Lecture 6: MCMC for hierarchical models

... or "an entire MCMC course in 90 minutes"

David Bolin
Chalmers University of Technology
February 9, 2015

Recap

So far in the course, we have looked at

1. Analytical and numerical properties of GMRFs
2. How we can construct various popular GMRF models

We will now start investigating how we can use these in statistics.

The focus will be on GMRF-based hierarchical models

We will look at

Today	MCMC-based inference
Next lecture	Approximate inference

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

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 $x_i = \text{logit}(p_i)$.

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

- precision parameter κ
- "strength" of dependency.

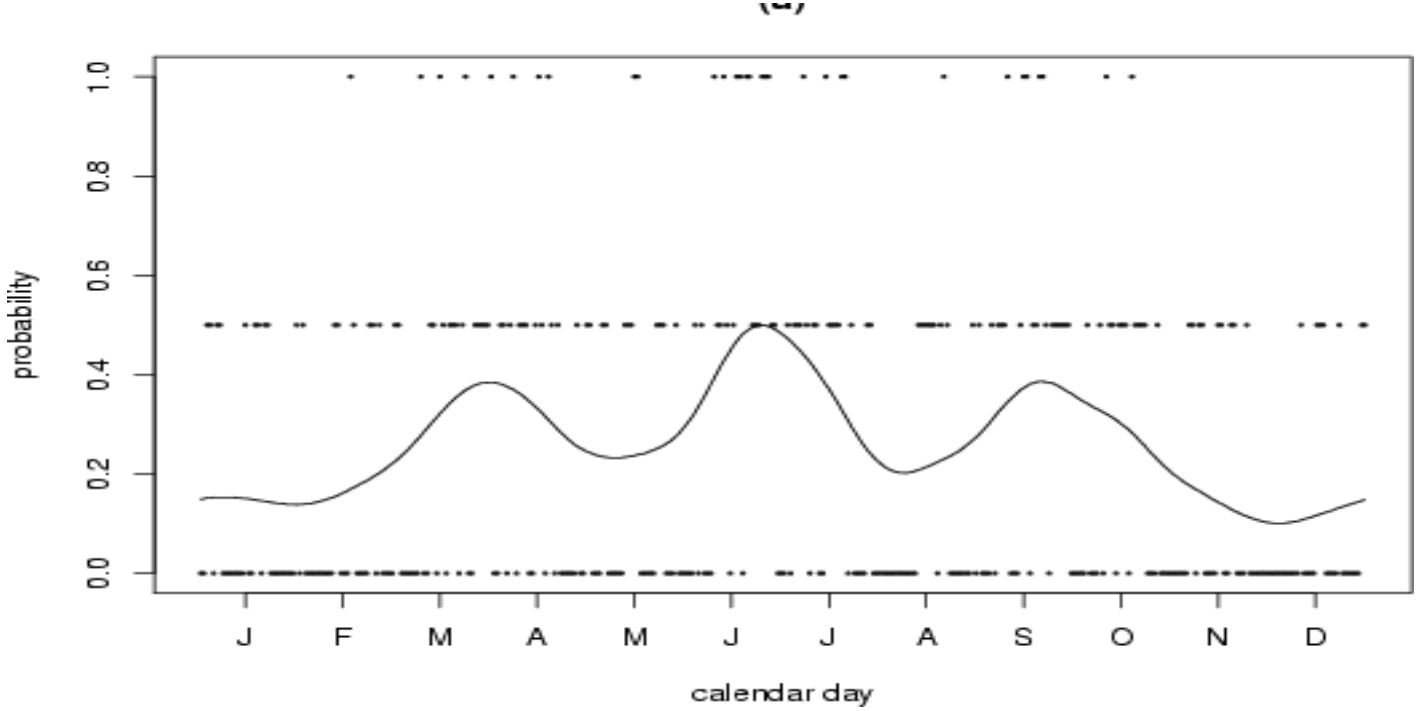
Further stages if needed.

A simple example: Tokyo rainfall

A much analysed binomial time series from Kitawaga (1987).

- Each day during the years 1983 and 1984, it was recorded whether there was more than 1 mm rainfall in Tokyo.
- Of interest is to study the underlying probability p_i of rainfall at calendar day $i = 1, \ldots, 366$, which is assumed to gradually change over time.
- For every day we have two observations, except for $i = 60$, which corresponds to February 29.
- Thus, in total we have $m = 366 + 365 = 731$ binary observations.

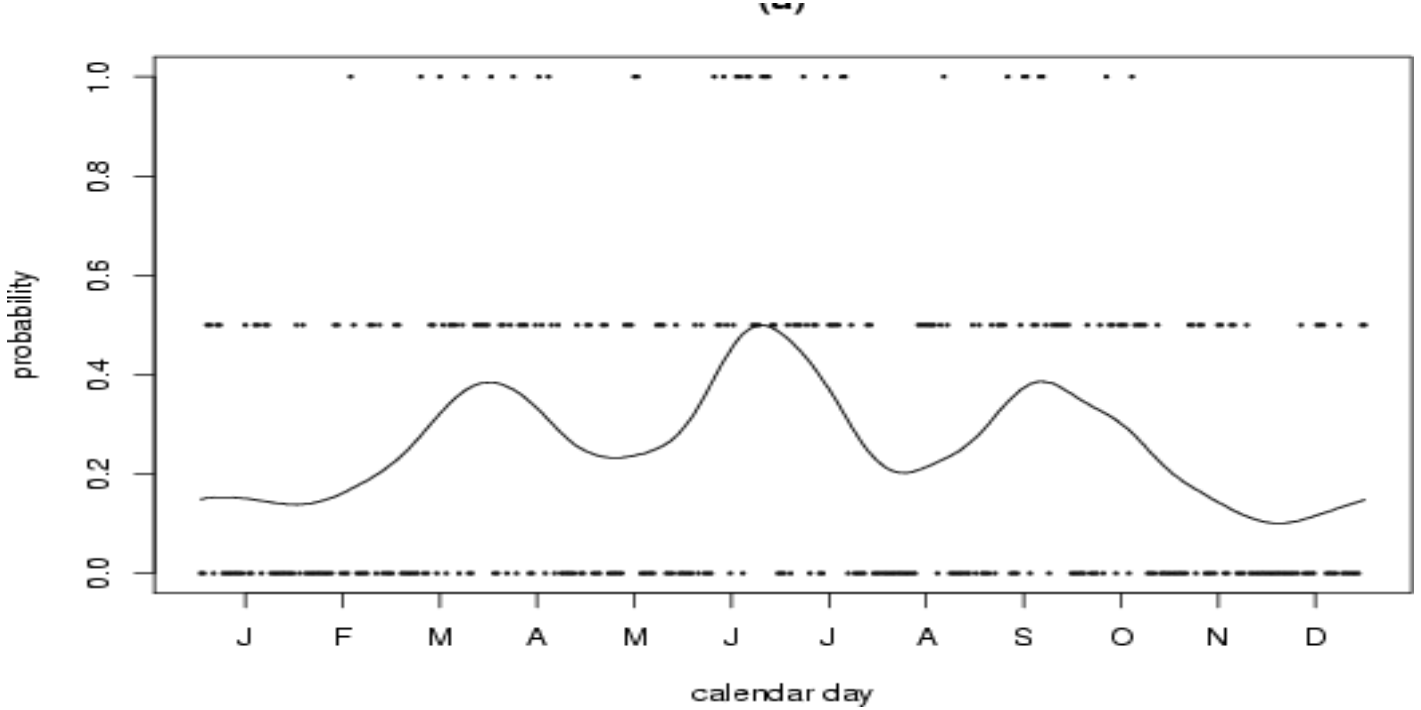
Tokyo rainfall data



Stage 1 Binomial data

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

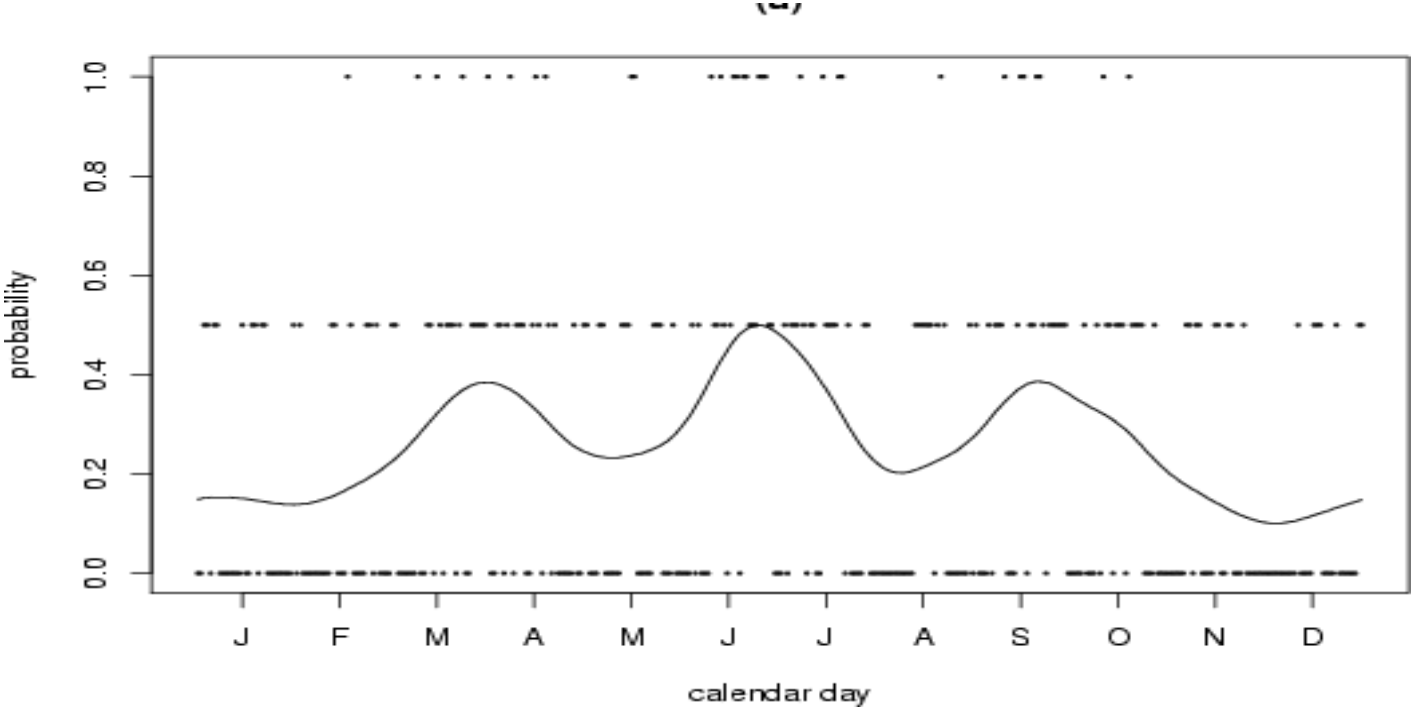
Tokyo rainfall data



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

We can divide these models into three classes with increasing difficulty in terms of estimation

1. Normal data
2. Non-normal data that allows for a normal-mixture representation
   - Student-t distribution
   - Logistic and Laplace (Binary regression)
3. Non-normal data
   - Poisson
   - and others...

What do we care about?

Mathematically, we usually want to know about two things:

$$\pi(x_i|\mathbf{y}) \propto \int_{x_{\{-i\}}} \int_{\theta} \pi(y|\mathbf{x}, \boldsymbol{\theta})\pi(\mathbf{x}|\boldsymbol{\theta})\pi(\theta)\, d\theta\, dx_{\{-i\}}$$

and

$$\pi(\theta_i|\mathbf{y}) \propto \int_{x} \int_{\theta_{\{-i\}}} \pi(y|\mathbf{x}, \boldsymbol{\theta})\pi(\mathbf{x}|\boldsymbol{\theta})\pi(\theta)\, d\theta_{\{-i\}}\, dx$$

These are very high dimensional integrals and are not typically analytically tractable.

Sampling based inference

Monte Carlo integration

$$\int_{\mathbb{R}^d} f(\mathbf{x})\pi(\mathbf{x})\,d\mathbf{x} \approx \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{x}_i),$$

where $\mathbf{x}_i$ are drawn randomly from a distribution with pdf $\pi(\mathbf{x})$.

The idea is that if we can sample from $\pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})$ then we can do all the required Bayesian computations.

One small problem: We can't!

So what can we do?

It turns out that we don't need to sample from $\pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})$ directly to make Monte Carlo methods work.

It's enough to construct *a Markov Chain* that has $\pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})$ as a stationary distribution.

- It turns out that this is *easy* to do.
- (But it is hard to do well...)
- Dates back to the 1950's with two key papers being:
  - Equations of State Calculations by Fast Computing Machines (1953)
  - Monte Carlo Sampling Methods Using Markov Chains and Their Applications (1970)
- This has been the focus of computational statistics for the last 25 years.

Markov Chain Monte Carlo

To construct a Markov chain with stationary distribution, $\pi(x)$:

- Start with a proposal kernel, $q(x, y)$, and accept the proposed jumps with probability $\alpha(x, y)$.
- The resulting combined proposal kernel is given by

$$\widetilde{q}(x, y) = \alpha(x, y)q(x, y) + \left(1 - \int_{\mathcal{X}} \alpha(x, z)q(x, z)dz\right)\delta_x(y).$$

- We now want an $\alpha(x, y)$ that gives detailed balance for the combined Markov chain,

$$\pi(x)\widetilde{q}(x, y) = \pi(y)\widetilde{q}(y, x).$$

  Because this ensures that $\pi(x)$ is a stationary distribution.
- Choose

$$\alpha(x, y) = \min\left(1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right).$$

The Metropolis Hastings algorithm

Given a density $\pi(x)$ and a proposal kernel $q(x, y)$
Start the chain with some $x^{(0)}$, and loop over $t = 1, \ldots, T$.

1. Given $x^{(t)}$, draw a proposal y from $q(x^{(t)}, y)$.
2. Calculate the acceptance probability

$$\alpha(x^{(t)}, y) = \min\left(1, \frac{\pi(y)q(y, x^{(t)})}{\pi(x^{(t)})q(x^{(t)}, y)}\right)$$

3. With probability $\alpha(x^{(t)}, y)$ accept the proposal, otherwise keep the old value, $x^{(t)}$:
   1. Draw $u \in U(0, 1)$.
   2. Take

$$x^{(t+1)} = \begin{cases} y, & \text{if } u < \alpha(x^{(t)}, y) \\ x^{(t)}, & \text{if } u \geq \alpha(x^{(t)}, y) \end{cases}$$

Ensuring convergence

The following are sufficient requirements for convergence of the Markov chain:

1. Detailed balance (by construction).
2. Irreducible chain
   The chain should be able to reach any point $\{x : \pi(x) \neq 0\}$ regardless of the starting point.
3. Aperiodic chain
   The existens of a unique stationary distribution is ensured by 1. and 2.; aperiodic chain is needed to ensure convergence.

If the above requirements are fulfilled, then for any set $A \subseteq \mathcal{X}$,

$$\mathrm{P}(X^{(t)} \in A) \to \int_A \pi(x)dx, \quad t \to \infty,$$

independently of starting point, $x^{(0)}$.

Different proposals

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 $\boldsymbol{\theta}$ conditionally 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 $\pi(\boldsymbol{\theta})$.
- The case when we can sample exactly from the full conditionals is of particular importance.
  - The resulting algorithm is then called the Gibbs sampler
  - Often thought of as a separate algorithm

Gibbs sampling: An introductory example

- Suppose we want to sample from the joint distribution of $x = 1, \ldots, n$, $0 \leq y \leq 1$ given by

$$\pi(x, y) = \frac{n!}{(n-x)!x!} y^{x+\alpha-1}(1-y)^{n-x+\beta-1}$$

- This joint density is complex and not easy to sample from.
- The conditional distributions are, however, simple:

$$x|y \in Bin(n, y)$$
$$y|x \in Beta(x+\alpha, n-x+\beta)$$

- The idea in Gibbs sampling is to construct a Markov chain by sampling from the simpler conditional distributions.

The Gibbs sampling algorithm

1. Choose a starting value $\boldsymbol{\theta}^{(0)}$.
2. Repeat for $i = 1, \ldots, N$:
   - i.1 Draw $\theta_1^{(i)}$ from $\pi(\theta_1 | \theta_2^{(i-1)}, \ldots, \theta_m^{(i-1)})$.
   - i.2 Draw $\theta_2^{(i)}$ from $\pi(\theta_2 | \theta_1^{(i)}, \theta_3^{(i-1)}, \ldots, \theta_m^{(i-1)})$.
   - i.3 Draw $\theta_3^{(i)}$ from $\pi(\theta_3 | \theta_1^{(i)}, \theta_2^{(i)}, \theta_4^{(i-1)}, \ldots, \theta_m^{(i-1)})$.
   - $\vdots$
   - i.m Draw $\theta_m^{(i)}$ from $\pi(\theta_m | \theta_1^{(i)}, \theta_2^{(i)}, \ldots, \theta_{m-1}^{(i)})$.
3. $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots \boldsymbol{\theta}^{(N)}$, is now a sequence of dependent draws approximately from π.

The chain is not reversible if the variables are sampled cyclically:

- This may cause problems with slow convergence.
- Sample using a random order or using a backward forward scheme.

Single site methods

One can rarely sample directly from all full conditionals.

In this case, a common strategy is to sample those variables using a MH step, using for example a random walk proposal.

These single site methods are simple to implement, 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 $\boldsymbol{\theta}$.

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 $r > \rho$

$$\lim_{k\to\infty} \mathrm{E}\left[\left(\mathrm{E}\left(h(\boldsymbol{\theta}^{(k)}) \mid \boldsymbol{\theta}^{(0)}\right) - \mathrm{E}\left(h(\boldsymbol{\theta})\right)\right)^2 r^{-2k}\right] = 0. \qquad (1)$$

if $\rho < 1$ we say that the chain is geometrically ergodic.

Example

Let $\mathbf{x}$ be a first-order autoregressive process

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

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

$$\rho = 4\frac{\gamma^2}{(1+\gamma^2)^2}. \tag{3}$$

- 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.

Example 2

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,

$$\begin{aligned} \mu^{(k)}|\mathbf{x}^{(k)} &\sim \mathcal{N}\left(\frac{\mathbf{1}^T\mathbf{Q}\mathbf{x}^{(k-1)}}{1+\mathbf{1}^T\mathbf{Q}\mathbf{1}},\ (1+\mathbf{1}^T\mathbf{Q}\mathbf{1})^{-1}\right) \\ \mathbf{x}^{(k)}|\mu^{(k)} &\sim \mathcal{N}(\mu^{(k)}\mathbf{1},\ \mathbf{Q}^{-1}). \end{aligned} \tag{4}$$

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

For our model

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

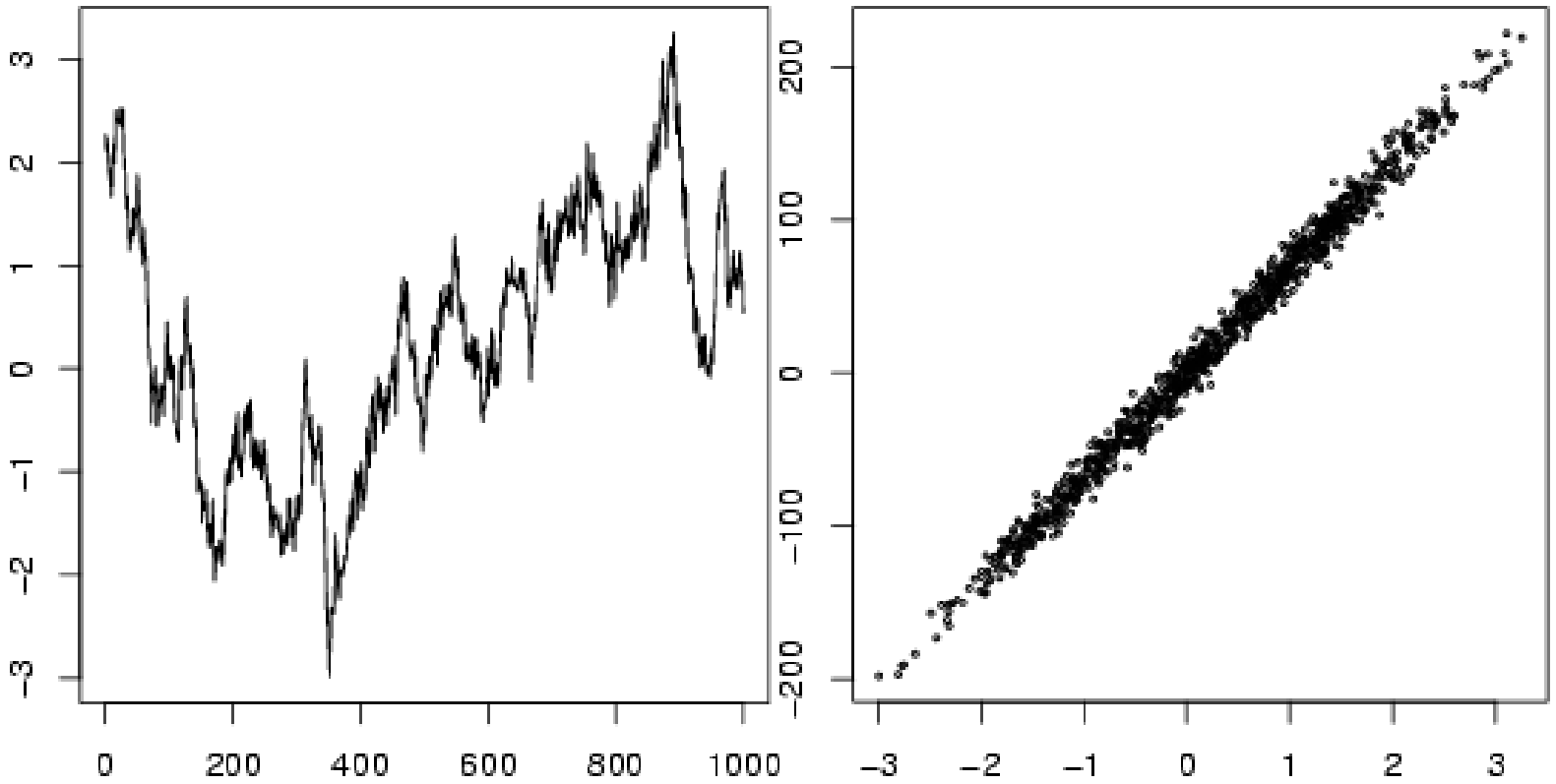
When n is large, ϕ is close to 1 and the chain will both mix and converge slowly even though we use a two-block Gibbs sampler.

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.

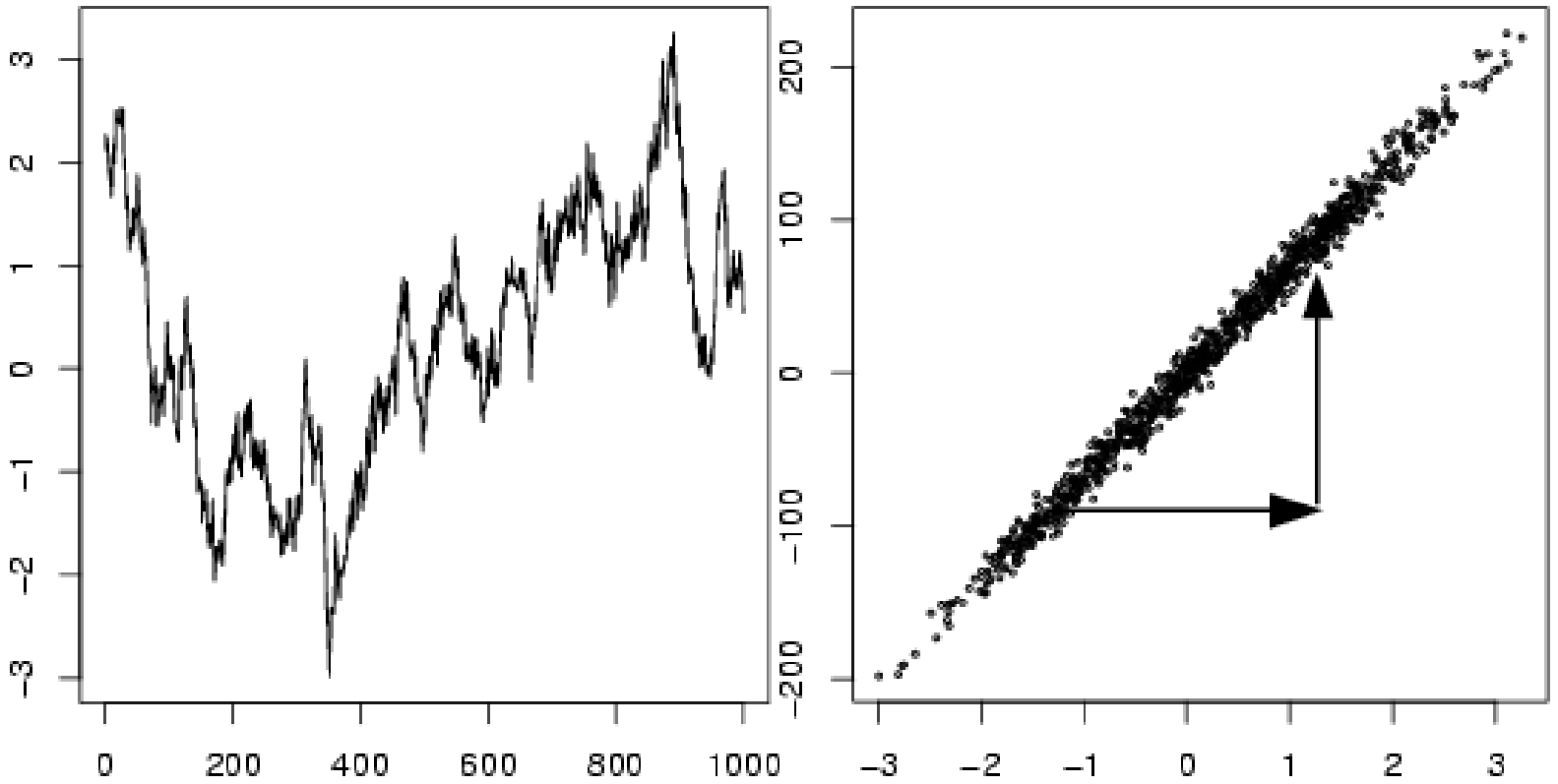
Update $(\mu, \mathbf{x})$ jointly by delaying the accept/reject step until $\mathbf{x}$ also is updated.

$$\begin{aligned} \mu^* &\sim q(\mu^* \mid \mu^{(k-1)}) \\ \mathbf{x}^* | \mu^* &\sim \mathcal{N}(\mu^* \mathbf{1}, \mathbf{Q}^{-1}) \end{aligned} \tag{5}$$

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,

$$\alpha = \min\left\{1,\ \exp(-\frac{1}{2}((\mu^*)^2 - (\mu^{(k-1)})^2))\right\}. \qquad (6)$$

- Only the *marginal density* of μ is needed in (6): 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:

$$\begin{aligned} \boldsymbol{\theta}^* &\sim q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(k-1)}) \\ \mathbf{x}^* &\sim \pi(\mathbf{x} \mid \boldsymbol{\theta}^*, \mathbf{y}). \end{aligned} \tag{7}$$

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

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

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

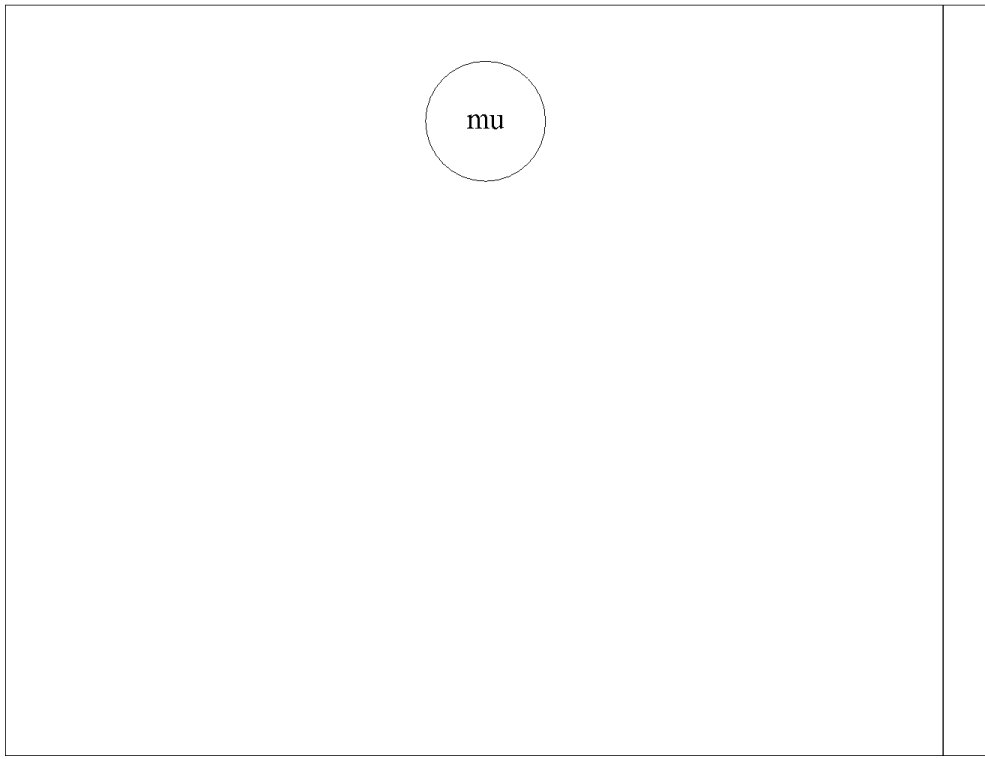
The dimension of $\boldsymbol{\theta}$ is typically low (1-5, say).

The proposed algorithm should not experience any serious mixing problems.

Merging GMRFs using conditioning (I)

We can merge GMRFs for use in the one-block algorithm!

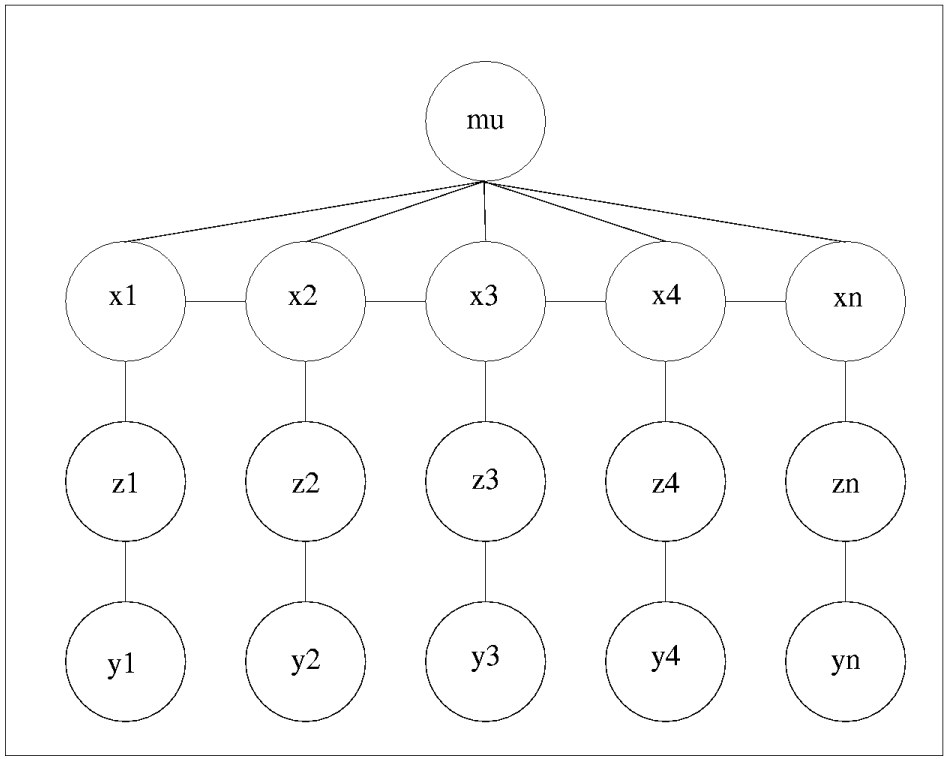
$$
\begin{aligned}
\mu &\sim \mathcal{N}(0, 1) \\
\mathbf{x} - \mu \mid \mu &\sim \text{AR}(1) \\
\mathbf{z} \mid \mathbf{x} &\sim \mathcal{N}(\mathbf{x}, \mathbf{I}) \\
\mathbf{y} \mid \mathbf{z} &\sim \mathcal{N}(\mathbf{z}, \mathbf{I})
\end{aligned}
$$



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

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.

Key lesson:The first rule of Bayes club...

There is no reason to separate "fixed" and "random" effects in Bayesian models!
They just have different priors!

- This implies that (Gaussian) fixed and random effects should be treated *together* in any inference method.
- The general procedure is to find all of the (jointly) Gaussian bits of your model and block!

However...

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 (7) 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

$$(\boldsymbol{\theta}_a, \mathbf{x}_a), \ (\boldsymbol{\theta}_b, \mathbf{x}_b) \quad \text{and} \quad (\boldsymbol{\theta}_c, \mathbf{x}_c), \tag{8}$$

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

Normal data: Munich rental guide

Response variable y_i: rent/m^2

Covariates:

- 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)

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

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

The dependency structure

The covariates $\mathbf{x}^S, \mathbf{x}^C, \mathbf{x}^L, \boldsymbol{\beta}$ are a priori independent.

Conditionally on the data, they are dependent.

Every combination of the covariates we observe in the data introduces dependence within this combination and makes the corresponding term in the joint precision matrix non-zero.

We have observations of a lot of different combinations so the joint posterior precision for the covariates will be non-sparse

However, their respective full conditionals are still GMRFs, with the same sparsity properties as the priors.

Inference using MCMC

Sub-block approach

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

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 $F > 1$.

This density has the nice property that $f(x/y)/y = f(y/x)/x$, so this is a symmetric proposal:

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

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

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

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

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

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 $|L(i)|$.

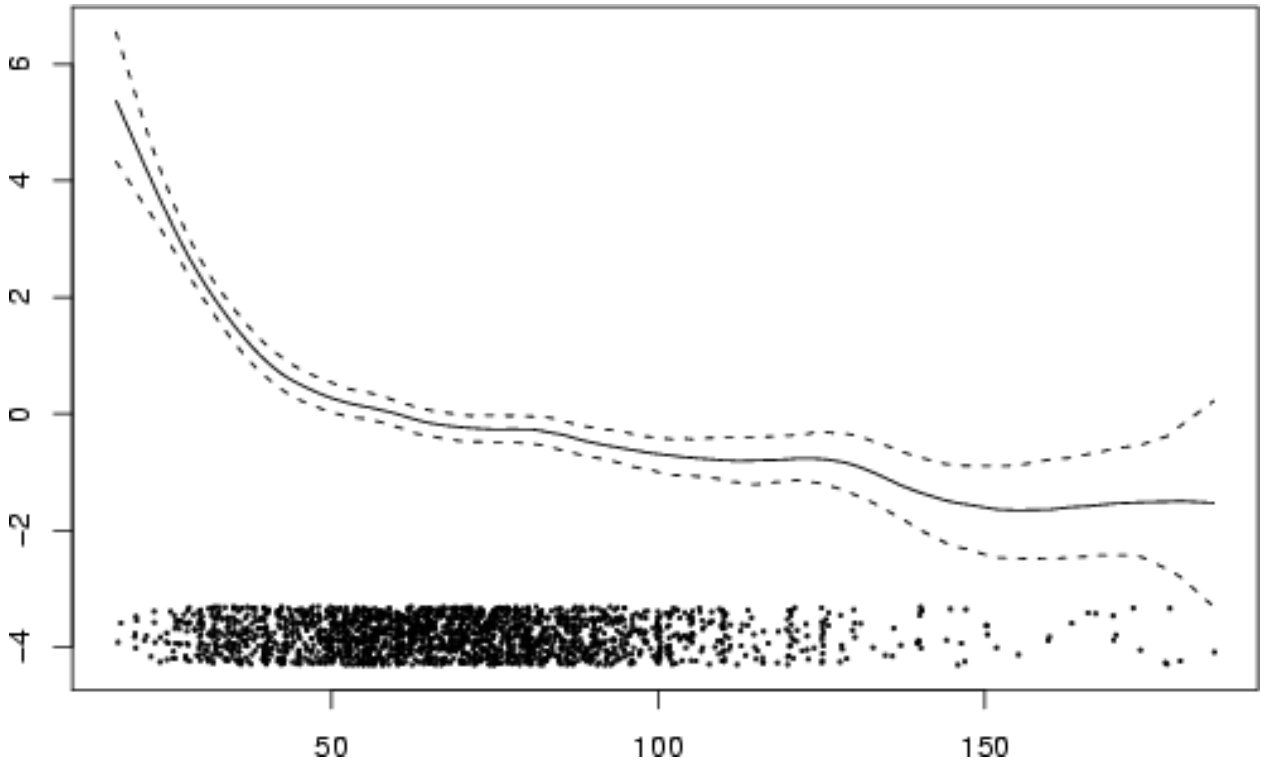
The full conditional of $\mathbf{x}^L$ is a GMRF with parameters $(\mathbf{Q}^{-1}\mathbf{b}, \mathbf{Q})$, where

$$b_i \ = \ \kappa_{\mathbf{y}} \sum_{k \in L_i} \tilde{y}_k \quad \text{and}$$

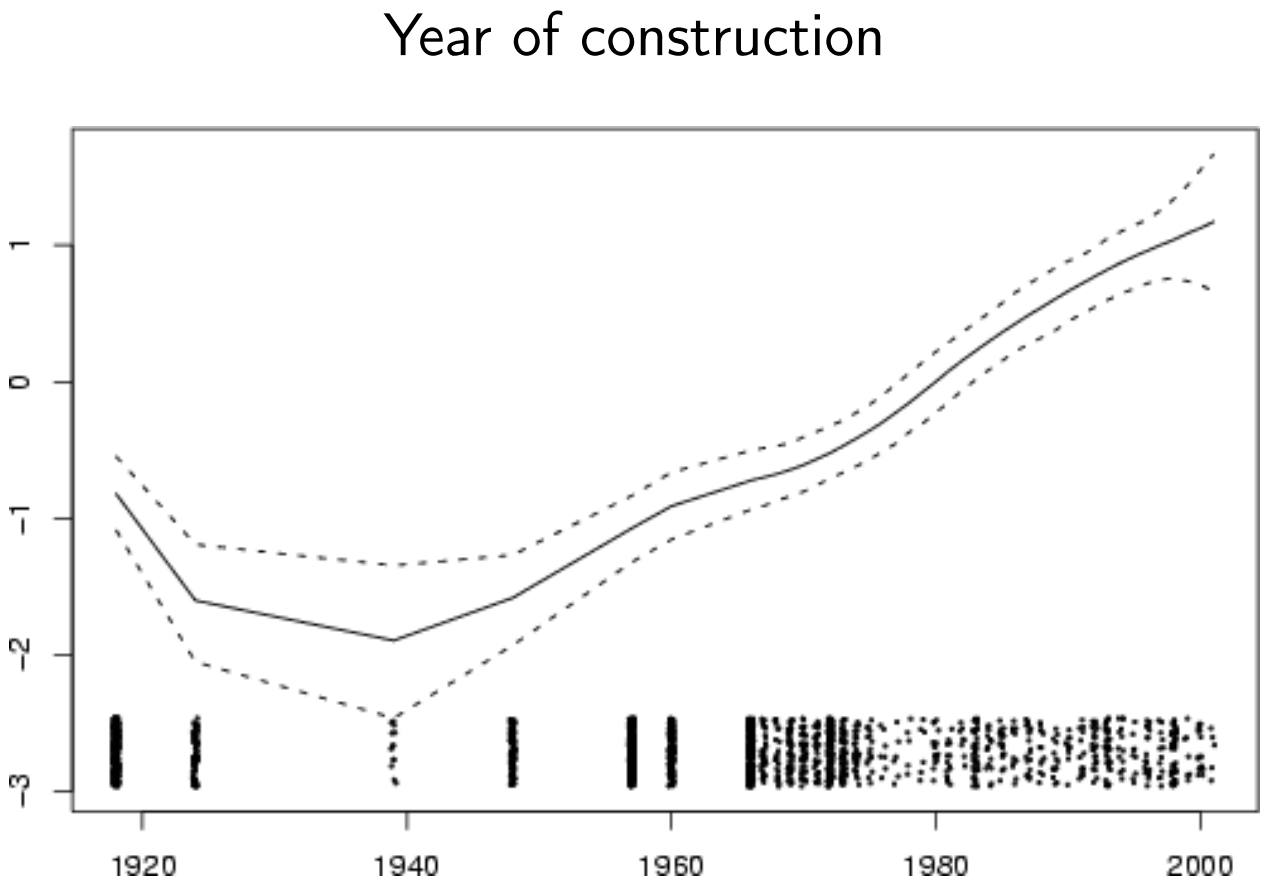
$$Q_{ij} \ = \ \begin{cases} \kappa_L n_i + \kappa_{\mathbf{y}} |L(i)| & \text{if } i = j \\ -\kappa_L & \text{if } i \sim j \\ 0 & \text{otherwise.} \end{cases}$$

Results

Floor space

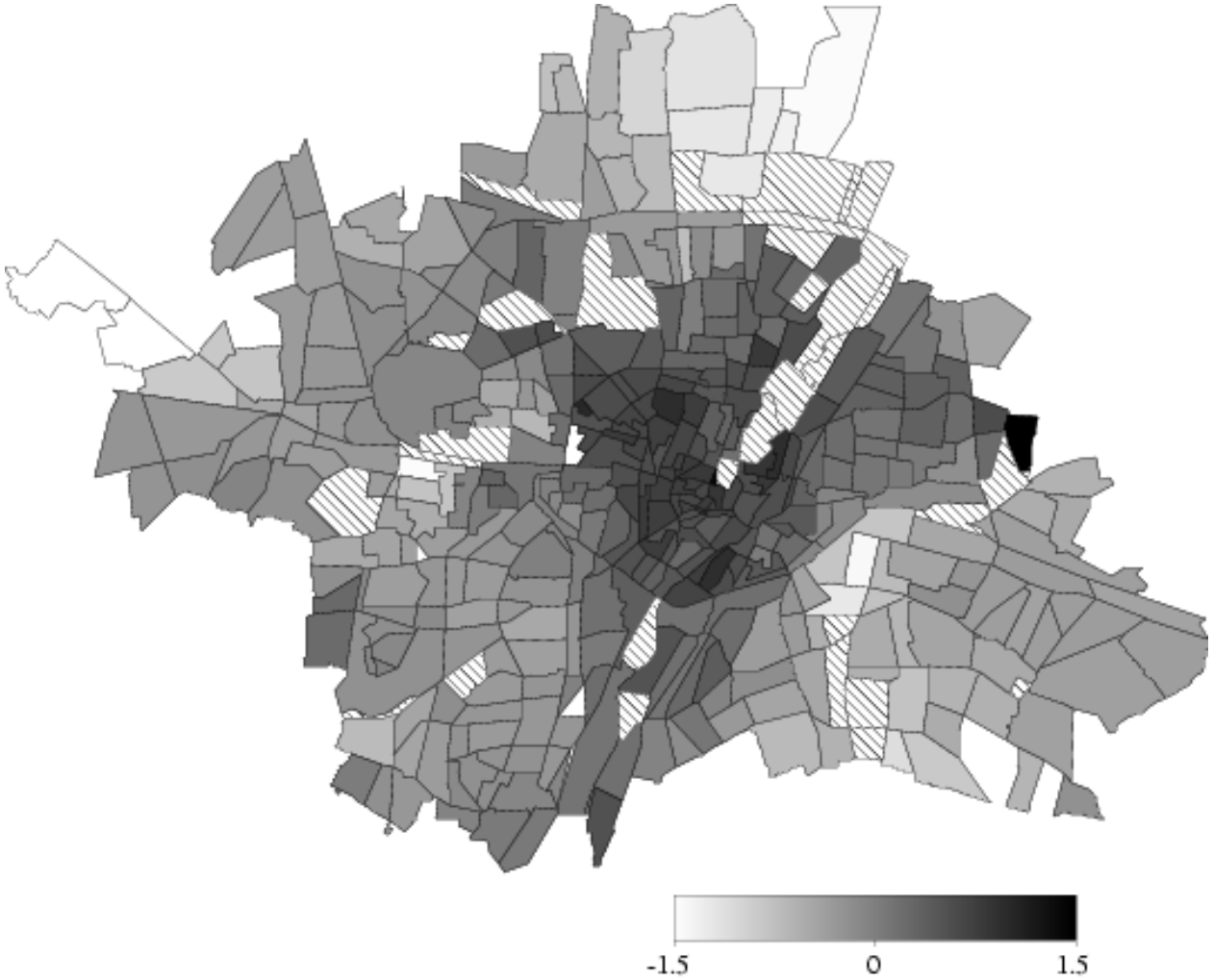


Results



Results

Location



Normal mixtures

The Tokyo model

$$\pi(\mathbf{x} \mid \tau)\, \pi(\tau) \prod_i \pi(y_i \mid x_i)$$

- $y_i|x_i$ is Binomial with $p_i = \Phi(x_i)$
- $\mathbf{x} \mid \tau$ is Gaussian with dimension 366
- τ is Gamma

Because of the likelihood, the full posterior is non-Gaussian, so we cannot use the methods we have discussed so far directly.

In some cases, we can retrieve Gaussian conditionals by introducing auxiliary variables.

For example in scale mixture distributions:

$x|\tau \sim \mathsf{N}(0, \tau^{-1}), \tau \sim \pi(\tau)$

Here, $\pi(x)$ is non-Gaussian even though $x|\tau$ is Gaussian.

Theorem (Kelker, 1971)

If x has density $\pi(x)$ 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 $\pi(x)$ satisfy

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

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

Examples:

- Student t distribution: $v \sim \Gamma(\nu/2, \nu/2)$.
- Logistic distribution: $v = 1/(2K)^2$, where K is Kolmogorov-Smirnov distributed.
- Laplace distribution: $v = 1/(2E)$, where E is exponential distributed.

Example: 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 $\Gamma(\nu/2, \nu/2)$ 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 $(\boldsymbol{\theta}, \mathbf{x})$ and $\boldsymbol{\lambda}$

Binary regression models

Another important example where Auxiliary variables are useful:

Example: Binary regression

Gaussian $\mathbf{x}$ and Bernoulli data

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

Equivalent representation using auxiliary variables $\mathbf{w}$

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

Auxiliary variables for the Tokyo data

We have binomial data, where each day

$$y_i \sim \begin{cases} y_{i,1} + y_{i,2}, i \neq 60 \\ y_{i,1}, i = 60 \end{cases}$$

where $y_{i,\bullet} \sim \mathcal{B}(p_i)$.
The data only contain information about y_i, so if $y_i = 1$ we let $y_{i,1} = 1$ and $y_{i,2} = 0$.
Let n_i be the number of observations for day i and introduce one auxiliary variable $w_{i,j}$ for each $y_{i,j}$.
With

$$w_i = \sum_{j=1}^{n_i} w_{i,j}$$

we have that $\mathbf{x}|$ the rest $\sim N(\boldsymbol{\mu}, \hat{\mathbf{Q}}^{-1})$ where $\hat{\mathbf{Q}} = \tau\mathbf{R} + \text{diag}(\mathbf{n})$ and $\boldsymbol{\mu} = \hat{\mathbf{Q}}^{-1}\mathbf{w}$.

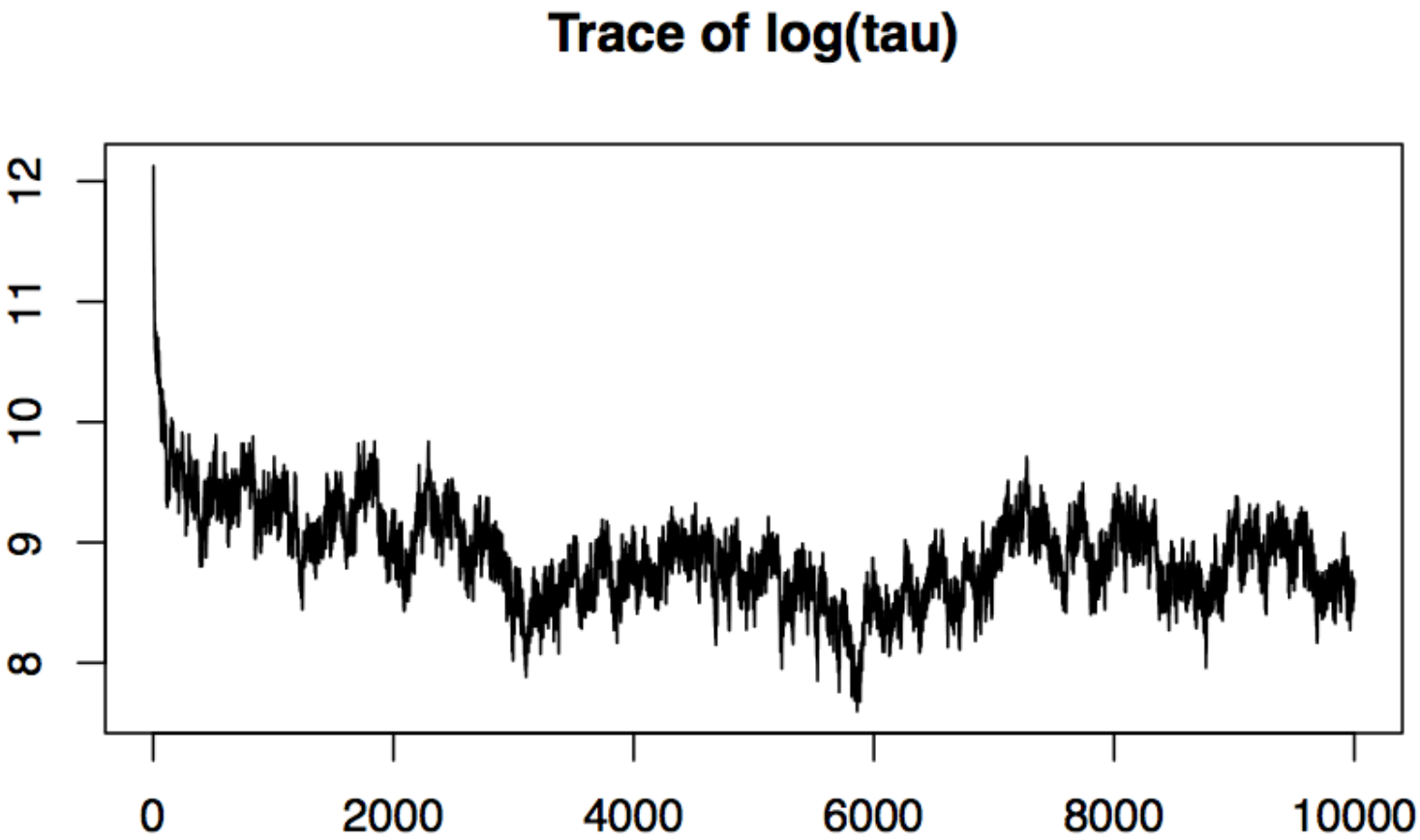
Single-site Gibbs sampling

Using the auxiliary variable formulation, we obtain the following Gibbs sampler for the problem:

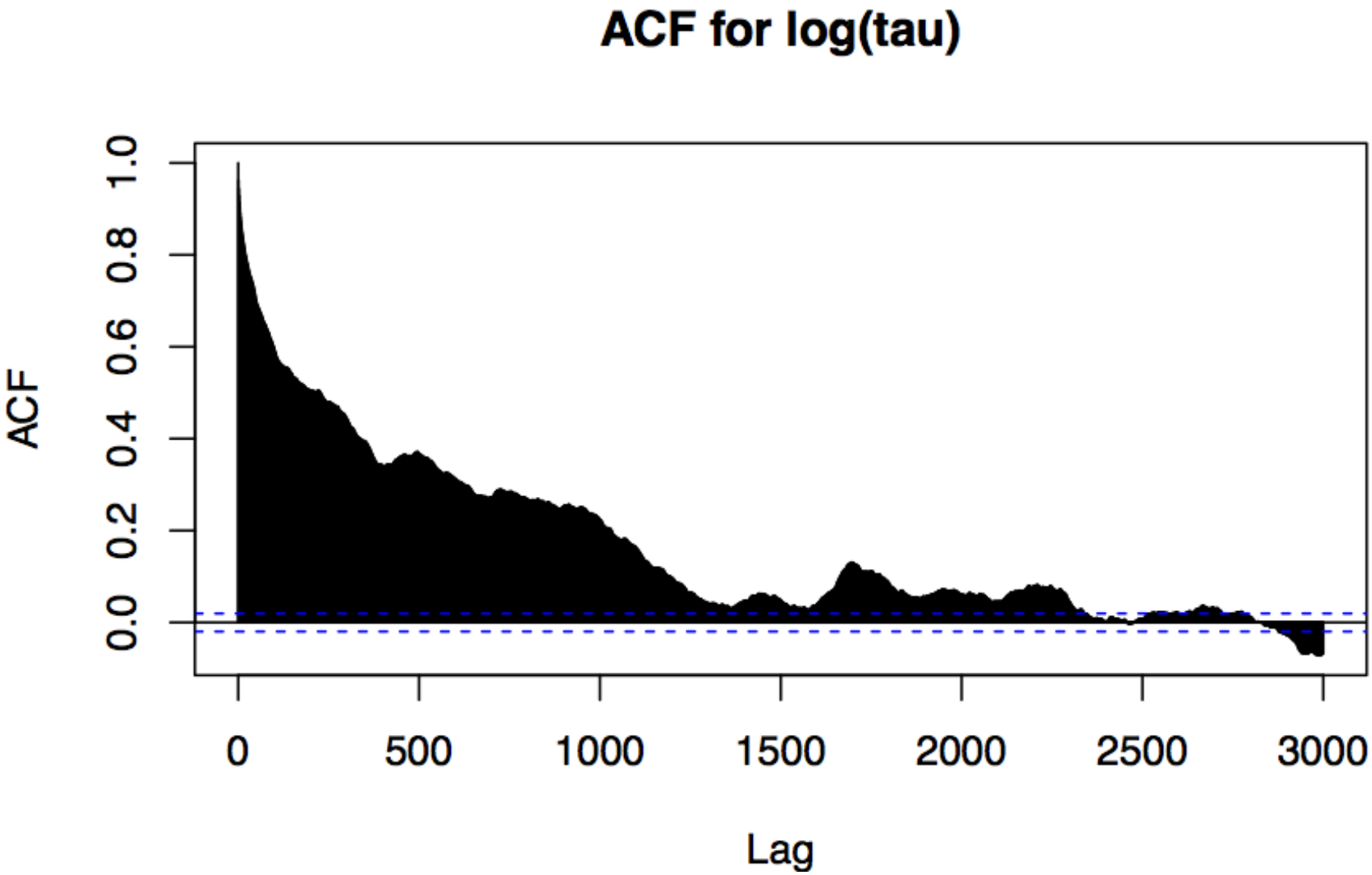
- $\tau \sim \Gamma\left(\frac{366-1}{2} + \alpha, \frac{1}{2}\mathbf{x}^\top \mathbf{R}\mathbf{x} + \beta\right)$
- for each i
  - $x_i \sim \mathcal{N}(\mu_i - \hat{Q}_{i,i}^{-1} \sum_{j \neq i} \hat{Q}_{ij}(x_i - \mu_i), \hat{Q}_{i,i}^{-1})$
- for each i
  - $w_i \sim \mathcal{W}(\cdot)$

The distribution for w_i is $N(x_i, 1)$ truncated to be positive if $y_i = 1$ and truncated to be negative if $y_i = 0$.

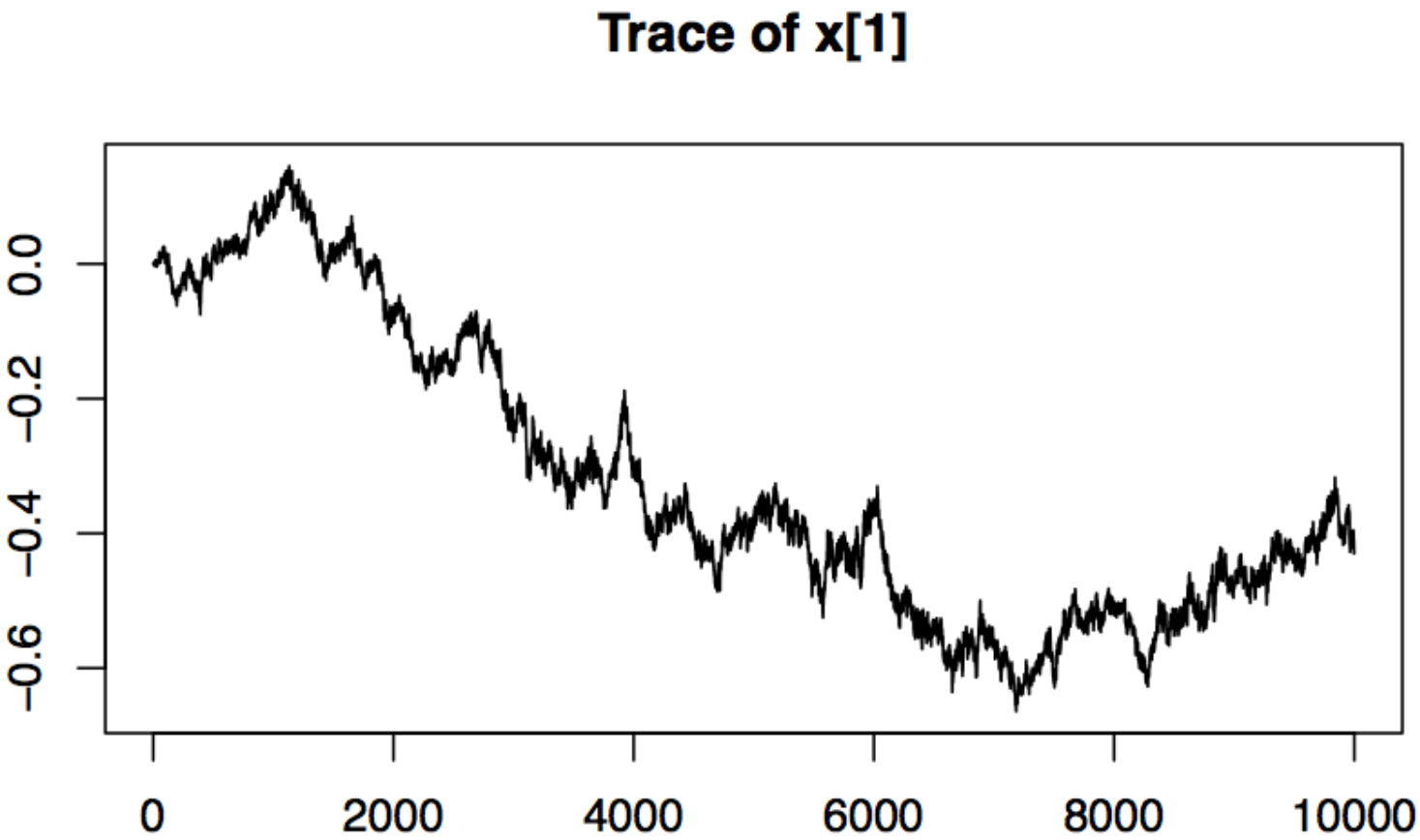
Results: hyper-parameter $\log(\tau)$



Results: hyper-parameter $\log(\tau)$



Results: x_1



We continue next time!