Gaussian Markov Random Fields: Theory and Applications

Håvard Rue[1]

[1]Department of Mathematical Sciences
NTNU, Norway

September 26, 2008

Part I

Definition and basic properties

Outline I

Why is it a good idea to learn about Gaussian Markov random fields (GMRFs)?
That is a good question!

What's there to learn? Isn't all just a Gaussian?

That is also a good question!

Why?

- Gaussians (and most often in the form of a GMRF) are extensively used in statistical models.
- However, its excellent computational properties are not often explored.
- Lack of knowledge of general purpose and near optimal numerical algorithms

Why?

- Fast(er) MCMC based inference

Recent developments (the *really* nice stuff...)

- Near instant (compare to MCMC) approximate Bayesian inference is often possible
- Deterministic
- Relative error
- In practice: no "error"

Longitudinal mixed effects model: Epil-example from BUGS

Patient	y_1	y_2	y_3	y_4	Trt	Base	Age
1	5	3	3	3	0	11	31
2	3	5	3	3	0	11	30
3	2	4	0	5	0	6	25
4	4	4	1	4	0	8	36
....							
8	40	20	21	12	0	52	42
9	5	6	6	5	0	12	37
....							
59	1	4	3	2	1	12	37

$y_{jk} \sim \text{Poisson}(m_{jk})$

Longitudinal mixed effects model: Epil-example from BUGS...

In this example

$$\mathbf{x} = (a_0, a_{\text{Base}}, a_{\text{Trt}}, a_{\text{BT}}, a_{\text{Age}}, a_{\text{V4}}, \{b1_j\}, \{b_{jk}\})$$

is a GMRF.

Two (hyper-)parameters

$$\text{Precision}(b1) \qquad \text{and} \qquad \text{Precision}(b)$$

Poisson-observations

Outline of this course: Day I

Lecture 1 GMRFs: definition and basic properties

Lecture 2 Simulation algorithms for GMRFs

Lecture 3 Numerical methods for sparse matrices

Lecture 4 Intrinsic GMRFs

Lecture 5 Hierarchical GMRF models and MCMC algorithms for such models

Lecture 6 Case-studies

Outline of this course: Day II

Lecture 1 Motivation for Approximate Bayesian inference

Lecture 2 INLA: Integrated Nested Laplace Approximations

Lecture 3 The *inla*-program: examples

Lecture 4 Using R-interface to the *inla*-program (Sara Martino)

Lecture 5 Case studies with R (Sara Martino)

Lecture 6 Discussion

What is a Gaussian Markov random field (GMRF)?

A GMRF is a simple construct

- A normal distributed random vector

$$\mathbf{x} = (x_1, \ldots, x_n)^T$$

- Additional Markov properties:

$$x_i \perp x_j \mid \mathbf{x}_{-ij}$$

x_i and x_j are conditional independent (CI).

If $x_i \perp x_j \mid \mathbf{x}_{-ij}$ for a set of $\{i, j\}$, then we need to constrain the parametrisation of the GMRF.

- Covariance matrix: difficult
- Precision matrix: easy

Conditional independence and the precision matrix

The density of a zero mean Gaussian

$$\pi(\mathbf{x}) \propto |\mathbf{Q}|^{1/2} \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x}\right)$$

Constraining the parametrisation to obey CI properties

Theorem

$$x_i \perp x_j \mid \mathbf{x}_{-ij} \quad \Longleftrightarrow \quad Q_{ij} = 0$$

Use a (undirected) graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to represent the CI properties,

$\mathcal{V}$ Vertices: $1, 2, \ldots, n$.

$\mathcal{E}$ Edges $\{i, j\}$

- No edge between i and j if $x_i \perp x_j \mid \mathbf{x}_{-ij}$.
- An edge between i and j if $x_i \not\perp x_j \mid \mathbf{x}_{-ij}$.

Definition of a GMRF

Definition (GMRF)

A random vector $\mathbf{x} = (x_1, \ldots, x_n)^T$ is called a GMRF wrt the graph $\mathcal{G} = (\mathcal{V} = \{1, \ldots, n\}, \mathcal{E})$ with mean $\boldsymbol{\mu}$ and precision matrix $\mathbf{Q} > 0$, iff its density has the form

$$\pi(\mathbf{x}) = (2\pi)^{-n/2} |\mathbf{Q}|^{1/2} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \mathbf{Q} (\mathbf{x} - \boldsymbol{\mu}) \right)$$

and

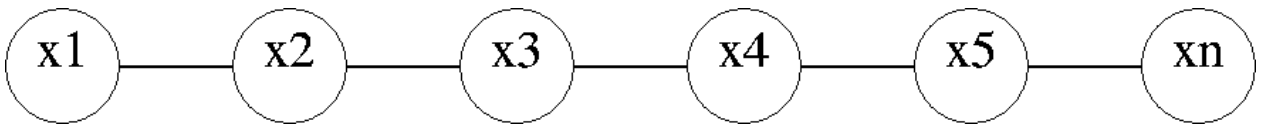
$$Q_{ij} \neq 0 \quad \Longleftrightarrow \quad \{i, j\} \in \mathcal{E} \quad \text{for all} \quad i \neq j.$$

Simple example of a GMRF

Auto-regressive process of order 1

$$x_t \mid x_{t-1}, \ldots, x_1 \sim \mathcal{N}(\phi x_{t-1}, 1), \quad t = 2, \ldots, n$$

and $x_1 \sim \mathcal{N}(0, (1-\phi^2)^{-1})$.



Tridiagonal precision matrix

$$\mathbf{Q} = \begin{pmatrix} 1 & -\phi & & & & \\ -\phi & 1+\phi^2 & -\phi & & & \\ & \ddots & \ddots & \ddots & & \\ & & -\phi & 1+\phi^2 & -\phi & \\ & & & -\phi & 1 & \end{pmatrix}$$

Usage of GMRFs (I)

Structural time-series analysis

- Autoregressive models.
- Gaussian state-space models.
- Computational algorithms based on the Kalman filter and its variants.

Analysis of longitudinal and survival data

- temporal GMRF priors
- state-space approaches
- spatial GMRF priors

used to analyse longitudinal and survival data.

Usage of GMRFs (II)

Graphical models

- A key model
- Estimate **Q** and its (associated) graph from data.
- Often used in a larger context.

Semiparametric regression and splines

- Model a smooth curve in time or a surface in space.
- Intrinsic GMRF models and random walk models
- Discretely observed integrated Wiener processes are GMRFs
- GMRFs models for coefficients in B-splines.

Usage of GMRFs (III)

Image analysis

- The first main area for spatial models
- Image restoration using the Wiener filter.
- Texture modelling and texture discrimination.
- Segmentation
- Deformable templates
- Object identification
- 3D reconstruction
- Restoring ultrasound images

Usage of GMRFs (IV)

Spatial statistics

- Latent GMRF model analysis of spatial binary data
- Geostatistics using GMRFs
- Analysis of data in social sciences and spatial econometrics
- Spatial and space-time epidemiology
- Environmental statistics
- Inverse problems

Main features

- Analytical tractable
- Modelling using conditional independence
- Merging GMRFs using conditioning (hierarchical models)
- Unified framework for
  - understanding
  - representation
  - computation using numerical methods for sparse matrices
- Fits nicely into the MCMC world
- Can construct fast and reliable block-MCMC algorithms.

Constraining the parametrisation to obey CI properties

Theorem

$$x_i \perp x_j \mid \mathbf{x}_{-ij} \quad \Longleftrightarrow \quad Q_{ij} = 0$$

Proof

Start with the *factorisation theorem*.

Theorem

$$x \perp y \mid z \quad \Longleftrightarrow \quad \pi(x, y, z) = f(x, z)g(y, z) \tag{1}$$

for some functions f and g, and for all z with $\pi(z) > 0$.

Example

For

$$\pi(x, y, z) \propto \exp(x + xz + yz)$$

on some bounded region, we see that $x \perp y|z$. However, this is not the case for

$$\pi(x, y, z) \propto \exp(xyz)$$

Proof...

$$\pi(x_i, x_j, \mathbf{x}_{-ij}) \propto \exp\Big(-\frac{1}{2}\sum_{k,l} x_k Q_{kl} x_l\Big)$$

$$\propto \exp\Big(-\frac{1}{2}\underbrace{x_i x_j (Q_{ij} + Q_{ji})}_{\text{term 1}} - \frac{1}{2}\underbrace{\sum_{\{k,l\}\neq\{i,j\}} x_k Q_{kl} x_l}_{\text{term 2}}\Big).$$

Term 1 involves $x_i x_j$ iff $Q_{ij} \neq 0$. Term 2 does not involve $x_i x_j$.

We now see that $\pi(x_i, x_j, \mathbf{x}_{-ij}) = f(x_i, \mathbf{x}_{-ij}) g(x_j, \mathbf{x}_{-ij})$ for some functions f and g, iff $Q_{ij} = 0$.
Use the factorisation theorem.

Interpretation of elements of **Q**

Let $\mathbf{x}$ be a GMRF wrt $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with mean $\boldsymbol{\mu}$ and precision matrix $\mathbf{Q} > 0$, then

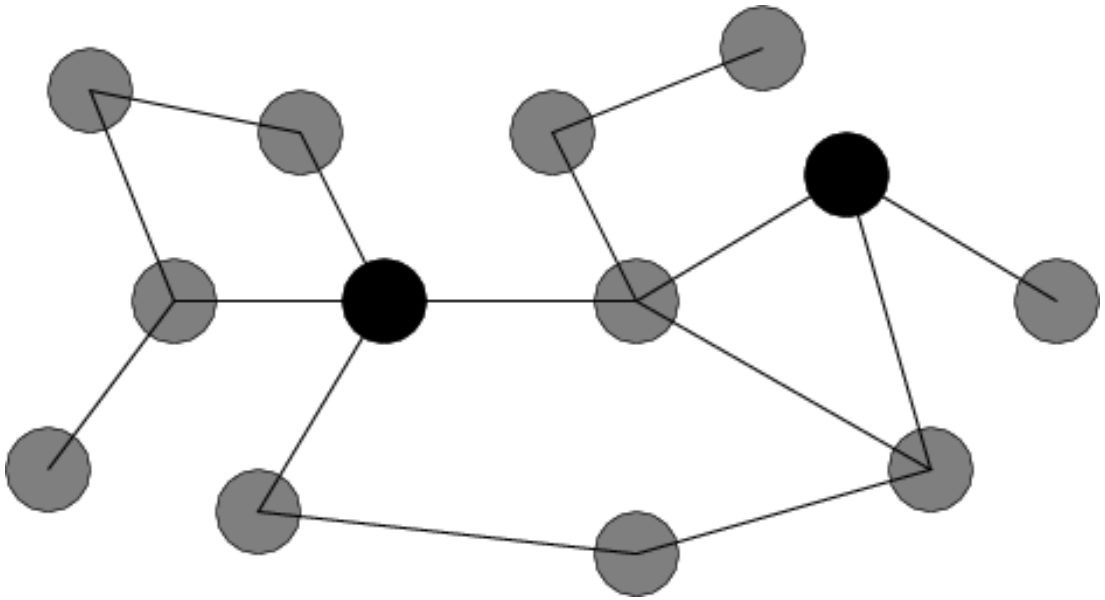
$$\begin{aligned}
\mathsf{E}(x_i \mid \mathbf{x}_{-i}) &= \mu_i - \frac{1}{Q_{ii}} \sum_{j:j\sim i} Q_{ij}(x_j - \mu_j), \\
\mathsf{Prec}(x_i \mid \mathbf{x}_{-i}) &= Q_{ii} \quad \text{and} \\
\mathsf{Corr}(x_i, x_j \mid \mathbf{x}_{-ij}) &= -\frac{Q_{ij}}{\sqrt{Q_{ii}Q_{jj}}}, \quad i \neq j.
\end{aligned}$$

Markov properties

Let $\mathbf{x}$ be a GMRF wrt $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Then the following are equivalent.

The pairwise Markov property:

$$x_i \perp x_j \mid \mathbf{x}_{-ij} \quad \text{if } \{i, j\} \notin \mathcal{E} \text{ and } i \neq j.$$

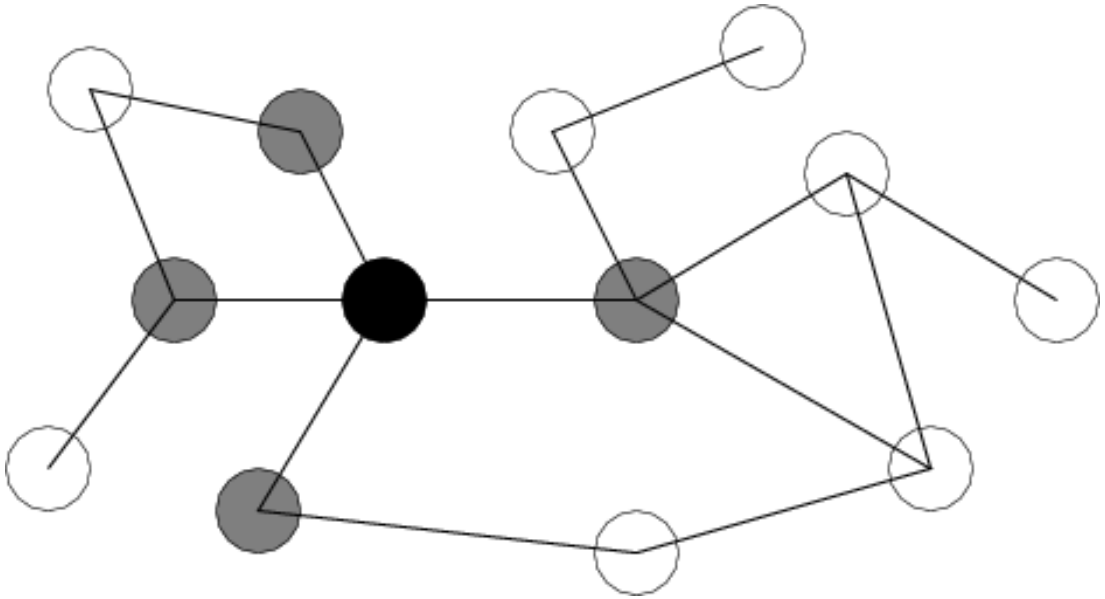


Markov properties

Let $\mathbf{x}$ be a GMRF wrt $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Then the following are equivalent.

The local Markov property:

$$x_i \perp \mathbf{x}_{-\{i,\mathsf{ne}(i)\}} \mid \mathbf{x}_{\mathsf{ne}(i)} \quad \text{for every } i \in \mathcal{V}.$$



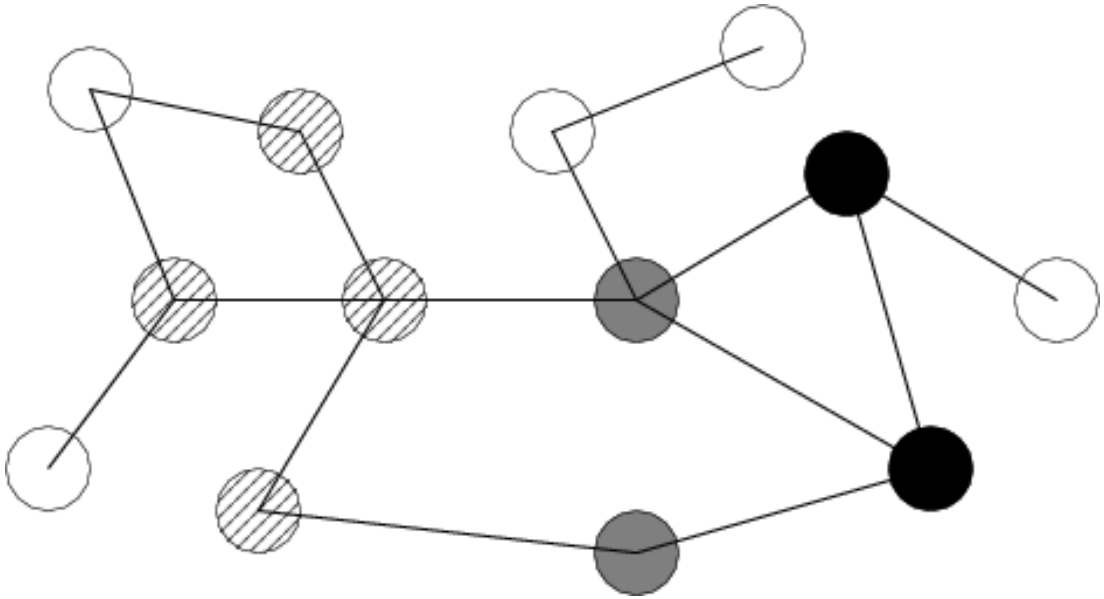
Markov properties

Let $\mathbf{x}$ be a GMRF wrt $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Then the following are equivalent.

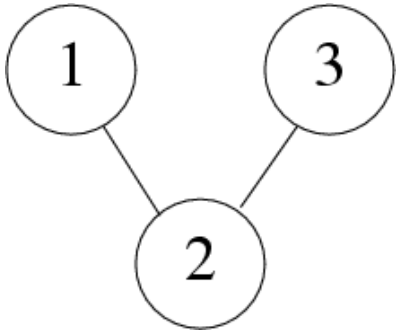
The global Markov property:

$$\mathbf{x}_A \perp \mathbf{x}_B \mid \mathbf{x}_C$$

for all disjoint sets A, B and C where C separates A and B, and A and B are non-empty.



(Induced) subgraph



Let $A \subset \mathcal{V}$

$\mathcal{G}^A$ denote the graph restricted to A.

- remove all nodes not belonging to A, and
- all edges where at least one node does not belong to A

Example

$\mathcal{A} = \{1, 2\}$, then

$$\mathcal{V}^A = \{1, 2\} \quad \text{and} \quad \mathcal{E}^A = \{\{1, 2\}\}$$

Conditional density I

Let $\mathcal{V} = A \cup B$ where $A \cap B = \emptyset$, and

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_A \\ \mathbf{x}_B \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_A \\ \boldsymbol{\mu}_B \end{pmatrix}, \quad \mathbf{Q} = \begin{pmatrix} \mathbf{Q}_{AA} & \mathbf{Q}_{AB} \\ \mathbf{Q}_{BA} & \mathbf{Q}_{BB} \end{pmatrix}.$$

Result: $\mathbf{x}_A | \mathbf{x}_B$ is then a GMRF wrt the subgraph $\mathcal{G}^A$ with parameters $\boldsymbol{\mu}_{A|B}$ and $\mathbf{Q}_{A|B} > 0$, where

$$\boldsymbol{\mu}_{A|B} = \boldsymbol{\mu}_A - \mathbf{Q}_{AA}^{-1}\mathbf{Q}_{AB}(\mathbf{x}_B - \boldsymbol{\mu}_B) \quad \text{and} \quad \mathbf{Q}_{A|B} = \mathbf{Q}_{AA}.$$

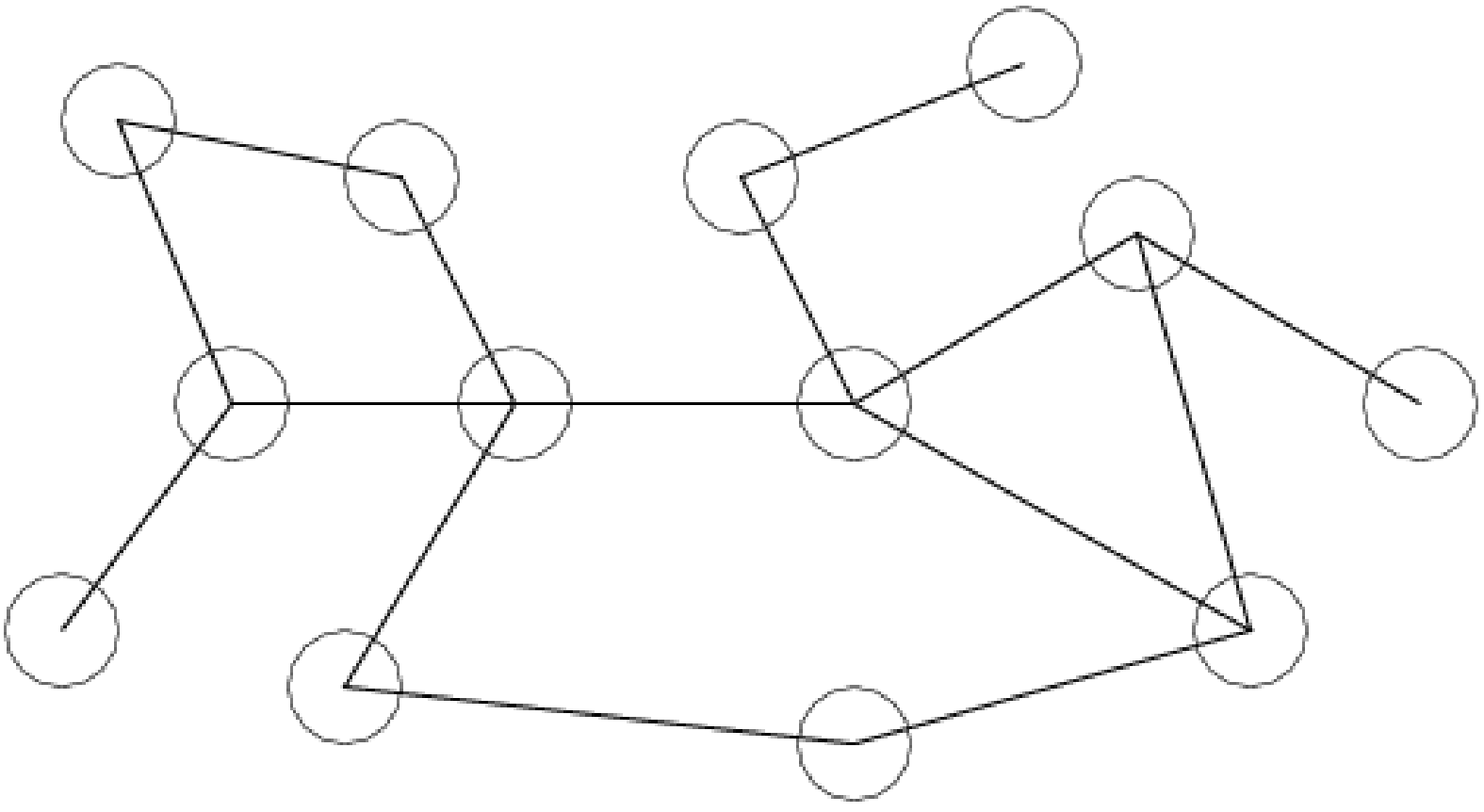
Conditional density II

$$\boldsymbol{\mu}_{A|B} = \boldsymbol{\mu}_A - \mathbf{Q}_{AA}^{-1}\mathbf{Q}_{AB}(\mathbf{x}_B - \boldsymbol{\mu}_B) \quad \text{and} \quad \mathbf{Q}_{A|B} = \mathbf{Q}_{AA}.$$

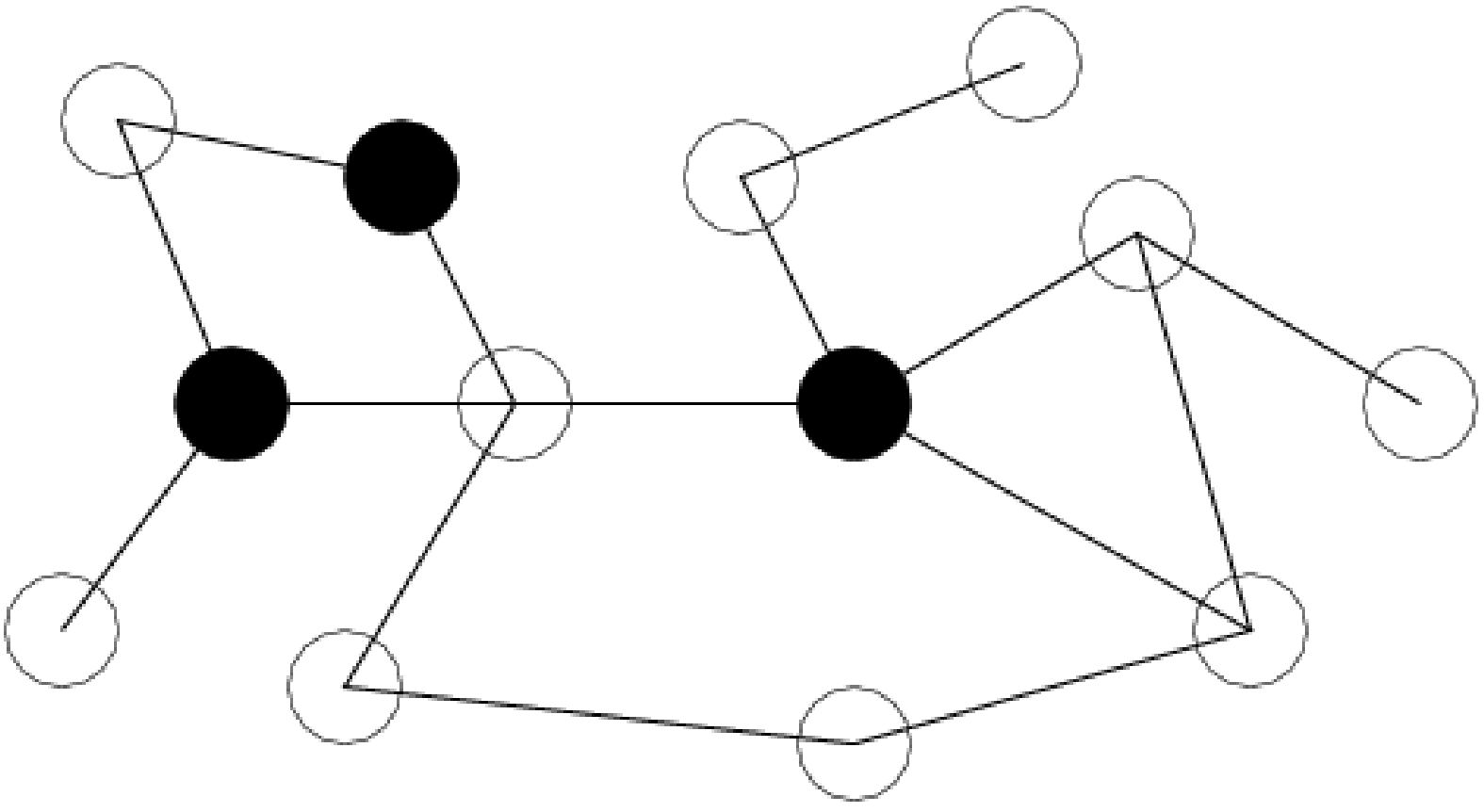
- Have explicit knowledge $\mathbf{Q}_{A|B}$ as the principal matrix $\mathbf{Q}_{AA}$.
- The subgraph $\mathcal{G}^A$ does not change the structure, only removes nodes and edges in A.
- The conditional mean only depends on nodes in $A \cup \text{ne}(A)$.
- If $\mathbf{Q}_{AA}$ is sparse, then $\boldsymbol{\mu}_{A|B}$ is the solution of a sparse linear system

$$\mathbf{Q}_{AA}(\boldsymbol{\mu}_{A|B} - \boldsymbol{\mu}_A) = -\mathbf{Q}_{AB}(\mathbf{x}_B - \boldsymbol{\mu}_B)$$

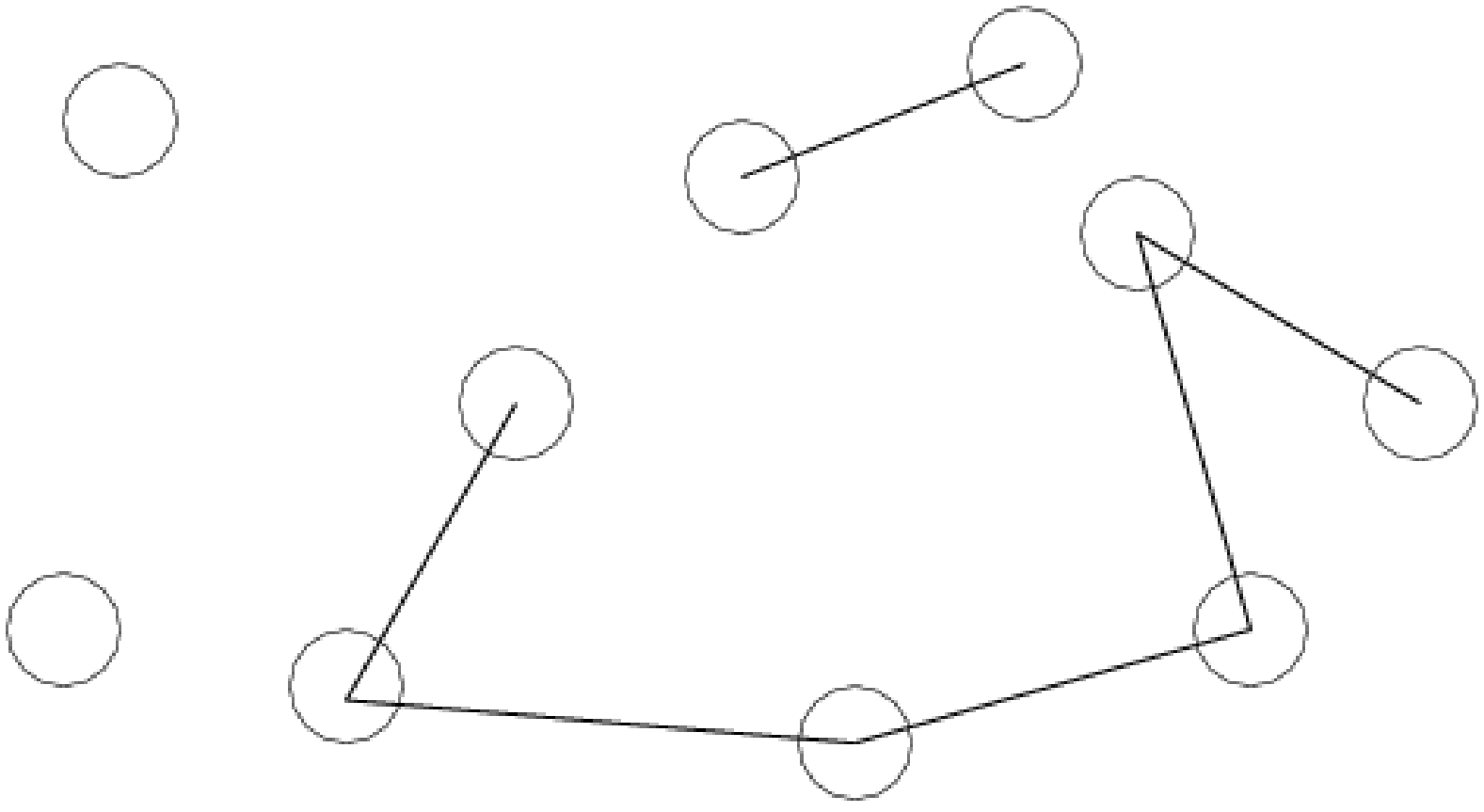
Conditional density III



Conditional density III



Conditional density III



Specification through full conditionals

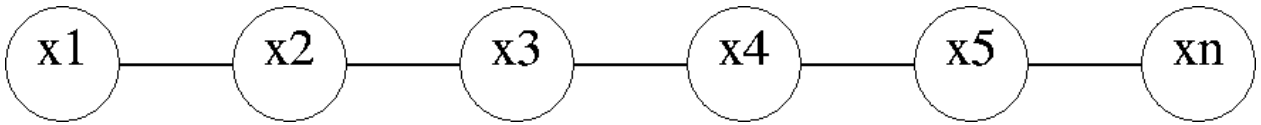
An alternative to specifying a GMRF by its mean and precision matrix, is to specify it implicitly through the full conditionals

$$\{\pi(x_i|\mathbf{x}_{-i}),\ i = 1, \ldots, n\}$$

This approach was pioneered by Besag (1974,1975)

Also known as *conditional autoregressions* or *CAR*-models.

Example: AR-process

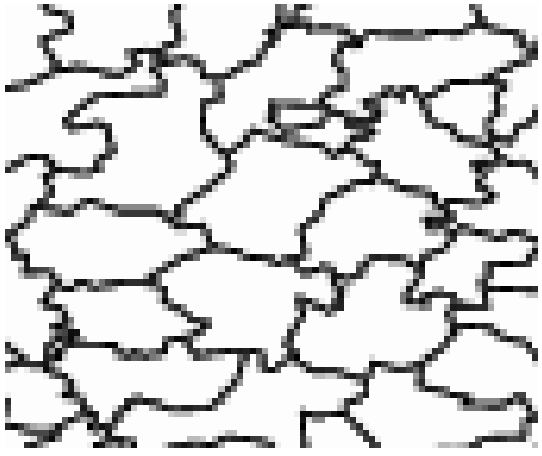


Specify full conditionals

$$E(x_1 \mid \mathbf{x}_{-1}) = 0.3x_2 \qquad Prec(x_1 \mid \mathbf{x}_{-1}) = 1$$
$$E(x_3 \mid \mathbf{x}_{-3}) = 0.2x_2 + 0.4x_4 \qquad Prec(x_3 \mid \mathbf{x}_{-3}) = 2$$

and so on.

Example: Spatial model



Specify full conditionals for each i (assuming zero mean)

$$E(x_i \mid \mathbf{x}_{-i}) = \sum_j w_{ij} x_j \qquad Prec(x_i \mid \mathbf{x}_{-i}) = \kappa_i$$

Potential problems

- Even though the "full conditionals" are well defined, are we sure that the *joint density* exists and is unique?
- What are the compatibility conditions required for a joint GMRF to exists with the prescribed Markov properties.

In general

Specify the full conditionals as normals with

$$\mathsf{E}(x_i \mid \mathbf{x}_{-i}) = \mu_i - \sum_{j:j\sim i} \beta_{ij}(x_j - \mu_j) \quad \text{and} \tag{2}$$

$$\mathsf{Prec}(x_i \mid \mathbf{x}_{-i}) = \kappa_i > 0 \tag{3}$$

for $i = 1, \ldots, n$, for some $\{\beta_{ij}, i \neq j\}$, and vectors $\boldsymbol{\mu}$ and $\boldsymbol{\kappa}$.

Clearly, $\sim$ is defined implicitly by the nonzero terms of $\{\beta_{ij}\}$.

$$\mathsf{E}(x_i \mid \mathbf{x}_{-i}) = \mu_i - \sum_{j:j\sim i} \beta_{ij}(x_j - \mu_j) \quad \text{and} \quad \mathsf{Prec}(x_i \mid \mathbf{x}_{-i}) = \kappa_i > 0$$

Must be a joint density $\pi(\mathbf{x})$ with these as the full conditionals.

Comparing term by term with (2):

$$Q_{ii} = \kappa_i, \quad \text{and} \quad Q_{ij} = \kappa_i \beta_{ij}$$

$\mathbf{Q}$ is symmetric, i.e.,

$$\kappa_i \beta_{ij} = \kappa_j \beta_{ji},$$

We have a candidate for a joint density provided $\mathbf{Q} > 0$.

Lemma (Brook's lemma)

Let $\pi(\mathbf{x})$ *be the density for* $\mathbf{x} \in \mathbb{R}^n$ *and define* $\Omega = \{\mathbf{x} \in \mathbb{R}^n : \pi(\mathbf{x}) > 0\}$. *Let* $\mathbf{x}, \mathbf{x}' \in \Omega$, *then*

$$\begin{aligned}
\frac{\pi(\mathbf{x})}{\pi(\mathbf{x}')} &= \prod_{i=1}^{n} \frac{\pi(x_i | x_1, \ldots, x_{i-1}, x'_{i+1}, \ldots, x'_n)}{\pi(x'_i | x_1, \ldots, x_{i-1}, x'_{i+1}, \ldots, x'_n)} \\
&= \prod_{i=1}^{n} \frac{\pi(x_i | x'_1, \ldots, x'_{i-1}, x_{i+1}, \ldots, x_n)}{\pi(x'_i | x'_1, \ldots, x'_{i-1}, x_{i+1}, \ldots, x_n)}.
\end{aligned}$$

Assume $\boldsymbol{\mu} = \mathbf{0}$ and fix $\mathbf{x}' = \mathbf{0}$. Then

$$\log \frac{\pi(\mathbf{x})}{\pi(\mathbf{0})} = -\frac{1}{2}\sum_{i=1}^{n} \kappa_i x_i^2 - \sum_{i=2}^{n}\sum_{j=1}^{i-1} \kappa_i \beta_{ij} x_i x_j. \tag{4}$$

and

$$\log \frac{\pi(\mathbf{x})}{\pi(\mathbf{0})} = -\frac{1}{2}\sum_{i=1}^{n} \kappa_i x_i^2 - \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \kappa_i \beta_{ij} x_i x_j. \tag{5}$$

Since (4) and (5) must be identical then $\kappa_i \beta_{ij} = \kappa_j \beta_{ji}$ for $i \neq j$, and

$$\log \pi(\mathbf{x}) = \text{const} - \frac{1}{2}\sum_{i=1}^{n} \kappa_i x_i^2 - \frac{1}{2}\sum_{i \neq j} \kappa_i \beta_{ij} x_i x_j;$$

hence $\mathbf{x}$ is zero mean multivariate normal provided $\mathbf{Q} > 0$.

Part II

Simulation algorithms for GMRFs

Outline I

Outline II

Sparse precision matrix

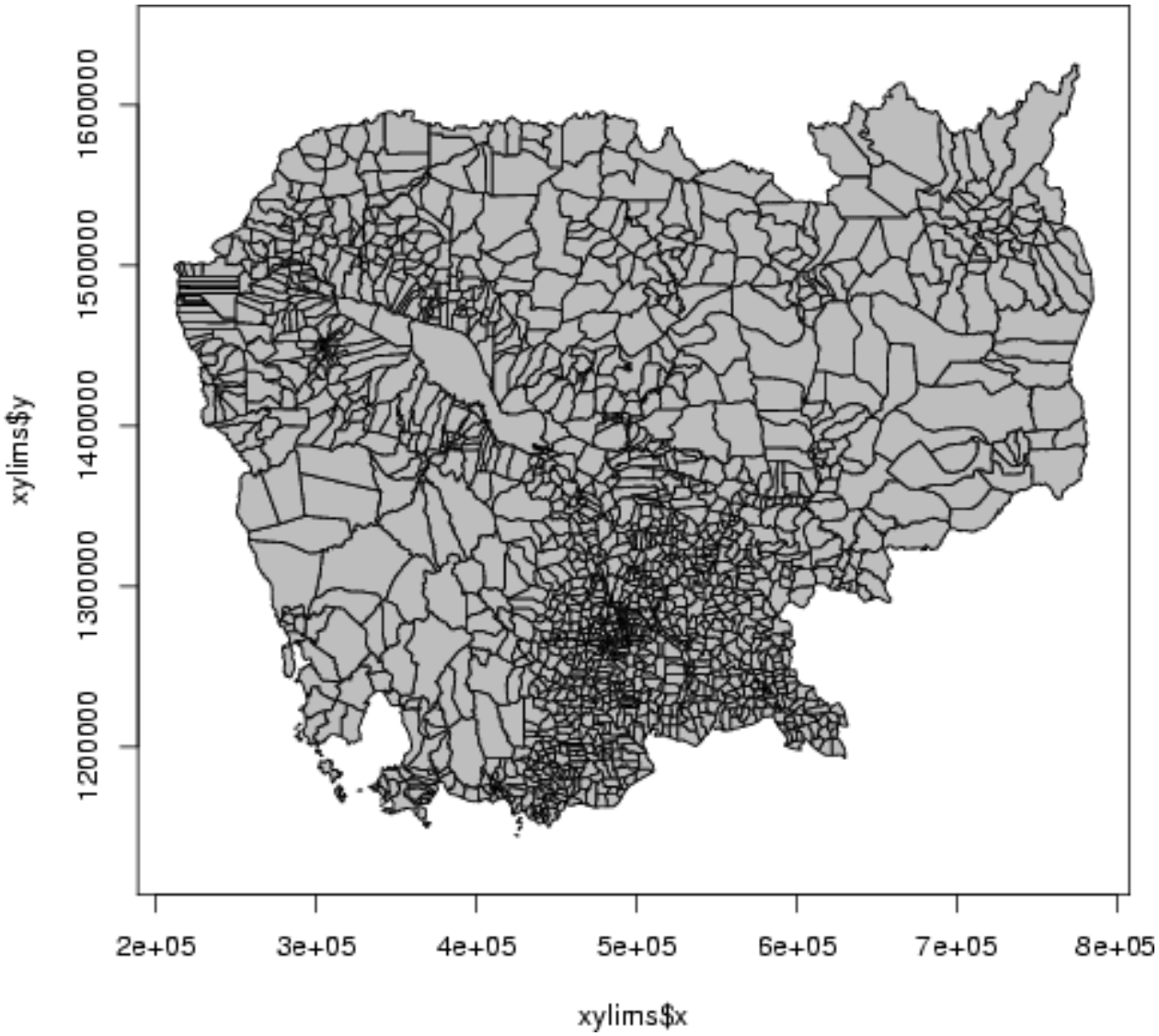
Recall that

$$E(x_i - \mu_i \mid \mathbf{x}_{-i}) = -\frac{1}{Q_{ii}} \sum_{j \sim i} Q_{ij}(x_i - \mu_j), \qquad \text{Prec}(x_i \mid \mathbf{x}_{-i}) = Q_{ii}$$

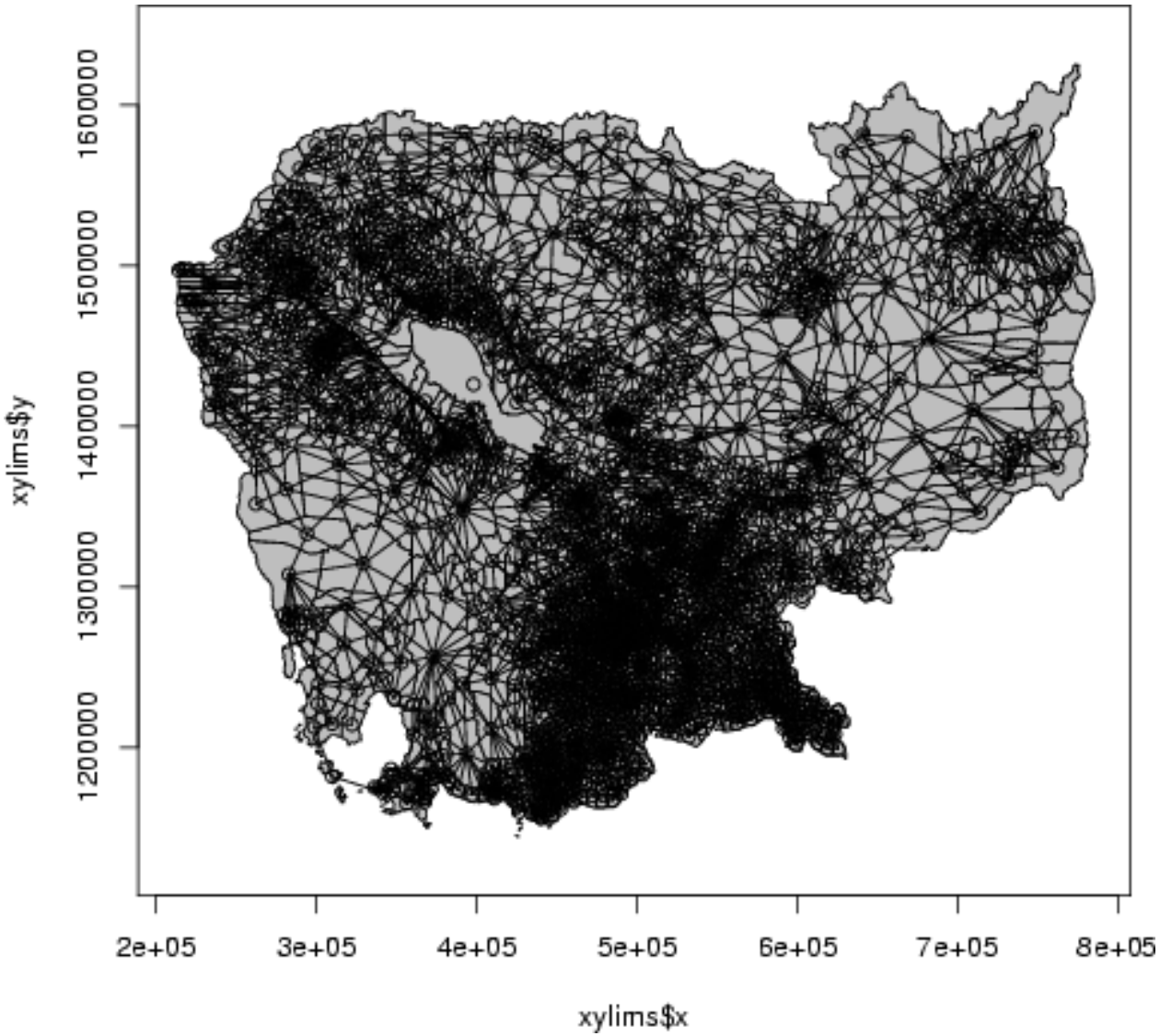
In most cases:

- Total number of neighbours is $\mathcal{O}(n)$.
- Only $\mathcal{O}(n)$ of the n^2 terms in $\mathbf{Q}$ will be non-zero.
- Use this to construct exact simulation algorithms for GMRFs, using numerical algorithms for sparse matrices.

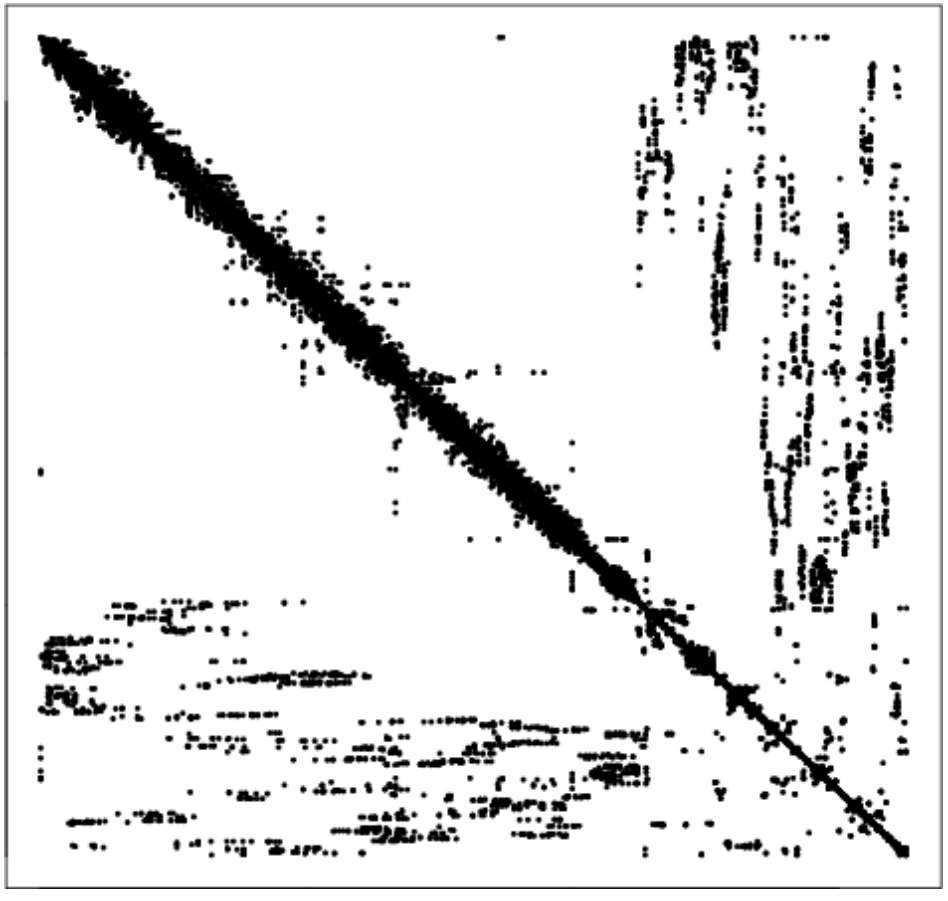
Example of a typical precision matrix



Example of a typical precision matrix



Example of a typical precision matrix



Each node have on the average 7 neighbours.

Simulation algorithms for GMRFs

Can we take advantage of the sparse structure of $\mathbf{Q}$?

- It is faster to factorise a sparse $\mathbf{Q}$ compared to a dense $\mathbf{Q}$.
- The speedup depends on the "pattern" in $\mathbf{Q}$, not only the number of non-zero terms.

Our task

- Formulate all algorithms to use only sparse matrices.
- Unconditional simulation
- Conditional simulation
  - Condition on a subset of variables
  - Condition on linear constraints
  - Condition on linear constraints with normal noise
- Evaluation of the log-density in all cases.

The result

In most cases, the cost is

- $\mathcal{O}(n)$ for temporal GMRFs
- $\mathcal{O}(n^{3/2})$ for spatial GMRFs
- $\mathcal{O}(n^2)$ for spatio-temporal GMRFs

including evaluation of the log-density.

Condition on k linear constraints, add $\mathcal{O}(k^3)$.

These are general algorithms only depending on the graph $\mathcal{G}$ not the numerical values in $\mathbf{Q}$.

The core is numerical algorithms for sparse matrices.

Cholesky factorisation

If $\mathbf{A} > 0$ be a $n \times n$ positive definite matrix, then there exists a unique Cholesky triangle $\mathbf{L}$, such that $\mathbf{L}$ is a lower triangular matrix, and

$$\mathbf{A} = \mathbf{L}\mathbf{L}^T$$

Computing $\mathbf{L}$ costs $n^3/3$ flops.

This factorisation is the basis for *solving* systems like

$$\mathbf{A}\mathbf{x} = \mathbf{b} \quad \text{or} \quad \mathbf{A}\mathbf{X} = \mathbf{B}$$

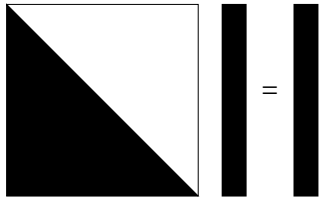
for k right hand sides, or equivalently, computing

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} \quad \text{or} \quad \mathbf{X} = \mathbf{A}^{-1}\mathbf{B}$$

Algorithm 1 Solving $\mathbf{Ax} = \mathbf{b}$ where $\mathbf{A} > 0$

1: Compute the Cholesky factorisation, $\mathbf{A} = \mathbf{L}\mathbf{L}^T$
2: Solve $\mathbf{Lv} = \mathbf{b}$
3: Solve $\mathbf{L}^T\mathbf{x} = \mathbf{v}$
4: **Return x**

Step 2 is called *forward-substitution* and cost $\mathcal{O}(n^2)$ flops.



The solution $\mathbf{v}$ is computed in a forward-loop

$$v_i = \frac{1}{L_{ii}}(b_i - \sum_{j=1}^{i-1} L_{ij} v_j), \quad i = 1, \ldots, n \tag{6}$$

Algorithm 1 Solving $\mathbf{Ax} = \mathbf{b}$ where $\mathbf{A} > 0$

1: Compute the Cholesky factorisation, $\mathbf{A} = \mathbf{LL}^T$
2: Solve $\mathbf{Lv} = \mathbf{b}$
3: Solve $\mathbf{L}^T\mathbf{x} = \mathbf{v}$
4: **Return x**

Step 3 is called *back-substitution* and costs $\mathcal{O}(n^2)$ flops.

The solution $\mathbf{x}$ is computed in a backward-loop

$$x_i = \frac{1}{L_{ii}}(v_i - \sum_{j=i+1}^{n} L_{ji}x_j), \quad i = n, \ldots, 1 \tag{7}$$

To compute $\mathbf{A}^{-1}\mathbf{B}$ where $\mathbf{B}$ is a $n \times k$ matrix, we do this by computing the solution $\mathbf{X}$ of

$$\mathbf{A}\mathbf{X}_j = \mathbf{B}_j$$

for each of the k columns of $\mathbf{X}$.

Algorithm 2 Solving $\mathbf{AX} = \mathbf{B}$ where $\mathbf{A} > 0$

1: Compute the Cholesky factorisation, $\mathbf{A} = \mathbf{L}\mathbf{L}^T$
2: **for** $j = 1$ to k **do**
3: Solve $\mathbf{L}\mathbf{v} = \mathbf{B}_j$
4: Solve $\mathbf{L}^T\mathbf{X}_j = \mathbf{v}$
5: **end for**
6: **Return X**

Sample $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{Q}^{-1})$

If $\mathbf{Q} = \mathbf{L}\mathbf{L}^T$ and $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, then $\mathbf{x}$ defined by

$$\mathbf{L}^T\mathbf{x} = \mathbf{z}$$

has covariance

$$\text{Cov}(\mathbf{x}) = \text{Cov}(\mathbf{L}^{-T}\mathbf{z}) = (\mathbf{L}\mathbf{L}^T)^{-1} = \mathbf{Q}^{-1}$$

Algorithm 3 Sampling $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{Q}^{-1})$

1: Compute the Cholesky factorisation, $\mathbf{Q} = \mathbf{L}\mathbf{L}^T$
2: Sample $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
3: Solve $\mathbf{L}^T\mathbf{v} = \mathbf{z}$
4: Compute $\mathbf{x} = \boldsymbol{\mu} + \mathbf{v}$
5: **Return** $\mathbf{x}$

The log-density is

$$\log \pi(\mathbf{x}) = -\frac{n}{2} \log 2\pi + \sum_{i=1}^{n} \log L_{ii} - \frac{1}{2} \underbrace{(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{Q} (\mathbf{x} - \boldsymbol{\mu})}_{=q}$$

If $\mathbf{x}$ is sampled, then

$$q = \mathbf{z}^T \mathbf{z}$$

otherwise, compute this term as

- $\mathbf{u} = \mathbf{x} - \boldsymbol{\mu}$
- $\mathbf{v} = \mathbf{Q}\mathbf{u}$
- $q = \mathbf{u}^T \mathbf{v}$

The canonical parameterisation

A GMRF $\mathbf{x}$ wrt $\mathcal{G}$ having a canonical parameterisation $(\mathbf{b}, \mathbf{Q})$, has density

$$\pi(\mathbf{x}) \propto \exp\Big(-\frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{b}^T\mathbf{x}\Big) \tag{8}$$

ie, precision matrix $\mathbf{Q}$ and mean $\mu = \mathbf{Q}^{-1}\mathbf{b}$.

Write this as

$$\mathbf{x} \sim \mathcal{N}_C(\mathbf{b}, \mathbf{Q}). \tag{9}$$

The relation to the Gaussian distribution, is that

$$\mathcal{N}(\boldsymbol{\mu}, \mathbf{Q}^{-1}) = \mathcal{N}_C(\mathbf{Q}\boldsymbol{\mu}, \mathbf{Q})$$

Conditional simulation of a GMRF

Decompose $\mathbf{x}$ as

$$\begin{pmatrix} \mathbf{x}_A \\ \mathbf{x}_B \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_A \\ \boldsymbol{\mu}_B \end{pmatrix}, \begin{pmatrix} \mathbf{Q}_{AA} & \mathbf{Q}_{AB} \\ \mathbf{Q}_{BA} & \mathbf{Q}_{BB} \end{pmatrix}^{-1} \right)$$

Then

$$\mathbf{x}_A \mid \mathbf{x}_B$$

has canonical parameterisation

$$\mathbf{x}_A - \boldsymbol{\mu}_A \mid \mathbf{x}_B \sim \mathcal{N}_C(-\mathbf{Q}_{AB}(\mathbf{x}_B - \boldsymbol{\mu}_B), \mathbf{Q}_{AA}) \tag{10}$$

Simulate using an algorithm for a canonical parameterisation.

Sample $\mathbf{x} \sim \mathcal{N}_C(\mathbf{b}, \mathbf{Q}^{-1})$

Recall that

$$\text{“}\mathcal{N}_C(\mathbf{b}, \mathbf{Q}) \quad = \quad \mathcal{N}(\mathbf{Q}^{-1}\mathbf{b}, \mathbf{Q}^{-1})\text{”}$$

so we need to compute the mean as well.

Algorithm 4 Sampling $\mathbf{x} \sim \mathcal{N}_C(\mathbf{b}, \mathbf{Q})$

1: Compute the Cholesky factorisation, $\mathbf{Q} = \mathbf{L}\mathbf{L}^T$
2: Solve $\mathbf{L}\mathbf{w} = \mathbf{b}$
3: Solve $\mathbf{L}^T\boldsymbol{\mu} = \mathbf{w}$
4: Sample $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Solve $\mathbf{L}^T\mathbf{v} = \mathbf{z}$
6: Compute $\mathbf{x} = \boldsymbol{\mu} + \mathbf{v}$
7: **Return** $\mathbf{x}$

Sampling from $\mathbf{x} \mid \mathbf{Ax} = \mathbf{e}$

- $\mathbf{A}$ is a $k \times n$ matrix, $0 < k < n$, with rank k,
- $\mathbf{e}$ is a vector of length k.

This case occurs quite frequently:

> *A sum-to-zero constraint corresponds to* $k = 1$, $\mathbf{A} = \mathbf{1}^T$ *and* $\mathbf{e} = 0$.

We will denote this problem sampling under a *hard constraint*.

Note that $\pi(\mathbf{x} \mid \mathbf{Ax})$ is singular with rank $n - k$.

The linear constraint makes the conditional distribution Gaussian

- but it is singular as the rank of the constrained covariance matrix is $n - k$
- more care must be exercised when sampling from this distribution.

We have that

$$\mathrm{E}(\mathbf{x} \mid \mathbf{A}\mathbf{x} = \mathbf{e}) = \boldsymbol{\mu} - \mathbf{A}\mathbf{Q}^{-1}(\mathbf{A}\mathbf{Q}^{-1}\mathbf{A}^T)^{-1}(\mathbf{A}\boldsymbol{\mu} - \mathbf{e}) \quad (11)$$

$$\mathrm{Cov}(\mathbf{x} \mid \mathbf{A}\mathbf{x} = \mathbf{e}) = \mathbf{Q}^{-1} - \mathbf{Q}^{-1}\mathbf{A}^T(\mathbf{A}\mathbf{Q}^{-1}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{Q}^{-1} \quad (12)$$

This is typically a dense-matrix case, which must be solved using general $\mathcal{O}(n^3)$ algorithms (see next frame for details).

We can sample from this distribution as follows. As the covariance matrix is singular we compute the eigenvalues and eigenvectors, and write the covariance matrix as $\mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$ where $\mathbf{V}$ have the eigenvectors on each row and $\mathbf{\Lambda}$ is a diagonal matrix with the eigenvalues on the diagonal. This corresponds to a different factorisation than the Cholesky triangle, but any matrix $\mathbf{C}$ which satisfy $\mathbf{C}\mathbf{C}^T = \mathbf{\Sigma}$, will do. Note that k of the eigenvalues are zero. We can produce a sample by computing $\mathbf{v} = \mathbf{C}\mathbf{z}$, where $\mathbf{C} = \mathbf{V}\mathbf{\Lambda}^{1/2}$, $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and then add the conditional mean. We can compute the log-density as

$$\begin{aligned}\log(\pi(\mathbf{x}|\mathbf{A}\mathbf{x} = \mathbf{e})) &= -\frac{n-k}{2}\log 2\pi - \frac{1}{2}\sum_{i:\Lambda_{ii}>0}\log \Lambda_{ii} \\ &\quad -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}^*)^T\mathbf{\Sigma}^-(\mathbf{x}-\boldsymbol{\mu}^*) \qquad (13)\end{aligned}$$

where $\boldsymbol{\mu}^*$ is (11), and $\mathbf{\Sigma}^- = \mathbf{V}\mathbf{\Lambda}^-\mathbf{V}^T$ where $(\mathbf{\Lambda}^-)_{ii}$ is Λ_{ii}^{-1} if $\Lambda_{ii} > 0$ and zero otherwise. In total, this is a quite computational demeaning procedure, as the algorithm is not able to take advantage of the sparse structure of $\mathbf{Q}$.

Conditioning via Kriging

There is an alternative algorithm which *correct* for the constraints, at nearly no costs if $k \ll n$.

Let $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{Q})$, then compute

$$\mathbf{x}^* = \mathbf{x} - \mathbf{Q}^{-1}\mathbf{A}^T(\mathbf{A}\mathbf{Q}^{-1}\mathbf{A}^T)^{-1}(\mathbf{A}\mathbf{x} - \mathbf{e}). \tag{14}$$

Now $\mathbf{x}^*$ has the correct conditional distribution!

$\mathbf{A}\mathbf{Q}^{-1}\mathbf{A}^T$ is a $k \times k$ matrix, hence its factorisation is fast to compute for small k.

Algorithm 5 Sampling $\mathbf{x} \mid \mathbf{A}\mathbf{x} = \mathbf{e}$ when $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{Q}^{-1})$

1: Compute the Cholesky factorisation, $\mathbf{Q} = \mathbf{L}\mathbf{L}^T$
2: Sample $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
3: Solve $\mathbf{L}^T\mathbf{v} = \mathbf{z}$
4: Compute $\mathbf{x} = \boldsymbol{\mu} + \mathbf{v}$
5: Compute $\mathbf{V}_{n\times k} = \mathbf{Q}^{-1}\mathbf{A}^T$ with Algorithm 2
6: Compute $\mathbf{W}_{k\times k} = \mathbf{A}\mathbf{V}$
7: Compute $\mathbf{U}_{k\times n} = \mathbf{W}^{-1}\mathbf{V}^T$ using Algorithm 2
8: Compute $\mathbf{c} = \mathbf{A}\mathbf{x} - \mathbf{e}$
9: Compute $\mathbf{x}^* = \mathbf{x} - \mathbf{U}^T\mathbf{c}$
10: **Return** $\mathbf{x}^*$

If $\mathbf{z} = \mathbf{0}$ in Algorithm 5, then $\mathbf{x}^*$ is the conditional mean.
Extra cost is only $\mathcal{O}(k^3)$ for large k!

Example

Let $x_1, \ldots, x_n$ be independent Gaussian variables with variance σ_i^2 and mean μ_i.
To sample $\mathbf{x}$ conditioned on

$$\sum x_i = 0$$

we sample first

$$x_i \sim \mathcal{N}(\mu_i, \sigma_i^2), \quad i = 1, \ldots, n$$

and compute the constrained sample $\mathbf{x}^*$, as

$$x_i^* = x_i - c\, \sigma_i^2 \tag{15}$$

where

$$c = \frac{\sum_j x_j}{\sum_j \sigma_j^2}$$

The log-density can be rapidly computed using the following identity,

$$\pi(\mathbf{x} \mid \mathbf{A}\mathbf{x}) = \frac{\pi(\mathbf{x})\pi(\mathbf{A}\mathbf{x} \mid \mathbf{x})}{\pi(\mathbf{A}\mathbf{x})}, \tag{16}$$

We can compute each term on the rhs easier than the lhs.

$\pi(\mathbf{x})$-term. This is a GMRF and the log-density is easy to compute using $\mathbf{L}$ computed in Algorithm 5 step 1.

The log-density can be rapidly computed using the following identity,

$$\pi(\mathbf{x} \mid \mathbf{A}\mathbf{x}) = \frac{\pi(\mathbf{x})\pi(\mathbf{A}\mathbf{x} \mid \mathbf{x})}{\pi(\mathbf{A}\mathbf{x})}, \tag{16}$$

We can compute each term on the rhs easier than the lhs.

$\pi(\mathbf{A}\mathbf{x} \mid \mathbf{x})$-term. This is a degenerate density, which is either zero or a constant,

$$\log \pi(\mathbf{A}\mathbf{x} \mid \mathbf{x}) = -\frac{1}{2} \log |\mathbf{A}\mathbf{A}^T| \tag{17}$$

The determinant of a $k \times k$ matrix can be found its Cholesky factorisation.

The log-density can be rapidly computed using the following identity,

$$\pi(\mathbf{x} \mid \mathbf{A}\mathbf{x}) = \frac{\pi(\mathbf{x})\pi(\mathbf{A}\mathbf{x} \mid \mathbf{x})}{\pi(\mathbf{A}\mathbf{x})}, \tag{17}$$

We can compute each term on the rhs easier than the lhs.

$\pi(\mathbf{Ax})$-term. $\mathbf{Ax}$ is Gaussian with mean $\mathbf{A}\boldsymbol{\mu}$ and covariance matrix $\mathbf{A}\mathbf{Q}^{-1}\mathbf{A}^T$ with Cholesky triangle $\widetilde{\mathbf{L}}$ available from Algorithm 5 step 7.

Sampling from $\mathbf{x} \mid \mathbf{Ax} = \boldsymbol{\epsilon}$

Let $\mathbf{x}$ be a GMRF which is observed by $\mathbf{e}$, where $\mathbf{e} \mid \mathbf{x} \sim \mathcal{N}(\mathbf{Ax}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$.

- $\mathbf{e}$ is a vector of length $k < n$
- $\mathbf{A}$ a $k \times n$ matrix rank k
- $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} > 0$ is the covariance matrix for the noise.

The conditional for $\mathbf{x}$, is

$$\log \pi(\mathbf{x} \mid \mathbf{e}) \doteq -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{Q}(\mathbf{x}-\boldsymbol{\mu}) - \frac{1}{2}(\mathbf{e}-\mathbf{Ax})^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}(\mathbf{e}-\mathbf{Ax}) \quad (18)$$

$$\mathbf{x} \mid \mathbf{e} \sim \mathcal{N}_C(\mathbf{Q}\boldsymbol{\mu} + \mathbf{A}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\mathbf{e}, \mathbf{Q} + \mathbf{A}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\mathbf{A}) \quad (19)$$

This precision matrix is most often a full matrix and the nice sparse structure of $\mathbf{Q}$ is lost.

Example

Observe the sum of x_i with unit variance noise, the posterior precision is

$$\mathbf{Q} + \mathbf{1}\mathbf{1}^T$$

which is a dense matrix.

In general, we have to sample from (18) using a general algorithm which is computational expensive for large n.

Alternative approach

Similar problem to sampling under a hard constraint $\mathbf{Ax} = \mathbf{e}$, but now "$\mathbf{e}$" is stochastic

$$\mathbf{Ax} = \boldsymbol{\epsilon} \quad \text{where} \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{e}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}). \tag{20}$$

We denote this case sampling under a *soft constraint*.

If we extend (14) to

$$\mathbf{x}^* = \mathbf{x} - \mathbf{Q}^{-1}\mathbf{A}^T(\mathbf{A}\mathbf{Q}^{-1}\mathbf{A}^T + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})^{-1}(\mathbf{Ax} - \boldsymbol{\epsilon}), \tag{21}$$

where $\boldsymbol{\epsilon}$ and $\mathbf{x}$ are samples from their respective distributions,

....then $\mathbf{x}^*$ has the correct conditional distribution.

Algorithm 6 Sampling $\mathbf{x} \mid \mathbf{Ax} = \boldsymbol{\epsilon}$ when $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{e}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$ and $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{Q}^{-1})$

1: Compute the Cholesky factorisation, $\mathbf{Q} = \mathbf{L}\mathbf{L}^T$
2: Sample $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
3: Solve $\mathbf{L}^T\mathbf{v} = \mathbf{z}$
4: Compute $\mathbf{x} = \boldsymbol{\mu} + \mathbf{v}$
5: Compute $\mathbf{V}_{n\times k} = \mathbf{Q}^{-1}\mathbf{A}^T$ using Algorithm 2 using $\mathbf{L}$
6: Compute $\mathbf{W}_{k\times k} = \mathbf{AV} + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$
7: Compute $\mathbf{U}_{k\times n} = \mathbf{W}^{-1}\mathbf{V}^T$ using Algorithm 2
8: Sample $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{e}, \mathbf{W})$ using the factorisation from step 7
9: Compute $\mathbf{c} = \mathbf{Ax} - \boldsymbol{\epsilon}$
10: Compute $\mathbf{x}^* = \mathbf{x} - \mathbf{U}^T\mathbf{c}$
11: **Return** $\mathbf{x}^*$

When $\mathbf{z} = \mathbf{0}$ and $\boldsymbol{\epsilon} = \mathbf{e}$ then $\mathbf{x}^*$ is the conditional mean.

The log-density is computed via

$$\pi(\mathbf{x} \mid \mathbf{e}) = \frac{\pi(\mathbf{x})\pi(\mathbf{e} \mid \mathbf{x})}{\pi(\mathbf{e})} \tag{22}$$

All Cholesky triangles required are available from the simulation algorithm.

$\pi(\mathbf{x})$-*term.* This is a GMRF.

$\pi(\mathbf{e} \mid \mathbf{x})$-*term.* $\mathbf{e} \mid \mathbf{x}$ is Gaussian with mean $\mathbf{Ax}$ and covariance $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$.

$\pi(\mathbf{e})$-*term.* $\mathbf{e}$ is Gaussian with mean $\mathbf{A}\boldsymbol{\mu}$ and covariance matrix $\mathbf{A}\mathbf{Q}^{-1}\mathbf{A}^T + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$.

Example

Let $x_1, \ldots, x_n$ be independent Gaussian variables with variance σ_i^2 and mean μ_i. We now observe

$$e \sim \mathcal{N}(\sum x_i, \sigma_\epsilon^2)$$

To sample from $\pi(\mathbf{x} \mid e)$ we sample first

$$x_i \sim \mathcal{N}(\mu_i, \sigma_i^2), \quad i = 1, \ldots, n$$

and then

$$\epsilon \sim \mathcal{N}(e, \sigma_\epsilon^2)$$

A conditional sample $\mathbf{x}^*$, is then

$$x_i^* = x_i - c\,\sigma_i^2, \quad \text{where} \quad c = \frac{\sum_j x_j - \epsilon}{\sum_j \sigma_j^2 + \sigma_\epsilon^2}$$

Part III

Numerical methods for sparse matrices

Outline I

Outline II

A numerical case-study

Numerical methods for sparse matrices

Have shown that computations on GMRFs can be expressed such that the main tasks are

1. compute the Cholesky factorisation of $\mathbf{Q} = \mathbf{L}\mathbf{L}^T$, and
2. solve $\mathbf{L}\mathbf{v} = \mathbf{b}$ and $\mathbf{L}^T\mathbf{x} = \mathbf{z}$.

The second task is much-faster than the first, but sparsity will be of advantage also here.

The goal is to explain

- **why** a sparse **Q** allow for fast factorisation
- **how** we can take advantage of it,
- **why** we gain if we permute the vertics before factorising the matrix,
- **how** statisticians can benefit for recent research in this area by the numerical mathematicians.

At the end we present a small case study factorising some typical matrices for GMRFs, using a classical and more recent methods for factorising matrices.

How to compute the Cholesky factorisation

$$Q_{ij} = \sum_{k=1}^{j} L_{ik} L_{jk}, \quad i \geq j.$$

$$v_i = Q_{ij} - \sum_{k=1}^{j-1} L_{ik} L_{jk}, \quad i \geq j,$$

Then

- $L_{jj}^2 = v_j$, and
- $L_{ij} L_{jj} = v_i$ for $i > j$.

If we know $\{v_i\}$ for fixed j, then

$$L_{jj} = \sqrt{v_j} \quad \text{and} \quad L_{ij} = v_i / \sqrt{v_j}, \quad \text{for} \quad i = j+1, \ldots, n.$$

This gives the jth column in $\mathbf{L}$.

Cholesky factorization of $\mathbf{Q} > 0$

Algorithm 7 Computing the Cholesky triangle **L** of **Q**

1: **for** $j = 1$ **to** n **do**
2: $\quad v_{j:n} = Q_{j:n,j}$
3: $\quad$ **for** $k = 1$ **to** $j - 1$ **do** $v_{j:n} = v_{j:n} - L_{j:n,k} L_{jk}$
4: $\quad L_{j:n,j} = v_{j:n} / \sqrt{v_j}$
5: **end for**
6: **Return L**

The overall process involves $n^3/3$ flops.

Interpretation of **L** *(I)*

Let $\mathbf{Q} = \mathbf{L}\mathbf{L}^T$, then the solution of

$$\mathbf{L}^T\mathbf{x} = \mathbf{z} \quad \text{where} \quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

is $\mathcal{N}(\mathbf{0}, \mathbf{Q}^{-1})$ distributed.

Since $\mathbf{L}$ is lower triangular then

$$x_n = \frac{1}{L_{nn}} z_n$$

$$x_{n-1} = \frac{1}{L_{n-1,n-1}} \left(z_{n-1} - L_{n,n-1} x_n\right)$$

$$\dots$$

Interpretation of **L** *(II)*

Theorem

Let $\mathbf{x}$ *be a GMRF wrt to the labelled graph* $\mathcal{G}$*, with mean* $\boldsymbol{\mu}$ *and precision matrix* $\mathbf{Q} > 0$*. Let* $\mathbf{L}$ *be the Cholesky triangle of* $\mathbf{Q}$*. Then for* $i \in \mathcal{V}$*,*

$$E(x_i \mid \mathbf{x}_{(i+1):n}) = \mu_i - \frac{1}{L_{ii}} \sum_{j=i+1}^{n} L_{ji}(x_j - \mu_j) \qquad \text{and}$$

$$Prec(x_i \mid \mathbf{x}_{(i+1):n}) = L_{ii}^2.$$

Determine the zero-pattern in **L** *(I)*

Theorem

Let $\mathbf{x}$ *be a GMRF wrt* $\mathcal{G}$*, with mean* $\boldsymbol{\mu}$ *and precision matrix* $\mathbf{Q} > 0$*. Let* $\mathbf{L}$ *be the Cholesky triangle of* $\mathbf{Q}$ *and define for* $1 \leq i < j \leq n$ *the set*

$$F(i,j) = \{i+1, \ldots, j-1, j+1, \ldots, n\},$$

which is the future of i *except* j*. Then*

$$x_i \perp x_j \mid \mathbf{x}_{F(i,j)} \quad \Longleftrightarrow \quad L_{ji} = 0.$$

If we can verify that L_{ji} is zero, we do not have to compute it when factorising $\mathbf{Q}$

Proof

Assume $\boldsymbol{\mu} = \mathbf{0}$ and fix $1 \leq i < j \leq n$. Theorem 11 gives that

$$\begin{aligned} \pi(\mathbf{x}_{i:n}) &\propto \exp\left(-\frac{1}{2}\sum_{k=i}^{n} L_{kk}^2 \left(x_k + \frac{1}{L_{kk}}\sum_{j=k+1}^{n} L_{jk}x_j\right)^2\right) \\ &= \exp\left(-\frac{1}{2}\mathbf{x}_{i:n}^T \mathbf{Q}^{(i:n)}\mathbf{x}_{i:n}\right), \end{aligned}$$

where $Q_{ij}^{(i:n)} = L_{ii}L_{ji}$. Then

$$x_i \perp x_j \mid \mathbf{x}_{F(i,j)} \quad \Longleftrightarrow \quad L_{ii}L_{ji} = 0,$$

which is equivalent to $L_{ji} = 0$ since $L_{ii} > 0$ as $\mathbf{Q}^{(i:n)} > 0$.

Determine the zero-pattern in **L** *(II)*

The global Markov property provide a simple and sufficient criteria for checking if $L_{ji} = 0$.

Corollary

If $F(i, j)$ separates $i < j$ in $\mathcal{G}$, then $L_{ji} = 0$.

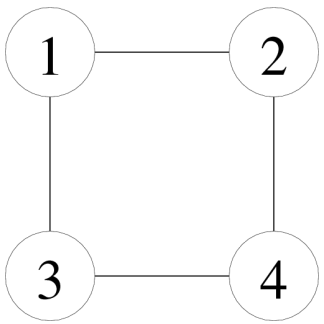
Corollary

If $i \sim j$ then $F(i, j)$ does not separates $i < j$.

The idea is simple

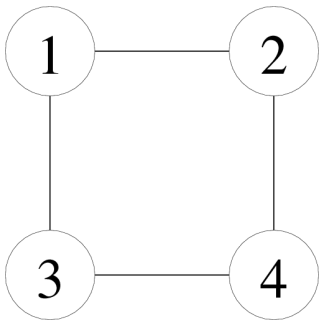
- Use the *global Markov property* to check if $L_{ji} = 0$.
- Compute only the non-zero terms in $\mathbf{L}$, so that $\mathbf{Q} = \mathbf{L}\mathbf{L}^T$.

Example



$$\mathbf{Q} = \begin{pmatrix} \times & \times & \times & \\ \times & \times & & \times \\ \times & & \times & \times \\ & \times & \times & \times \end{pmatrix} \qquad \mathbf{L} = \begin{pmatrix} \times & & & \\ \times & \times & & \\ \times & ? & \times & \\ ? & \times & \times & \times \end{pmatrix}$$

Example



$$\mathbf{Q} = \begin{pmatrix} \times & \times & \times & \\ \times & \times & & \times \\ \times & & \times & \times \\ & \times & \times & \times \end{pmatrix} \qquad \mathbf{L} = \begin{pmatrix} \times & & & \\ \times & \times & & \\ \times & \surd & \times & \\ & \times & \times & \times \end{pmatrix}$$

Example: AR(1)-process

$$x_t \mid \mathbf{x}_{1:(t-1)} \sim \mathcal{N}(\phi x_{t-1}, \sigma^2), \quad t = 1, \ldots, n$$

$$\mathbf{Q} = \begin{pmatrix} \times & \times & & & & & \\ \times & \times & \times & & & & \\ & \times & \times & \times & & & \\ & & \times & \times & \times & & \\ & & & \times & \times & \times & \\ & & & & \times & \times & \times \\ & & & & & \times & \times \end{pmatrix} \quad \mathbf{L} = \begin{pmatrix} \times & & & & & & \\ \times & \times & & & & & \\ & \times & \times & & & & \\ & & \times & \times & & & \\ & & & \times & \times & & \\ & & & & \times & \times & \\ & & & & & \times & \times \end{pmatrix}$$

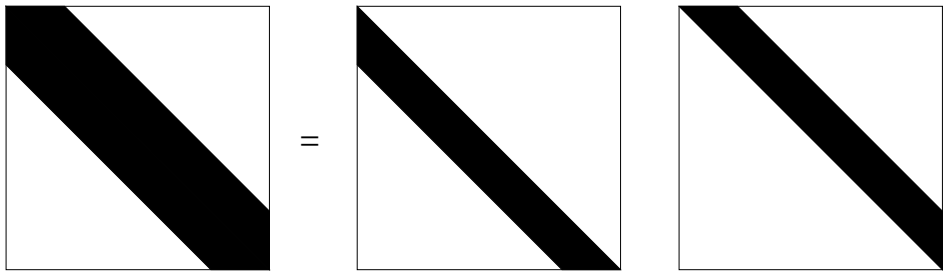
Bandwidth is preserved

Similarly, for an AR(p)-process

- $\mathbf{Q}$ have bandwidth p.
- $\mathbf{L}$ have lower-bandwidth p.

Theorem

Let $\mathbf{Q} > 0$ *be a band matrix with bandwidth* p *and dimension* n*, then the Cholesky triangle of* $\mathbf{Q}$ *has (lower) bandwidth* p*.*



...easy to modify existing Cholesky-factorisation code to use only entries where $|i - j| \leq p$.

Avoid computing L_{ij} and reading Q_{ij} for $|i-j| > p$.

Algorithm 8 Band-Cholesky factorization of $\mathbf{Q}$ with bandwidth p

for $j = 1$ **to** n **do**
  $\lambda = \min\{j+p, n\}$
  $v_{j:\lambda} = Q_{j:\lambda,j}$
  for $k = \max\{1, j-p\}$ **to** $j-1$ **do**
    $i = \min\{k+p, n\}$
    $v_{j:i} = v_{j:i} - L_{j:i,k} L_{jk}$
  end for
  $L_{j:\lambda,j} = v_{j:\lambda} / \sqrt{v_j}$
end for
Return L

Cost is now $n(p^2 + 3p)$ flops assuming $n \gg p$.

Reorder the vertices

We can permute the vertexes;

> *select one of the* $n!$ *possible permutations, define the corresponding permutation matrix* $\mathbf{P}$*, such that* $\mathbf{i}^P = \mathbf{P}\mathbf{i}$*, where* $\mathbf{i} = (1, \ldots, n)^T$*, is the new ordering of the vertexes.*

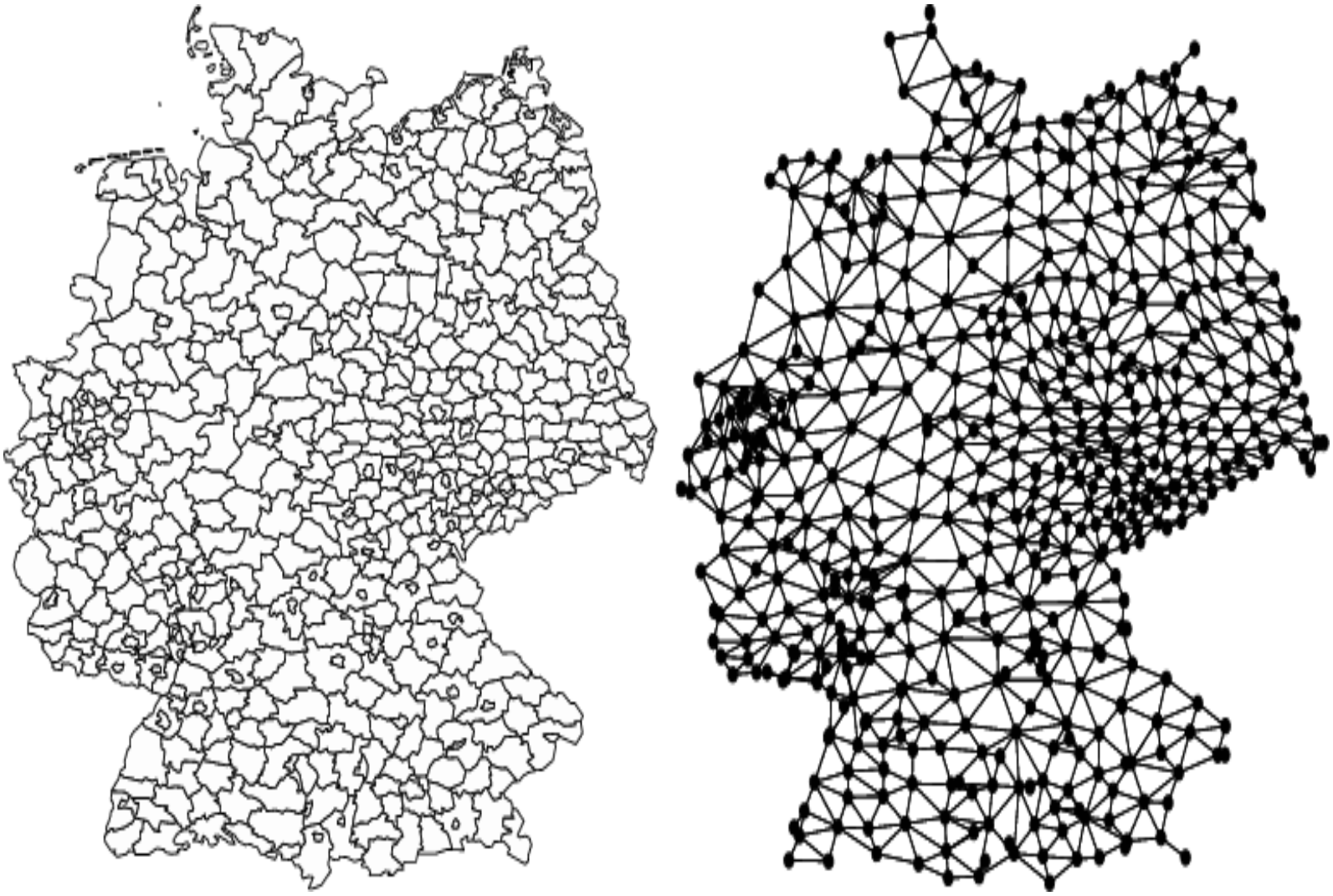
Chose $\mathbf{P}$, if possible, such that

$$\mathbf{Q}^P = \mathbf{P}\mathbf{Q}\mathbf{P}^T \tag{23}$$

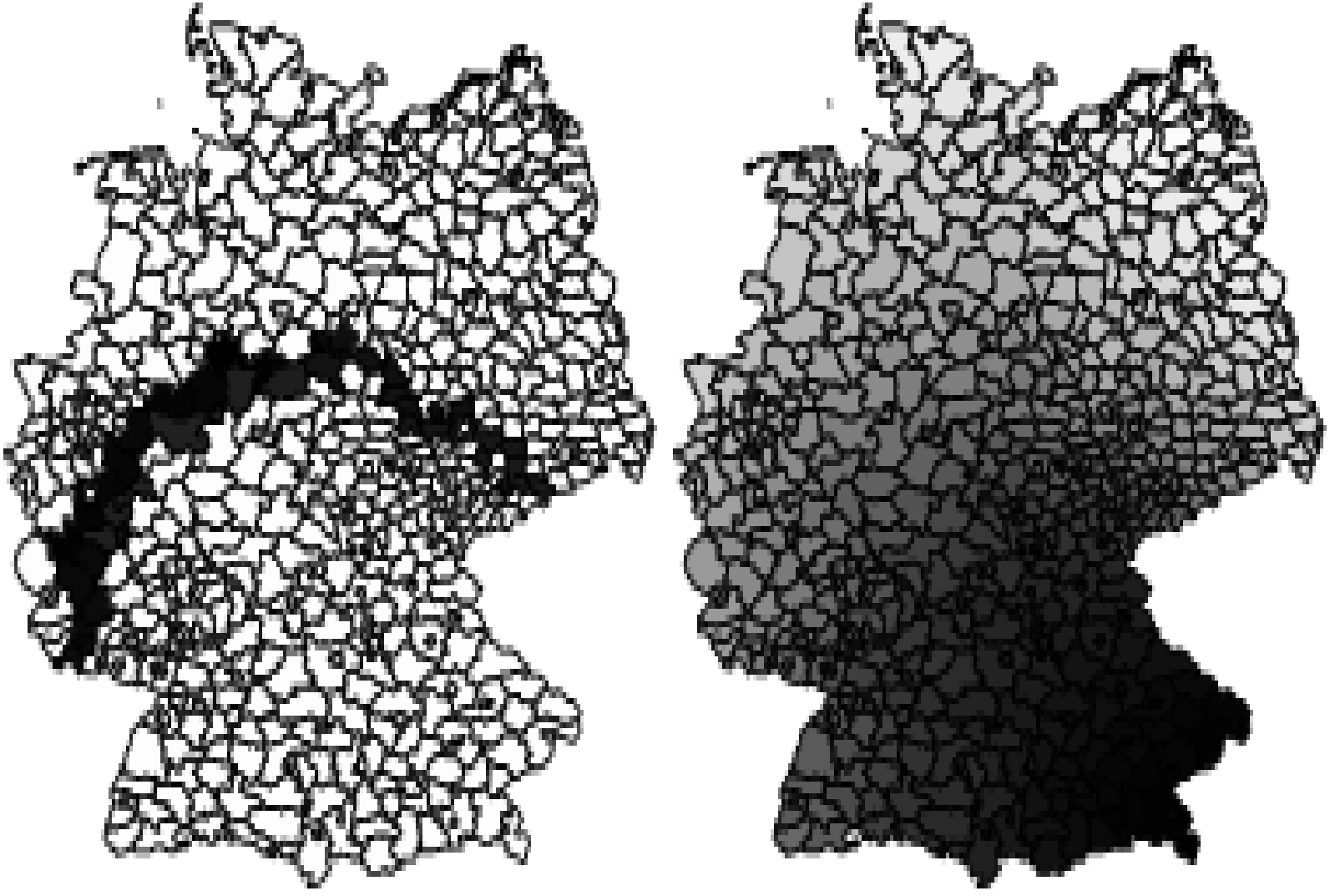
is a band-matrix with a small bandwidth.

- Impossible in general to obtain the optimal permutation, $n!$ is to large!
- A sub-optimal ordering will do as well.
- Solve $\mathbf{Q}\boldsymbol{\mu} = \mathbf{b}$ as follows:
  - $\mathbf{b}^P = \mathbf{P}\mathbf{b}$.
  - Solve $\mathbf{Q}^P \boldsymbol{\mu}^P = \mathbf{b}^P$
  - Map the solution back, $\boldsymbol{\mu} = \mathbf{P}^T \boldsymbol{\mu}^P$.

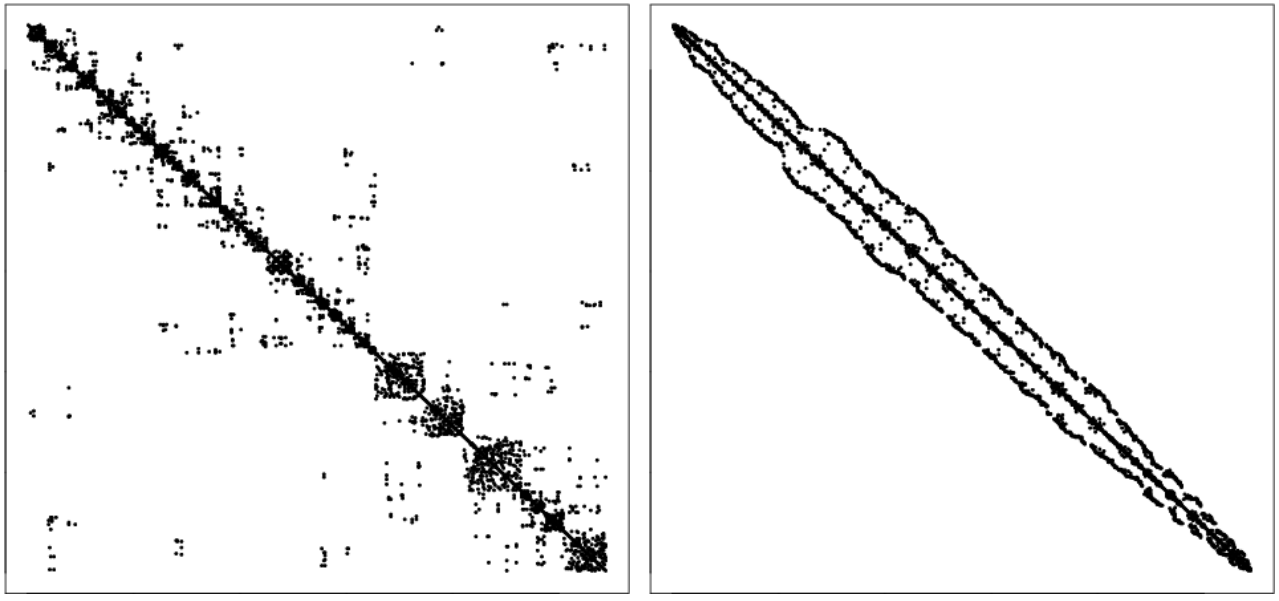
Reordering to band-matrices



Reordering to band-matrices

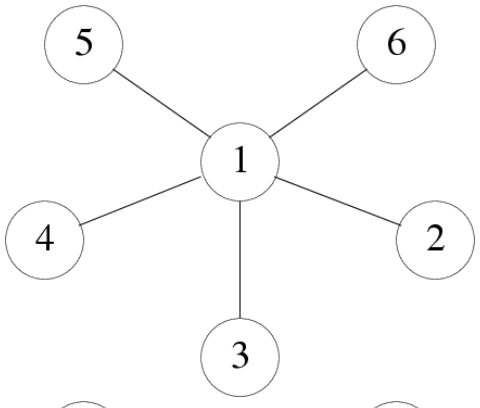


Reordering to band-matrices

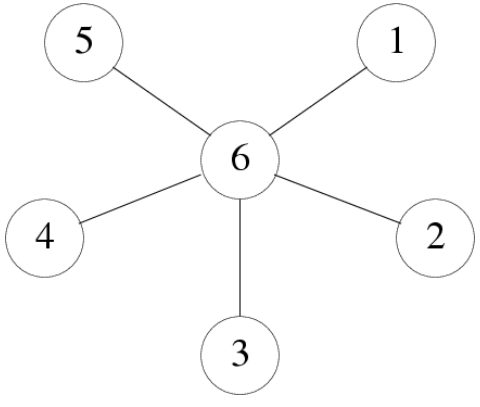


- factorisation of $\mathbf{Q}^P$ required about 0.0018 seconds
- Solving $\mathbf{Q}\boldsymbol{\mu} = \mathbf{b}$ required about 0.0006 seconds.
- on a 1 200MHz laptop.

More optimal reordering schemes



$$\begin{pmatrix} \times & \times & \times & \times & \times & \times \\ \times & \times & & & & \\ \times & & \times & & & \\ \times & & & \times & & \\ \times & & & & \times & \\ \times & & & & & \times \end{pmatrix} \quad \begin{pmatrix} \times & & & & & \\ \times & \times & & & & \\ \times & \surd & \times & & & \\ \times & \surd & \surd & \times & & \\ \times & \surd & \surd & \surd & \times & \\ \times & \surd & \surd & \surd & \surd & \times \end{pmatrix}$$



$$\begin{pmatrix} \times & & & & & \times \\ & \times & & & & \times \\ & & \times & & & \times \\ & & & \times & & \times \\ & & & & \times & \times \\ \times & \times & \times & \times & \times & \times \end{pmatrix} \quad \begin{pmatrix} \times & & & & & \\ & \times & & & & \\ & & \times & & & \\ & & & \times & & \\ & & & & \times & \\ \times & \times & \times & \times & \times & \times \end{pmatrix}$$

Nested dissection reordering (I)

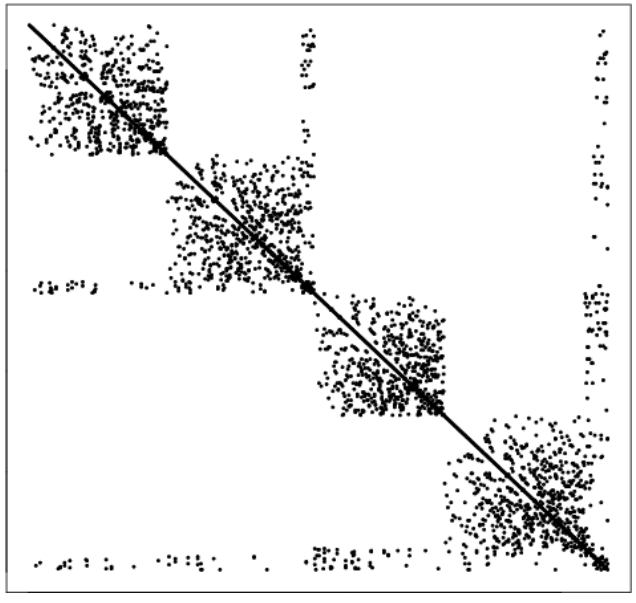
The idea generalise as follows.

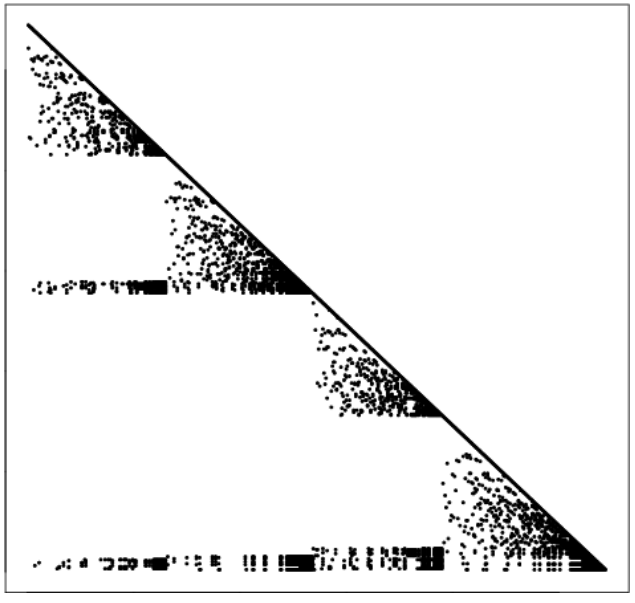
- Select a (small) set of nodes whose removal divides the graph into two disconnected subgraphs of almost equal size.
- Order the nodes chosen *after* ordering all the nodes in both subgraphs.
- Apply this procedure recursively to the nodes in each subgraph.

Costs in the spatial case

- Factorisation $\mathcal{O}(n^{3/2})$
- Fill-in $\mathcal{O}(n \log n)$
- Optimal in the order sense.

Nested dissection reordering (II)





Statisticians and numerical methods for sparse matrices

Gupta (2002) summarises his findings on recent advances for sparse linear solvers:

> *. . . recent sparse solvers have significantly improved the state of the art of the direct solution of general sparse systems.*
> *. . . recent years have seen some remarkable advances in the general sparse direct-solver algorithms and software.*

- Good news
- We just use their software

A numerical case-study of typical GMRFs

Divide GMRF models into three categories.

1. GMRF models in time or on a line; auto-regressive models and models for smooth functions.
   Neighbours to x_t are then those $\{x_s\}$ such that $|s - t| \leq p$.
2. Spatial GMRF models; regular lattice, or irregular lattice induced by a tessellation or regions of land.
   Neighbours to x_i (spatial index i), are those j spatially "close" to i, where "close" is defined from its context.
3. Spatio-temporal GMRF models. Often an extension of spatial models to also include dynamic changes.

Also include some "global" nodes:

Let $\mathbf{x}$ be a GMRF with a common mean μ

$$\mathbf{x}|\mu \sim \mathcal{N}(\mu\mathbf{1}, \mathbf{Q}^{-1}) \tag{24}$$

Assume $\mu \sim \mathcal{N}(0, \sigma^2)$

...then $(\mathbf{x}, \mu)$ is also a GMRF where the node μ is neighbour with all x_i's.

Two different algorithms

1. The band-Cholesky factorisation (BCF) as in Algorithm 8. Here we use the LAPACK-routines `DPBTRF` and `DTBSV`, for the factorisation and the forward/back-substitution, respectively, and the Gibbs-Poole-Stockmeyer algorithm for bandwidth reduction.
2. The Multifrontal Supernodal Cholesky factorisation (MSCF) implementation in the library `TAUCS` using the nested dissection reordering from the library `METIS`.

Both solvers are available in the `GMRFLib`-library

The tasks we want to investigate are

1. factorising $\mathbf{Q}$ into $\mathbf{L}\mathbf{L}^T$, and
2. solving $\mathbf{L}\mathbf{L}^T\boldsymbol{\mu} = \mathbf{b}$.

Producing a random sample from the GMRF, is half the cost of solving the linear system in step 2.

All tests reported here, we conducted on a $1\,200 MHz$ laptop with $512 Mb$ memory running Linux.

New machines are nearly $3 - 10$ times as fast...

GMRF models in time

Let **Q** be a band-matrix with bandwidth p and dimension n.
For such a problem, using BCF will be (theoretically) optimal, as the fillin will be be zero.

	$n = 10^3$		$n = 10^4$		$n = 10^5$	
CPU-time	$p = 5$	$p = 25$	$p = 5$	$p = 25$	$p = 5$	$p = 25$
Factorise	0.0005	0.0019	0.0044	0.0271	0.0443	0.2705
Solve	0.0000	0.0004	0.0031	0.0109	0.0509	0.1052

"Long and thin" are fast!
The MSCF is less optimal for band-matrices: fillin and more complicated data-structures.

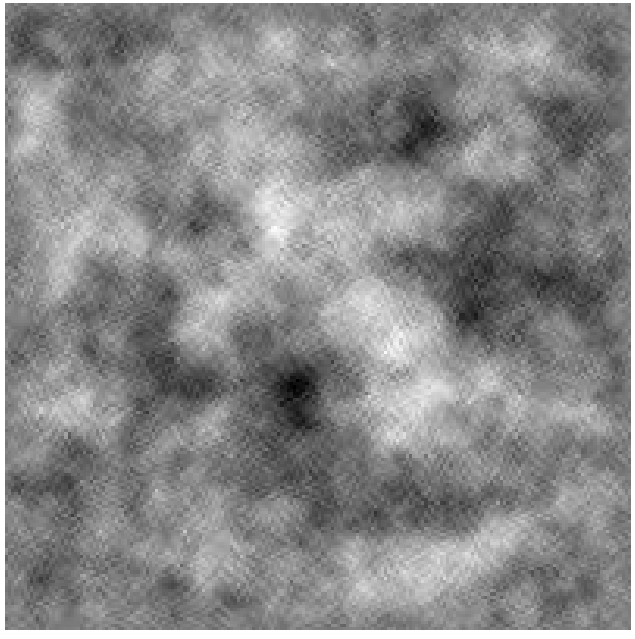
Add 10 global nodes:

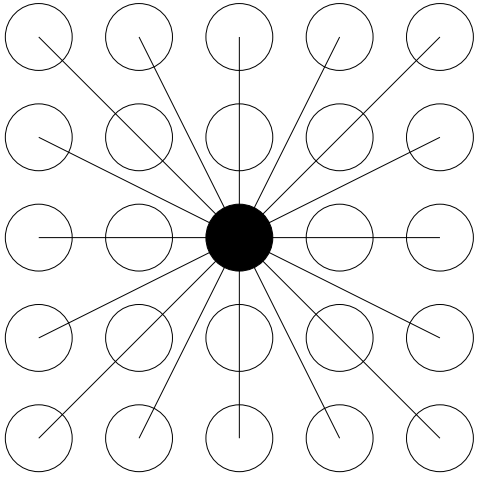
	$n = 10^3$		$n = 10^4$		$n = 10^5$	
CPU-time	$p = 5$	$p = 25$	$p = 5$	$p = 25$	$p = 5$	$p = 25$
Factorise	0.0119	0.0335	0.1394	0.4085	1.6396	4.1679
Solve	0.0007	0.0035	0.0138	0.0306	0.1541	0.3078

- The fillin $\approx pn$.
- The nested dissection ordering give good results in all cases considered so far.

Spatial GMRF models

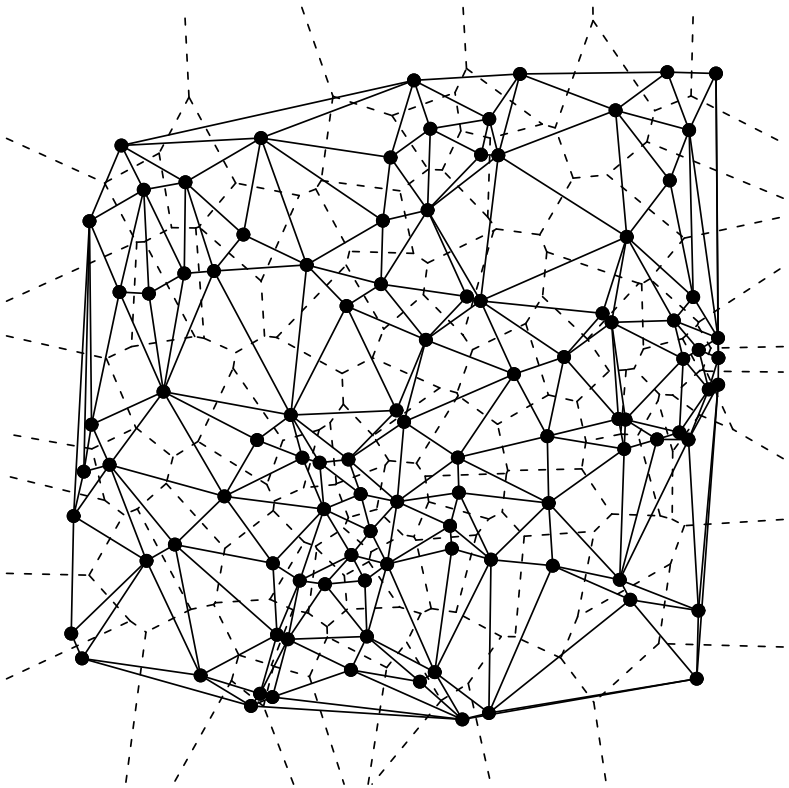
Regular lattice





Spatial GMRF models

Irregular lattice

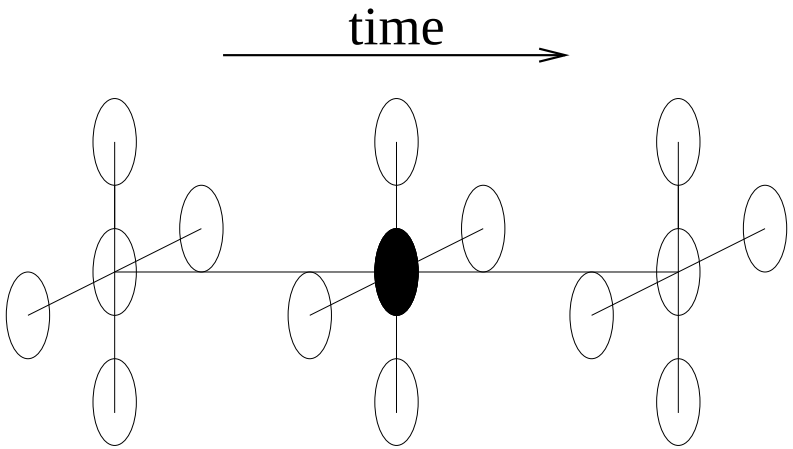


		$n = 100^2$		$n = 150^2$		$n = 200^2$	
CPU-time	Method	3×3	5×5	3×3	5×5	3×3	5×5
Factorise	BCF	0.51	1.02	2.60	4.93	13.30	38.12
	MSCF	0.17	0.62	0.55	1.92	1.91	4.90
Solve	BCF	0.03	0.05	0.10	0.16	0.24	0.43
	MSCF	0.01	0.04	0.04	0.11	0.08	0.21

- For the largest lattice the MSCF really outperform the BCF.
- The reason is the $\mathcal{O}(n^{3/2})$ cost for MSCF compared to $\mathcal{O}(n^2)$ for the BCF.

Spatio-temporal GMRF models

Spatio-temporal GMRF models is often an extension of spatial GMRF models to account to time variation.



Use this model and the graph of Germany, for $T = 10$ and $T = 100$.

			10 global nodes	
CPU-time	$T = 10$	$T = 100$	$T = 10$	$T = 100$
Factorise	0.25	39.96	0.31	39.22
Solve	0.02	0.42	0.02	0.42

- The results shows a quite heavy dependency on T.
- Spatio-Temporal GMRFs are more demanding than spatial ones, $\mathcal{O}(n^2)$ for a $n^{1/3} \times n^{1/3} \times n^{1/3}$-cube.

Part IV

Intrinsic GMRFs

Outline I

Intrinsic GMRFs (IGMRF)

- IGMRFs are *improper*, i.e., they have precision matrices not of full rank.
- Often used as prior distributions in various applications.
- Of particular importance are IGMRFs that are invariant to any trend that is a polynomial of the locations of the nodes up to a specific order.

Improper GMRF

Definition

Let $\mathbf{Q}$ be an $n \times n$ SPSD matrix with rank $n - k > 0$. Then $\mathbf{x} = (x_1, \ldots, x_n)^T$ is an improper GMRF of rank $n - k$ with parameters $(\boldsymbol{\mu}, \mathbf{Q})$, if its density is

$$\pi(\mathbf{x}) = (2\pi)^{\frac{-(n-k)}{2}} (|\mathbf{Q}|^*)^{1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{Q}(\mathbf{x} - \boldsymbol{\mu})\right). \quad (25)$$

Further, $\mathbf{x}$ is an improper GMRF wrt to the labelled graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where

$$Q_{ij} \neq 0 \Longleftrightarrow \{i, j\} \in \mathcal{E} \quad \text{for all} \quad i \neq j.$$

$|\cdot|^*$ denote the generalised determinant: product of the non-zero eigenvalues.

Markov properties

The Markov properties are interpreted as those obtained from the limit of a proper density.

Let the columns of $\mathbf{A}^T$ span the null space of $\mathbf{Q}$

$$\mathbf{Q}(\gamma) = \mathbf{Q} + \gamma \mathbf{A}^T \mathbf{A}. \tag{26}$$

$\mathbf{Q}(\gamma)$ tends to the corresponding one in $\mathbf{Q}$ as $\gamma \to 0$.

$$\mathsf{E}(x_i \mid \mathbf{x}_{-i}) = \mu_i - \frac{1}{Q_{ii}} \sum_{j \sim i} Q_{ij}(x_j - \mu_j)$$

is interpreted as $\gamma \to 0$.

IGMRF of first order

An intrinsic GMRF of first order is an improper GMRF of rank $n-1$ *where* $\mathbf{Q1} = \mathbf{0}$.

The condition $\mathbf{Q1} = \mathbf{0}$ means that

$$\sum_j Q_{ij} = 0, \quad i = 1, \ldots, n$$

Let $\boldsymbol{\mu} = \mathbf{0}$ so

$$\mathrm{E}(x_i \mid \mathbf{x}_{-i}) = -\frac{1}{Q_{ii}} \sum_{j:j\sim i} Q_{ij} x_j \tag{27}$$

$$-\sum_{j:j\sim i} Q_{ij}/Q_{ii} = 1$$

- The conditional mean of x_i is a weighted mean of its neighbours.
- No shrinking towards an overall level.
- Many IGMRFs are constructed such that the *deviation* from the overall level is a smooth curve in time or a smooth surface in space.

Definition (Forward difference)

Define the first-order forward difference of a function $f(\cdot)$ as

$$\Delta f(z) = f(z+1) - f(z).$$

Higher-order forward differences are defined recursively:

$$\Delta^k f(z) = \Delta \Delta^{k-1} f(z)$$

so

$$\Delta^2 f(z) = f(z+2) - 2f(z+1) + f(z) \tag{28}$$

and in general for $k = 1, 2, \ldots,$

$$\Delta^k f(z) = (-1)^k \sum_{j=0}^{k} (-1)^j \binom{k}{j} f(z+j).$$

For a vector $\mathbf{z} = (z_1, z_2, \ldots, z_n)^T$, $\Delta\mathbf{z}$ has elements

$$\Delta z_i = z_{i+1} - z_i, \quad i = 1, \ldots, n-1$$

The forward difference of kth order as an approximation to the kth derivative of $f(z)$,

$$f'(z) = \lim_{h \to 0} \frac{f(z+h) - f(z)}{h}$$

IGMRFs of first order on the line

Location of the node i is i (think "time").

Assume *independent increments*

$$\Delta x_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \kappa^{-1}), \quad i = 1, \ldots, n-1, \qquad \text{so} \tag{29}$$

$$x_j - x_i \sim \mathcal{N}(0, (j-i)\kappa^{-1}) \text{ for } i < j. \tag{30}$$

If the intersection between $\{i, \ldots, j\}$ and $\{k, \ldots, l\}$ is empty for $i < j$ and $k < l$, then

$$\text{Cov}(x_j - x_i, x_l - x_k) = 0. \tag{31}$$

Properties coincide with those of a *Wiener process*.

The density for $\mathbf{x}$ is

$$\pi(\mathbf{x} \mid \kappa) \propto \kappa^{(n-1)/2} \exp\left(-\frac{\kappa}{2}\sum_{i=1}^{n-1}(\Delta x_i)^2\right) \tag{32}$$

$$= \kappa^{(n-1)/2} \exp\left(-\frac{\kappa}{2}\sum_{i=1}^{n-1}(x_{i+1} - x_i)^2\right), \tag{33}$$

or

$$\pi(\mathbf{x}) \propto \kappa^{(n-1)/2} \exp\left(-\frac{\kappa}{2}\mathbf{x}^T\mathbf{R}\mathbf{x}\right) \tag{34}$$

The *structure matrix*

$$\mathbf{R} = \begin{pmatrix} 1 & -1 & & & & & \\ -1 & 2 & -1 & & & & \\ & -1 & 2 & -1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & -1 & 2 & -1 & \\ & & & & -1 & 2 & -1 \\ & & & & & -1 & 1 \end{pmatrix}. \tag{35}$$

- The $n-1$ independent increments ensure that the rank of $\mathbf{Q}$ is $n-1$.
- Denote this model by RW1(κ) or short RW1.

Full conditionals

$$x_i \mid \mathbf{x}_{-i}, \kappa \sim \mathcal{N}(\frac{1}{2}(x_{i-1} + x_{i+1}),\ 1/(2\kappa)), \quad 1 < i < n, \tag{36}$$

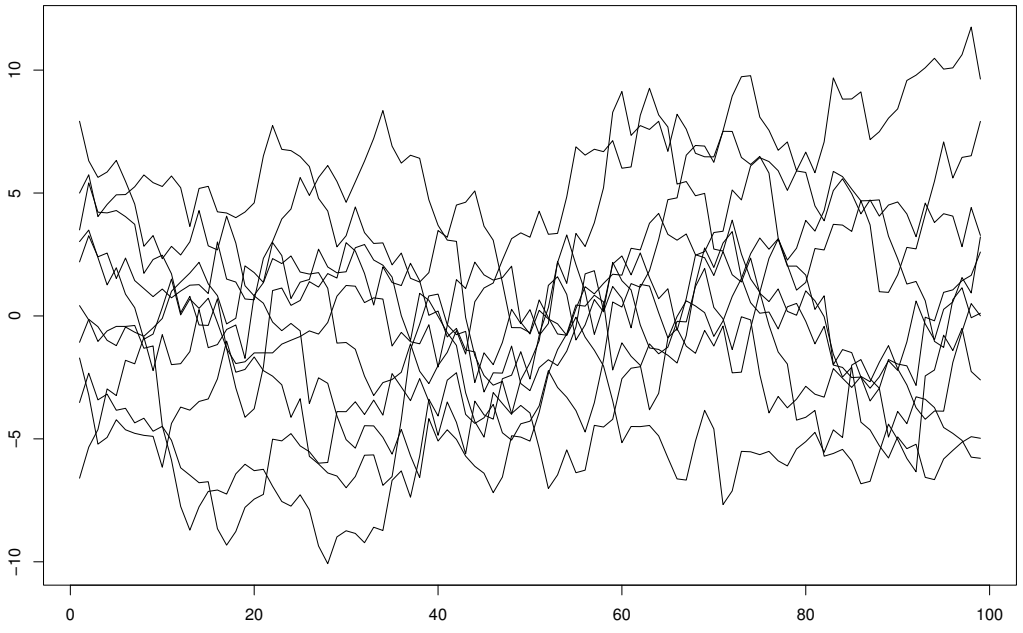
Alternative interpretation:

- fit a first order polynomial

$$p(j) = \beta_0 + \beta_1 j, \tag{37}$$

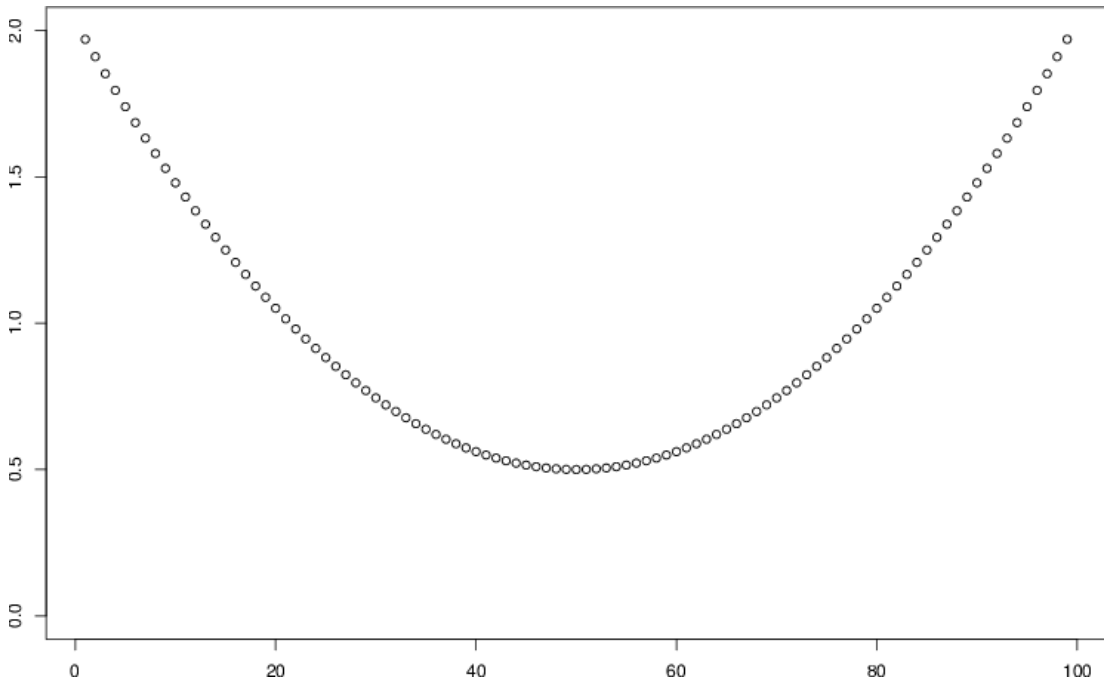
  locally through the points $(i-1, x_{i-1})$ and $(i+1, x_{i+1})$.
- The conditional mean is $p(i)$.

Example



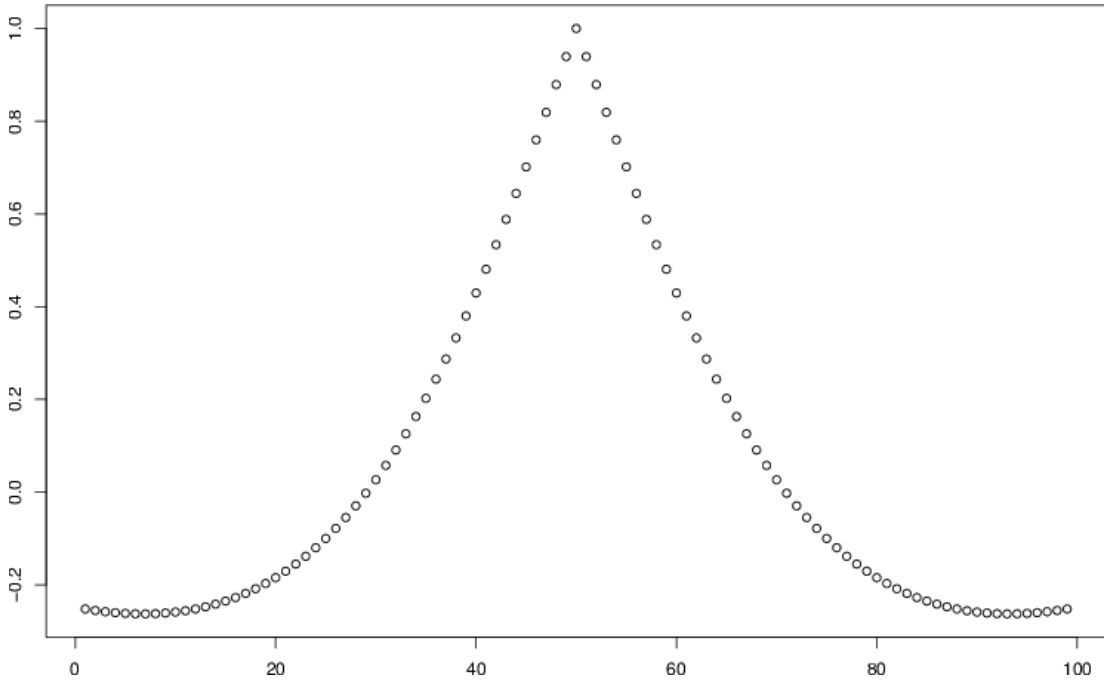
Samples with $n = 99$ and $\kappa = 1$ by conditioning on the constraint $\sum x_i = 0$.

Example



Marginal variances $\mathrm{Var}(x_i)$ for $i = 1, \ldots, n$.

Example



$\mathrm{Corr}(x_{n/2}, x_i)$ for $i = 1, \ldots, n$

The first order RW for irregular locations

Location of x_i is s_i but $s_{i+1} - s_i$ is not constant.
Assume

$$s_1 < s_2 < \cdots < s_n$$

and let

$$\delta_i \stackrel{\text{def}}{=} s_{i+1} - s_i. \tag{38}$$

Wiener process

Consider x_i as the realisation of an integrated Brownian motion in continuous time, i.e. a Wiener process $W(t)$, at time s_i.

Definition (Wiener process)

- A Wiener process with precision κ is a continuous-time stochastic process $W(t)$ for $t \geq 0$ with $W(0) = 0$ and such that the increments $W(t) - W(s)$ are Gaussian with mean 0 and variance $(t - s)/\kappa$ for any $0 \leq s < t$.
- Furthermore, increments for non overlapping time intervals are independent.
- For $\kappa = 1$, this process is called a standard Wiener process.

Brownian bridge

The full conditional is in this case known as the the *Brownian bridge*:

$$\mathsf{E}(x_i \mid \mathbf{x}_{-i}, \kappa) = \frac{\delta_i}{\delta_{i-1} + \delta_i} x_{i-1} + \frac{\delta_{i-1}}{\delta_{i-1} + \delta_i} x_{i+1} \tag{39}$$

and

$$\mathsf{Prec}(x_i \mid \mathbf{x}_{-i}, \kappa) = \kappa \left(\frac{1}{\delta_{i-1}} + \frac{1}{\delta_i} \right). \tag{40}$$

κ is a precision parameter.

The precision matrix is now

$$Q_{ij} = \kappa \begin{cases} \frac{1}{\delta_{i-1}} + \frac{1}{\delta_i} & j = i \\ -\frac{1}{\delta_i} & j = i+1 \\ 0 & \text{otherwise} \end{cases} \tag{41}$$

for $1 < i < n$.

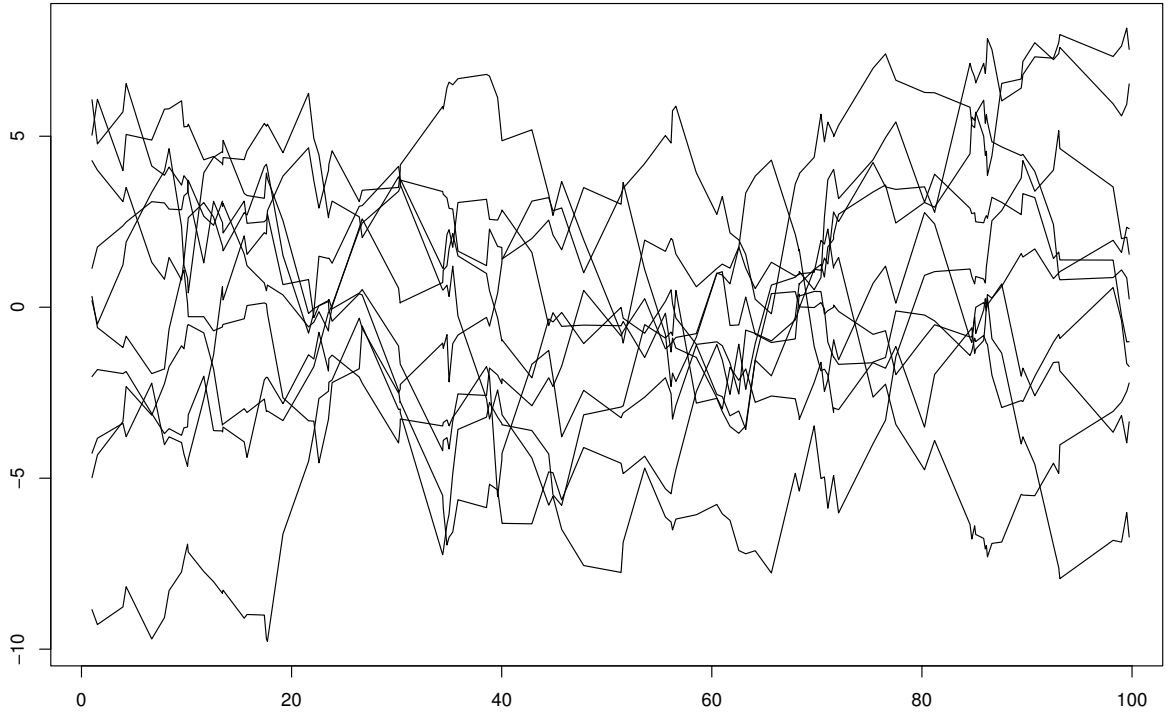
A proper correction at the boundary gives the remaining diagonal terms $Q_{11} = \kappa/\delta_1$, $Q_{nn} = \kappa/\delta_{n-1}$.

The joint density is

$$\pi(\mathbf{x} \mid \kappa) \propto \kappa^{(n-1)/2} \exp\left(-\frac{\kappa}{2}\sum_{i=1}^{n-1}(x_{i+1} - x_i)^2/\delta_i\right), \tag{42}$$

and is invariant to the addition of a constant.

- The interpretation of a RW1 model as a discretely observed Wiener-process, justifies the corrections needed.
- The underlying model *is the same*, it is only observed differently.



When the mean is only locally constant

Alternative IGMRF

$$\pi(\mathbf{x}) \propto \kappa^{(n-1)/2} \exp\left(-\frac{\kappa}{2}\sum_{i=1}^{n}(x_i - \bar{\mathbf{x}})^2\right) \tag{43}$$

where $\bar{\mathbf{x}}$ is the empirical mean of $\mathbf{x}$.

Assume n is even and

$$x_i = \begin{cases} 0, & 1 \leq i \leq n/2 \\ 1, & n/2 < i \leq n \end{cases}. \tag{44}$$

Thus $\mathbf{x}$ is locally constant with two levels.

Evaluate the density at this configuration under the RW1 model and the alternative (43), we obtain

$$\kappa^{(n-1)/2} \exp\left(-\frac{\kappa}{2}\right) \quad \text{and} \quad \kappa^{(n-1)/2} \exp\left(-n\frac{\kappa}{8}\right) \tag{45}$$

The log ratio of the densities is then of order $\mathcal{O}(n)$.

- The RW1 model only penalises the *local* deviation from a constant level.
- The alternative penalises the *global* deviation from a constant level.

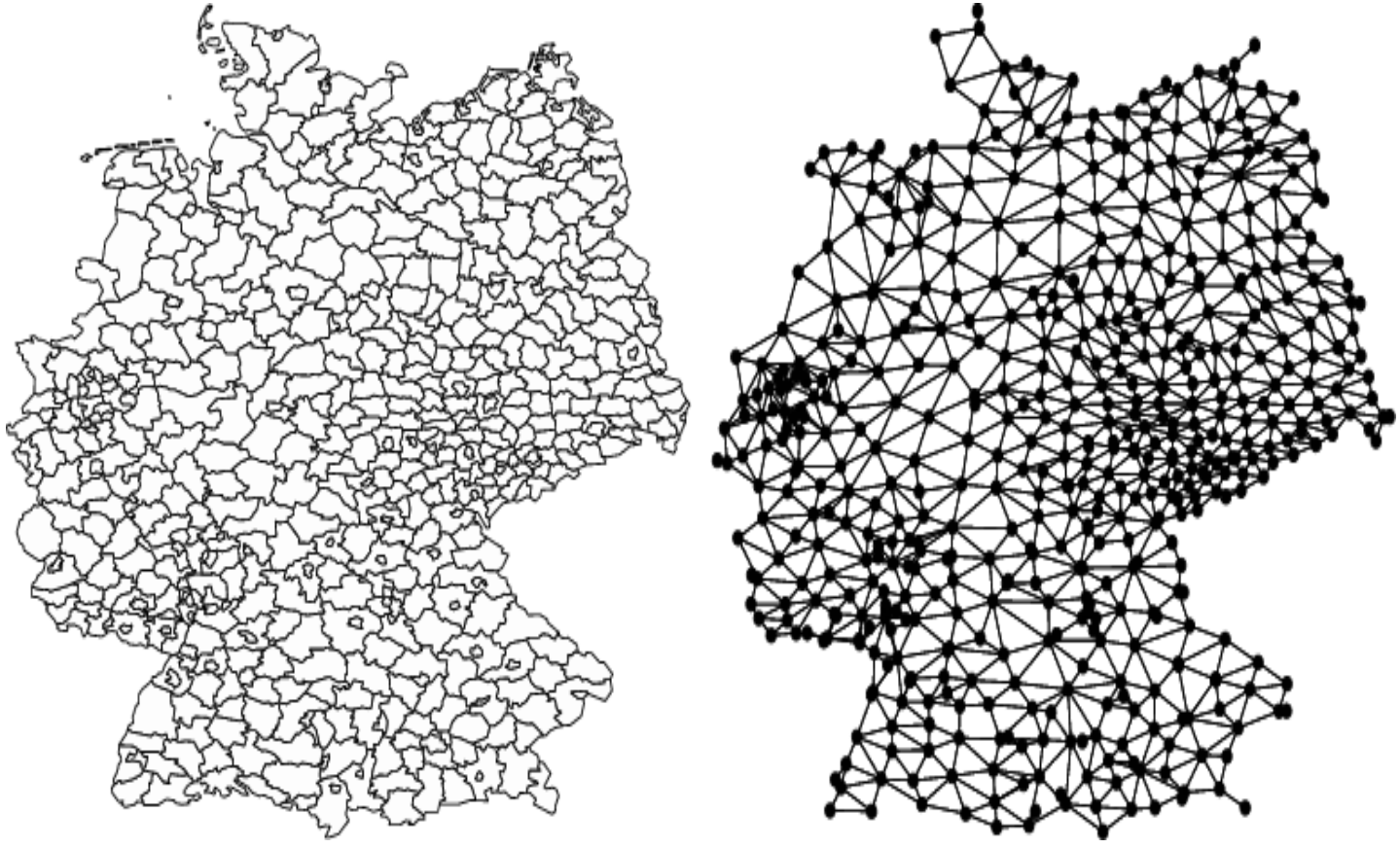
This *local* behaviour is advantageous in applications if the mean level of $\mathbf{x}$ is approximately or locally constant.

A similar argument also apply for polynomial IGMRFs of higher order constructed using forward differences of order k as independent Gaussian increments.

First order IGMRFs on irregular lattices

Consider the map of the 544 regions in Germany.

Two regions are *neighbours* if they share a common border.



Between neighbouring regions i and j, say, we define a "independent" Gaussian increment

$$x_i - x_j \sim \mathcal{N}(0, \kappa^{-1}) \tag{46}$$

Assume independent increments yields

$$\pi(\mathbf{x}) \propto \kappa^{(n-1)/2} \exp\left(-\frac{\kappa}{2}\sum_{i\sim j}(x_i - x_j)^2\right). \tag{47}$$

"$i \sim j$" denotes the set of all *unordered* pairs of neighbours.

Number of increments $|i \sim j|$ is larger than n, but the rank of the corresponding precision matrix is still $n - 1$.

There are hidden constraints in the increments due to the more complicated geometry on a lattice than on the line.

Example

Let $n = 3$ where all nodes are neighbours. Then $x_1 - x_2 = \epsilon_1$, $x_2 - x_3 = \epsilon_2$, and $x_3 - x_1 = \epsilon_3$, where ϵ_1, ϵ_2 and ϵ_3 are the increments.

This implies that

$$\epsilon_1 + \epsilon_2 + \epsilon_3 = 0$$

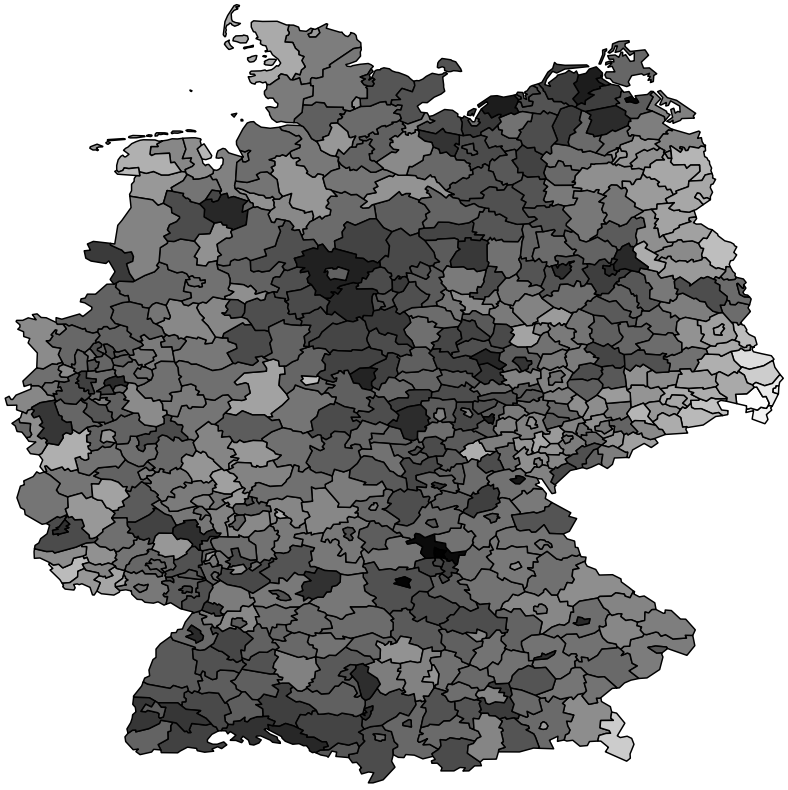
which is the 'hidden' linear constraint.

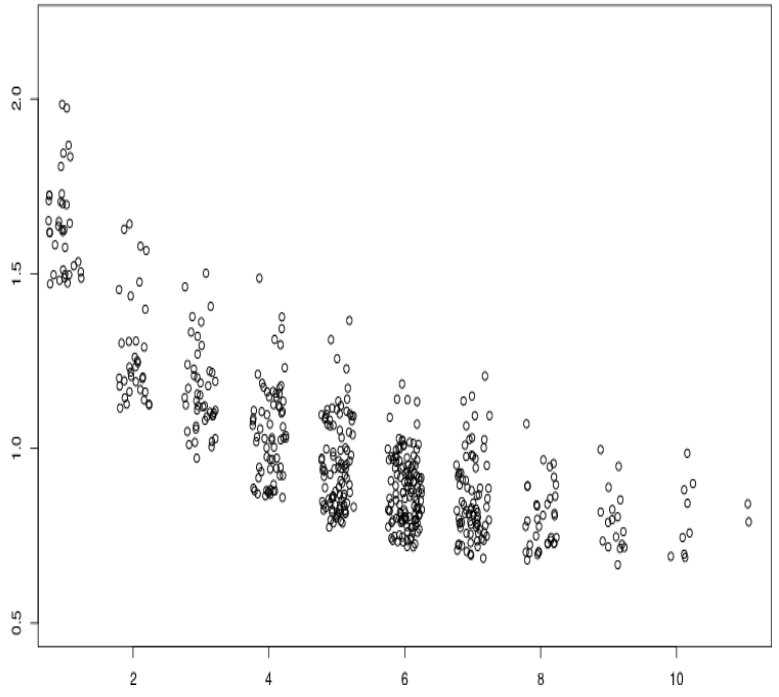
Let n_i denote the number of neighbours of region i.
The precision matrix $\mathbf{Q}$ is

$$Q_{ij} = \kappa \begin{cases} n_i & i = j, \\ -1 & i \sim j, \\ 0 & \text{otherwise}, \end{cases} \tag{48}$$

$$x_i \mid \mathbf{x}_{-i}, \kappa \sim \mathcal{N}(\frac{1}{n_i} \sum_{j:j\sim i} x_j, \frac{1}{n_i \kappa}). \tag{49}$$

Example





- In general there is no longer an underlying continuous stochastic process that we can relate to this density.
- If we change the spatial resolution or split a region into two new ones, we change the model.

Weighted variants

Incorporate symmetric weights w_{ij} for each pair of adjacent nodes i and j.

For example, $w_{ij} = 1/d(i,j)$

Assuming independent increments

$$x_i - x_j \sim \mathcal{N}(0, 1/(w_{ij}\kappa)), \tag{50}$$

then

$$\pi(\mathbf{x}) \propto \kappa^{(n-1)/2} \exp\left(-\frac{\kappa}{2}\sum_{i\sim j} w_{ij}(x_i - x_j)^2\right). \tag{51}$$

The precision matrix is

$$Q_{ij} = \kappa \begin{cases} \sum_{k:k\sim i} w_{ik} & i = j, \\ -1/w_{ij} & i \sim j, \\ 0 & \text{otherwise}, \end{cases} \tag{52}$$

and with mean and precision

$$\frac{\sum_{j:j\sim i} x_j w_{ij}}{\sum_{j:j\sim i} w_{ij}} \quad \text{and} \quad \kappa \sum_{j:j\sim i} w_{ij}, \tag{53}$$

respectively.

First order IGMRFs on regular lattices

For a lattice $\mathcal{I}_{\mathbf{N}}$ with $n = n_1 n_2$ nodes, let $i = (i_1, i_2)$ denote the node in the i_1th row and i_2th column.

Use the nearest four sites of i as its neighbours

$$(i_1 + 1, i_2),\ (i_1 - 1, i_2),\ (i_1, i_2 + 1),\ (i_1, i_2 - 1).$$

The precision matrix is

$$Q_{ij} = \kappa \begin{cases} n_i & i = j, \\ -1 & i \sim j, \\ 0 & \text{otherwise}, \end{cases} \tag{54}$$

and the full conditionals for x_i are

$$x_i \mid \mathbf{x}_{-i}, \kappa \ \sim\ \mathcal{N}\Big(\frac{1}{n_i} \sum_{j:j\sim i} x_j,\ \frac{1}{n_i \kappa}\Big). \tag{55}$$

Limiting behaviour

This model have an important property

The process converge to the de Wijs-process: a Gaussian process with variogram

$$\log(\textit{distance})$$

This is important

- there is a limiting process
- The de Wijs process has dense precision matrix whereas the IGMRF is very sparse!

The RW2 model for regular locations

Let $s_i = i$ for $i = 1, \ldots, n$, with a constant distance between consecutive nodes.

Use the second order increments

$$\Delta^2 x_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \kappa^{-1}) \tag{56}$$

for $i = 1, \ldots, n-2$, to define the joint density of $\mathbf{x}$

$$\begin{aligned} \pi(\mathbf{x}) &\propto \kappa^{(n-2)/2} \exp\left(-\frac{\kappa}{2}\sum_{i=1}^{n-2}(x_i - 2x_{i+1} + x_{i+2})^2\right) && (57) \\ &= \kappa^{(n-2)/2} \exp\left(-\frac{\kappa}{2}\mathbf{x}^T\mathbf{R}\mathbf{x}\right) && (58) \end{aligned}$$

$$\mathbf{R} = \begin{pmatrix} 1 & -2 & 1 & & & & & & \\ -2 & 5 & -4 & 1 & & & & & \\ 1 & -4 & 6 & -4 & 1 & & & & \\ & 1 & -4 & 6 & -4 & 1 & & & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & & \\ & & & 1 & -4 & 6 & -4 & 1 & \\ & & & & 1 & -4 & 6 & -4 & 1 \\ & & & & & 1 & -4 & 5 & -2 \\ & & & & & & 1 & -2 & 1 \end{pmatrix}. \quad (59)$$

- Verify directly that $\mathbf{QS}_1 = \mathbf{0}$ and that the rank of $\mathbf{Q}$ is $n - 2$.
- IGMRF of second order: invariant to the adding line to $\mathbf{x}$.
- Known as the second order random walk model, denoted by RW2(κ) or simply RW2 model.

Remarks

- This is the RW2 model defined and used in the literature.
- We cannot extend it consistently to the case where the locations are irregular.
- Similar problems occur if we increase the resolution from n to $2n$ locations, say.
- This is in contrast to the RW1 model.
- There exists an alternative (somewhat more involved) formulation with the desired continuous time interpretation.

The conditional mean and precision is

$$\begin{aligned} \mathrm{E}(x_i \mid \mathbf{x}_{-i}, \kappa) &= \frac{4}{6}(x_{i+1} + x_{i-1}) - \frac{1}{6}(x_{i+2} + x_{i-2}), \quad (60) \\ \mathrm{Prec}(x_i \mid \mathbf{x}_{-i}, \kappa) &= 6\kappa, \quad (61) \end{aligned}$$

respectively for $2 < i < n - 2$.

Consider the second order polynomial

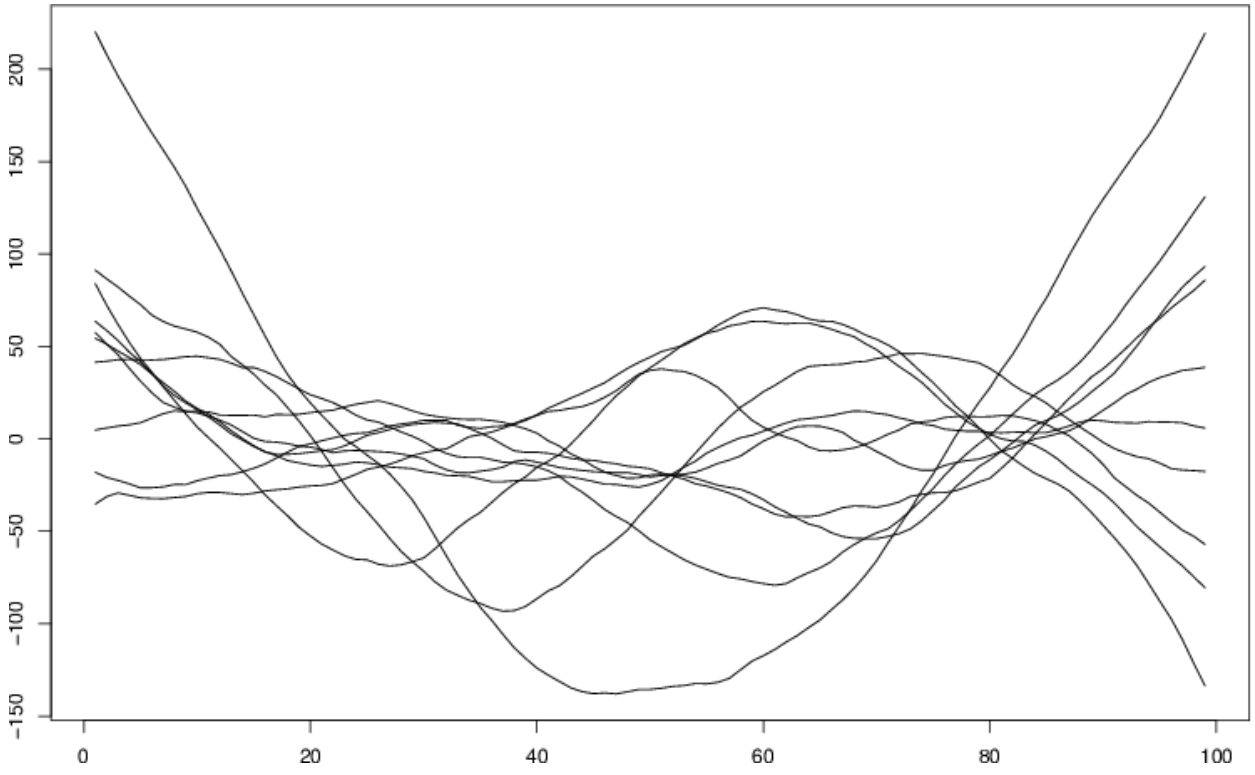
$$p(j) = \beta_0 + \beta_1 j + \frac{1}{2}\beta_2 j^2 \tag{62}$$

and compute the coefficients by a local least squares fit to the points

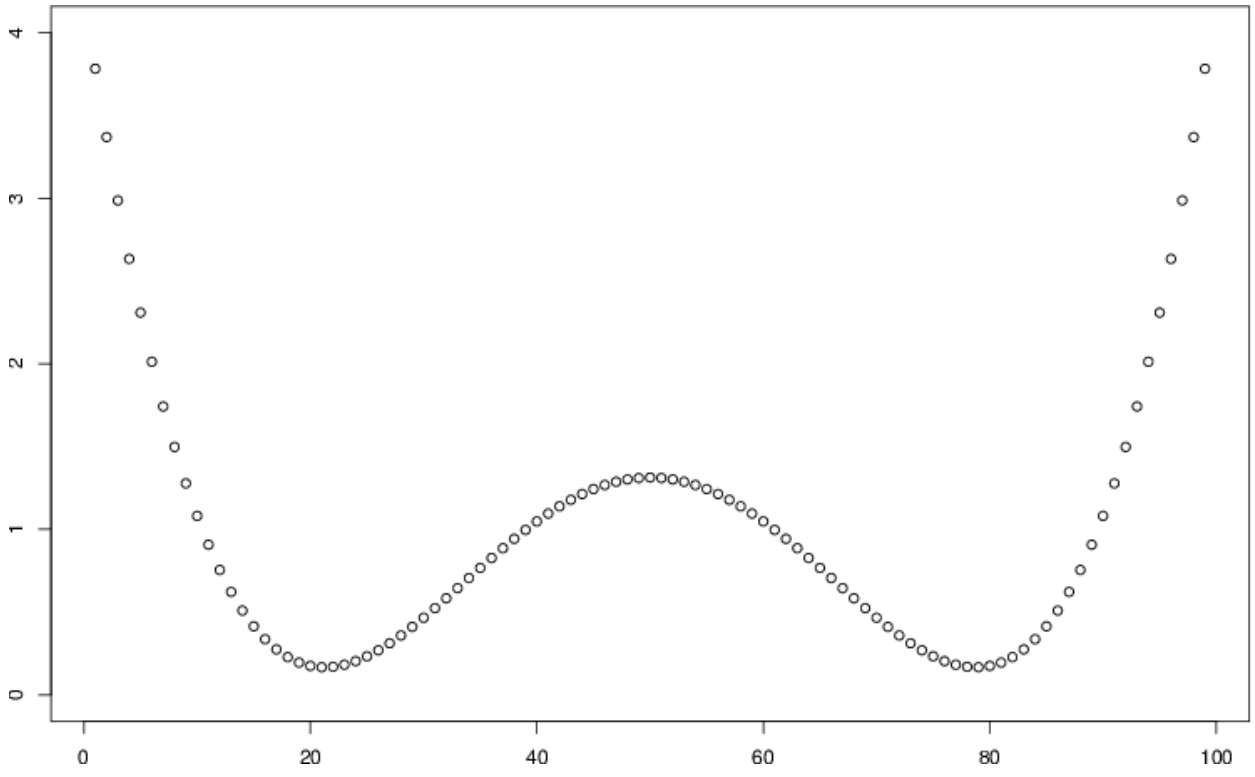
$$(i-2, x_{i-2}),\ (i-1, x_{i-1}),\ (i+1, x_{i+1}),\ (i+2, x_{i+2}). \tag{63}$$

Turns out that

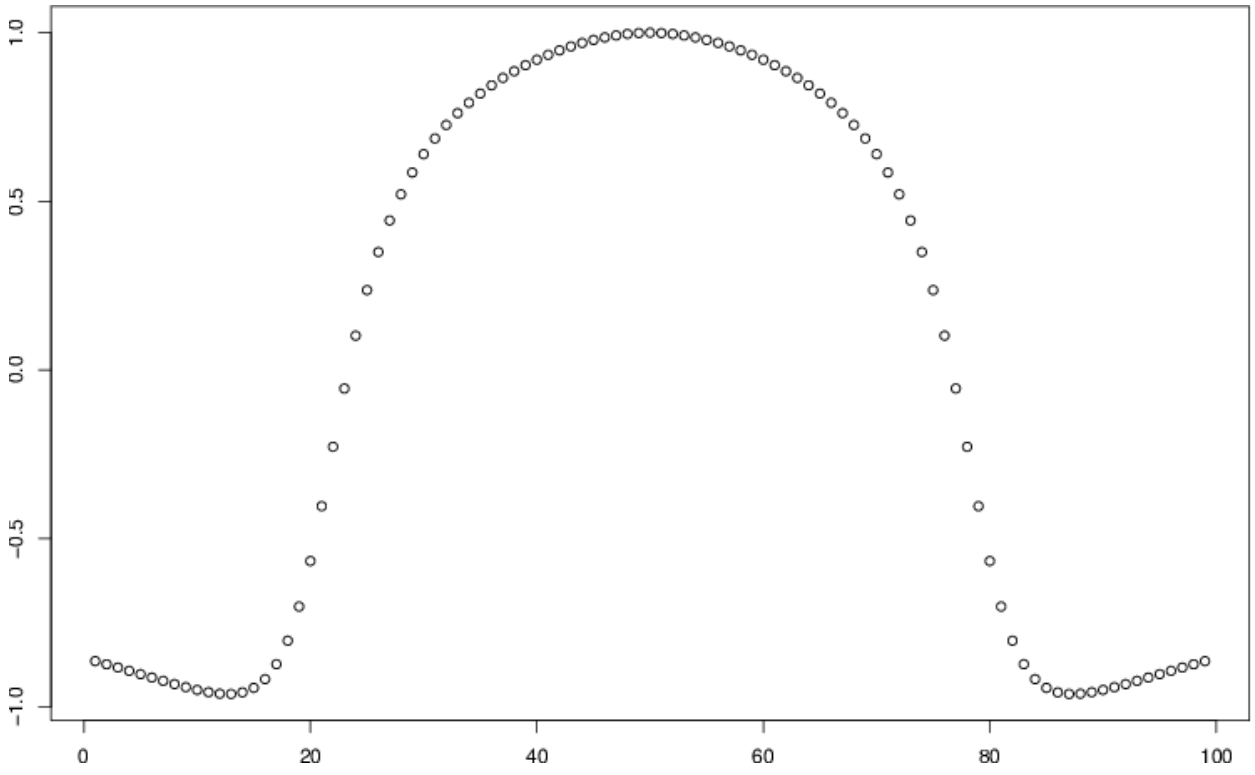
$$\mathrm{E}(x_i \mid \mathbf{x}_{-i}, \kappa) = p(i)$$



Samples with $n = 99$ and $\kappa = 1$ by conditioning on the constraints.



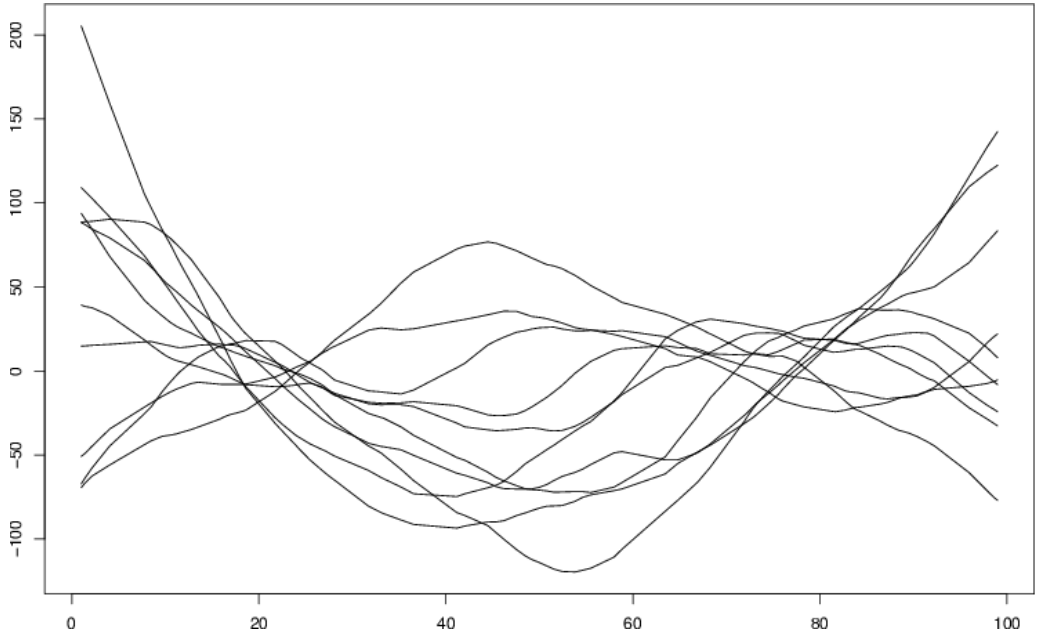
Marginal variances $\mathrm{Var}(x_i)$ for $i = 1, \ldots, n$.



$\text{Corr}(x_{n/2}, x_i)$ for $i = 1, \ldots, n$

Random walk models: irregular locations

- Locations $s_1 < s_2 < \ldots < s_n$.
- Discretely observed (integrated) Wiener process
- Augment with the velocities
- Valid for any order
- Connection to spline-models



Random walk models: irregular locations

- Locations $s_1 < s_2 < \ldots < s_n$.
- Discretely observed (integrated) Wiener process
- Augment with the velocities
- Valid for any order
- Connection to spline-models

$$\begin{pmatrix} \mathbf{A}_1 & \mathbf{B}_1 & & & & & \\ \mathbf{B}_1^T & \mathbf{A}_2 + \mathbf{C}_1 & \mathbf{B}_2 & & & & \\ & \mathbf{B}_2^T & \mathbf{A}_3 + \mathbf{C}_2 & \mathbf{B}_3 & & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & & \mathbf{B}_{n-1} & \\ & & & & \mathbf{B}_{n-1}^T & \mathbf{C}_{n-1} & \end{pmatrix}.$$

Random walk models: irregular locations

- Locations $s_1 < s_2 < \ldots < s_n$.
- Discretely observed (integrated) Wiener process
- Augment with the velocities
- Valid for any order
- Connection to spline-models

$$\mathbf{A}_i = \begin{pmatrix} 12/\delta_i^3 & 6/\delta_i^2 \\ 6/\delta_i^2 & 4/\delta_i \end{pmatrix} \quad \mathbf{B}_i = \begin{pmatrix} -12/\delta_i^3 & 6/\delta_i^2 \\ -6/\delta_i^2 & 2/\delta_i \end{pmatrix} \quad \mathbf{C}_i = \begin{pmatrix} 12/\delta_i^3 & -6/\delta_i^2 \\ -6/\delta_i^2 & 4/\delta_i \end{pmatrix}$$

where

$$\delta_i = s_{i+1} - s_i.$$

IGMRFs of higher order on regular lattices

The precision matrix is orthogonal to a polynomial design matrix of a certain degree.

Definition (IGMRFs of order k in dimension d)

An IGMRF of order k in dimension d, is an improper GMRF of rank $n - m_{k-1,d}$ where $\mathbf{Q}\mathbf{S}_{k-1,d} = \mathbf{0}$.

A second order IGMRF in two dimensions

Consider a regular lattice $\mathcal{I}_{\mathbf{N}}$ in $d = 2$ dimensions where $s_{i_1} = i_1$ and $s_{i_2} = i_2$.

Choose the independent increments

$$\left(x_{(i_1+1,i_2)} + x_{(i_1-1,i_2)} + x_{(i_1,i_2+1)} + x_{(i_1,i_2-1)}\right) - 4x_{i_1,i_2} \tag{64}$$

The motivation for this choice is that (64) is

$$\left(\Delta_{i_1}^2 + \Delta_{i_2}^2\right) x_{i_1-1,i_2-1} \tag{65}$$

Invariant to adding a first order polynomial

$$p_{1,2}(i_1, i_2) = \beta_{00} + \beta_{10} i_1 + \beta_{01} i_2,$$

The precision matrix (apart from boundary effects) should have non-zero elements

$$-\left(\Delta_{i_1}^2+\Delta_{i_2}^2\right)^2=-\left(\Delta_{i_1}^4+2\Delta_{i_1}^2\Delta_{i_2}^2+\Delta_{i_2}^4\right) \tag{66}$$

which is a negative difference approximation to the *biharmonic* differential operator

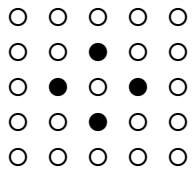
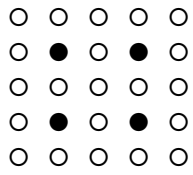
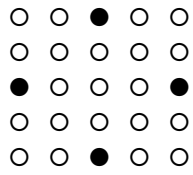
$$\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)^2=\frac{\partial^4}{\partial x^4}+2\frac{\partial^4}{\partial x^2\partial y^2}+\frac{\partial^4}{\partial y^4}. \tag{67}$$

The fundamental solution of the biharmonic equation

$$\left(\frac{\partial^4}{\partial x^4}+2\frac{\partial^4}{\partial x^2\partial y^2}+\frac{\partial^4}{\partial y^4}\right)\phi(x,y)=0 \tag{68}$$

is the *thin plate spline*.

Full conditionals in the interior

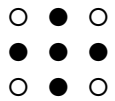
$$\mathsf{E}(x_i \mid \mathbf{x}_{-i}) = \frac{1}{20}\left(8 \; \raisebox{-0.5em}{  } \; - 2 \; \raisebox{-0.5em}{  } \; - 1 \; \raisebox{-0.5em}{  } \right) \tag{69}$$

and

$$\mathsf{Prec}(x_i \mid \mathbf{x}_{-i}) = 20\kappa. \tag{70}$$

Alternative IGMRFs in two dimensions

Using only the terms



(71)

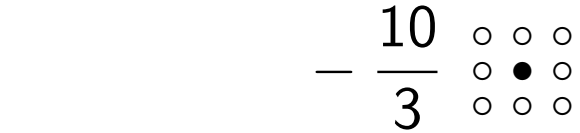
to obtain an difference approximation to

$$\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \tag{72}$$

is not optimal.

The discretization error different 45 degrees to the main directions, hence we should expect a "directional" effect.

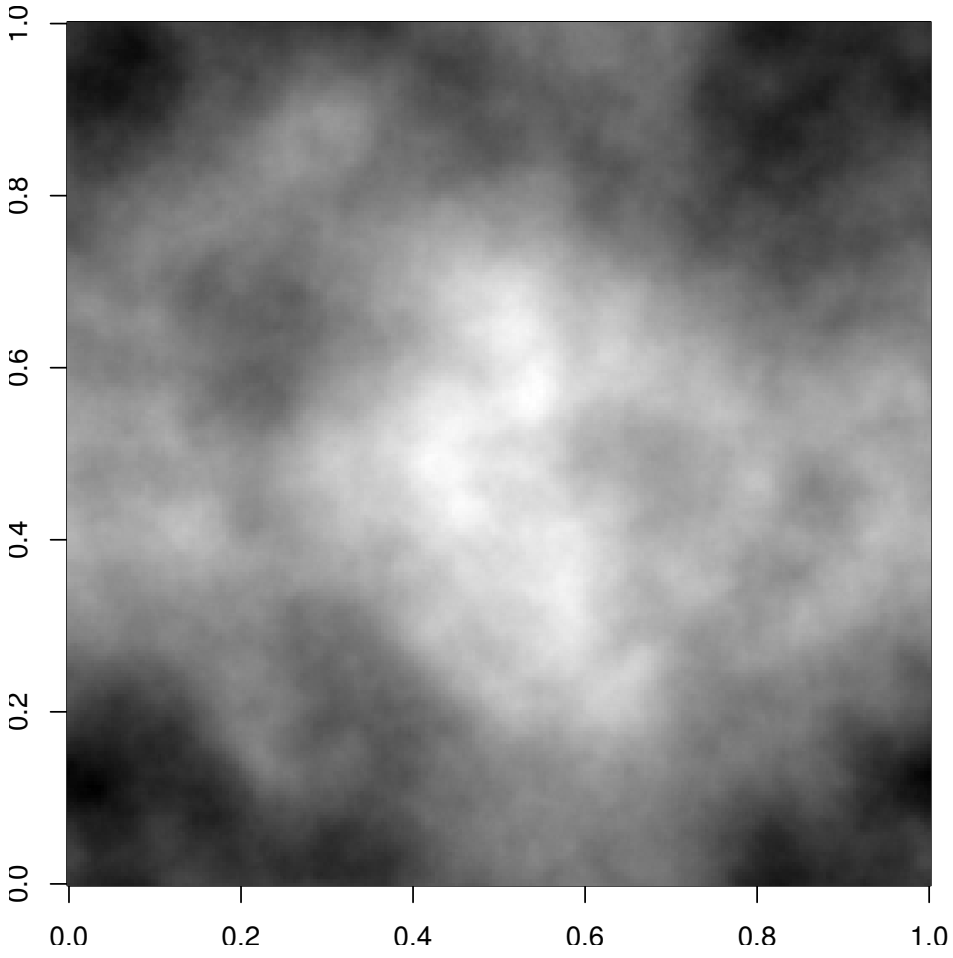
Use an isotropic approximations (ex. "Mehrstellen-stencil")

$$-\frac{10}{3}$$  $$+\frac{2}{3}$$ $$+\frac{1}{6}$$ (73)

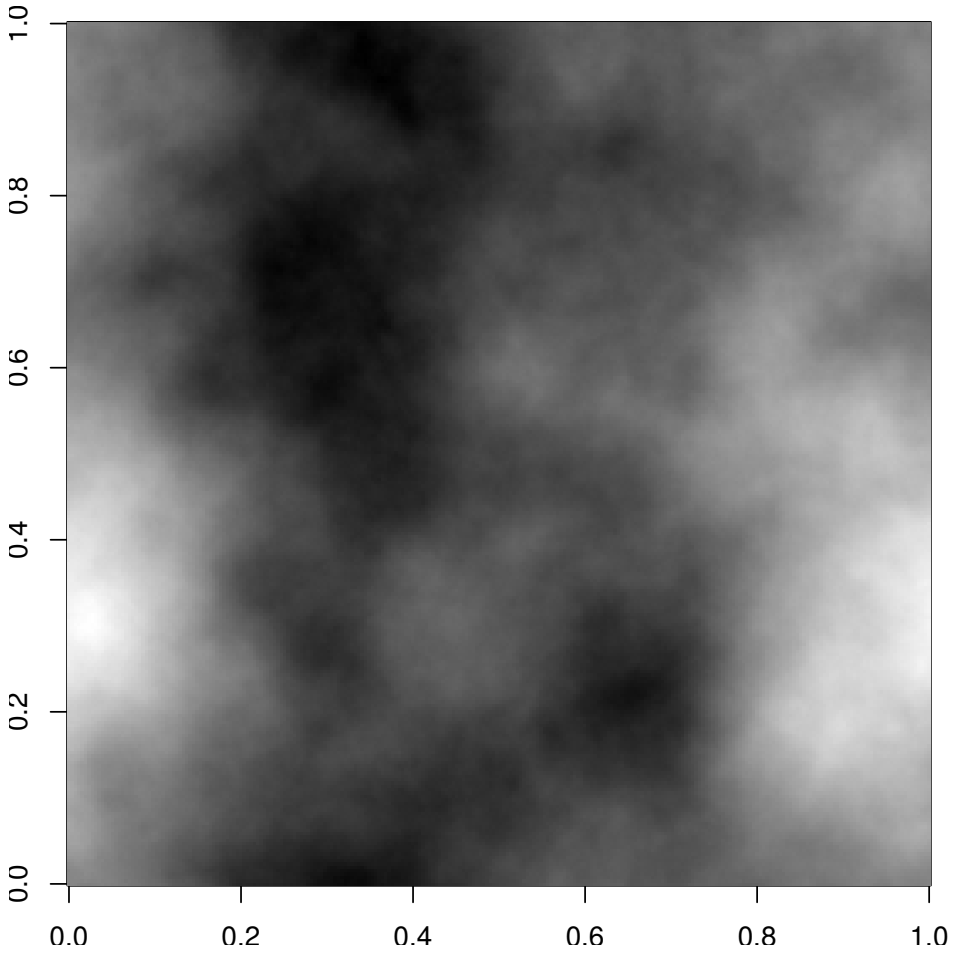
$$\mathsf{E}(x_i \mid \mathbf{x}_{-i}) = \frac{1}{468}\Bigg(144$$ $$- 18$$

$$+8$$ $$- 8$$ $$- 1$$ $$\Bigg)$$ (74)

$$\mathsf{Prec}(x_i \mid \mathbf{x}_{-i}) = 13\kappa.$$ (75)



Samples.



Samples.

Part V

Hierarchical GMRF models

Outline I

Outline II

Why is it so?

Hierarchical GMRFs models

Characterised through several *stages* of observables and parameters.

A typical scenario is as follows.

Stage 1 Formulate a distributional assumption for the observables, dependent on latent parameters.

- Time series of binary observations $\mathbf{y}$, we may assume

$$y_i, \quad i = 1, \ldots, n \ : \ y_i \sim \mathcal{B}(p_i)$$

- We assume the observations to be *conditionally independent*

Hierarchical GMRFs models

Characterised through several *stages* of observables and parameters.

A typical scenario is as follows.

Stage 2 Assign a prior model, i.e. a GMRF, for the unknown parameters, here p_i.

- Chose an autoregressive model for the logit-transformed probabilities $x_i = \text{logit}(p_i)$.

Hierarchical GMRFs models

Characterised through several *stages* of observables and parameters.

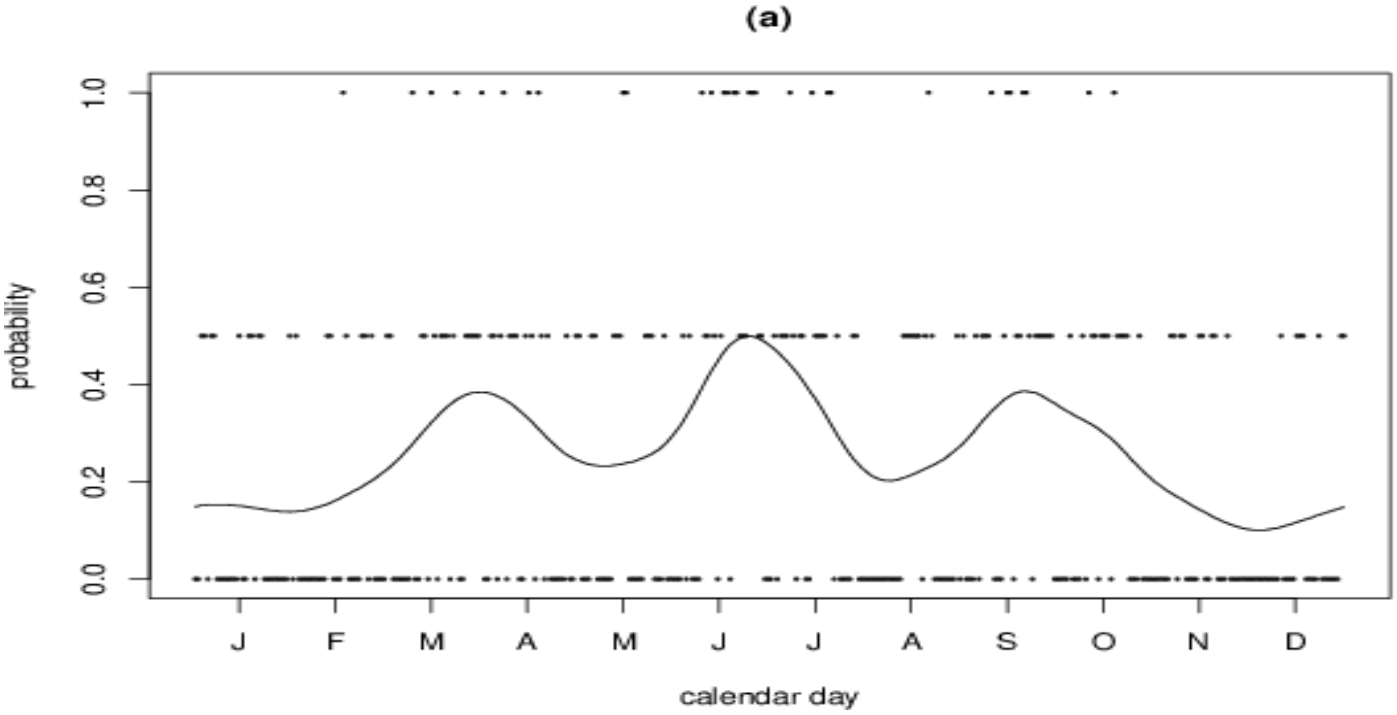
A typical scenario is as follows.

Stage 3 Assign to unknown parameters (or hyperparameters) of the GMRF

- precision parameter κ
- "strength" of dependency.

Further stages if needed.

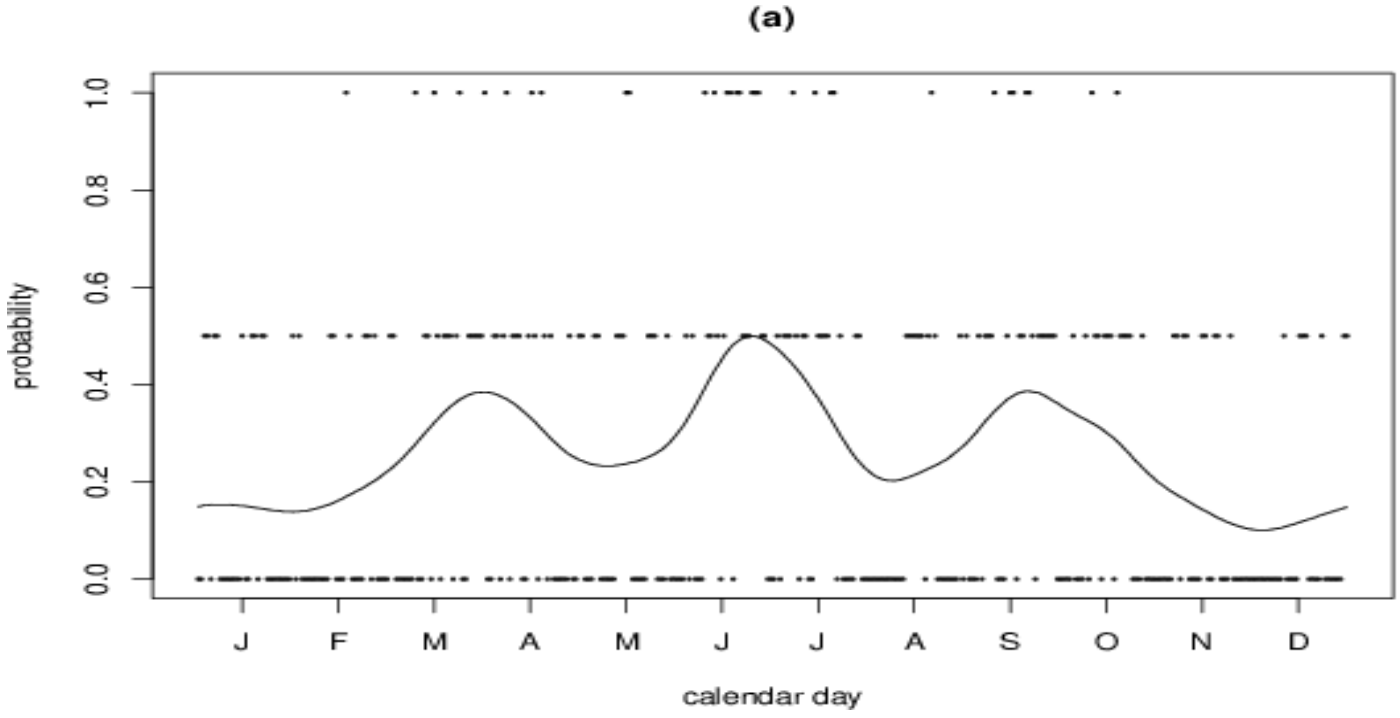
Tokyo rainfall data



Stage 1 Binomial data

$$y_i \sim \begin{cases} \text{Binomial}(2, p(x_i)) \\ \text{Binomial}(1, p(x_i)) \end{cases}$$

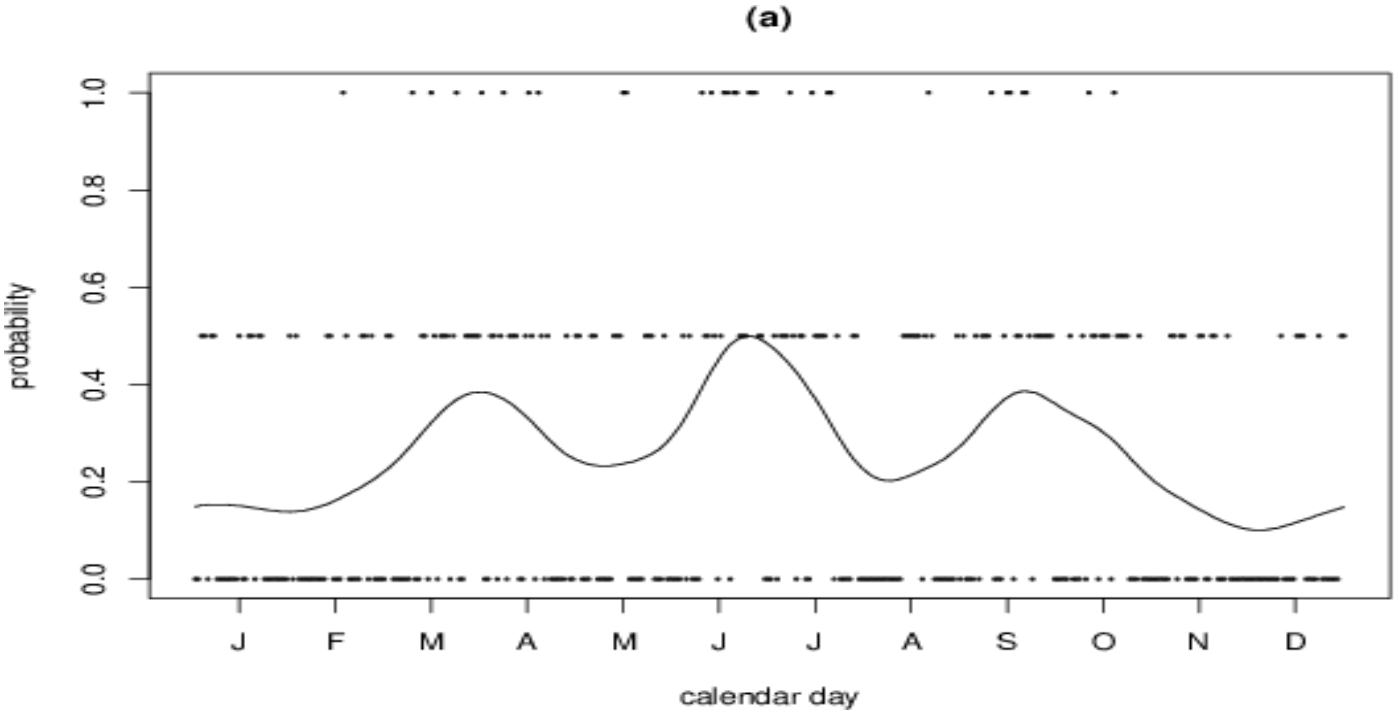
Tokyo rainfall data



Stage 2 Assume a smooth latent $\mathbf{x}$,

$$\mathbf{x} \sim RW2(\kappa), \qquad \text{logit}(p_i) = x_i$$

Tokyo rainfall data



Stage 3 Gamma(α, β)-prior on κ

Classes of hierarchical GMRF models

- Normal data
  - Block-MCMC
- Non-normal data that allows for a normal-mixture representation
  - Student-t distribution
  - Logistic and Laplace (Binary regression)
  - Block-MCMC with auxillary variables
- Non-normal data
  - Poisson
  - and others...
  - Block-MCMC with GMRF-approximations

MCMC

Construct a Markov chain

$$\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(k)}, \ldots$$

that converges (under some conditions) to $\pi(\boldsymbol{\theta})$.

1. Start with $\boldsymbol{\theta}^{(0)}$ where $\pi(\boldsymbol{\theta}^{(0)}) > 0$. Set $k = 1$.
2. Generate a *proposal* θ^* from some *proposal kernel* $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(k-1)})$.
   Set $\boldsymbol{\theta}^{(k)} = \boldsymbol{\theta}^*$ with probability

$$\alpha = \min\left\{1, \frac{\pi(\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta}^{(k-1)})} \frac{q(\boldsymbol{\theta}^{(k-1)}|\boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(k-1)})}\right\};$$

   otherwise set $\boldsymbol{\theta}^{(k)} = \boldsymbol{\theta}^{(k-1)}$.
3. Set $k = k + 1$ and go back to 2

Depending on the specific choice of the proposal kernel $q(\boldsymbol{\theta}^*|\boldsymbol{\theta})$, very different algorithms result.

- When $q(\boldsymbol{\theta}^*|\boldsymbol{\theta})$ does not depend on the current value of $\boldsymbol{\theta}$ proposal is called an *independence proposal*.
- When $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}) = q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)$ we have a *Metropolis proposal*. These includes so-called *random-walk proposals*.

The rate of convergence toward $\pi(\boldsymbol{\theta})$ and the degree of dependence between successive samples of the Markov chain (*mixing*) will depend on the chosen proposal.

Single-site algorithms

- Most MCMC algorithms have been based on updating each scalar component

$$\theta_i, \quad i = 1, \ldots, p$$

  of $\boldsymbol{\theta}$ conditional on $\boldsymbol{\theta}_{-i}$.
- Apply the MH-algorithm in turn to every component θ_i of $\boldsymbol{\theta}$ with arbitrary proposal kernels

$$q_i(\theta_i^* | \theta_i, \boldsymbol{\theta}_{-i})$$

- As long as we update each component of $\boldsymbol{\theta}$, this algorithm will converge to the target distribution $\pi(\boldsymbol{\theta})$.

However...

- Such *single-site* updating can be disadvantageous if parameters are highly dependent in the posterior distribution $\pi(\boldsymbol{\theta})$.
- The problem is that the Markov chain may move around very slowly in its target (posterior) distribution.
- A general approach to circumvent this problem is to update parameters in larger blocks, $\boldsymbol{\theta}_j$: a vector of components of $\boldsymbol{\theta}$.
- The choice of blocks are often controlled by what is possible to do in practise.
- Ideally, we should choose a small number of blocks with large dependence within the blocks but with less dependence between blocks.

Blocking

Assume the following

- $\boldsymbol{\theta} = (\kappa, \mathbf{x})$ where
  - κ is the precision and
  - $\mathbf{x}$ is a GMRF.

Two-block approach

- sample $\mathbf{x} \sim \pi(\mathbf{x}|\kappa)$, and
- sample $\kappa \sim \pi(\kappa|\mathbf{x})$

Often strong dependence between κ and $\mathbf{x}$ in the posterior; resolved using a joint update of $(\mathbf{x}, \kappa)$.

Why this modification is important and why it works is discussed next.

Rate of convergence

Let $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots$ denote a Markov chain with target distribution $\pi(\boldsymbol{\theta})$ and initial value $\boldsymbol{\theta}^{(0)} \sim \pi(\boldsymbol{\theta})$.

Rate of convergence ρ: how quickly $\mathrm{E}(h(\boldsymbol{\theta}^{(t)})|\boldsymbol{\theta}^{(0)})$ approaches the stationary value $\mathrm{E}(h(\boldsymbol{\theta}))$ for all square π-integrable functions $h(\cdot)$.

Let ρ be the minimum number such that for all $h(\cdot)$ and for all $r > \rho$

$$\lim_{k \to \infty} \mathrm{E}\left[\left(\mathrm{E}\left(h(\boldsymbol{\theta}^{(k)}) \mid \boldsymbol{\theta}^{(0)}\right) - \mathrm{E}\left(h(\boldsymbol{\theta})\right)\right)^2 r^{-2k}\right] = 0. \qquad (76)$$

Example

Let **x** be a first-order autoregressive process

$$x_t - \mu = \gamma(x_{t-1} - \mu) + \nu_t, \quad t = 2, \ldots, n, \tag{77}$$

where $|\gamma| < 1$, $\{\nu_t\}$ are iid normals with zero mean and variance σ^2, and $x_1 \sim \mathcal{N}(\mu, \frac{\sigma^2}{1-\gamma^2})$.

Let γ, σ^2, and μ be fixed parameters.

At each iteration a single-site Gibbs sampler will sample x_t from the full conditional $\pi(x_t|\mathbf{x}_{-t})$ for $t = 1, \ldots, n$,

$$x_t \mid \mathbf{x}_{-t} \sim \begin{cases} \mathcal{N}(\mu + \gamma(x_2 - \mu),\ \sigma^2) & t = 1, \\ \mathcal{N}(\mu + \frac{\gamma}{1+\gamma^2}(x_{t-1} + x_{t+1} - 2\mu),\ \frac{\sigma^2}{1+\gamma^2}) & t = 2, \ldots, n-1, \\ \mathcal{N}(\mu + \gamma(x_{n-1} - \mu),\ \sigma^2) & t = n. \end{cases}$$

- For large n, the rate of convergence is

$$\rho = 4\frac{\gamma^2}{(1+\gamma^2)^2}. \tag{78}$$

- For $|\gamma|$ close to one the rate of convergence can be slow: If $\gamma = 1 - \delta$ for small $\delta > 0$, then $\rho = 1 - \delta^2 + \mathcal{O}(\delta^3)$.

To circumvent this problem, we may update $\mathbf{x}$ in one block. This is possible as $\mathbf{x}$ is a GMRF.

This yields immediate convergence.

Relax the assumptions of fixed hyperparameters.

Consider a hierarchical formulation where the mean of x_t, μ, is unknown and assigned with a standard normal prior,

$$\mu \sim \mathcal{N}(0, 1) \quad \text{and} \quad \mathbf{x} \mid \mu \sim \mathcal{N}(\mu \mathbf{1}, \mathbf{Q}^{-1}),$$

where $\mathbf{Q}$ is the precision matrix of the GMRF $\mathbf{x}|\mu$.

The joint density of $(\mu, \mathbf{x})$ is normal.

We have two natural blocks, μ and $\mathbf{x}$.

A two-block Gibbs sampler update μ and $\mathbf{x}$ with samples from their full conditionals,

$$\begin{aligned} \mu^{(k)}|\mathbf{x}^{(k)} &\sim \mathcal{N}\left(\frac{\mathbf{1}^T\mathbf{Q}\mathbf{x}^{(k-1)}}{1+\mathbf{1}^T\mathbf{Q}\mathbf{1}},\ (1+\mathbf{1}^T\mathbf{Q}\mathbf{1})^{-1}\right) \\ \mathbf{x}^{(k)}|\mu^{(k)} &\sim \mathcal{N}(\mu^{(k)}\mathbf{1},\ \mathbf{Q}^{-1}). \end{aligned} \tag{79}$$

The presence of the hyperparameter μ will slow down the convergence compared to the case when μ is fixed.

Due to the nice structure of (79) we can characterise explicitly the marginal chain of $\{\mu^{(k)}\}$.

Theorem

The marginal chain $\mu^{(1)}, \mu^{(2)}, \ldots$ from the two-block Gibbs sampler defined in (79) and started in equilibrium, is a first-order autoregressive process

$$\mu^{(k)} = \phi \mu^{(k-1)} + \epsilon_k,$$

where

$$\phi = \frac{\mathbf{1}^T \mathbf{Q} \mathbf{1}}{1 + \mathbf{1}^T \mathbf{Q} \mathbf{1}}$$

and $\epsilon_k \overset{iid}{\sim} \mathcal{N}(0, 1 - \phi^2)$.

Proof. It follows directly that the marginal chain $\mu^{(1)}, \mu^{(2)}, \ldots$ is a first-order autoregressive process.

The coefficient ϕ is found by computing the covariance at lag 1,

$$\begin{aligned}
\mathrm{Cov}(\mu^{(k)}, \mu^{(k+1)}) &= \mathrm{E}\left(\mu^{(k)}\mu^{(k+1)}\right) \\
&= \mathrm{E}\left(\mu^{(k)}\mathrm{E}\left(\mu^{(k+1)} \mid \mu^{(k)}\right)\right) \\
&= \mathrm{E}\left(\mu^{(k)}\mathrm{E}\left(\mathrm{E}\left(\mu^{(k+1)} \mid \mathbf{x}^{(k)}\right) \mid \mu^{(k)}\right)\right) \\
&= \mathrm{E}\left(\mu^{(k)}\mathrm{E}\left(\frac{\mathbf{1}^T\mathbf{Q}\mathbf{x}^{(k)}}{1+\mathbf{1}^T\mathbf{Q}\mathbf{1}} \mid \mu^{(k)}\right)\right) \\
&= \frac{\mathbf{1}^T\mathbf{Q}\mathbf{1}}{1+\mathbf{1}^T\mathbf{Q}\mathbf{1}}\mathrm{Var}(\mu^{(k)}),
\end{aligned}$$

which is known to be ϕ times the variance $\mathrm{Var}(\mu^{(k)})$ for a first-order autoregressive process.

For our model

$$\phi = \frac{n(1-\gamma)^2/\sigma^2}{1+n(1-\gamma)^2/\sigma^2} = 1 - \frac{\mathrm{Var}(x_t)}{n}\frac{1-\gamma^2}{(1-\gamma)^2} + \mathcal{O}(1/n^2).$$

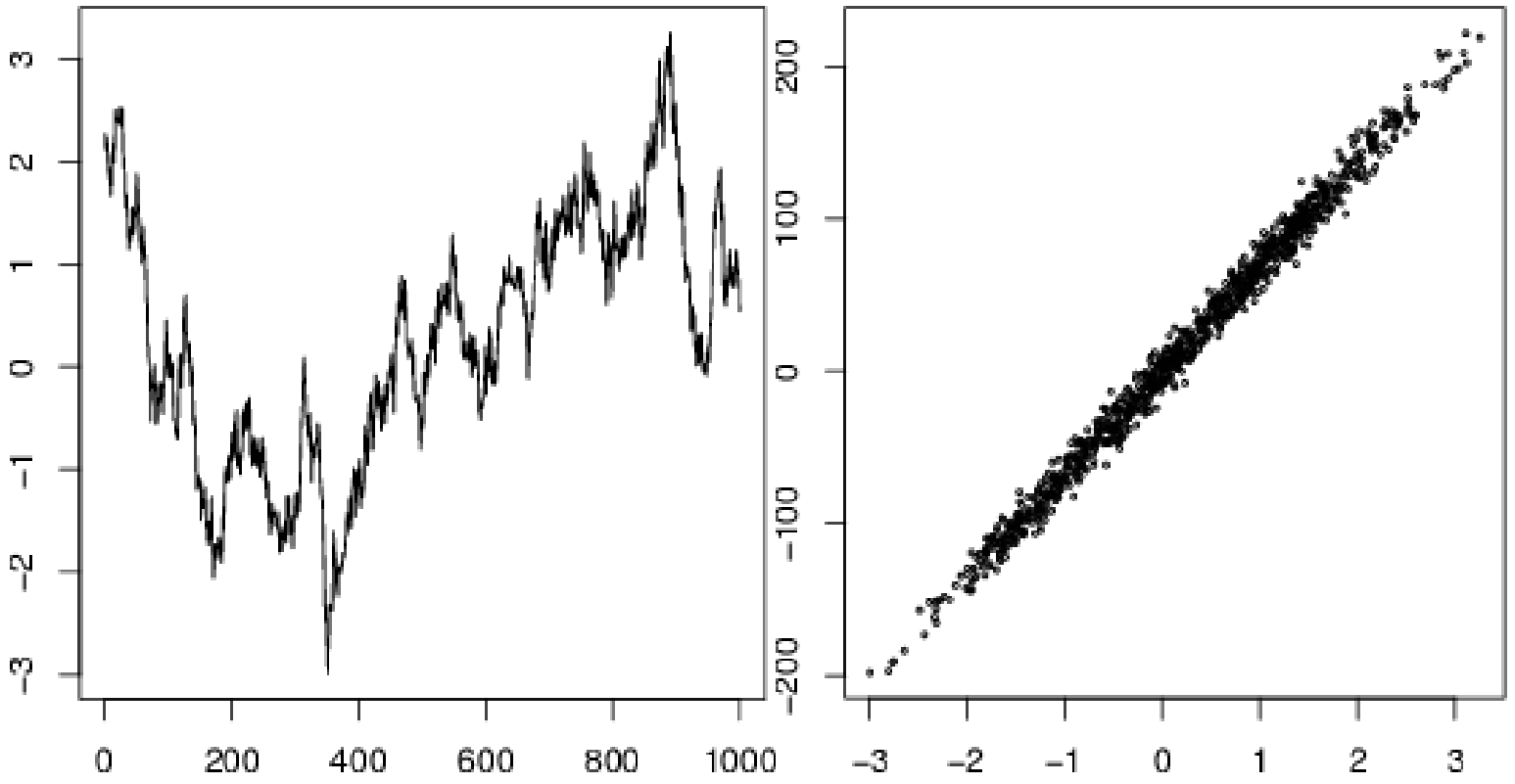
When n is large, ϕ is close to 1 and the chain will both mix and converge slowly even though we use a two-block Gibbs sampler.

Note that

- increasing variance of x_t improves the convergence.

However, $\phi \to 1$ as $n \to \infty$, which is bad.

What is going on?



(a) Trace of $\mu^{(k)}$, and (b) the pairs $(\mu^{(k)}, \mathbf{1}^T\mathbf{Q}\mathbf{x}^{(k)})$, with $\mu^{(k)}$ on the horizontal axis.

One-block algorithm

So far: blocking improves mainly within the block. If there is strong dependence *between* blocks, the MCMC algorithm may still suffer from slow convergence.

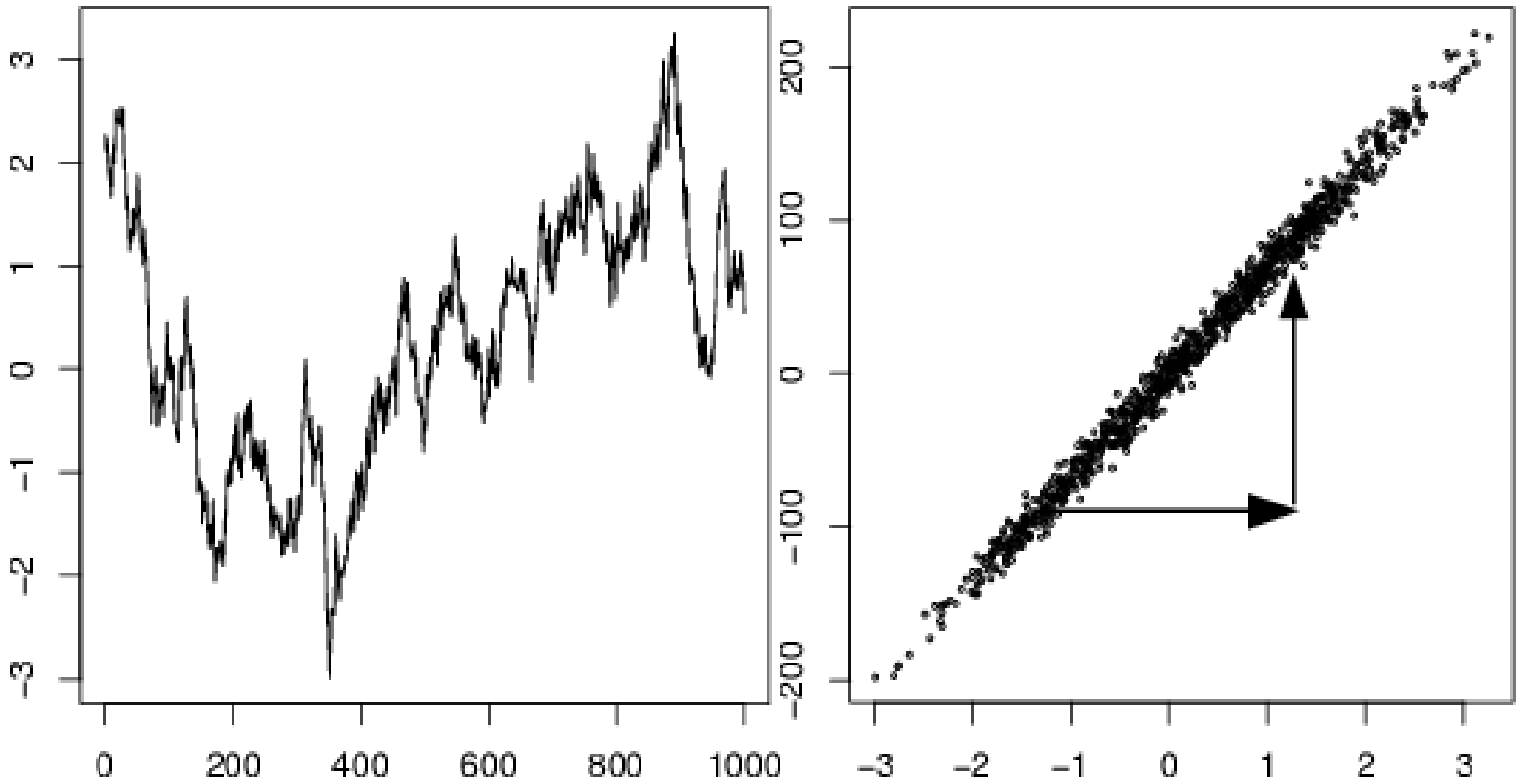
Update $(\mu, \mathbf{x})$ jointly by delaying the accept/reject step until $\mathbf{x}$ also is updated.

$$\begin{aligned} \mu^* &\sim q(\mu^* \mid \mu^{(k-1)}) \\ \mathbf{x}^* | \mu^* &\sim \mathcal{N}(\mu^* \mathbf{1}, \mathbf{Q}^{-1}) \end{aligned} \tag{80}$$

then accept/reject $(\mu^*, \mathbf{x}^*)$ jointly.

Here, $q(\mu^* | \mu^{(k-1)})$ can be a simple random-walk proposal or some other suitable proposal distribution.

What is going on?



(a) Trace of $\mu^{(k)}$, and (b) the pairs $(\mu^{(k)}, \mathbf{1}^T\mathbf{Q}\mathbf{x}^{(k)})$, with $\mu^{(k)}$ on the horizontal axis.

With a symmetric μ-proposal,

$$\alpha = \min\left\{1,\ \exp(-\frac{1}{2}((\mu^*)^2 - (\mu^{(k-1)})^2))\right\}. \qquad (81)$$

- Only the *marginal density* of μ is needed in (81): we effectively integrate $\mathbf{x}$ out of the target density.
- The minor modification to delay the accept/reject step until $\mathbf{x}$ is updated as well can give a large improvement.
- In this case: random walk on a one-dimensional density.

A more general setup

- Hyperparameters $\boldsymbol{\theta}$ (low dimension)
- GMRF $\mathbf{x} \mid \boldsymbol{\theta}$ of size n
- Observe $\mathbf{x}$ with data $\mathbf{y}$.

The posterior is

$$\pi(\mathbf{x}, \boldsymbol{\theta} \mid \mathbf{y}) \propto \pi(\boldsymbol{\theta})\, \pi(\mathbf{x} \mid \boldsymbol{\theta})\, \pi(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta}).$$

Assume we are able to sample from $\pi(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$, i.e., the full conditional of $\mathbf{x}$ is a GMRF.

The one-block algorithm

The following proposal update $(\boldsymbol{\theta}, \mathbf{x})$ in one block:

$$\begin{aligned} \boldsymbol{\theta}^* &\sim q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(k-1)}) \\ \mathbf{x}^* &\sim \pi(\mathbf{x} \mid \boldsymbol{\theta}^*, \mathbf{y}). \end{aligned} \tag{82}$$

The proposal $(\boldsymbol{\theta}^*, \mathbf{x}^*)$ is then accepted/rejected jointly.

We denote this as the *one-block* algorithm.

Consider the $\boldsymbol{\theta}$-chain, then we are in fact sampling from the posterior marginal $\pi(\boldsymbol{\theta}|\mathbf{y})$ using the proposal

$$\boldsymbol{\theta}^* \sim q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(k-1)})$$

The dimension of $\boldsymbol{\theta}$ is typically low (1-5, say).

The proposed algorithm should not experience any serious mixing problems.

The one-block algorithm is not always feasible

The one-block algorithm is not always feasible for the following reasons:

1. The full conditional of $\mathbf{x}$ can be a GMRF with a precision matrix that is not sparse.
   This will prohibit a fast factorisation, hence a joint update is feasible but not computationally efficient.
2. The data can be non-normal so the full conditional of $\mathbf{x}$ is not a GMRF and sampling $\mathbf{x}^*$ using (82) is not possible (in general).

...not sparse precision matrix

These cases can often be approached using *sub-blocks* of $(\boldsymbol{\theta}, \mathbf{x})$, the *sub-block* algorithm.

Assume a natural splitting exists for both $\boldsymbol{\theta}$ and $\mathbf{x}$ into

$$(\boldsymbol{\theta}_a, \mathbf{x}_a),\ (\boldsymbol{\theta}_b, \mathbf{x}_b) \quad \text{and} \quad (\boldsymbol{\theta}_c, \mathbf{x}_c), \tag{83}$$

The sets a, b, and c do not need to be disjoint.

One class of examples where such an approach is fruitful is (geo-)additive models where a, b, and c represent three different covariate effects with their respective hyperparameters.

...non-Gaussian full conditionals

Auxillary variables

- can help achieving Gaussian full conditionals.
- logit and probit regression models for binary and multi-categorical data, and
- Student-t_ν distributed observations.

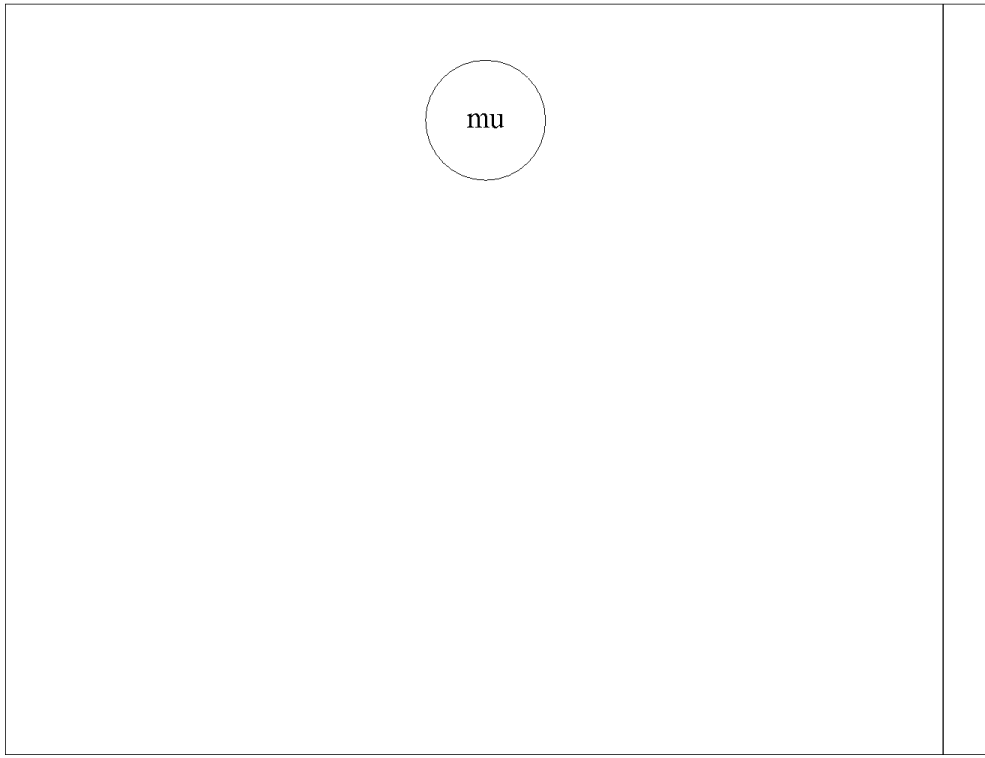
GMRF approximations

- can approximate $\pi(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ using a second-order Taylor expansion.
- Prominent example: Poisson-regression.
- Such approximations can be surprisingly accurate in many cases and can be interpreted as integrating $\mathbf{x}$ *approximately* out of $\pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})$.

Merging GMRFs using conditioning (I)

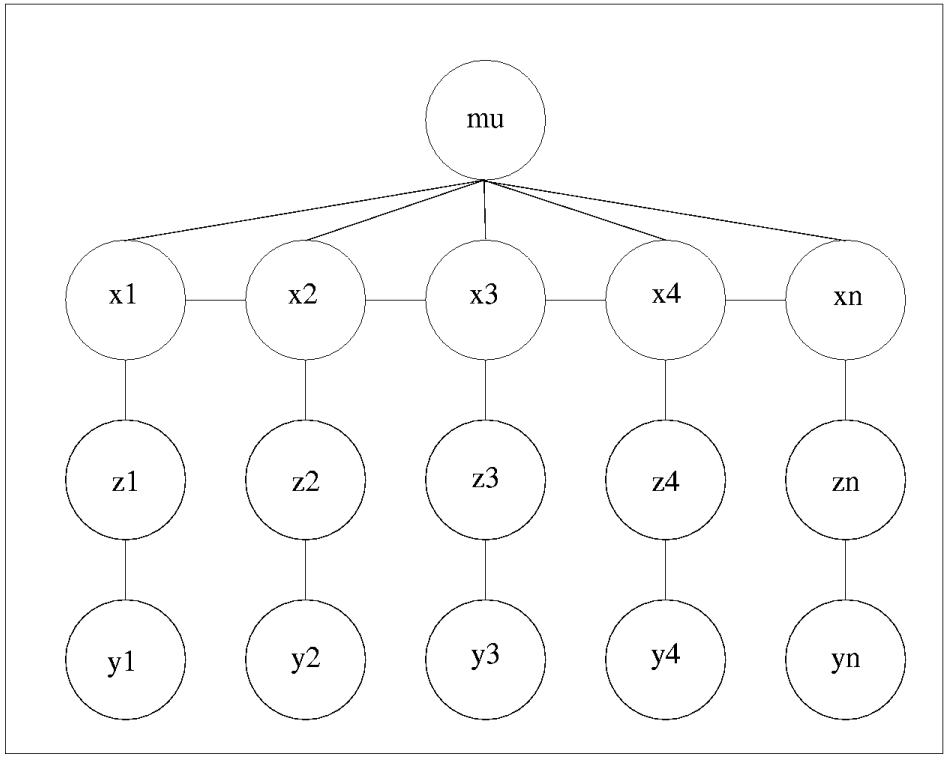
$$
\begin{aligned}
\mu &\sim \mathcal{N}(0, 1) \\
\mathbf{x} - \mu \mid \mu &\sim \text{AR}(1) \\
\mathbf{z} \mid \mathbf{x} &\sim \mathcal{N}(\mathbf{x}, \mathbf{I}) \\
\mathbf{y} \mid \mathbf{z} &\sim \mathcal{N}(\mathbf{z}, \mathbf{I})
\end{aligned}
$$

- $\mathbf{x}^* = (\mu, \mathbf{x}, \mathbf{z}, \mathbf{y})$ is a GMRF
- $\mathbf{x}^* \mid \mathbf{y}$ is a GMRF



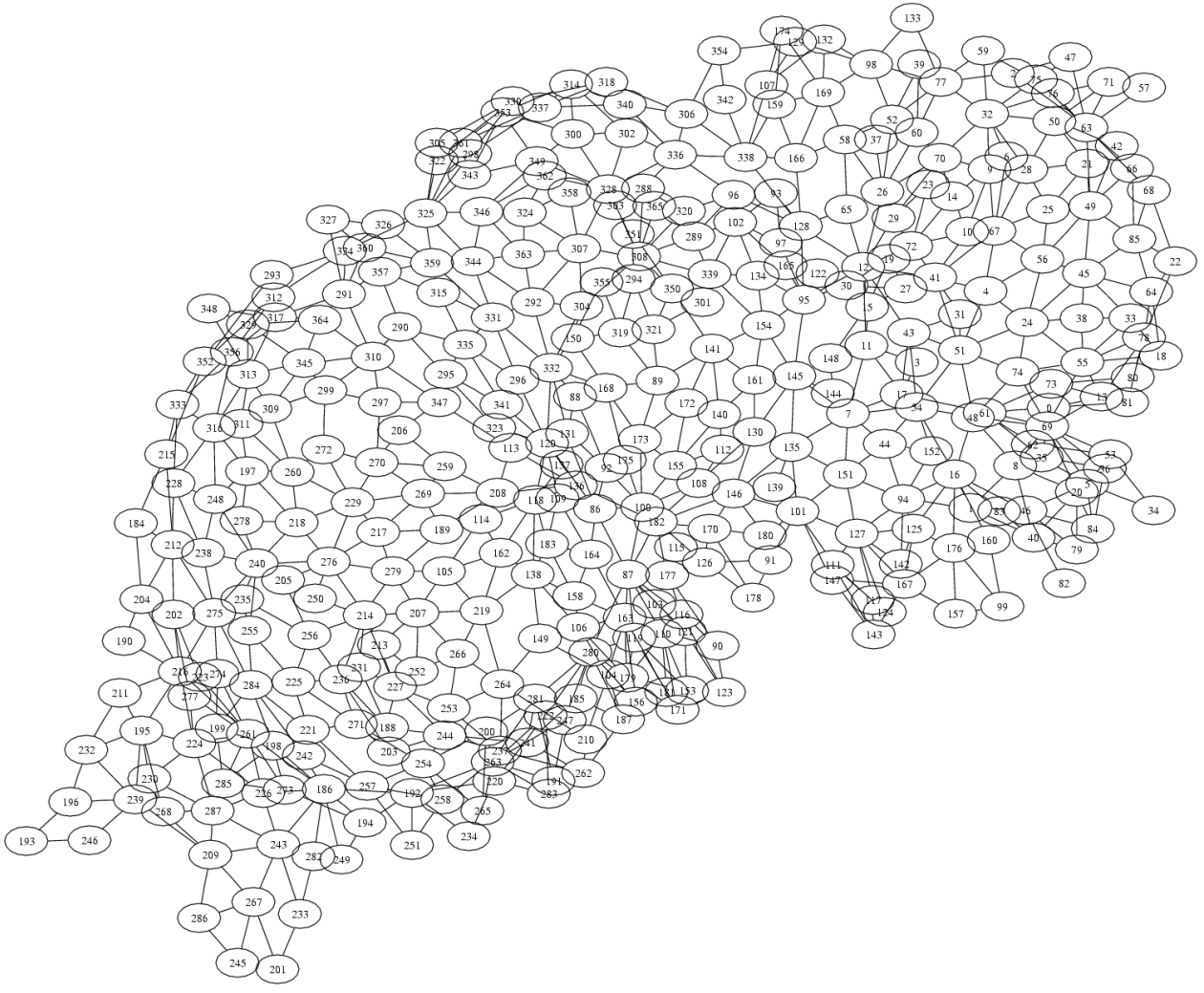
Merging GMRFs using conditioning (I)

$$
\begin{aligned}
\mu &\sim \mathcal{N}(0, 1) \\
\mathbf{x} - \mu \mid \mu &\sim \text{AR}(1) \\
\mathbf{z} \mid \mathbf{x} &\sim \mathcal{N}(\mathbf{x}, \mathbf{I}) \\
\mathbf{y} \mid \mathbf{z} &\sim \mathcal{N}(\mathbf{z}, \mathbf{I})
\end{aligned}
$$



- $\mathbf{x}^* = (\mu, \mathbf{x}, \mathbf{z}, \mathbf{y})$ is a GMRF
- $\mathbf{x}^* \mid \mathbf{y}$ is a GMRF
- Additional hyperparameters $\boldsymbol{\theta}$

Merging GMRFs using conditioning (II)



Merging GMRFs using conditioning (III)

If

$$\begin{aligned} \mathbf{x} &\sim \mathcal{N}(0, \mathbf{Q}^{-1}) \\ \mathbf{y} \mid \mathbf{x} &\sim \mathcal{N}(\mathbf{x}, \mathbf{K}^{-1}) \\ \mathbf{z} \mid \mathbf{x}, \mathbf{y} &\sim \mathcal{N}(\mathbf{y}, \mathbf{H}^{-1}) \end{aligned}$$

then

$$\text{Prec}(\mathbf{x}, \mathbf{y}, \mathbf{z}) = \begin{bmatrix} \mathbf{Q} + \mathbf{K} & -\mathbf{K} & \mathbf{0} \\ & \mathbf{K} + \mathbf{H} & -\mathbf{H} \\ & & \mathbf{H} \end{bmatrix}$$

which is sparse if $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{H}$ are.

Part VI

Applications

Outline I

Normal data: Munich rental guide

- Response variable y_i: rent pr m^2
- location
- floor space
- year of construction
- various indicator variables, such as
  - no central heating
  - no bathroom
  - large balcony

German law: increase in the rent is based on an 'average rent' of a comparable flat.

Spatial regression model

$$y_i \sim \mathcal{N}\left(\mu + x^S(i) + x^C(i) + x^L(i) + \mathbf{z}_i^T\boldsymbol{\beta},\ 1/\kappa_{\mathbf{y}}\right)$$

$x^S(i)$ Floor space (continuous RW2)

$x^C(i)$ Year of construction (continuous RW2)

$x^L(i)$ Location (first order IGMRF)

$\boldsymbol{\beta}$ Parameter for indicator variables

- 380 spatial locations
- 2035 observations
- sum-to-zero constraint on spatial IGMRF model and the year of construction covariate

Inference using MCMC

Sub-block approach

$$(\mathbf{x}^S, \kappa_S), \quad (\mathbf{x}^C, \kappa_C), \quad (\mathbf{x}^L, \kappa_L), \quad \text{and} \quad (\beta, \mu, \kappa_{\mathbf{y}}).$$

Update each block at a time, using

$$\begin{aligned} \kappa_L^* &\sim q(\kappa_L^* \mid \kappa_L) \\ \mathbf{x}^{L,*} &\sim \pi(\mathbf{x}^{L,*} \mid \text{the rest}) \end{aligned}$$

and then accepts/rejects $(\kappa_L^*, \mathbf{x}^{L,*})$ jointly.

Log-RW proposal for precisions

It's convenient to propose new precisions using

$$\kappa^{\text{new}} = \kappa^{\text{old}} \cdot f$$

$$f \sim \pi(f) \propto 1 + 1/f$$

for f in $[1/F, F]$ and $F > 1$.

With this choice

$$\frac{q(\kappa^{\text{old}} \mid \kappa^{\text{new}})}{q(\kappa^{\text{new}} \mid \kappa^{\text{old}})} = 1$$

$$\text{Var}(\kappa^{\text{new}} \mid \kappa^{\text{old}}) \propto (\kappa^{\text{old}})^2$$

Full conditional $\pi(\mathbf{x}^{L,*}|\textit{the rest})$

Introduce 'fake' data $\tilde{\mathbf{y}}$

$$\tilde{y}_i = y_i - \left(\mu + x^S(i) + x^C(i) + \mathbf{z}_i^T \boldsymbol{\beta}\right),$$

The full conditional of $\mathbf{x}^L$ is

$$\begin{aligned} \pi(\mathbf{x}^L \mid \text{the rest}) &\propto \exp(-\frac{\kappa_L}{2} \sum_{i \sim j} (x_i^L - x_j^L)^2) \\ &\times \exp(-\frac{\kappa_{\mathbf{y}}}{2} \sum_k \left(\tilde{y}_k - x^L(k)\right)^2). \end{aligned}$$

The data $\tilde{\mathbf{y}}$ do not introduce extra dependence between the x_i^L's, as $\tilde{y}_i$ acts as a noisy observation of x_i^L.

Denote by n_i the number of neighbors to location i and let $L(i)$ be

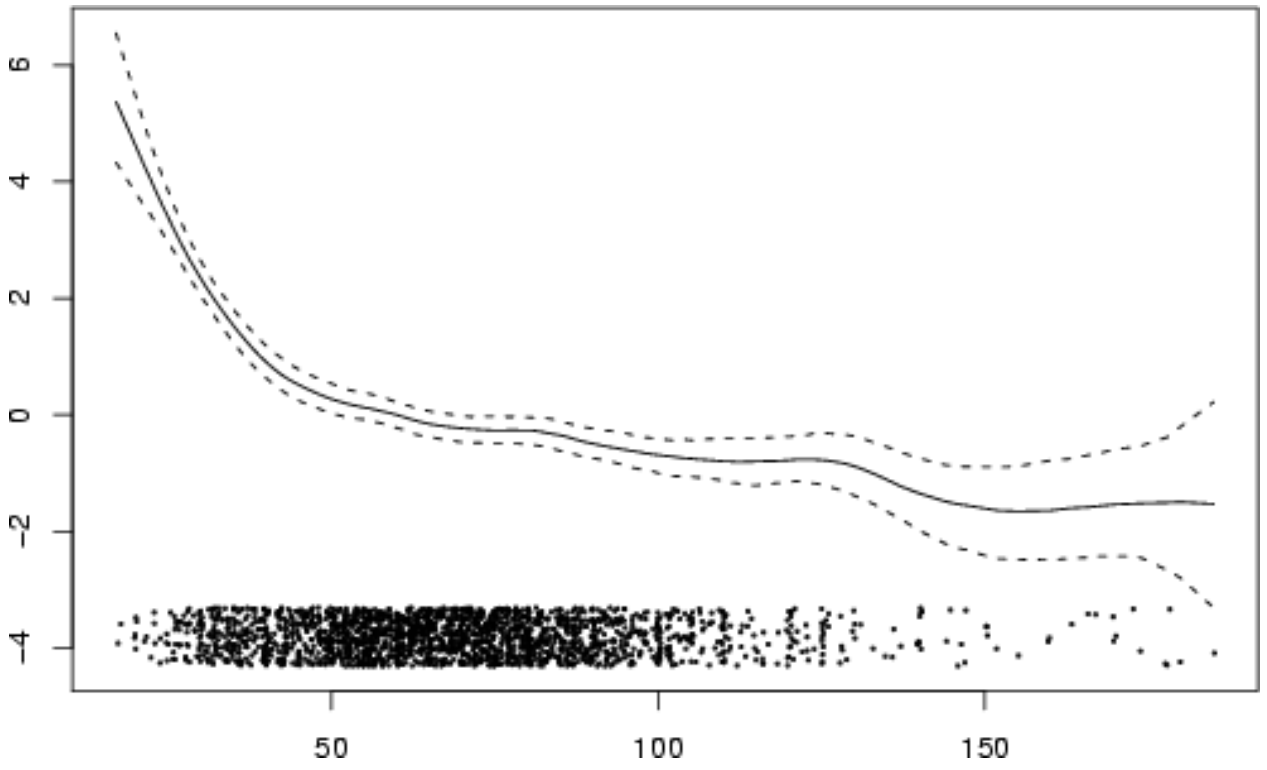
$$L(i) = \{k \ : \ x^L(k) = x_i^L\},$$

where its size is $|L(i)|$.

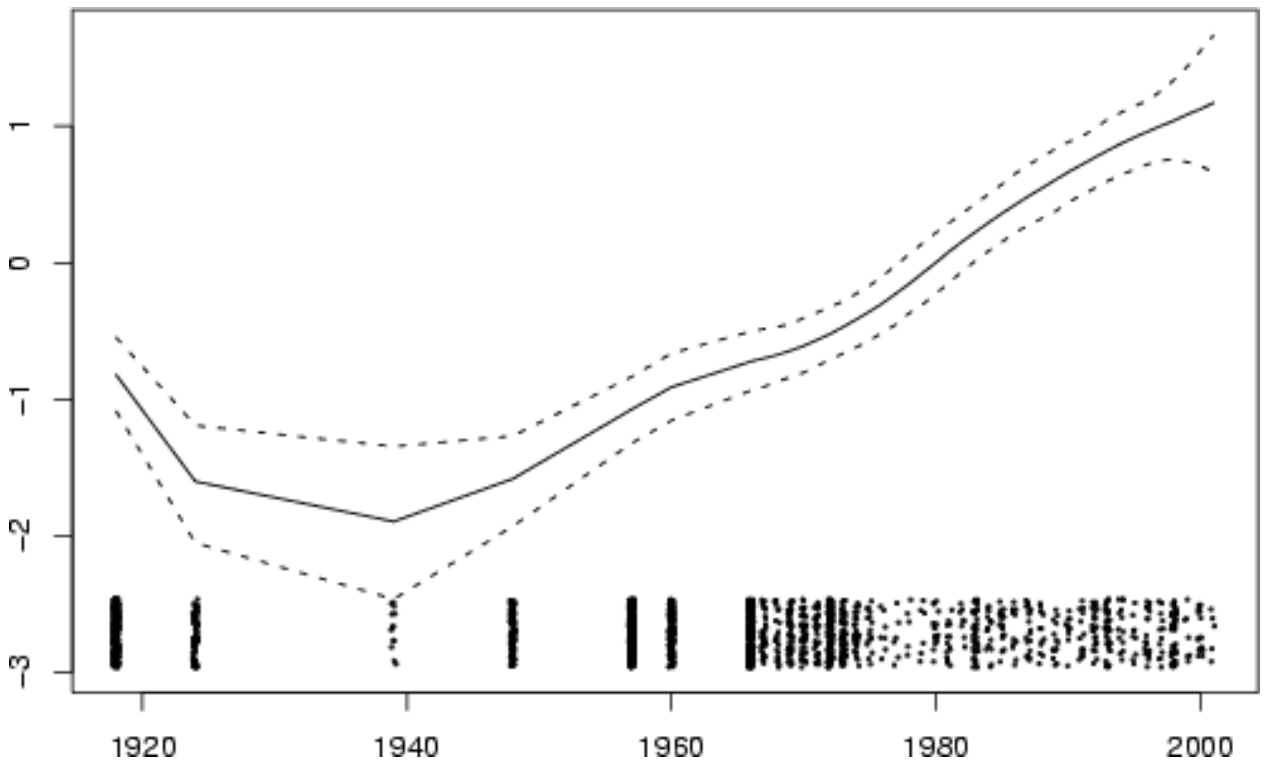
The full conditional of $\mathbf{x}^L$ is a GMRF with parameters $(\mathbf{Q}^{-1}\mathbf{b}, \mathbf{Q})$, where

$$\begin{aligned} b_i &= \kappa_{\mathbf{y}} \sum_{k \in L_i} \tilde{y}_k \quad \text{and} \\ Q_{ij} &= \begin{cases} \kappa_L n_i + \kappa_{\mathbf{y}} |L(i)| & \text{if } i = j \\ -\kappa_L & \text{if } i \sim j \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

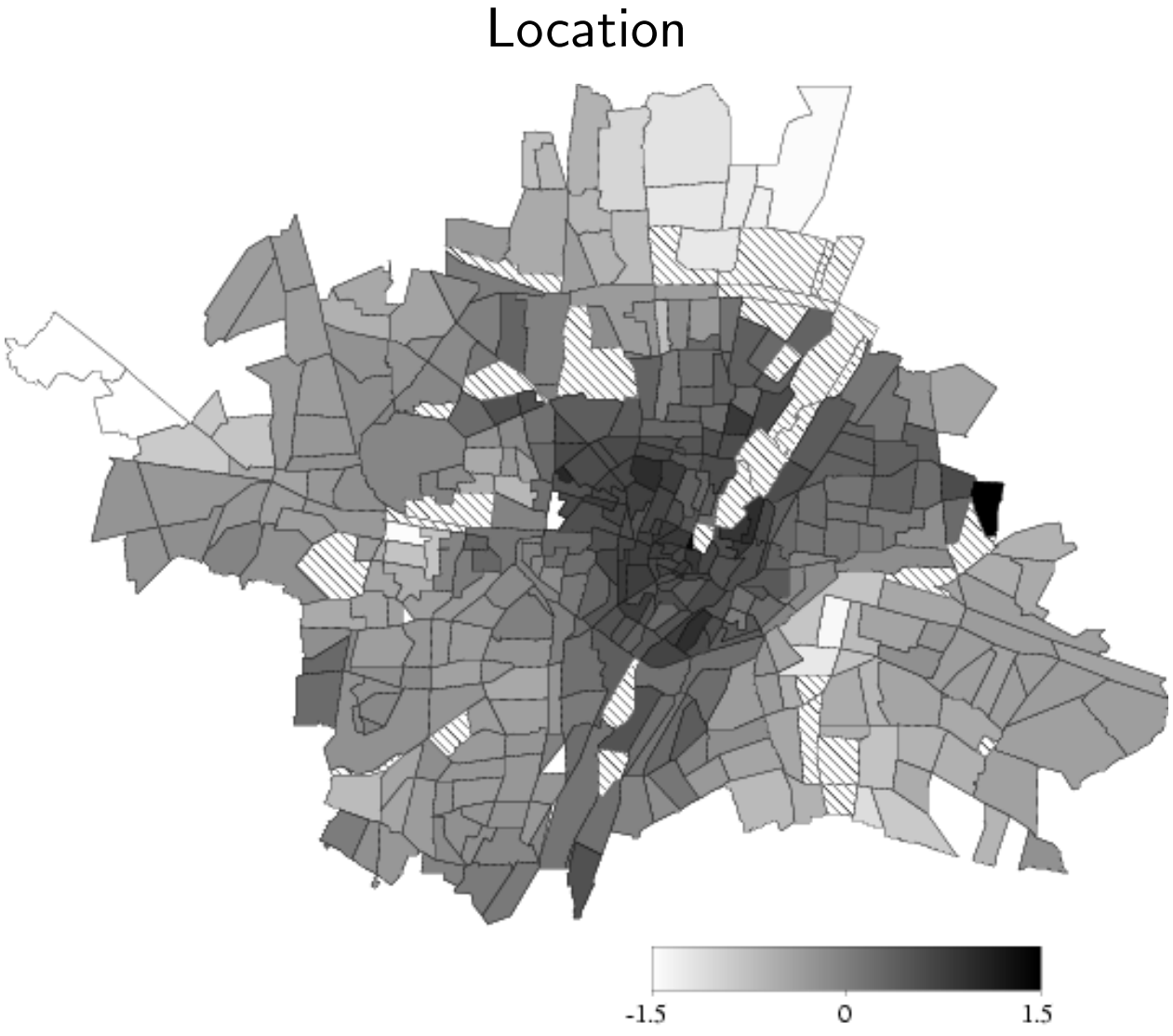
Results



Results



Results



Normal-mixture representation

Theorem (Kelker, 1971)

If x has density $f(x)$ symmetric around 0, then there exist independent random variables z and v, with z standard normal such that $x = z/v$ iff the derivatives of $f(x)$ satisfy

$$\left(-\frac{d}{dy}\right)^k f(\sqrt{y}) \geq 0$$

for $y > 0$ and for $k = 1, 2, \ldots$.

- Student-t
- Logistic and Laplace

Corresponding mixing distribution for the precision parameter λ that generates these distributions as scale mixtures of normals.

Distribution of x	Mixing distribution of λ
Student-t_ν	$\mathcal{G}(\nu/2, \nu/2)$
Logistic	$1/(2K)^2$ where K is Kolmogorov-Smirnov distributed
Laplace	$1/(2E)$ where E is exponential distributed

RW1 with t_ν – increments

Replace the assumption of normally distributed increments by a Student-t_ν distribution to allow for larger jumps in the sequence $\mathbf{x}$.

Introduce $n-1$ independent $\mathcal{G}(\nu/2, \nu/2)$ scale mixture variables λ_i:

$$\Delta x_i \mid \lambda_i \overset{\text{iid}}{\sim} \mathcal{N}(0, (\kappa\lambda_i)^{-1}), \quad i = 1, \ldots, n-1.$$

Observe data $y_i \sim \mathcal{N}(x_i, \kappa_{\mathbf{y}}^{-1})$ for $i = 1, \ldots, n$

The posterior density for $(\mathbf{x}, \boldsymbol{\lambda})$ is

$$\pi(\mathbf{x}, \boldsymbol{\lambda} \mid \mathbf{y}) \propto \pi(\mathbf{x} \mid \boldsymbol{\lambda})\, \pi(\boldsymbol{\lambda})\, \pi(\mathbf{y} \mid \mathbf{x}).$$

Note that

- $\mathbf{x}|(\mathbf{y}, \boldsymbol{\lambda})$ is now a GMRF, while
- $\lambda_1, \ldots, \lambda_{n-1}|(\mathbf{x}, \mathbf{y})$ are conditionally independent gamma distributed with parameters $(\nu + 1)/2$ and $(\nu + \kappa(\Delta x_i)^2)/2$.

Use the sub-block algorithm with blocks

$$(\boldsymbol{\theta}, \mathbf{x}), \quad \boldsymbol{\lambda}$$

Example: Binary regression

GMRF **x** and Bernoulli data

$$\begin{aligned} y_i &\sim \mathcal{B}(g^{-1}(x_i)) \\ g(p) &= \begin{cases} \log(p/(1-p)) & \text{logit link} \\ \Phi(p) & \text{probit link} \end{cases} \end{aligned}$$

Equivalent representation using auxiliary variables **w**

$$\begin{aligned} \epsilon_i &\overset{\text{iid}}{\sim} \mathcal{N}(0,1) \\ w_i &= x_i + \epsilon_i \\ y_i &= \begin{cases} 1 & \text{if } w_i > 0 \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

for the probit-link.

MCMC Inference

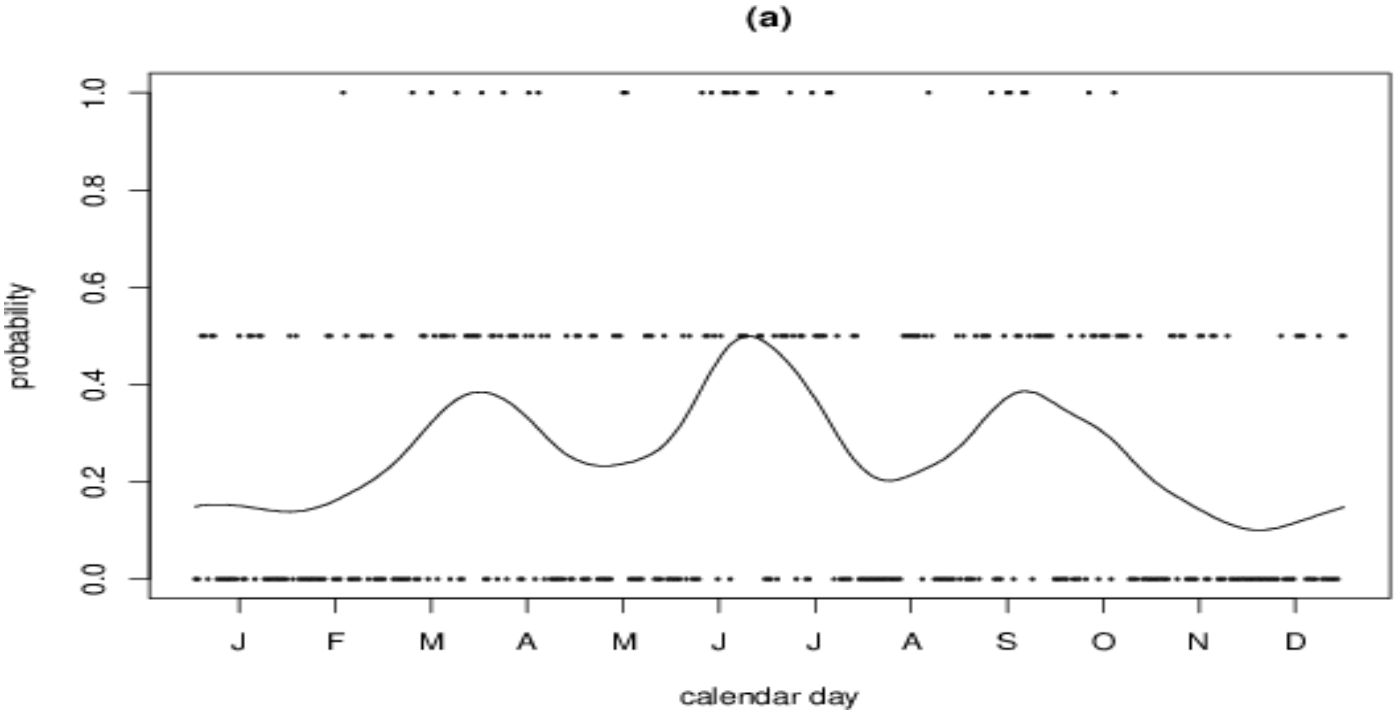
- Full conditional for the GMRF prior is a GMRF
- Full conditional for the auxiliary variables

$$\pi(\mathbf{w} \mid \ldots) = \prod_i \pi(w_i \mid \ldots)$$

Use the sub-block algorithm with blocks

$$(\boldsymbol{\theta}, \mathbf{x}), \quad \mathbf{w}$$

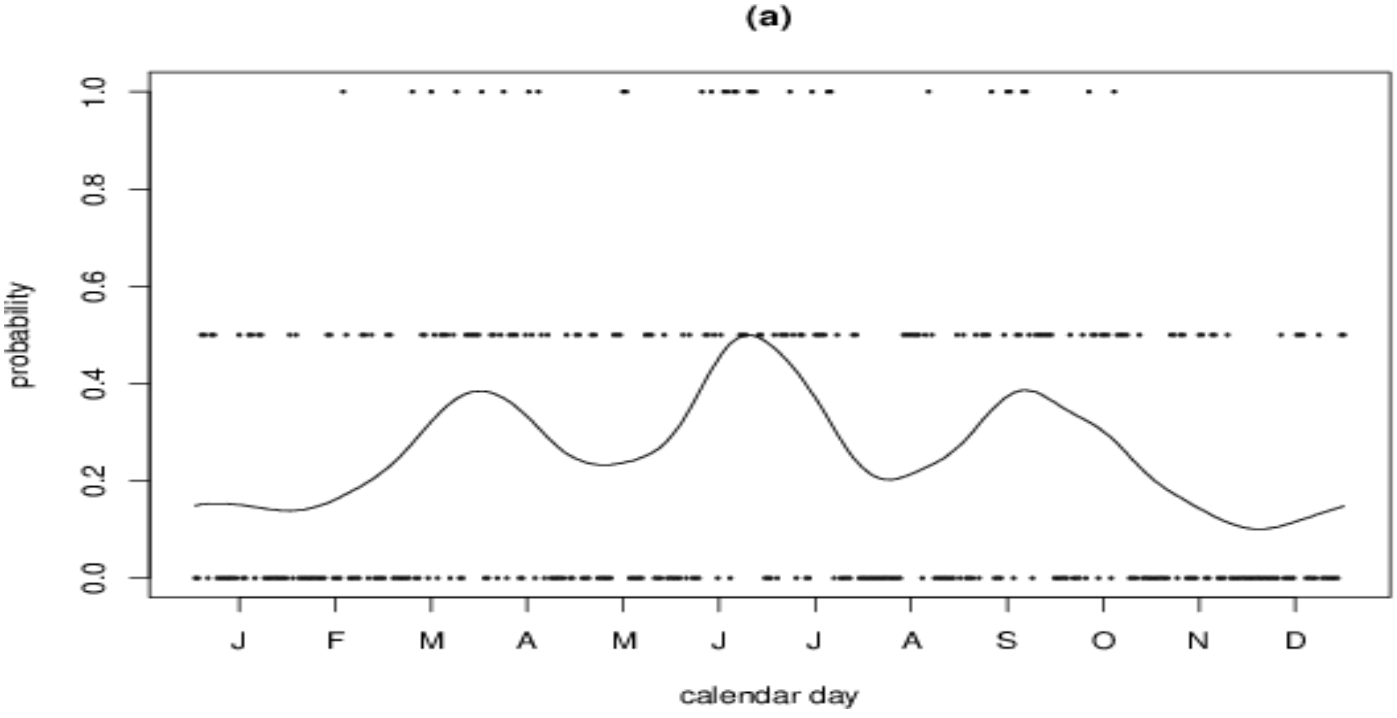
Simple example: Tokyo rainfall data



Stage 1 Binomial data

$$y_i \sim \begin{cases} \text{Binomial}(2, p(x_i)) \\ \text{Binomial}(1, p(x_i)) \end{cases}$$

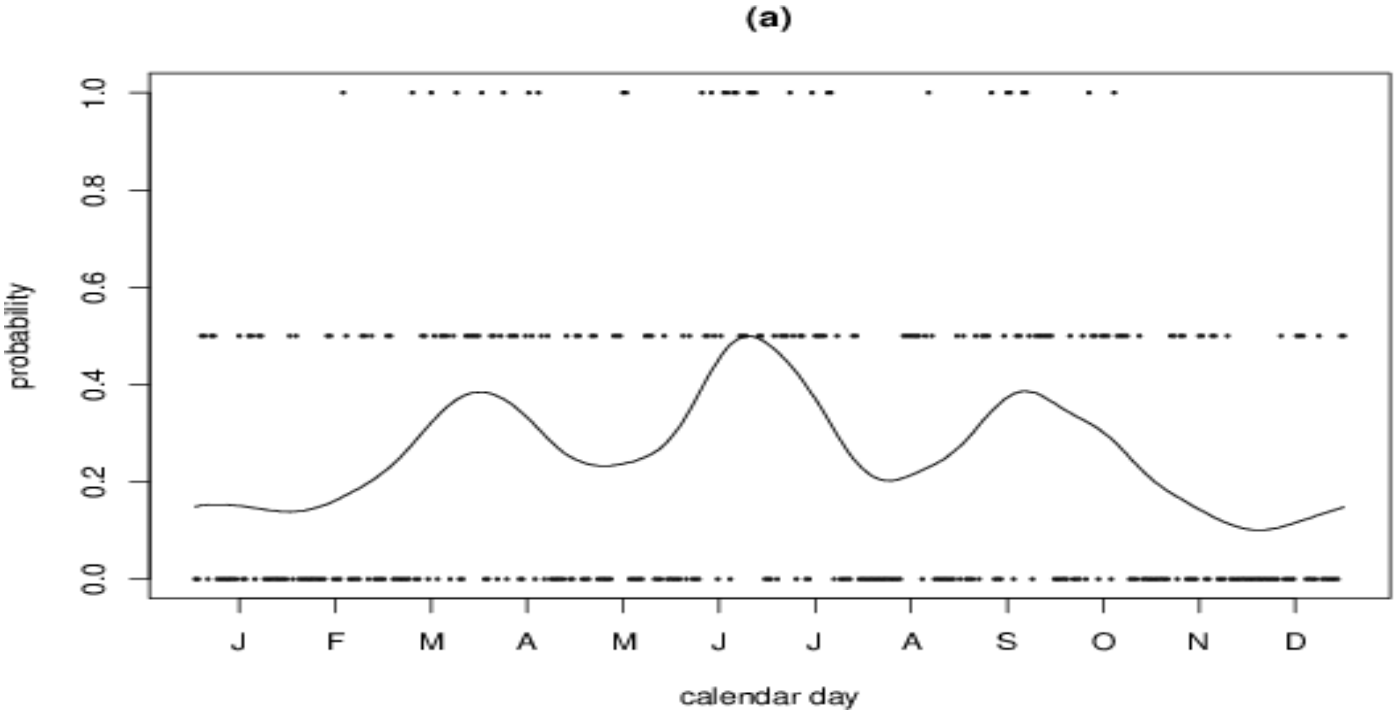
Simple example: Tokyo rainfall data



Stage 2 Assume a smooth latent $\mathbf{x}$,

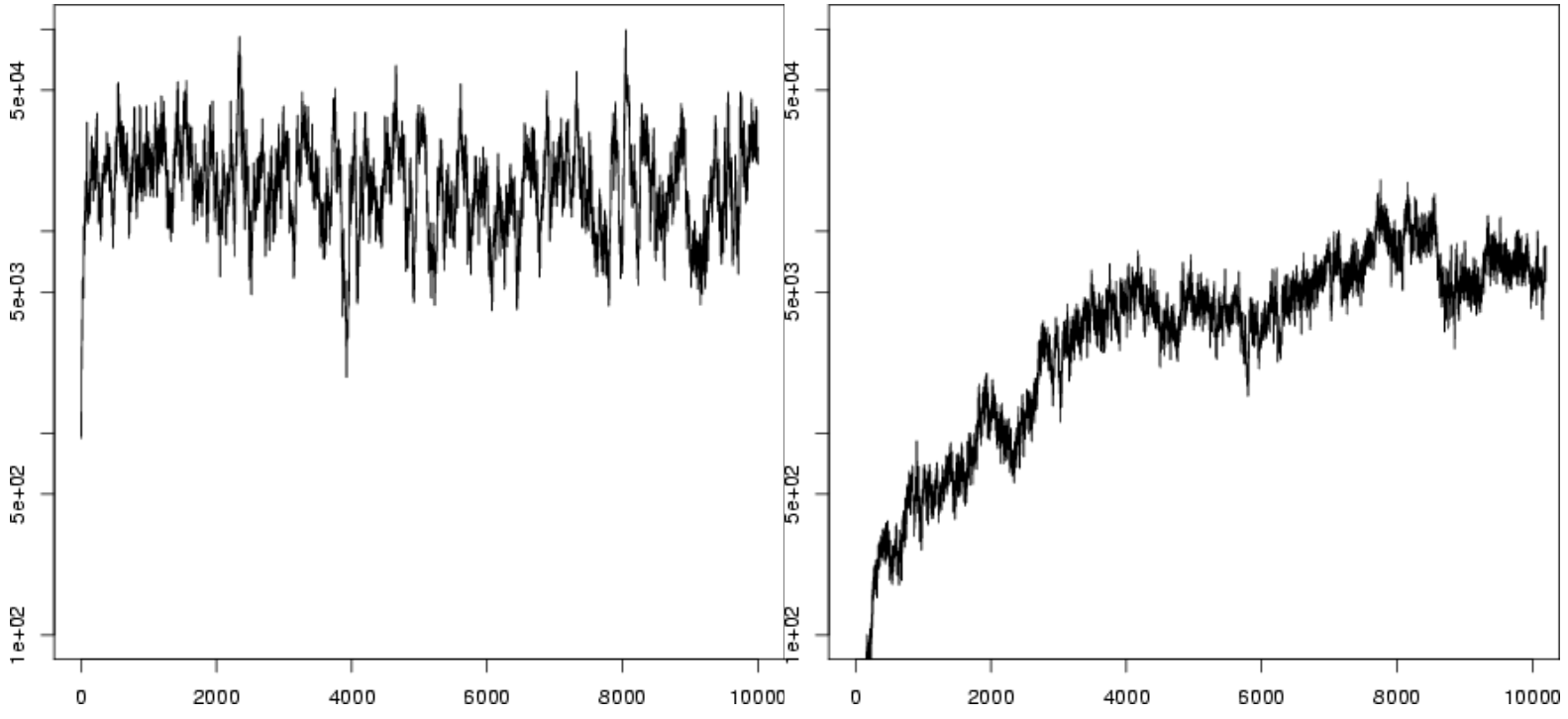
$$\mathbf{x} \sim RW2(\kappa), \qquad \text{logit}(p_i) = x_i$$

Simple example: Tokyo rainfall data

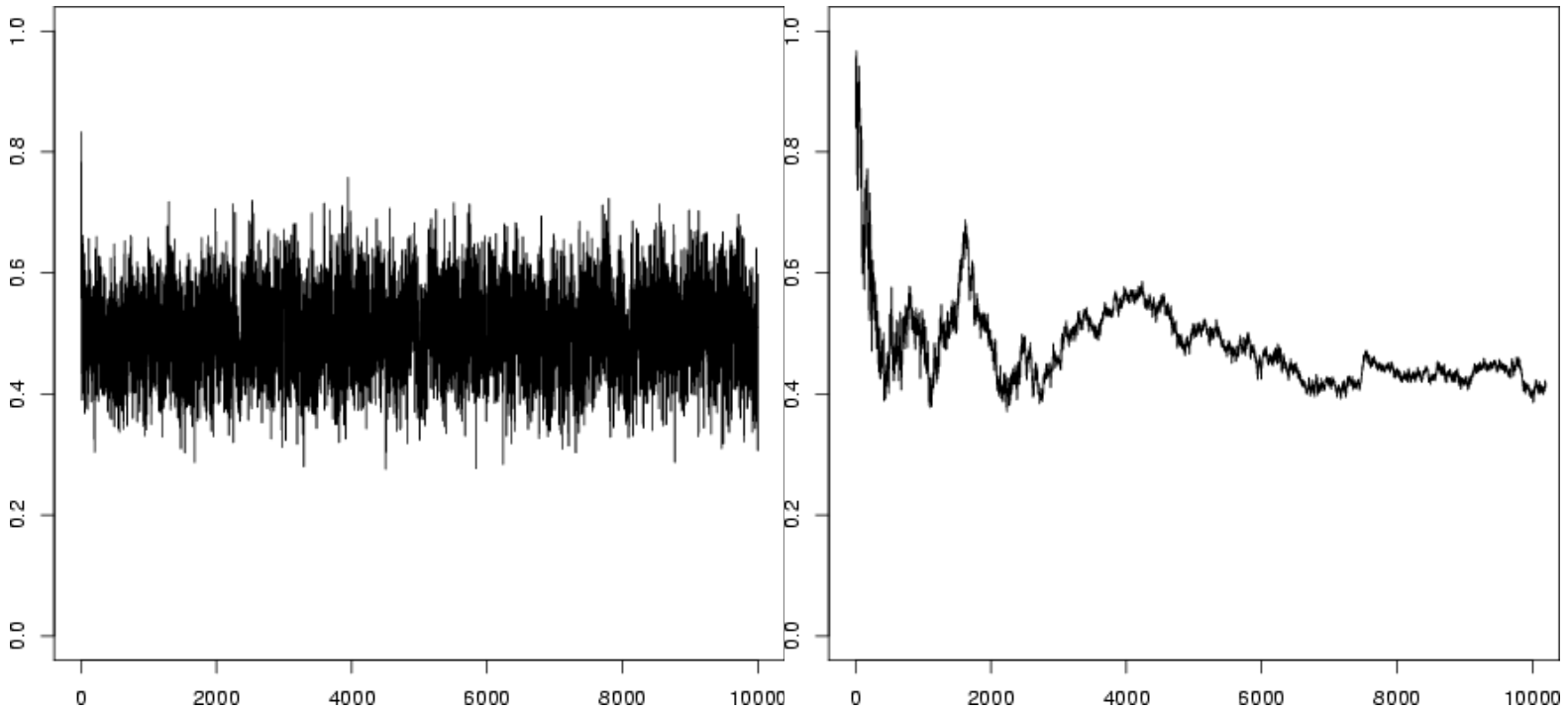


Stage 3 Gamma(α, β)-prior on κ

Simple example: Tokyo rainfall data



Simple example: Tokyo rainfall data



- These auxiliary variables methods works better than one might think.
- Not that strong dependency between the blocks:

$$(\boldsymbol{\theta}, \mathbf{x}), \quad \mathbf{w}$$

- Nearly as random normal data.

Non-normal response models

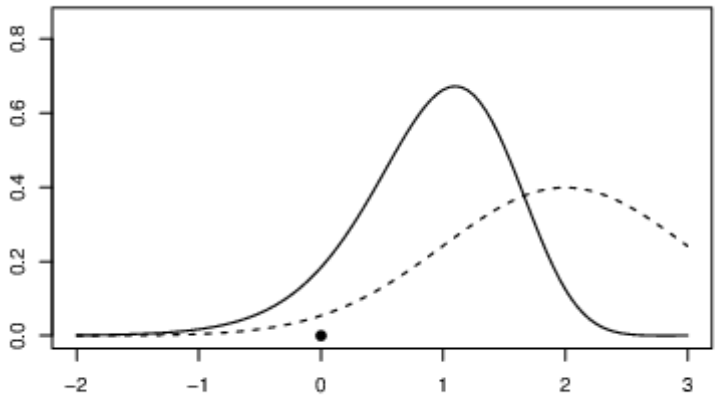
Example of full conditional

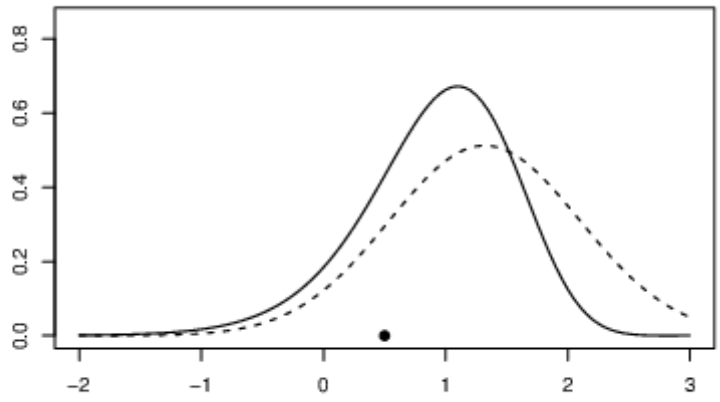
$$\pi(\mathbf{x} \mid \mathbf{y}) \propto \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} - \sum_i \exp(x_i)\right)$$
$$\approx \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T(\mathbf{Q} + \text{diag}(c_i))(\mathbf{x}-\boldsymbol{\mu})\right)$$

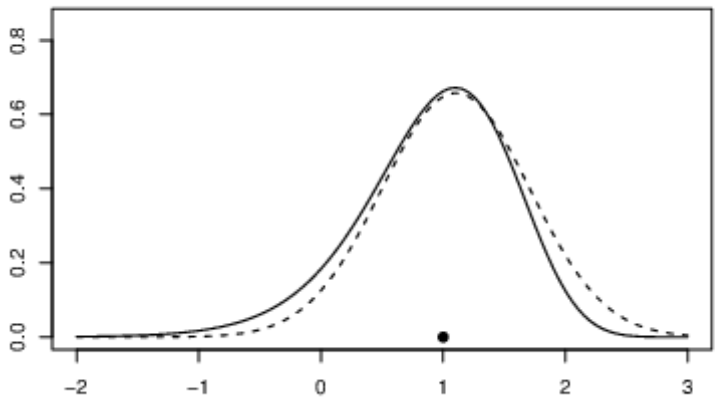
Construct GMRF approximation

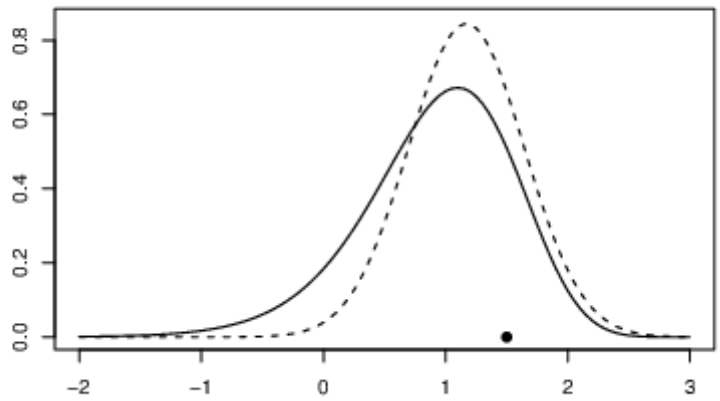
- Locate the mode $\mathbf{x}^*$
- Expand to second order
- To obtain the GMRF approximation
- The graph is unchanged

Why use the mode?









How to find the mode?

- Numerical gradient-search methods
  - evaluate the gradient and hessian in $\mathcal{O}(n)$ flops.
  - faster for large GMRFs
- Newton-Raphson method
  - Current $\mathbf{x}$ is initial value
  - Taylor-expand around $\mathbf{x}$
  - Compute the mean in the GMRF-approximation
  - Repeat this until convergence
- Further complications with linear constraints $\mathbf{Ax} = \mathbf{b}$.

"Taylor" or not? (I)

Consider a non-quadratic function $g(x)$

$$g(x) = g(x_0) + (x - x_0)g'(x_0) + \frac{1}{2}(x - x_0)^2 g''(x_0) + \dots$$

- error is zero in x_0
- error typically increase with $|x - x_0|$

"Taylor" or not? (II)

Alternative approximation

$$g(x) \approx \widehat{g}(x) = a + b + \frac{1}{2}cx^2 \tag{84}$$

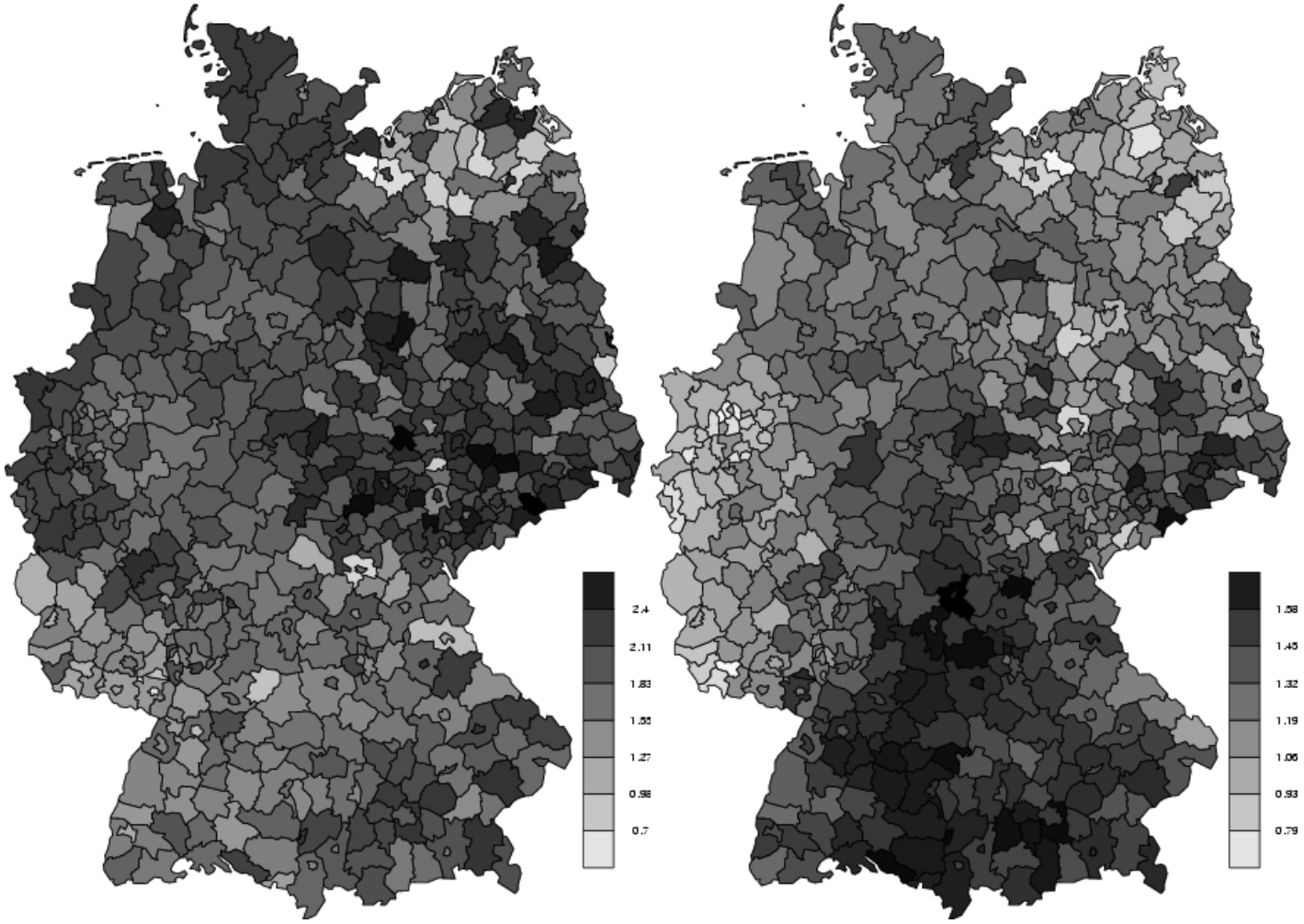
where a, b and c are such that

$$g(x_0) = \widehat{g}(x_0), \quad g(x_0 \pm h) = \widehat{g}(x_0 \pm h)$$

- More "uniformly distributed error"
- Better for distribution approximation
- ....if h is a rough guess of ROI.

Preference to numerical approxmations of derivatives compared to analytical expressions.

Joint analysis of diseases



SMR Oral

SMR Lung

Separate analysis: model

y_{ij}: number of cases in area i for cancer type j.

$$y_{ij} \sim \mathcal{P}(e_{ij} \exp(\eta_{ij})),$$

η_{ij}: log relative risk in area i for disease j.
e_{ij}: constants

- Decompose the log-relative risk $\boldsymbol{\eta}$ as

$$\boldsymbol{\eta} = \mu \mathbf{1} + \mathbf{u} + \mathbf{v}$$

- μ is the overall mean
- $\mathbf{u}$ a *spatially structured* component (IGMRF)
- $\mathbf{v}$ an *unstructured* component (random effects)

Separate analysis: model

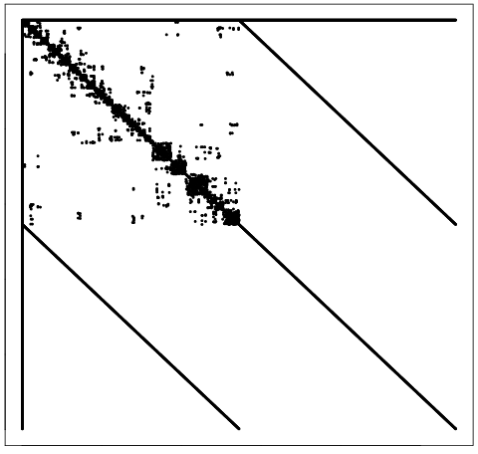
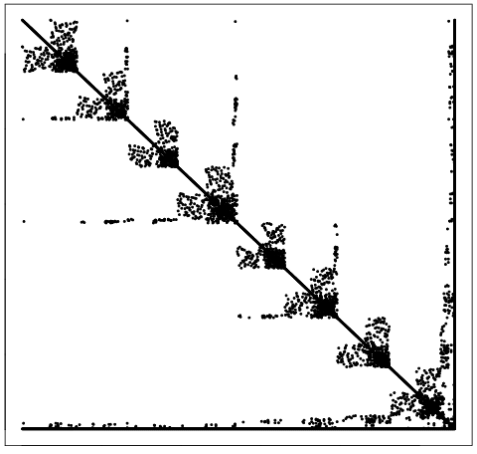
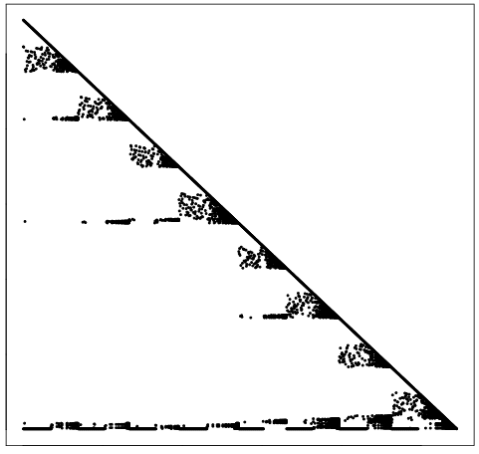
Posterior

$$\begin{aligned}
\pi(\mathbf{x}, \boldsymbol{\kappa} \mid \mathbf{y}) &\propto \kappa_{\mathbf{v}}^{n/2} \kappa_{\mathbf{u}}^{(n-1)/2} \exp\left(-\frac{1}{2}\mathbf{x}^T \mathbf{Q}\mathbf{x}\right) \\
&\times \exp\left(\sum_i y_i \eta_i - e_i \exp(\eta_i)\right) \pi(\boldsymbol{\kappa}) \\
\mathbf{Q} &= \begin{pmatrix} \kappa_\mu + n\kappa_{\mathbf{v}} & \kappa_{\mathbf{v}}\mathbf{1}^T & -\kappa_{\mathbf{v}}\mathbf{1}^T \\ \kappa_{\mathbf{v}}\mathbf{1} & \kappa_{\mathbf{u}}\mathbf{R} + \kappa_{\mathbf{v}}\mathbf{I} & -\kappa_{\mathbf{v}}\mathbf{I} \\ -\kappa_{\mathbf{v}}\mathbf{1} & -\kappa_{\mathbf{v}}\mathbf{I} & \kappa_{\mathbf{v}}\mathbf{I} \end{pmatrix}
\end{aligned}$$

$\mathbf{Q}$ is a $2n + 1 \times 2n + 1$ matrix where $n = 544$.

The spatially structured term $\mathbf{u}$ has a sum-to-zero constraint, $\mathbf{1}^T\mathbf{u} = 0$.

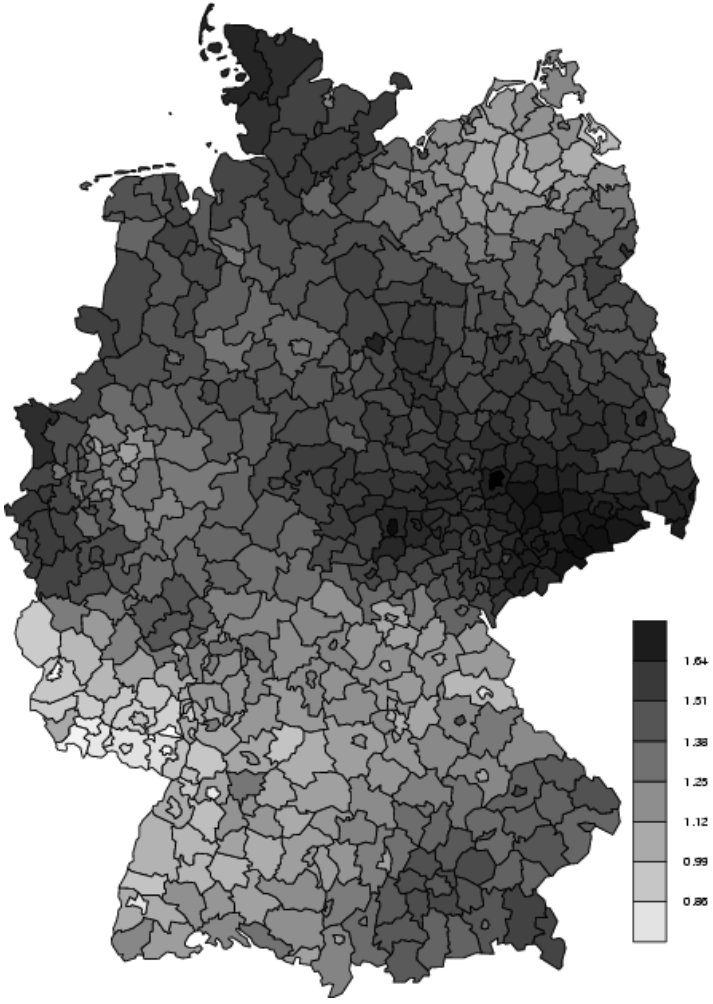
Separate analysis: MCMC

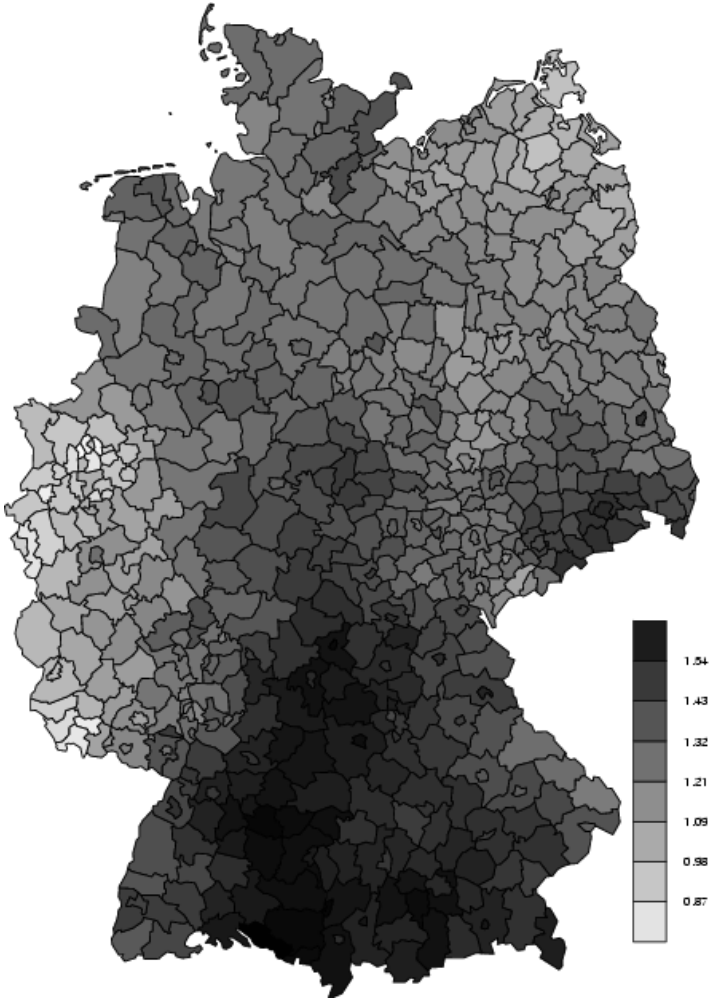
Joint update of all parameters:

$$\begin{aligned} \kappa_{\mathbf{u}}^{*} &\sim q\left(\kappa_{\mathbf{u}}^{*} \mid \kappa_{\mathbf{u}}\right) \\ \kappa_{\mathbf{v}}^{*} &\sim q\left(\kappa_{\mathbf{v}}^{*} \mid \kappa_{\mathbf{v}}\right) \\ \mathbf{x}^{*} &\sim \widetilde{\pi}\left(\mathbf{x} \mid \boldsymbol{\kappa}^{*}, \mathbf{y}\right) \end{aligned}$$

Separate analysis: Results



Oral



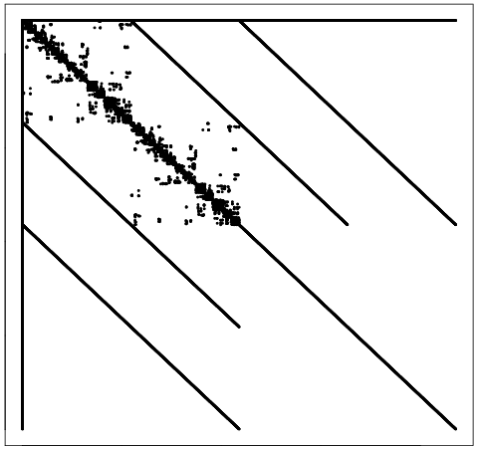
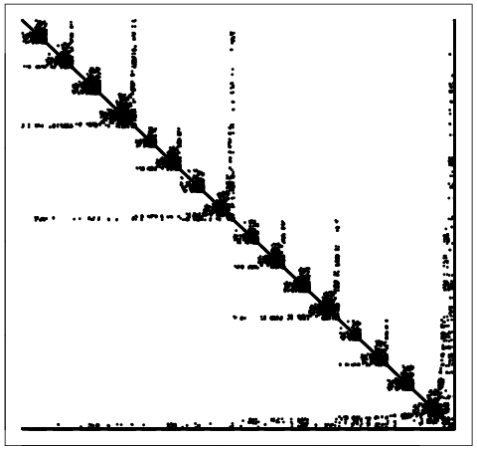
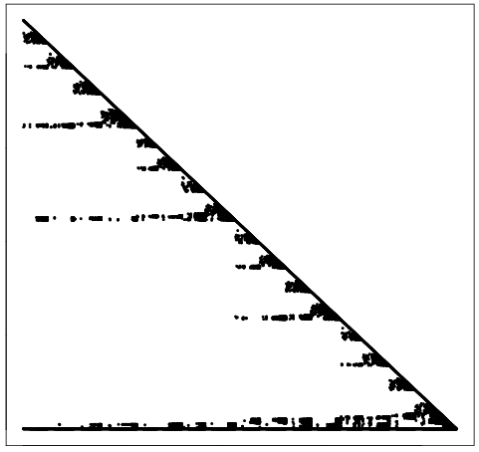
Lung

Joint analysis: Model

- Oral cavity and lung cancer relates to tobacco smoking ($\mathbf{u}_1$)
- Oral cancer relates to alcohol consumption ($\mathbf{u}_2$)
- Latent shared and specific spatial components:

$$\begin{aligned}\boldsymbol{\eta}_1 \mid \mathbf{u}_1, \mathbf{u}_2, \boldsymbol{\mu}, \boldsymbol{\kappa} &\sim \mathcal{N}(\mu_1 \mathbf{1} + \delta \mathbf{u}_1 + \mathbf{u}_2,\ \kappa_{\boldsymbol{\eta}_1}^{-1} \mathbf{I}) \\ \boldsymbol{\eta}_2 \mid \mathbf{u}_1, \mathbf{u}_2, \boldsymbol{\mu}, \boldsymbol{\kappa} &\sim \mathcal{N}(\mu_2 \mathbf{1} + \delta^{-1} \mathbf{u}_1,\ \kappa_{\boldsymbol{\eta}_2}^{-1} \mathbf{I}),\end{aligned}$$

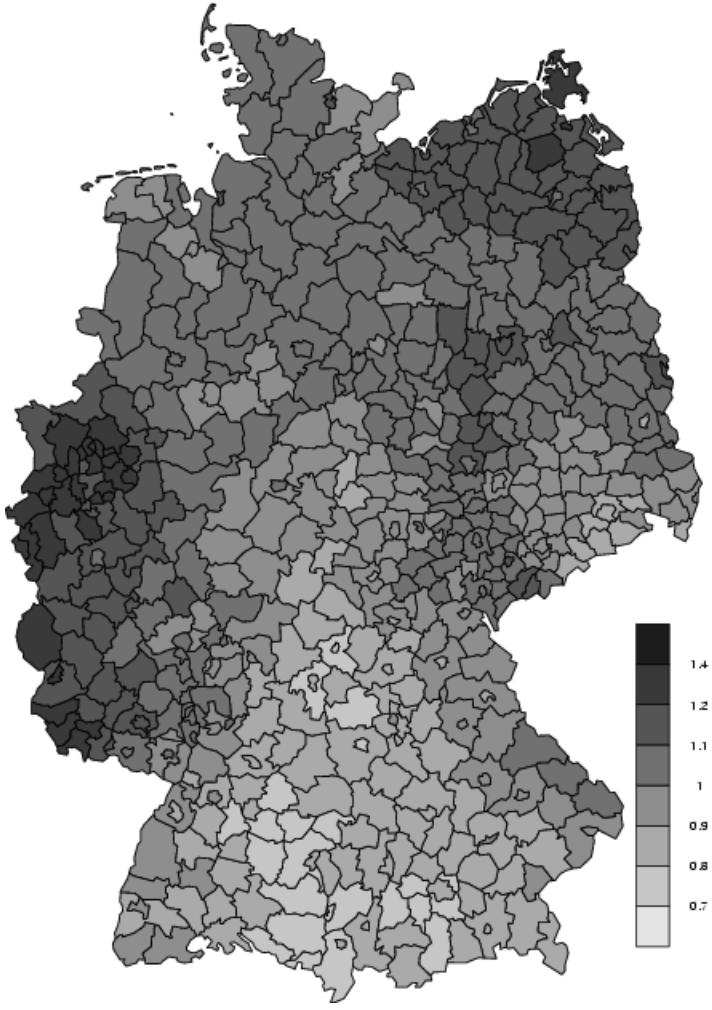
Joint analysis: MCMC

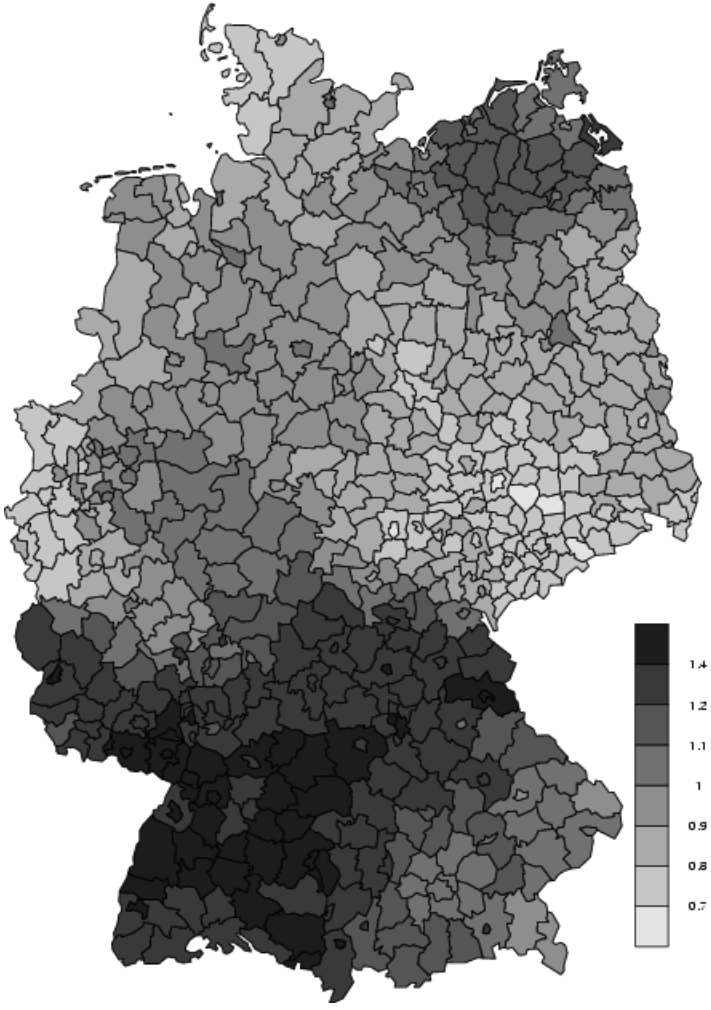
Update all parameters in one block

- Simple (log-)RW for the hyperparamteters
- Sample $(\mu_1, \boldsymbol{\eta}_1, \mathbf{u}_1, \mu_2, \boldsymbol{\eta}_2, \mathbf{u}_2)$ from the GMRF approximation.
- Accepts/rejects jointly.

Joint analysis: Results



"tobacco"



"alcohol"

Independence samplers

For many of the examples it's quite easy/possible to construct *independence samplers*

- (More or less) exact samples
- The basis to methods avoiding completely MCMC

...continue tomorrow!

Part VII

Advanced topics

Outline I

*Computing marginal variances for GMRFs**

Let

$$\mathbf{Q} = \mathbf{V}\mathbf{D}\mathbf{V}^T$$

- where $\mathbf{D}$ is a diagonal matrix, and
- $\mathbf{V}$ is a lower triangular matrix with ones on the diagonal.

The matrix identity

$$\mathbf{\Sigma} = \mathbf{D}^{-1}\mathbf{V}^{-1} + (\mathbf{I} - \mathbf{V}^T)\mathbf{\Sigma}$$

define *recursions* which can be used to compute

- $\text{Var}(x_i)$ and $\text{Cov}(x_i, x_j)$ for $i \sim j$

essentially without cost when the Cholesky triangle $\mathbf{L}$ is known. This is a result from 1973.

Statistical derivation

Recall for a zero mean GMRF that

$$x_i \mid x_{i+1}, \ldots, x_n \sim \mathcal{N}(-\frac{1}{L_{ii}} \sum_{k=i+1}^{n} L_{ki} x_k, 1/L_{ii}^2), \quad i = n, \ldots, 1.$$

provides a sequential representation of the GMRF backward in "time" i.

Multiply by x_j, $j \geq i$, and taking expectation yields

$$\Sigma_{ij} = \delta_{ij}/L_{ii}^2 - \frac{1}{L_{ii}} \sum_{k \in \mathcal{I}(i)}^{n} L_{ki}\Sigma_{kj}, \quad j \geq i,\ i = n, \ldots, 1,$$

where $\mathcal{I}(i)$ as those k where L_{ki} is non-zero,

$$\mathcal{I}(i) = \{k > i\ :\ L_{ki} \neq 0\}$$

and δ_{ij} is one if $i = j$ and zero otherwise.

We can use

$$\Sigma_{ij} = \delta_{ij}/L_{ii}^2 - \frac{1}{L_{ii}} \sum_{k \in \mathcal{I}(i)}^{n} L_{ki}\Sigma_{kj}, \quad j \geq i, \; i = n, \ldots, 1,$$

to compute Σ_{ij} for each ij:

- Outer loop $i = n, \ldots, 1$
- Inner loop $j = n, \ldots, i$

Example

Let $n = 3$, $\mathcal{I}(1) = \{2, 3\}$, $\mathcal{I}(2) = \{3\}$, then we get

$$\Sigma_{33} = \frac{1}{L_{33}^2} \qquad \Sigma_{23} = -\frac{1}{L_{22}}(L_{32}\Sigma_{33})$$

$$\Sigma_{22} = \frac{1}{L_{22}^2} - \frac{1}{L_{22}}(L_{32}\Sigma_{32}) \qquad \Sigma_{13} = -\frac{1}{L_{11}}(L_{21}\Sigma_{23} + L_{31}\Sigma_{33})$$

$$\Sigma_{12} = -\frac{1}{L_{11}}(L_{21}\Sigma_{22} + L_{31}\Sigma_{32}) \qquad \Sigma_{11} = \frac{1}{L_{11}^2} - \frac{1}{L_{11}}(L_{21}\Sigma_{21} + L_{31}\Sigma_{31})$$

where we also need to use that $\mathbf{\Sigma}$ is symmetric.

- Assume we want to compute all marginal variances.
- To do so, we need to compute Σ_{ij} (or Σ_{ji}) for all ij in some set $\mathcal{S}$.
- If the recursions can be solved by only computing Σ_{ij} for all $ij \in \mathcal{S}$ we say that the recursions are *solvable* using $\mathcal{S}$.

From

$$\Sigma_{ij} = \delta_{ij}/L_{ii}^2 - \frac{1}{L_{ii}} \sum_{k \in \mathcal{I}(i)}^{n} L_{ki}\Sigma_{kj}, \quad j \geq i,\ i = n, \ldots, 1, \tag{85}$$

it is evident that $\mathcal{S}$ must satisfy

$$ij \in \mathcal{S} \text{ and } k \in \mathcal{I}(i) \implies kj \in \mathcal{S} \tag{86}$$

We also need that $ii \in \mathcal{S}$ for $i = 1, \ldots, n$.

- $\mathcal{S} = \mathcal{V} \times \mathcal{V}$ is a valid set, but we want $|\mathcal{S}|$ to be minimal to avoid unnecessary computations.
- Such a minimal set depends however on the numerical values in $\mathbf{L}$, or $\mathbf{Q}$ implicitly.
- Denote by $\mathcal{S}(\mathbf{Q})$ a minimal set.

Theorem

The union of $\mathcal{S}(\mathbf{Q})$ *for all* $\mathbf{Q} > 0$ *with fixed graph* $\mathcal{G}$*, is a subset of*

$$\mathcal{S}^* = \{ij \in \mathcal{V} \times \mathcal{V} \ : \ j \geq i, \ i \text{ and } j \text{ are not separated by } F(i,j)\}$$

and $\mathcal{S}^*$ *is solvable.*

Proof. *We first note that* $ii \in \mathcal{S}^*$, *for* $i = 1, \ldots, n$, *since* i *and* i *are not separated by* $F(i, i)$. *We will now verify that the recursions are solvable using* $\mathcal{S}^*$. *The global Markov property ensure that if* $ij \notin \mathcal{S}^*$ *then* $L_{ji} = 0$ *for all* $\mathbf{Q} > 0$ *with fixed graph* $\mathcal{G}$. *We use this to replace* $\mathcal{I}(i)$ *with* $\mathcal{I}^*(i) = \{k > i : ik \in \mathcal{S}^*\}$ *in (86), which is legal since* $\mathcal{I}(i) \subseteq \mathcal{I}^*(i)$ *and the difference only identify terms* L_{ki} *which are zero. It is now sufficient to show that*

$$ij \in \mathcal{S}^* \text{ and } ik \in \mathcal{S}^* \implies kj \in \mathcal{S}^* \tag{87}$$

which implies (86). Eq. (87) is trivially true for $i \leq k = j$. *Fix now* $i < k < j$. *Then* $ij \in \mathcal{S}^*$ *says that there exists a path* $i, i_1, \ldots, i_n, j$, *where* $i_1, \ldots, i_n$ *are all smaller than* i, *and* $ik \in \mathcal{S}^*$ *says that exists a path* $i, i'_1, \ldots, i'_{n'}, k$, *where* $i'_1, \ldots, i'_{n'}$ *are all smaller than* i. *Then there is a path from* k *to* i *and from* i *to* j *where all nodes are less or equal to* i, *but then also less than* k *since* $i < k$. *Hence,* k *and* j *are not separated by* $F(k, j)$ *so* $kj \in \mathcal{S}^*$. *Finally, since* $\mathcal{S}^*$ *contains* $11, \ldots, nn$ *and only depend on* $\mathcal{G}$, *it must contain the union of all* $\mathcal{S}(\mathbf{Q})$ *since each* $\mathcal{S}(\mathbf{Q})$ *is minimal.* ■

Interpretation of $\mathcal{S}^*$

- $\mathcal{S}^*$ is the set of all possible non-zero elements in $\mathbf{L}$ based on $\mathcal{G}$ only.
- This is the set of L_{ji}'s that are computed when computing $\mathbf{Q} = \mathbf{L}\mathbf{L}^T$.
- Since $L_{ji} \neq 0$ in general when $i \sim j$, then we compute also $\text{Cov}(x_i, x_j)$ for $i \sim j$.
- Some of the L_{ij}'s might turn out to be zero depending on the conditional independence properties of the marginal density for $\mathbf{x}_{i:n}$ for $i = n, \ldots, 1$. This might cause a slight problem in practice (implementation dependent).

General algorithm

for $i = n, \ldots, 1$

for decreasing j *in* $\mathcal{I}(i)$

Compute $\Sigma_{i,j}$ *from Eq. (85)*

Band matrices

for $i = n, \ldots, 1$
  for $j = \min(i + b_w, n), \ldots, i$
    Compute $\Sigma_{i,j}$ *from Eq. (85).*

Equivalent to Kalman-recursions for smoothing.

Marginal variances under hard and soft constraints

Let $\widetilde{\mathbf{\Sigma}}$ be the covariance with the constraints and $\mathbf{\Sigma}$ be without.

Then $\widetilde{\mathbf{\Sigma}}$ relates to $\mathbf{\Sigma}$ as

$$\widetilde{\mathbf{\Sigma}} = \mathbf{\Sigma} - \mathbf{Q}^{-1}\mathbf{A}^T \left(\mathbf{A}\mathbf{Q}^{-1}\mathbf{A}^T\right)^{-1} \mathbf{A}\mathbf{Q}^{-1}.$$

hence

$$\widetilde{\Sigma}_{ii} = \Sigma_{ii} - \left(\mathbf{Q}^{-1}\mathbf{A}^T \left(\mathbf{A}\mathbf{Q}^{-1}\mathbf{A}^T\right)^{-1} \mathbf{A}\mathbf{Q}^{-1}\right)_{ii}, \quad i = 1, \ldots, n.$$

for the hard constraint $\mathbf{A}\mathbf{x} = \mathbf{b}$.

Computational costs

$$\begin{aligned} \textit{Time} \quad & \mathcal{O}(n) \\ \textit{Spatial} \quad & \mathcal{O}(n \log(n)^2) \\ \textit{Spatio-temporal} \quad & \mathcal{O}(n^{5/3}). \end{aligned}$$