, 1981):

$$\Delta^{\alpha/2} \boldsymbol{x}(\boldsymbol{s}) = \mathcal{W}$$

This corresponds to a Matérn (1960) covariance with range $\to \infty$, and the spline can be seen as a hierarchical model

$$\boldsymbol{y}|\boldsymbol{x} \in \mathrm{N}\left(\boldsymbol{A}\boldsymbol{x}, \sigma^2 \boldsymbol{I}\right) \qquad \boldsymbol{x} \in \mathrm{N}\left(0, \left(\tau^2 \boldsymbol{Q}\right)^{-1}\right).$$

Smoothing splines ≈ Gaussian processes

The spline smoothing of $\boldsymbol{y}$ can be seen as a conditional expectation

$$\begin{aligned}\widehat{\mu} = \mathsf{E}\left(\boldsymbol{x}|\boldsymbol{y};\tau^2,\sigma^2\right) &= \left(\tau^2\boldsymbol{Q} + \boldsymbol{A}^\top(\sigma^2\boldsymbol{I})^{-1}\boldsymbol{A}\right)^{-1}\left(\boldsymbol{A}^\top(\sigma^2\boldsymbol{I})^{-1}\boldsymbol{y}\right) \\ &= \left(\lambda\boldsymbol{Q} + \boldsymbol{A}^\top\boldsymbol{A}\right)^{-1}\boldsymbol{A}^\top\boldsymbol{y},\end{aligned}$$

where $\lambda = \tau^2\sigma^2$.

λ found by minimizing the GCV-error

$$GCV(\lambda) = \frac{\sum_i \|y_i - \widehat{\mu}_i\|_2^2}{(n - \mathsf{tr}(\boldsymbol{S}(\lambda)))^2}$$

with $\boldsymbol{S}(\lambda) = \boldsymbol{A}\left(\lambda\boldsymbol{Q} + \boldsymbol{A}^\top\boldsymbol{A}\right)^{-1}\boldsymbol{A}^\top$

Smoothing splines on a regular grid

For our splines ($\alpha = 2$) we have

$$\widehat{\mu} = \left(\lambda \boldsymbol{Q} + \boldsymbol{A}^\top \boldsymbol{A}\right)^{-1} \boldsymbol{A}^\top \boldsymbol{y}$$
$$\boldsymbol{Q} = \boldsymbol{G}^\top \boldsymbol{G}$$

where $\boldsymbol{G} \approx \Delta$.

For a regular grid in $1D$:

$$\boldsymbol{G} = \begin{bmatrix} 1 & -1 & 0 & \cdots & \\ -1 & 2 & -1 & 0 & \cdots \\ 0 & -1 & 2 & -1 & \cdots \end{bmatrix}$$

Neumann boundary condition

Smoothing splines on a regular grid

For our splines ($\alpha = 2$) we have

$$\widehat{\mu} = \left(\lambda \boldsymbol{Q} + \boldsymbol{A}^\top \boldsymbol{A}\right)^{-1} \boldsymbol{A}^\top \boldsymbol{y}$$
$$\boldsymbol{Q} = \boldsymbol{G}^\top \boldsymbol{G}$$

where $\boldsymbol{G} \approx \Delta$.

For a regular grid in $1D$:

$$\boldsymbol{G} = \begin{bmatrix} 2 & -1 & 0 & \cdots & 1 \\ -1 & 2 & -1 & 0 & \cdots \\ 0 & -1 & 2 & -1 & \cdots \end{bmatrix}$$

Cyclic boundary condition (Thon et al., 2012)

Preconditioner on the grid

Assume complete observations, $\boldsymbol{A} = \boldsymbol{I}$, and take $\boldsymbol{M} = \lambda \boldsymbol{Q} + \boldsymbol{I}$.

$$\boldsymbol{M} = \lambda \underbrace{\boldsymbol{G}^\top \boldsymbol{G}}_{=\boldsymbol{U}\boldsymbol{\Gamma}\boldsymbol{U}^\top} + \underbrace{\boldsymbol{I}}_{=\boldsymbol{U}\boldsymbol{U}^\top} = \boldsymbol{U} \underbrace{\left(\lambda \boldsymbol{\Gamma}^2 + \boldsymbol{I}\right)}_{=\boldsymbol{\Xi}} \boldsymbol{U}^\top,$$

where

$$\boldsymbol{U} = \begin{cases} \text{DCT}, & \text{Neumann} \\ \text{FFT}, & \text{Cyclic} \end{cases}$$

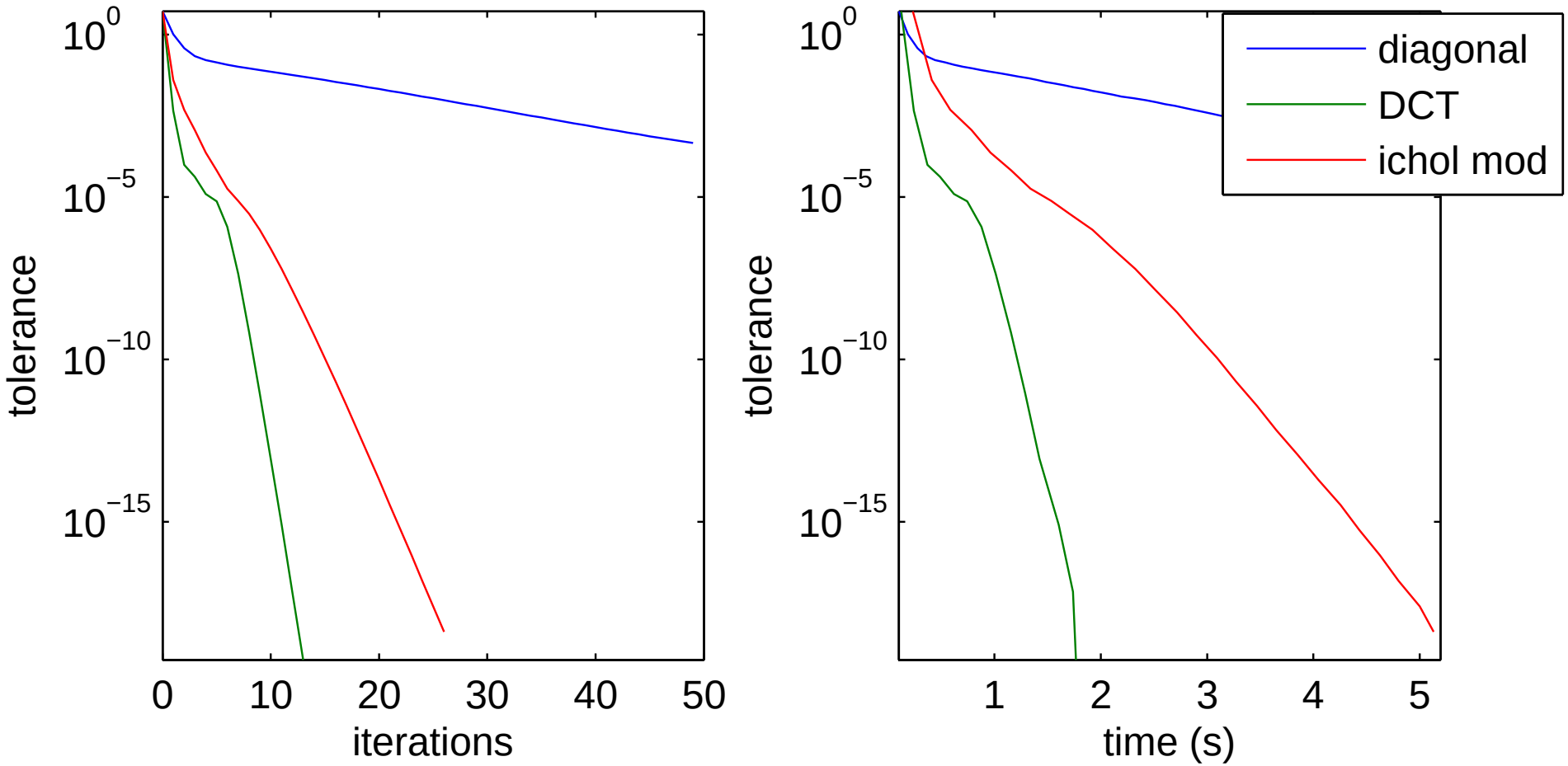
$$\boldsymbol{\Gamma}_{ii} = \begin{cases} 2 - 2\cos\left((i-1)\frac{\pi}{n}\right), & \text{Neumann} \\ 2 - 2\cos\left((i-1)\frac{2\pi}{n}\right), & \text{Cyclic} \end{cases}$$

For Neumann boundary conditions

$$\boldsymbol{M}^{-1}\boldsymbol{b} = \text{iDCT}\left(\boldsymbol{\Xi}^{-1}\,\text{DCT}(\boldsymbol{b})\right), \qquad \boldsymbol{\Xi}_{ii} = \lambda \boldsymbol{\Gamma}_{ii}^2 + 1$$

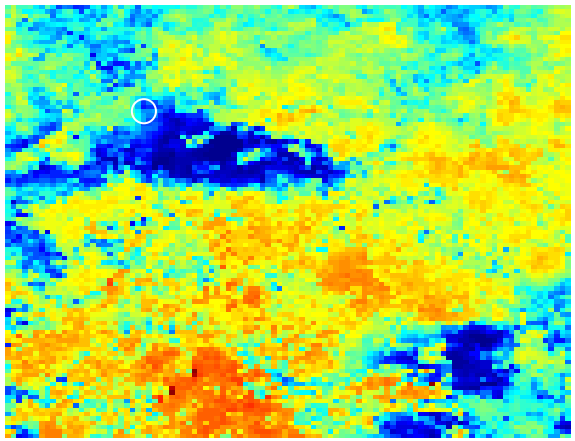
Convergence speed for PCG

Test run on a $2D$ 256×256 grid

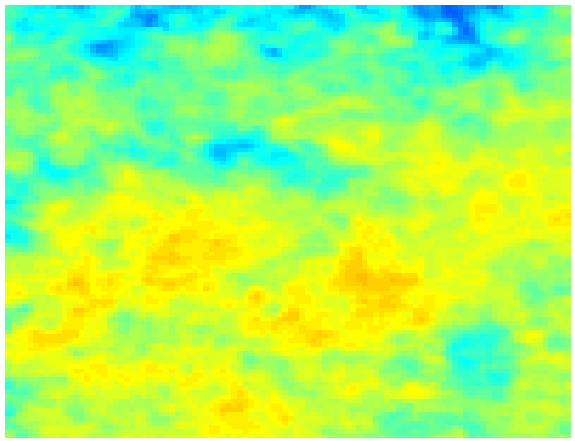


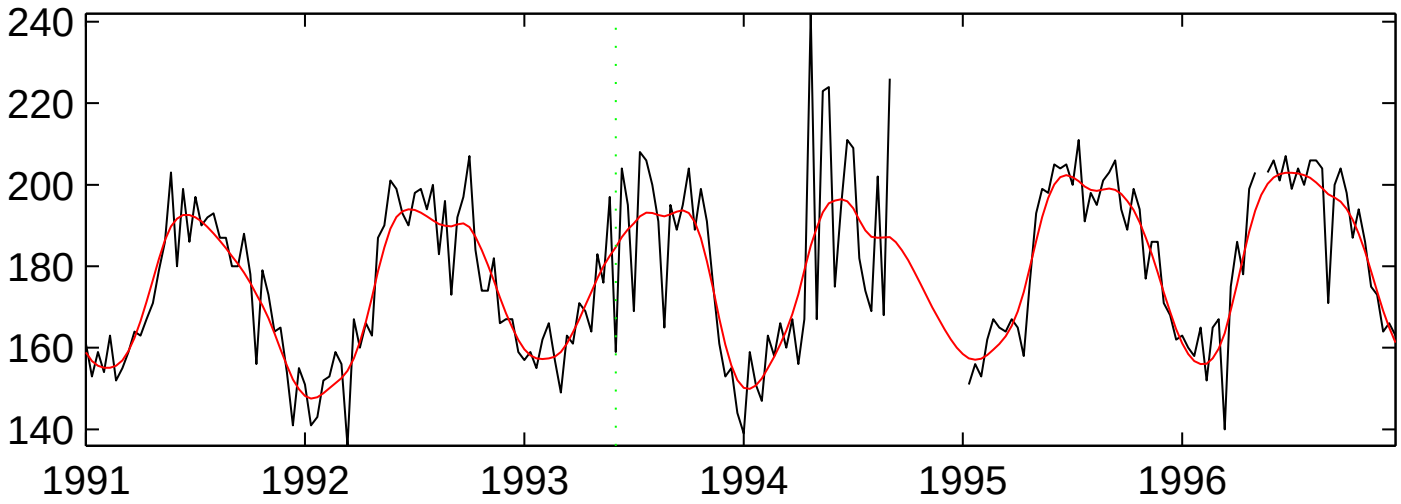
Preliminary results

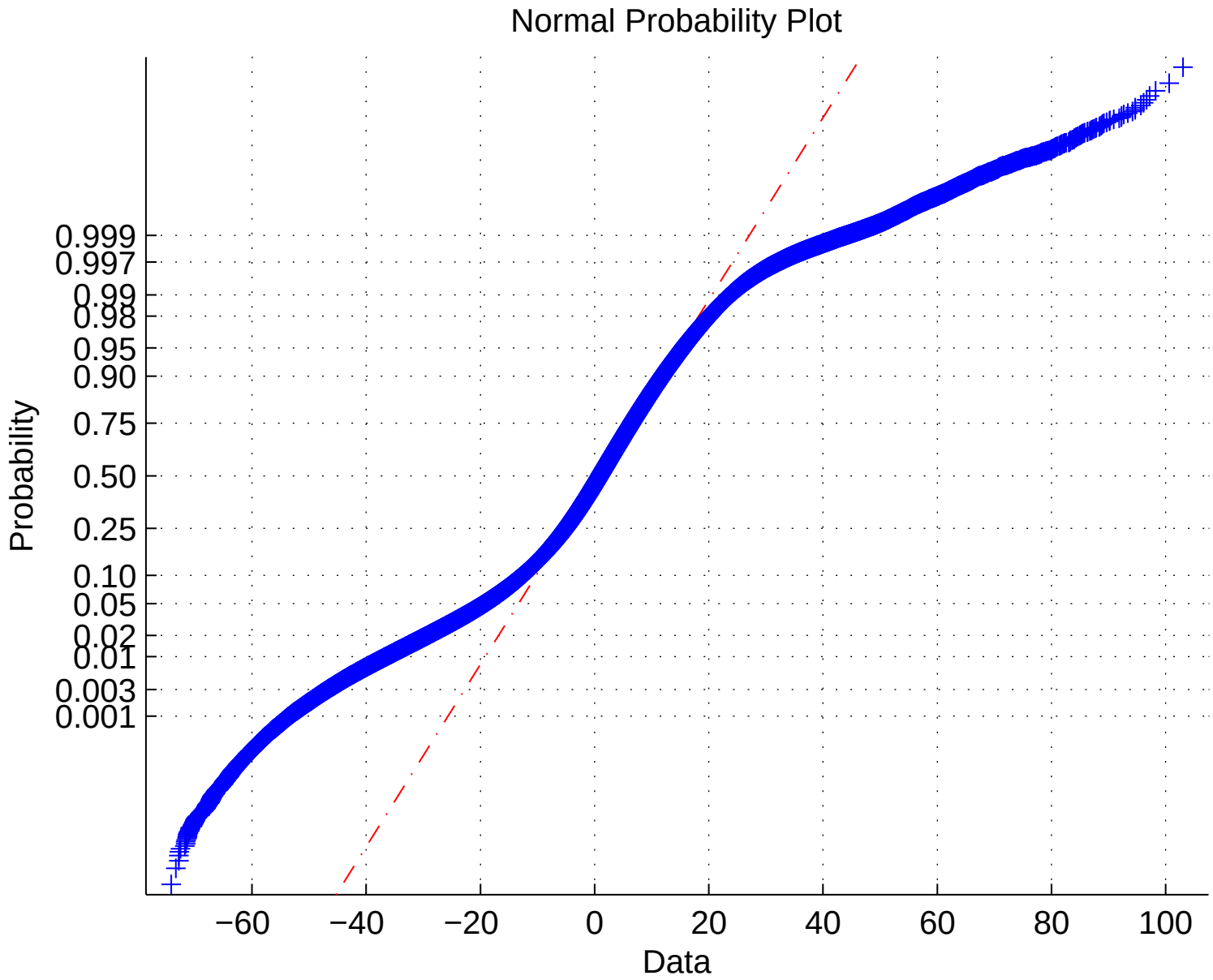
NDVI May 1993



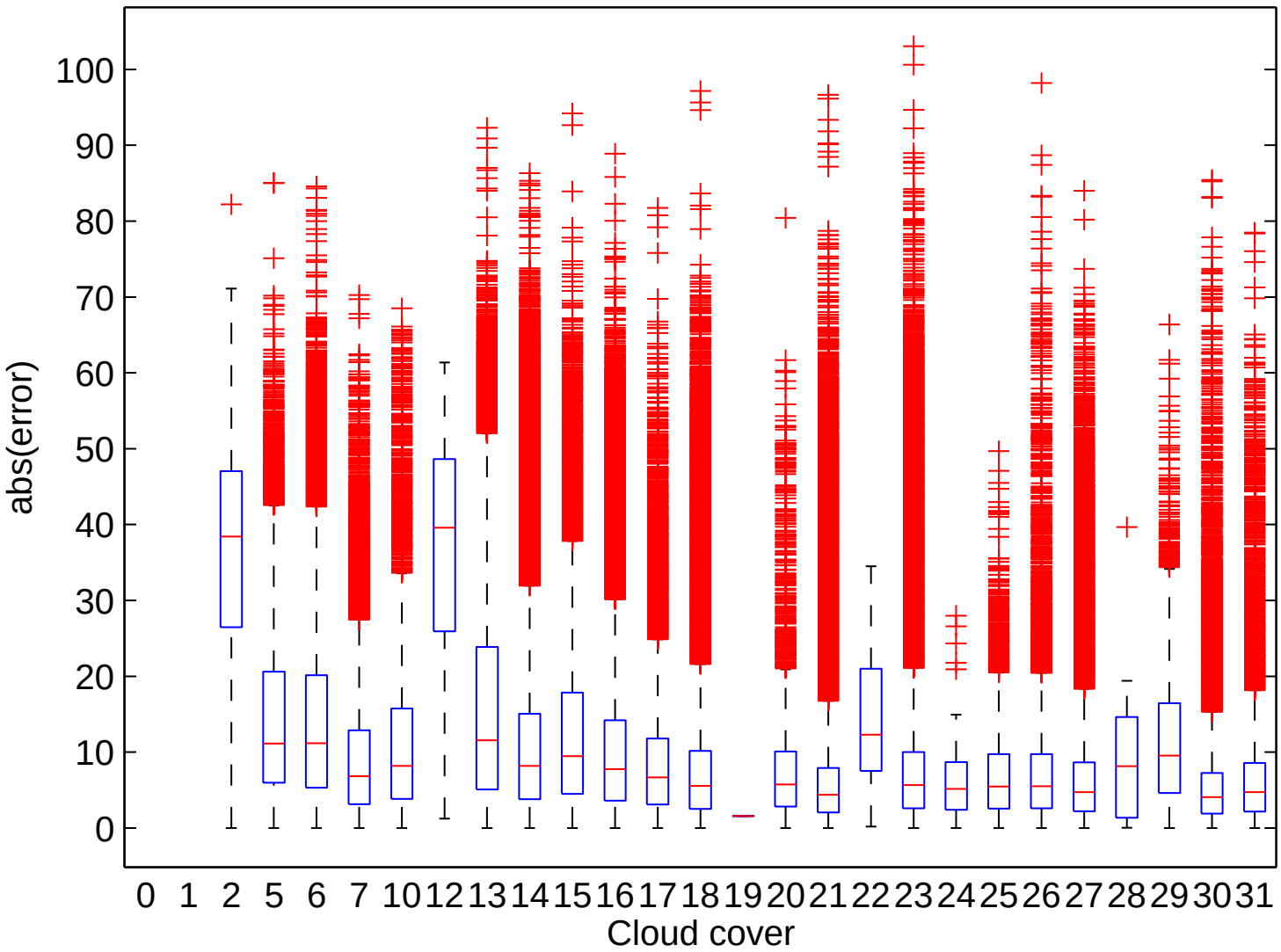
Smooth





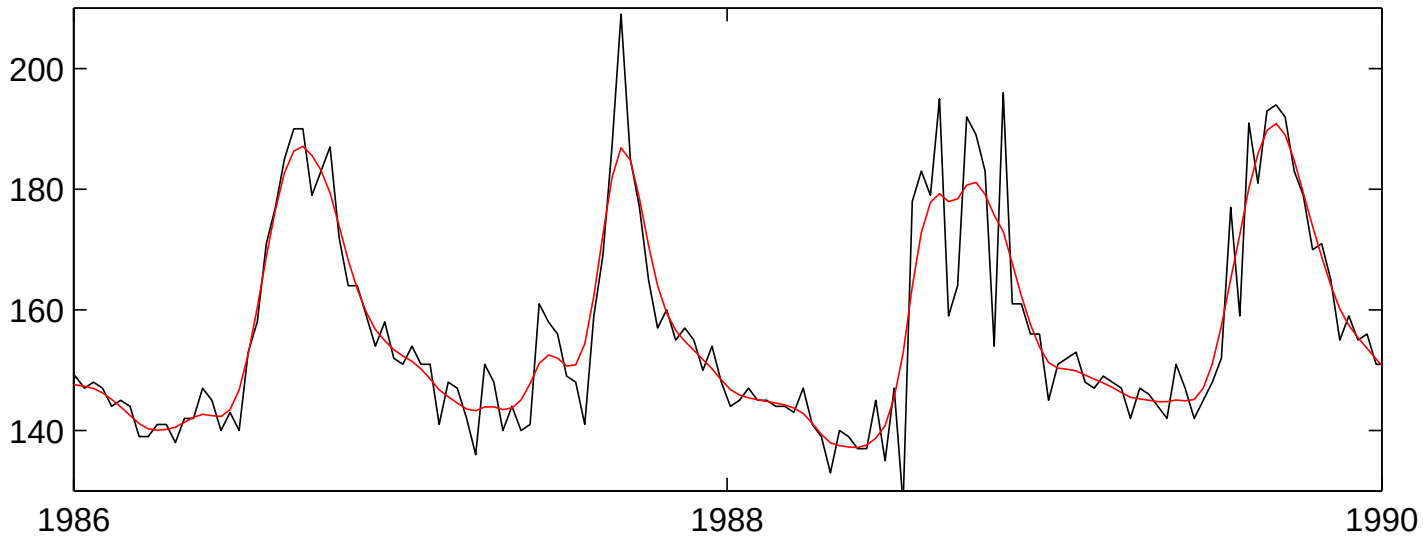

Normal Probability Plot
Probability
0.999
0.997
0.99
0.98
0.95
0.90
0.75
0.50
0.25
0.10
0.05
0.02
0.01
0.003
0.001
−60
−40
−20
0
20
40
60
80
100
Data

Can we use the cloud cover?



Further problems

- Non-Gaussian errors
- Axis anisotropy
- Non-stationarity in time
- Spatially correlated errors (due to clouds)
- Non-stationarity in space?



"When optimising function that involve high-dimensional determinant approximations, it is important to **use whatever structure is available** *in order to speed up computations."*

— Aune et al. (2014, pp. 256)

Bibliography I

Anitescu, M., Chen, J., and Wang, L. (2012), "A Matrix-free Approach for Solving the Parametric Gaussian Process Maximum Likelihood Problem," *SIAM J. Sci. Comput.*, 34, A240–A262.

Aune, E., Simpson, D. P., and Eidsvik, J. (2014), "Parameter estimation in high dimensional Gaussian distributions," *Stat. Comput.*, 24, 247–263.

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008), "Gaussian predictive process models for large spatial data sets," *J. R. Stat. Soc. B*, 70, 825–848.

Eidsvik, J., Shaby, B. A., Reich, B. J., Wheeler, M., and Niemi, J. (2014), "Estimation and Prediction in Spatial Models With Block Composite Likelihoods," *J. Comput. Graphical Stat.*, 23, 295–315.

Furrer, R., Genton, M. G., and Nychka, D. (2006), "Covariance Tapering for Interpolation of Large Spatial Datasets," *J. Comput. Graphical Stat.*, 15, 502–523.

Harville, D. A. (1997), *Matrix Algebra From a Statistician's Perspective*, Springer, 1st ed.

Higdon, D. (2001), "Space and space time modeling using process convolutions," Tech. rep., Institute of Statistics and Decision Sciences, Duke University, Durham, NC, USA.

Bibliography II

Hutchinson, M. (1990), "A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines," *Commun. Stat. Simul. Comput.*, 19, 433–450.

Lindgren, F., Rue, H., and Lindström, J. (2011), "An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach," *J. R. Stat. Soc. B*, 73, 423–498.

Matérn, B. (1960), "Spatial variation. Stochastic models and their application to some problems in forest surveys and other sampling investigations." Ph.D. thesis, Stockholm University, Stockholm, Sweden.

Petersen, K. B. and Pedersen, M. S. (2012), "The Matrix Cookbook," http://matrixcookbook.com.

Stein, M. L., Chen, J., and Anitescu, M. (2013), "Stochastic approximation of score functions for Gaussian processes," *Ann. Stat.*, 7, 1162–1191.

Thon, K., Rue, H., Skrøvseth, S. O., and Godtliebsen, F. (2012), "Bayesian multiscale analysis of images modeled as Gaussian Markov random fields," *Comput. Stat. Data Anal.*, 56, 49–61.

Wahba, G. (1981), "Spline interpolation and smoothing on the sphere," *SIAM J. Sci. Comput.*, 2, 5–16.