Introduction to Bayesian Distributional Regression

A Tutorial

Thomas Kneib & Johannes Brachem

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN IN PUBLICA COMMODA SEIT 1737

11.–13.2.2025

Florence

1 Introduction

1.1 Aims and Scope

- Flexible forms of regression modelling have recent a lot of attention in applied statistics and statistical modelling:
  - generalized linear mixed models for clustered and longitudinal data,
  - generalized additive models for nonlinear effects of continuous covariates,
  - geoadditive regression models for spatial data,
  - etc.

- Recent trend: Overcome the focus of classical regression models on the conditional mean and model the complete conditional distribution of the response.
- Prominent examples:
  - generalized additive models for location, scale and shape,
  - quantile and expectile regression,
  - conditional transformation models,
  - etc.

- Bayesian inference provides a particularly convenient way of estimating distributional regression models:
  - Markov chain Monte Carlo simulations (MCMC) are the de facto standard of Bayesian inference in applied statistics (although approximate forms such as variational inference are gaining in popularity).
  - MCMC was also the foundation for the success of Bayesian inference since the 1990s.
  - It comes with a number of specific challenges but also has distinct advantages for complex statistical models.
  - Advantages are particularly beneficial for semiparametric regression models that can be represented as directed acyclic graphs due to their hierarchical model formulation.

1.2 Outline

Tuesday: Introduction to Bayesian Inference

- Principles of Bayesian Inference
- Markov Chain Monte Carlo Simulations
- Monitoring Mixing and Convergence
- Posterior Summaries

Wednesday: Semiparametric Regression with Structured Additive Predictors

- Penalized Spline Smoothing
- A Generic Basis Function Framework
- Spatial Smoothing
- Random Effects Models
- Hyperpriors for the Smoothing Parameter
- Interactions and Identification

Thursday: Distributional Regression Models

- Generalized Additive Models for Location, Scale and Shape
- Applications with Continuous, Discrete and Multivariate Responses
- Other Frameworks for Distributional Regression

2 Bayesian Inference with Markov Chain Monte Carlo Simulations

2.1 Introduction

Aims of this section:

- Introduce the foundations of Bayesian inference and compare it to frequentist maximum likelihood.
- Motivate how Markov chain Monte Carlo (MCMC) simulations provide numerical access to the posterior distributions.
- Discuss practical aspects of working with MCMC simulations.

Bayes' theorem:

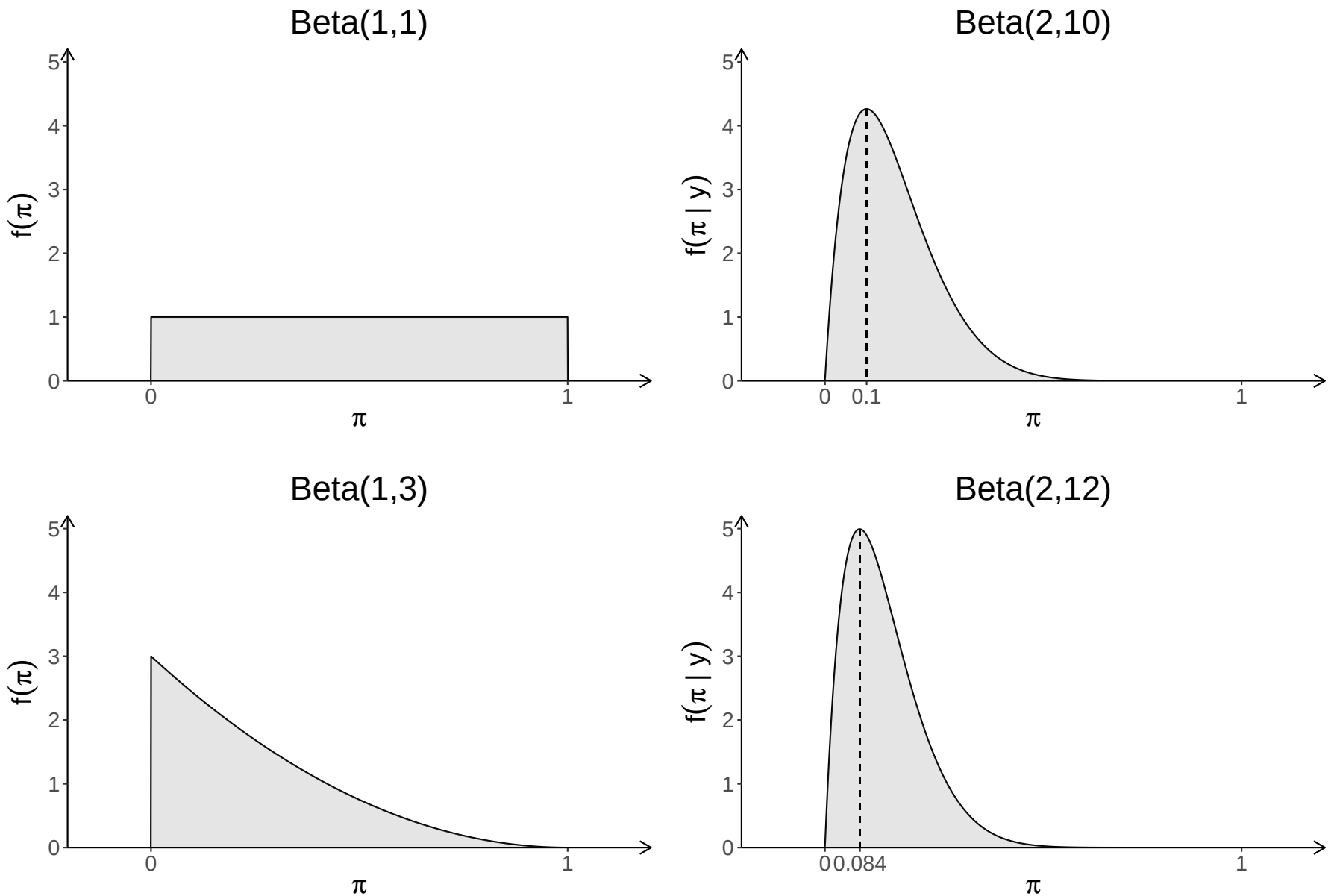
- Two central components of a Bayesian model formulation:
  - Observation model $f(\boldsymbol{y}|\boldsymbol{\vartheta})$ which describes how the data $\boldsymbol{y}$ are generated for given model parameters $\boldsymbol{\vartheta}$.
  - Prior distribution $f(\boldsymbol{\vartheta})$ representing prior beliefs about the parameter vector $\boldsymbol{\vartheta}$
- Bayesian learning updates prior beliefs on $\boldsymbol{\vartheta}$ based on information in the data $\boldsymbol{y}$ using Bayes' theorem

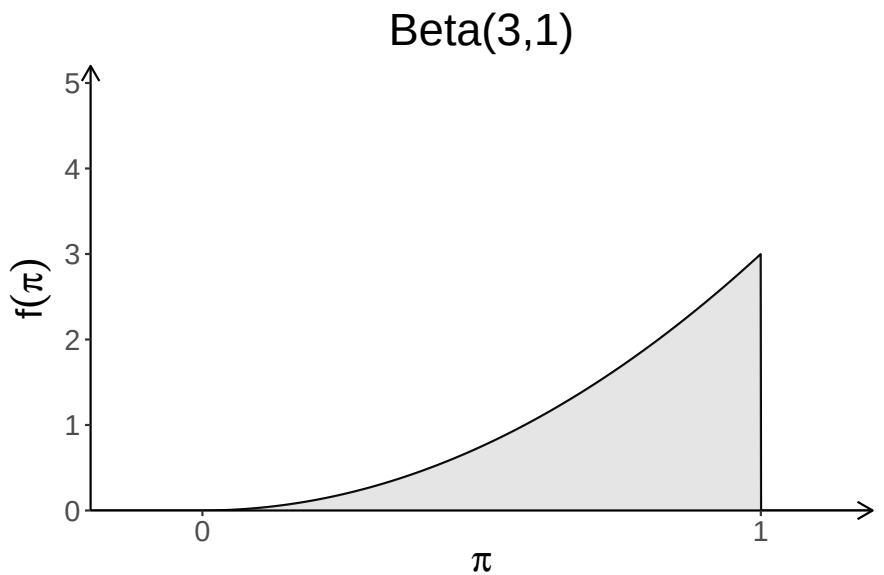
$$f(\boldsymbol{\vartheta}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\vartheta})f(\boldsymbol{\vartheta})}{f(\boldsymbol{y})} = \frac{f(\boldsymbol{y}|\boldsymbol{\vartheta})f(\boldsymbol{\vartheta})}{\int f(\boldsymbol{y}|\boldsymbol{\vartheta})f(\boldsymbol{\vartheta})d\boldsymbol{\vartheta}}$$

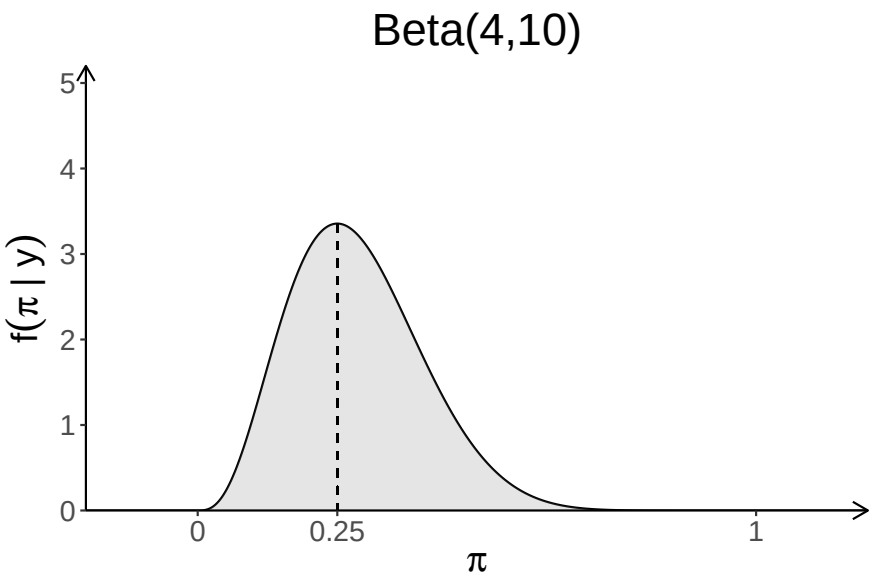
where $f(\boldsymbol{y})$ is the marginal density of the data.

Example: Bayesian inference for the success probability in a Bernoulli trial:

- Data $y_i \overset{\text{i.i.d.}}{\sim} \text{Be}(\pi)$ with unknown success probability $\pi \in (0, 1)$.
- We consider $n = 10$ trials with one success and nine failures such that the maximum likelihood estimate is
$$\hat{\pi} = \frac{1}{n} \sum_{i=1}^{n} y_i = \frac{1}{10}.$$
- As a prior distribution, we choose the beta distribution with parameters $a > 0$ and $b > 0$.
- In this case, the posterior an be worked out analytically and turns out to be a beta distribution with parameters
$$\tilde{a} = a + \sum_{i=1}^{n} y_i \qquad \tilde{b} = b + n - \sum_{i=1}^{n} y_i.$$


Beta(1,1)
f(π)
π
0
1
Beta(2,10)
f(π | y)
0
0.1
1
Beta(1,3)
Beta(2,12)
0.084


Beta(3,1)
f(π)
5
4
3
2
1
0
0
1
π


Beta(4,10)
f(π | y)
5
4
3
2
1
0
0
0.25
1
π

Relation to maximum likelihood estimation:

- If the prior distribution is flat, i.e.

$$f(\boldsymbol{\vartheta}) \propto \text{const},$$

the posterior is proportional to the likelihood:

$$f(\boldsymbol{\vartheta}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\vartheta})f(\boldsymbol{\vartheta})}{f(\boldsymbol{y})} \propto f(\boldsymbol{y}|\boldsymbol{\vartheta})f(\boldsymbol{\vartheta}) \propto f(\boldsymbol{y}|\boldsymbol{\vartheta}).$$

- Hence the mode of the posterior coincides with the maximum likelihood estimate.
- In general,
  - the likelihood is a central part of Bayes' theorem that quantifies the information coming from the data and
  - the posterior forms a compromise between data (likelihood) and prior beliefs (prior).

Prior beliefs and prior elicitation:

- Main conceptual difference between likelihood-based and Bayesian inference: Coming up with a sensible prior distribution.
- The prior should reflect your prior beliefs about the parameter of interest.
- Very common practice:
  - Pick a mathematically convenient class of distributions for the prior and
  - only decide on the parameter of this prior distribution.
- For example, one can formulate belief statements such as

$$\mathbb{P}(c_1 \leq \vartheta \leq c_2) = 1 - \alpha$$

  where c_1 and c_2 are pre-specified constants from which the prior parameters are determined.
- It is also very common to run analyses for a variety of different priors to study prior sensitivity.

Noninformative prior specifications:

- Flat priors

$$f(\boldsymbol{\vartheta}) \propto \text{const},$$

  are a popular choice to implement noninformative priors (no value of the parameter is favored a priori).

- Conceptual difficulties:
  - For non-bounded parameter spaces, flat priors are not actual probability distributions.
  - Flat priors are not invariant under transformations of the parameter of interest.

- An alternative are reference priors for which the prior has the smallest possible influence on the posterior (i.e. it maximizes the Kullback-Leibler discrepancy between the prior and the posterior for given data).
- Another option is Jeffreys' invariant prior

$$f(\boldsymbol{\vartheta}) \propto \sqrt{|\boldsymbol{F}(\boldsymbol{\vartheta})|}$$

  with expected Fisher information $\boldsymbol{F}(\boldsymbol{\vartheta})$.
- For scalar parameters, Jeffreys' prior is equivalent to the reference prior approach.

Priors for the success probability:

- The beta distribution is conjugate to the bernoulli observation model, i.e. the posterior is then also a beta distribution with updated parameters.
- Elicit the hyperparameters $a > 0$ and $b > 0$ based on prior statements, e.g. the prior expectation, variance, quantiles, probabilities, etc.
- A flat prior is $\pi \sim \mathrm{U}(0, 1)$ which is also a beta with $a = b = 1$.
- Jeffreys' prior is a beta distribution with $a = b = 0.5$.

- A typical discussion on Bayesian inference is that
  - frequentist inference assumes a true, fixed parameter value whereas
  - Bayesian inference assumes the parameter to be a random variable.
- This is, in general, misleading since the prior is merely used to reflect prior (un)certainty about the parameter of interest.
- The underlying philosophical question is, whether this can be done in a sensible way. . .

2.2 Markov Chain Monte Carlo Simulations

Numerically assessing the posterior:

- The ultimate outcome of a Bayesian data analysis is the posterior, reflecting posterior beliefs about the parameter of interest.
- This is often reduced to point estimates, credible intervals, etc.
- Unfortunately, in most models of reasonable complexity, the posterior is not analytically accessible.
- In particular, the normalizing constant

$$f(\boldsymbol{y}) = \int f(\boldsymbol{y}|\boldsymbol{\vartheta}) f(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta}$$

  is unknown and for models of at least moderate complexity it can also not easily be numerically determined.

- If we could obtain random samples $\boldsymbol{\vartheta}^{[t]}$, $t = 1, \ldots, T$ from the posterior, we could empirically estimate any quantity of interest at any desired level of precision:
  - Posterior expectations can be determined based on the law of large numbers via

$$\frac{1}{T} \sum_{t=1}^{T} g(\boldsymbol{\vartheta}^{[t]}) \rightarrow \mathbb{E}(g(\boldsymbol{\vartheta})|\boldsymbol{y}).$$

  - Similar statements exist for empirical quantiles.
  - Even the complete posterior could be estimated based on histograms or kernel density estimates.
- Markov chain Monte Carlo simulations are a way of simulating from the unknown and numerically untractable posterior!

Basic principles of MCMC:

- Generate a Markov chain that iteratively samples new values $\boldsymbol{\vartheta}^{[t]}$ given current values $\boldsymbol{\vartheta}^{[t-1]}$.
- The transition probabilities are chosen such that the Markov chain converges to the posterior as its stationarity distribution.
- Important consequences:
  - The samples $\boldsymbol{\vartheta}^{[t]}$, $t = 1, \ldots, T$ are not independent but feature serial correlation.
  - The Markov chain has to converge such that early values with small index t are not yet realisations from the posterior.

Some Markov chain theory:

- Consider a discrete time Markov chain with a discrete state space $\mathcal{S}$ of size $S = |\mathcal{S}|$
- Transitions from the current state to future states can then be characterized by a transition probability matrix $\boldsymbol{P}$.
- Under some regularity conditions, one can show that

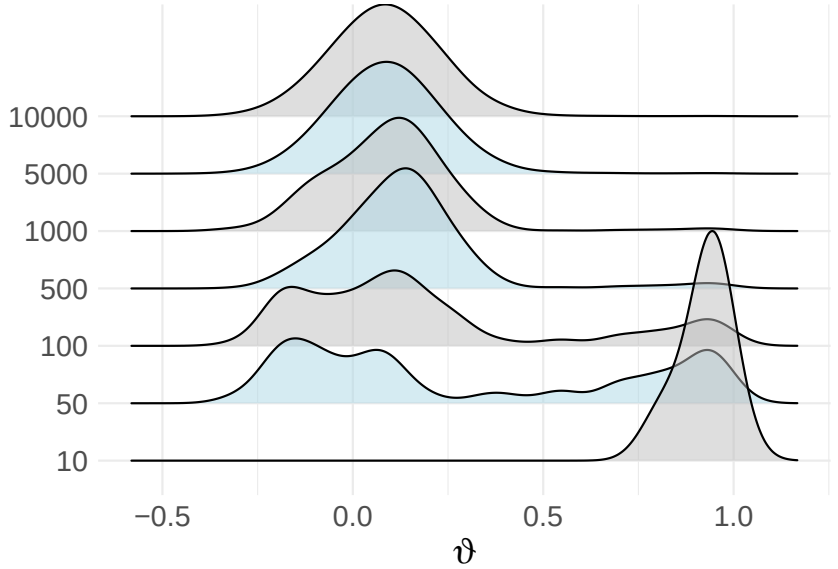
$$\lim_{t \to \infty} \boldsymbol{P}^t = \boldsymbol{P}^\infty$$

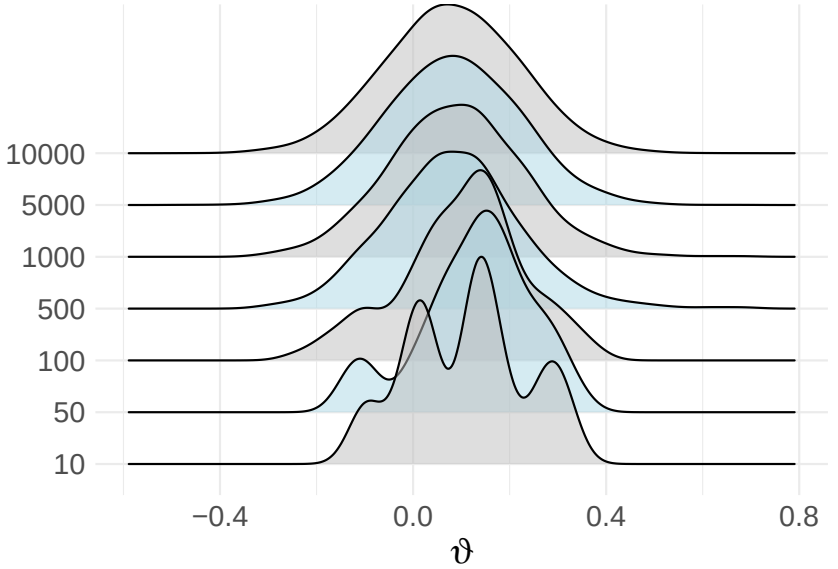
i.e. the repeated application of the transition probability matrix converges to a limiting matrix and each row in this matrix has exactly the same entries such that

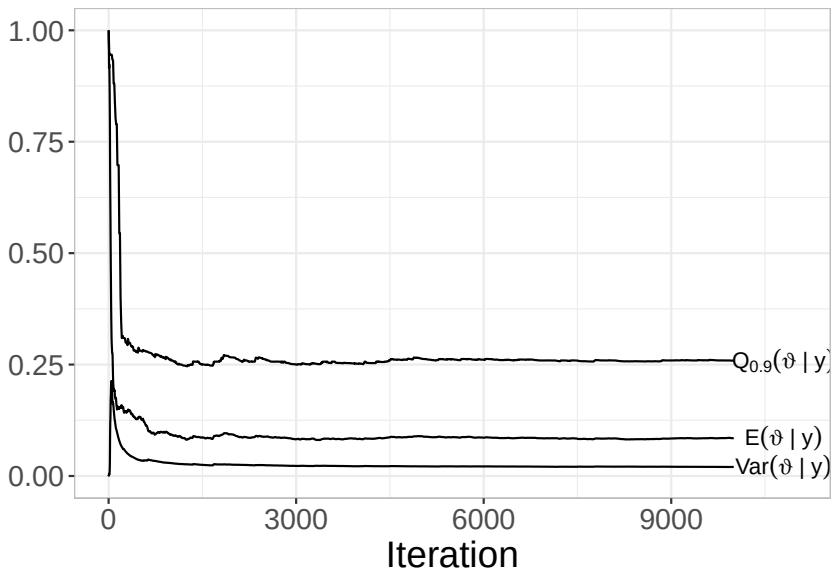
$$\boldsymbol{P}^\infty = \begin{pmatrix} \boldsymbol{\pi} \\ \vdots \\ \boldsymbol{\pi} \end{pmatrix}$$

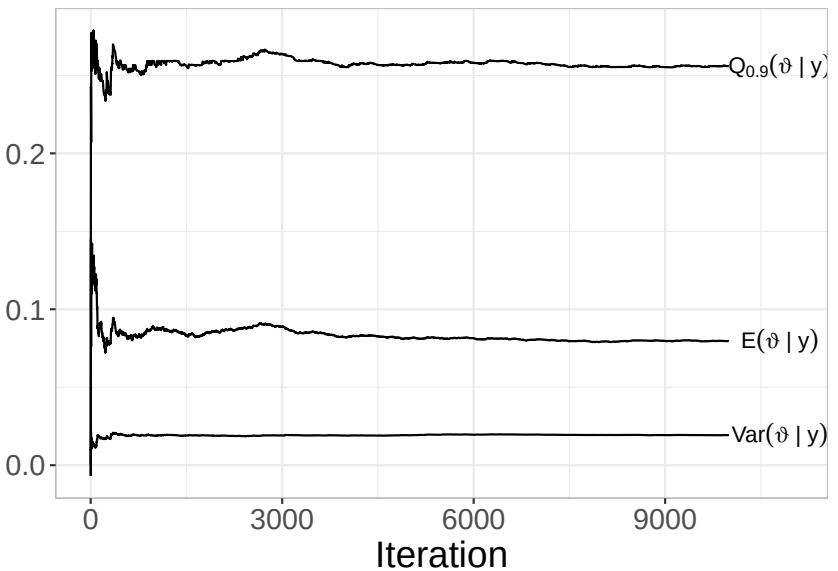
where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_S)$ is the stationary distribution of the Markov chain.

- In Bayesian inference, the stationary distribution should be the posterior distribution.
- Mathematical theory ensures convergence towards the stationary distribution in the limit, but in practice convergence has to be monitored appropriately.
- The convergence behaviour also depends on the starting values
  $\Rightarrow$ Remove burn in period.
- Samples from a Markov chain exhibit serial dependence that has to be accounted for
  $\Rightarrow$ thin out the Markov chain to achieve approximate independence.


10000
5000
1000
500
100
50
10
−0.5
0.0
0.5
1.0
ϑ


10000
5000
1000
500
100
50
10
−0.4
0.0
0.4
0.8
ϑ


1.00
0.75
0.50
0.25
0.00
Q0.9(ϑ | y)
E(ϑ | y)
Var(ϑ | y)
0
3000
6000
9000
Iteration


0.2
0.1
0.0
Q0.9(ϑ | y)
E(ϑ | y)
Var(ϑ | y)
0
3000
6000
9000
Iteration

A generic MCMC algorithm:

- Iterate the following two steps:
  - generate proposals for a new value of $\boldsymbol{\vartheta}^*$ from a so-called proposal density that depends on the data and the current state of all parameters.
  - accept the proposal only with a certain probability that depends on the posterior as well as proposal density. If the proposal is not accepted, the parameter remains in its current state.
- Instead of considering all parameters simultaneously, this is typically done in turn for sub-blocks of parameters, i.e. we split $\boldsymbol{\vartheta}$ into

$$\boldsymbol{\vartheta} = (\boldsymbol{\vartheta}_1', \ldots, \boldsymbol{\vartheta}_s', \ldots, \boldsymbol{\vartheta}_S')'.$$

- The blocks typically reflect structures from the model formulation.

- More precisely
  - propose a new value for $\boldsymbol{\vartheta}_s$ from a proposal density

$$q_s(\boldsymbol{\vartheta}_s^*|\boldsymbol{\vartheta}_1^{[t]},\ldots,\boldsymbol{\vartheta}_{s-1}^{[t]},\boldsymbol{\vartheta}_s^{[t-1]},\ldots,\boldsymbol{\vartheta}_S^{[t-1]},\boldsymbol{y})$$

  - accept the proposed new value $\boldsymbol{\vartheta}_s^*$ with probability

$$\alpha(\boldsymbol{\vartheta}_s^*|\boldsymbol{\vartheta}_s^{[t-1]}) = \min\left\{\frac{f(\boldsymbol{\vartheta}_s^*|\boldsymbol{\vartheta}_{-s}^{[t-1]},\boldsymbol{y})q_s(\boldsymbol{\vartheta}_s^{[t-1]}|\boldsymbol{\vartheta}_1^{[t]},\ldots,\boldsymbol{\vartheta}_{s-1}^{[t]},\boldsymbol{\vartheta}_s^*,\ldots,\boldsymbol{\vartheta}_S^{[t-1]})}{f(\boldsymbol{\vartheta}_s^{[t-1]}|\boldsymbol{\vartheta}_{-s}^{[t-1]},\boldsymbol{y})q_s(\boldsymbol{\vartheta}_s^*|\boldsymbol{\vartheta}_1^{[t]},\ldots,\boldsymbol{\vartheta}_{s-1}^{[t]},\boldsymbol{\vartheta}_s^{[t-1]},\ldots,\boldsymbol{\vartheta}_S^{[t-1]})},1\right\}$$

    otherwise set $\boldsymbol{\vartheta}_s^{[t]} = \boldsymbol{\vartheta}_s^{[t-1]}$.

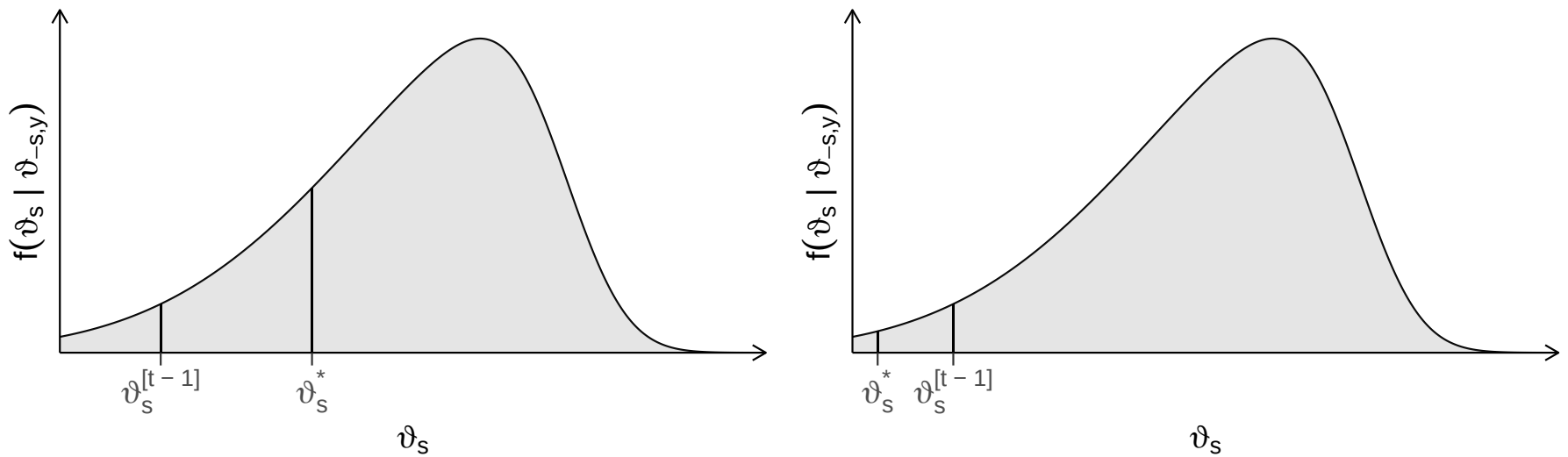
- The full conditional distribution

$$f(\boldsymbol{\vartheta}_s|\boldsymbol{\vartheta}_{-s}^{[t-1]}, \boldsymbol{y}) = f(\boldsymbol{\vartheta}_s|\boldsymbol{\vartheta}_1^{[t]}, \ldots, \boldsymbol{\vartheta}_{s-1}^{[t]}, \boldsymbol{\vartheta}_{s+1}^{[t-1]}, \ldots, \boldsymbol{\vartheta}_S^{[t-1]}, \boldsymbol{y})$$

  is proportional to the posterior, i.e.

$$f(\boldsymbol{\vartheta}_s|\boldsymbol{\vartheta}_{-s}^{[t-1]}, \boldsymbol{y}) \propto f(\boldsymbol{\vartheta}|\boldsymbol{y}).$$

- Intuition for the acceptance probability:



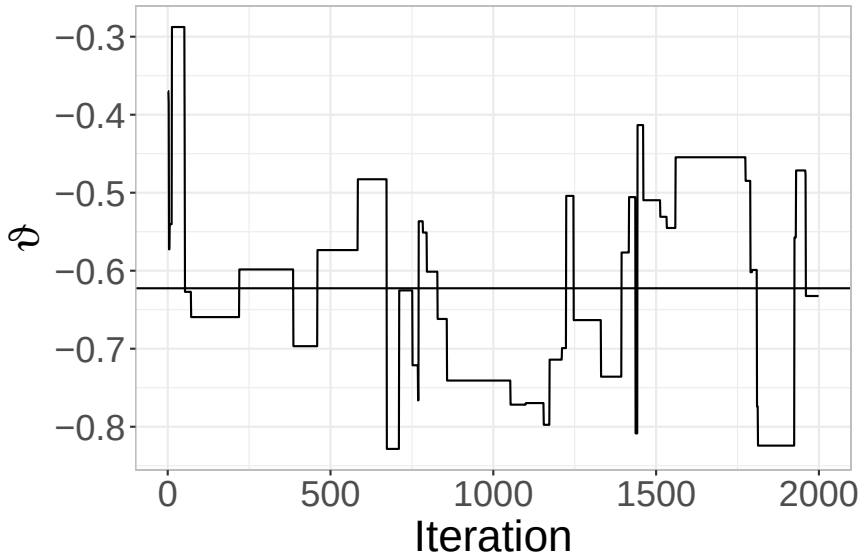
Monitoring MCMC:

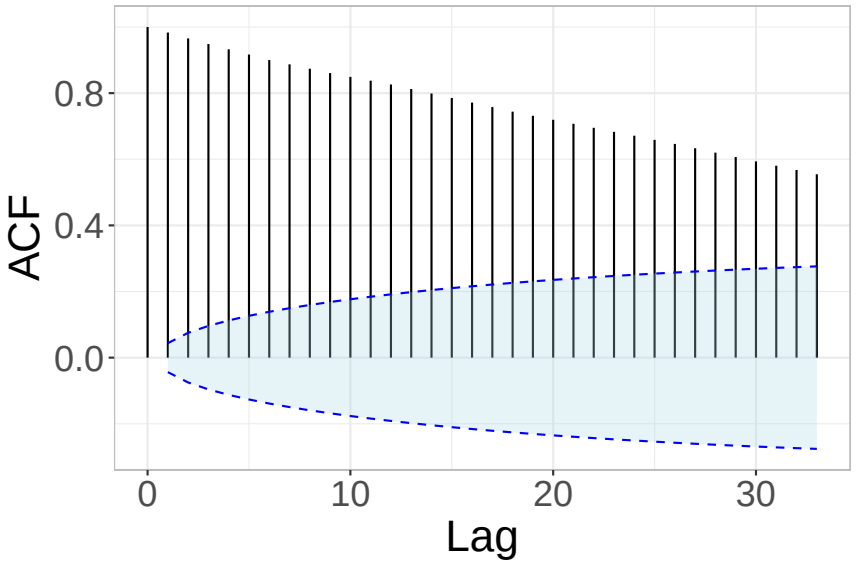
- Early versions of MCMC often relied on random walk proposals

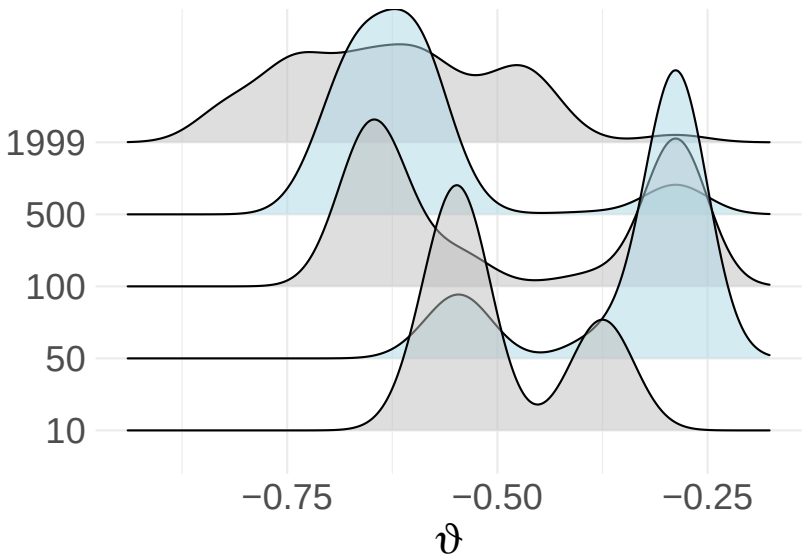
$$\boldsymbol{\vartheta}_s^* = \boldsymbol{\vartheta}_s^{[t-1]} + \boldsymbol{u}_t, \qquad \boldsymbol{u}_t \sim \mathrm{N}(\mathbf{0}, \tau_u^2 \boldsymbol{I})$$

  where the variance τ_u^2 is a tuning parameter that determines the mixing and convergence behaviour.

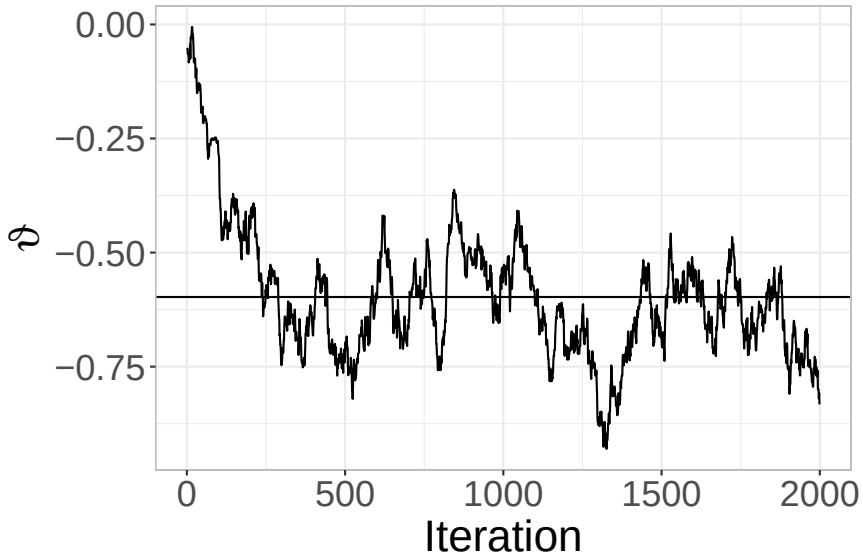
- Random walk proposal with a large variance τ_u^2:

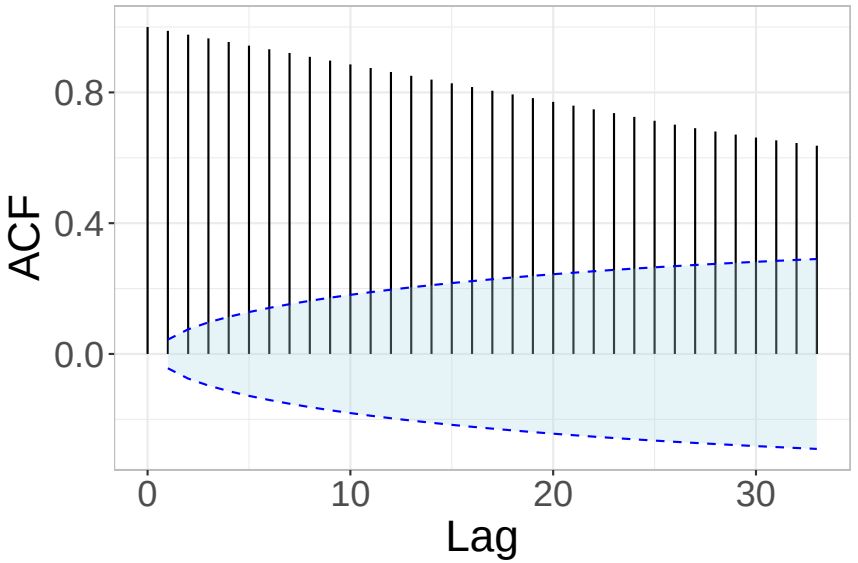


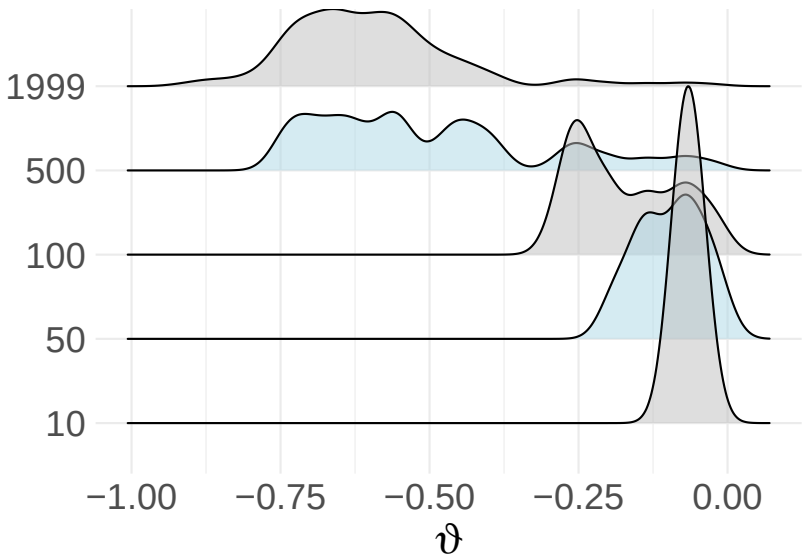




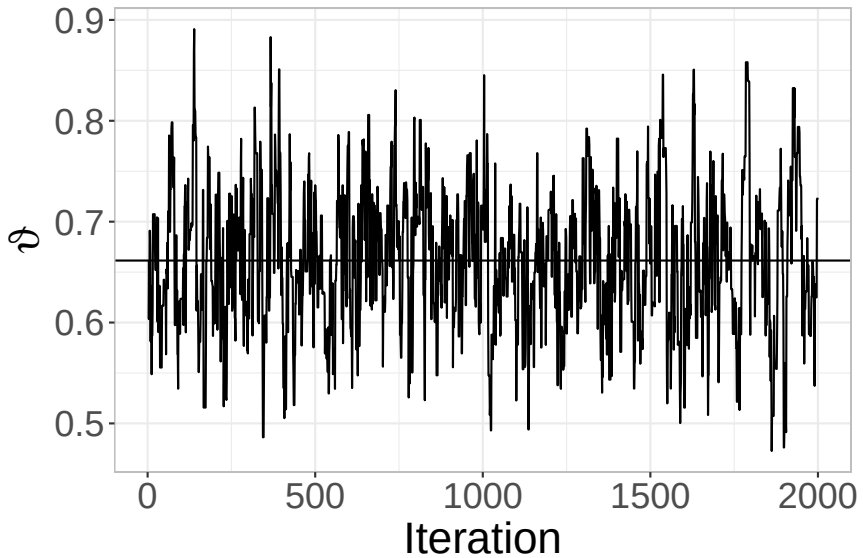
- Random walk proposal with a small variance τ_u^2:

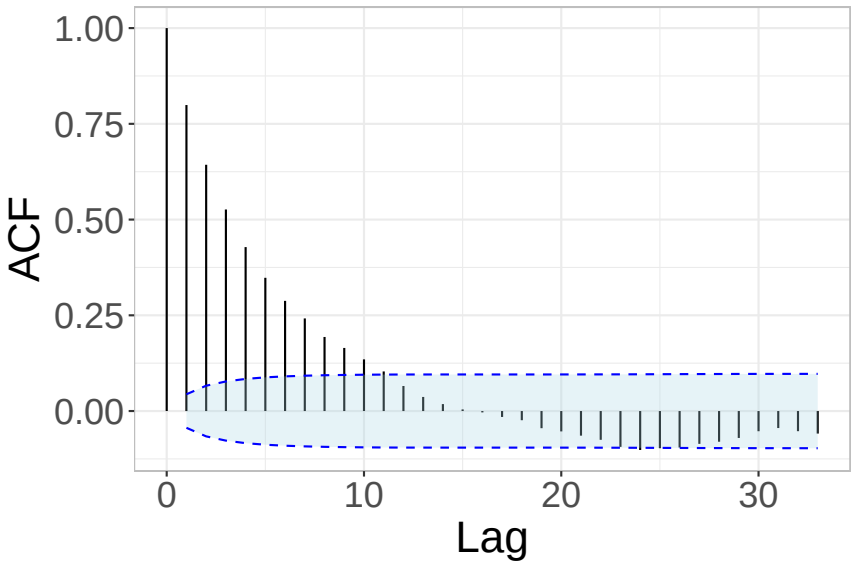


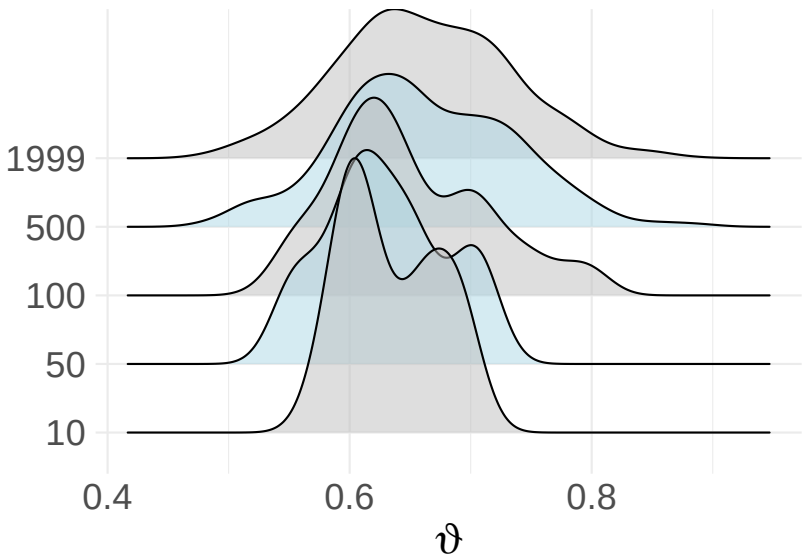




- Locally quadratic approximation of the log-full conditional:







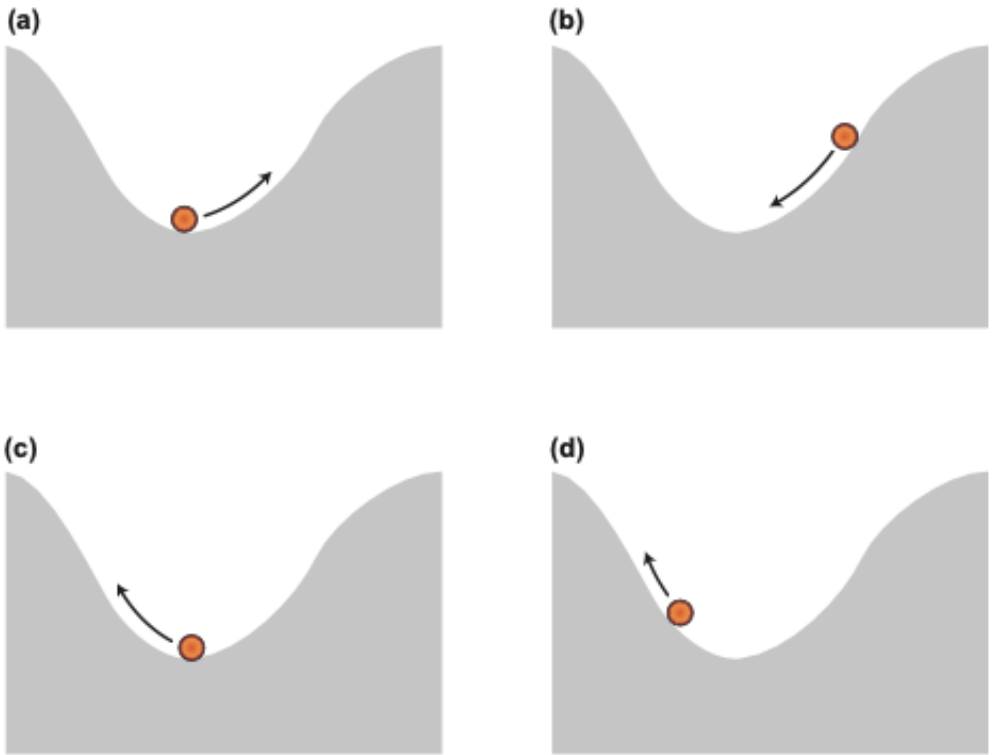
Locally quadratic approximation of the log-full conditional:

- Asymptotic theory suggests that posteriors tend to be normal for large samples.
- The moments of the asymptotic normal can be estimated as the mode and the negative curvature at the mode of the log-posterior.
- Idea: Apply the same idea to the log-full conditionals and propose from the resulting local quadratic approximation.
- Advantages:
  - Automatically adapts to the shape of the full conditionals.
  - Deviations from the asymptotic normal will be corrected for in the acceptance step.
- In the context of regression models, this is also referred to as iteratively weighted least squares (IWLS) updates since they resemble IWLS optimisation for finding maximum likelihood estimates.

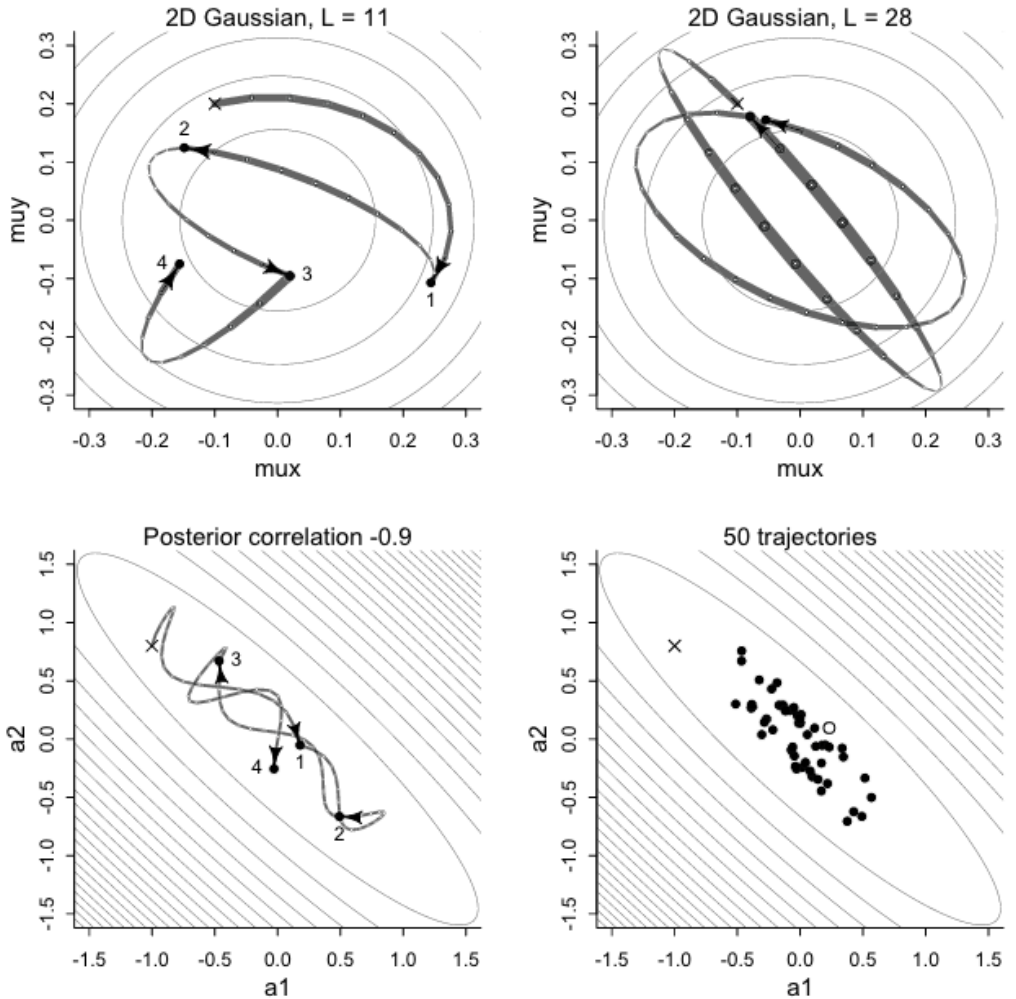
Hamiltonian Monte Carlo:

- Alternative approach for samples from the posterior that is motivated from a physical perspective.
- Consider the parameter vector $\boldsymbol{\vartheta}$ as a particle situated on a surface defined by $-\log(f(\boldsymbol{\vartheta}|\boldsymbol{y}))$.
- Assuming frictionless movement, the parameter would traverse towards the mode of the posterior along a path determined by the gradient $-\frac{\partial}{\partial \boldsymbol{\vartheta}} \log(f(\boldsymbol{\vartheta}|\boldsymbol{y}))$.
- Idea of HMC:
  - Push the current value of $\boldsymbol{\vartheta}$ with a random kinetic impulse $\boldsymbol{p} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{M})$ (the momentum induced by the mass matrix $\boldsymbol{M}$) and follow the resulting path over the surface $-\log(f(\boldsymbol{\vartheta}|\boldsymbol{y}))$.
  - Pick a random value from this path to make this the next value of the Markov chain.

- Univariate illustration (from Thomas & Tu, 2020, Learning Hamiltonian Monte Carlo in R):



- Bivariate illustration (from McElreath, 2020, Statistical rethinking: A Bayesian course with examples in R and Stan):



- More precisely:
  - The joint behaviour of $\boldsymbol{\vartheta}$ and $\boldsymbol{p}$ is described by the Hamiltonian function

$$H(\boldsymbol{\vartheta}, \boldsymbol{p}) = U(\boldsymbol{\vartheta}) + K(\boldsymbol{p})$$

    where $U(\boldsymbol{\vartheta})$ is the potential energy and $K(\boldsymbol{p})$ is the kinetic energy.

  - The potential energy is given by

$$U(\boldsymbol{\vartheta}) = -\log(f(\boldsymbol{\vartheta}|\boldsymbol{y}))$$

    and the kinetic energy is

$$K(\boldsymbol{p}) = \frac{1}{2}\boldsymbol{p}'\boldsymbol{M}^{-1}\boldsymbol{p}.$$

– The trajectory over time t is then described by the Hamiltonian equations

$$\begin{aligned}\frac{\partial \boldsymbol{p}}{\partial t} &= -\frac{H(\boldsymbol{\vartheta}, \boldsymbol{p})}{\partial \boldsymbol{\vartheta}} = \frac{\partial}{\partial \boldsymbol{\vartheta}} \log(f(\boldsymbol{\vartheta}|\boldsymbol{y})) \\ \frac{\partial \boldsymbol{\vartheta}}{\partial t} &= \frac{H(\boldsymbol{\vartheta}, \boldsymbol{p})}{\partial \boldsymbol{p}} = \boldsymbol{M}^{-1}\boldsymbol{p}.\end{aligned}$$

– These differential equations have to be solved numerically.

– In HMC, the common way of doing this is by the leapfrog iterations characterised by a step length ϵ:

$$\begin{aligned}\boldsymbol{p}^{[t+0.5\epsilon]} &= \boldsymbol{p}^{[t]} + \frac{\epsilon}{2}\frac{\partial}{\partial \boldsymbol{\vartheta}} \log(f(\boldsymbol{\vartheta}^{[t]}|\boldsymbol{y})) \\ \boldsymbol{\vartheta}^{[t+\epsilon]} &= \boldsymbol{\vartheta}^{[t]} + \epsilon \boldsymbol{M}^{-1}\boldsymbol{p}^{[t+0.5\epsilon]} \\ \boldsymbol{p}^{[t+\epsilon]} &= \boldsymbol{p}^{([+0.5\epsilon]} + \frac{\epsilon}{2}\frac{\partial}{\partial \boldsymbol{\vartheta}} \log(f(\boldsymbol{\vartheta}^{[t+\epsilon]}|\boldsymbol{y}))\end{aligned}$$

which are repeated L times in each MCMC iteration.

- To correct for the numerical approximation, an acceptance step similar as in the MH algorithm is added:

$$\alpha(\boldsymbol{\vartheta}^*|\boldsymbol{\vartheta}^{[t-1]}) = \min\left\{\frac{f(\boldsymbol{\vartheta}^*|\boldsymbol{y})f(\boldsymbol{p}^*)}{f(\boldsymbol{\vartheta}^{[t-1]}|\boldsymbol{y})f(\boldsymbol{p}^{[t-1]})}, 1\right\}$$

  where $\boldsymbol{\vartheta}^*$ and $\boldsymbol{p}^*$ are the values at the end of the leapfrog iterations and $\boldsymbol{\vartheta}^{[t-1]}$ and $\boldsymbol{p}^{[t-1]}$ are the values from the previous MCMC iteration.

- The normalizing constant of the posterior is not needed since only derivatives of $\log(f(\boldsymbol{\vartheta}|\boldsymbol{y}))$ are used.

- The main tuning parameters of HMC are
  - the distribution of the kinetic impulse $\boldsymbol{p}$ and in particular the covariance matrix $\boldsymbol{M}$.
  - the step size ϵ used in the leapfrog iterations (small values get the solution closer to the analytic solution but require more steps).
  - the number of leapfrog iterations L.
- The tuning parameters can be optimized in a warm up period and can also be made parameter-specific and/or stochastic.
- Orthogonalizing the parameter allows to split the parameter in blocks.

Advantages of MCMC:

- Access to the complete posterior distribution without requiring asymptotic considerations.
- Divide and conquer approach based on updating blocks of parameters separately allows handling very complex models having hundreds or thousands of parameters.
- Modular representation of hierarchically formulated statistical models where certain parts of the model can be replaced without affecting the other model components
- From the samples of the model parameters, we can determine not only inferences about these parameters themselves, but also inference for complex functionals of these parameters.

Comparison of MCMC approaches:

- IWLS requires second derivatives while HMC needs only the gradient.
- Both can use auto-diff functionality to determine these derivatives
- Parameters should be transformed to the real line first.
- IWLS and HMC use information from the full conditional to guide the sampler to regions of high posterior density (unlike random walk proposals).
- Advocates of HMC often do not use thinning of the Markov chain since it is supposed to produce uncorrelated or even anticorrelated draws.

2.3 Bayesian Inference with MCMC

Posterior summaries:

- While the ultimate outcome of Bayesian inference is the posterior, this is often compressed into posterior summaries, in particular
  - posterior point estimates and
  - posterior measures of uncertainty.
- Typical point estimates:
  - posterior mean (estimated by averages of samples),
  - posterior median (estimated by empirical median),
  - posterior mode (difficult to determine from samples).

- Typical measures of uncertainty:
  - posterior variance / standard deviation (estimated by empirical analogues),
  - posterior quantiles.

Credible intervals and bands:

- A pointwise Bayesian credible interval $[\vartheta_{s,\text{low}}, \vartheta_{s,\text{upp}}]$ for a scalar parameter ϑ_s is characterized by the posterior coverage probability

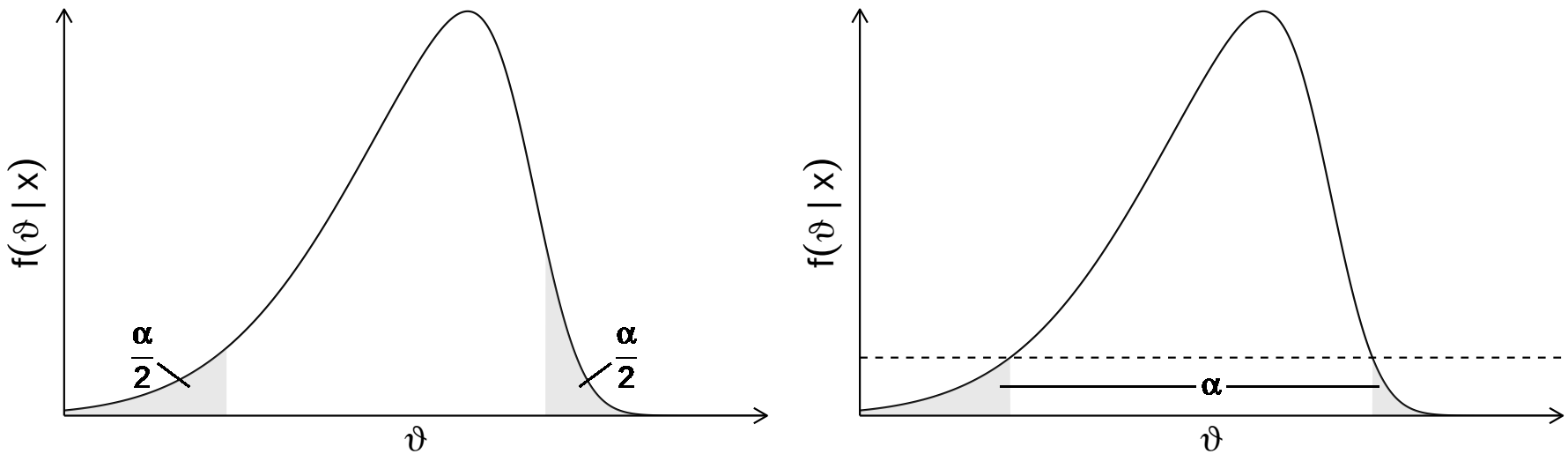
$$\mathbb{P}(\vartheta_{s,\text{low}} \leq \vartheta_s \leq \vartheta_{s,\text{upp}} | \boldsymbol{y}) \geq 1 - \alpha$$

  where $1 - \alpha$ denotes the desired coverage level.

- A simultaneous band for multiple parameters $\{\vartheta_s, s \in \mathcal{S}\}$ should have

$$\mathbb{P}(\vartheta_{s,\text{low}} \leq \vartheta_s \leq \vartheta_{s,\text{upp}}, s \in \mathcal{S} | \boldsymbol{y}) \geq 1 - \alpha$$

- Symmetric and highest posterior density credible intervals:



Bayesian Tests:

- To test the hypotheses

$$H_0 : \boldsymbol{\vartheta} \in \boldsymbol{\Theta}_0 \quad \text{vs.} \quad H_1 : \boldsymbol{\vartheta} \notin \boldsymbol{\Theta}_0$$

  we can compute the posterior probabilities

$$p_0 = P(\boldsymbol{\vartheta} \in \boldsymbol{\Theta}_0 | \boldsymbol{y}) \quad \text{and} \quad p_1 = P(\boldsymbol{\vartheta} \notin \boldsymbol{\Theta}_0 | \boldsymbol{y}).$$

- The decision can then be based on the ratio

$$\frac{p_1}{p_0}$$

  that measures the evidence in favor of H_1 as compared to H_0.

- H_1 and H_0 are therefore treated symmetrically in the Bayesian context.
- Unfortunately, point hypotheses can not meaningfully be tested in the Bayesian paradigm since then
$$P(\boldsymbol{\vartheta} = \boldsymbol{\vartheta}_0 | \boldsymbol{y}) = 0.$$
- Instead of formally testing hypothesis, the decision between H_0 and H_1 is often made based on model choice procedures in the Bayesian framework.
- As an alternative, one often considers a Bayesian credible interval for $\boldsymbol{\vartheta}$ and evaluates whether $\boldsymbol{\vartheta}_0$ is contained in the credible interval or not.

Inference for derived quantities:

- Goal: Conduct Bayesian inference for a derived quantity

$$\boldsymbol{\eta} = g(\boldsymbol{\vartheta}).$$

- Convenient feature of MCMC: If $\boldsymbol{\vartheta}^{[1]}, \ldots, \boldsymbol{\vartheta}^{[T]}$ is a sample from the posterior of $\boldsymbol{\vartheta}$, $g(\boldsymbol{\vartheta}^{[1]}), \ldots, g(\boldsymbol{\vartheta}^{[T]})$ will be a sample from the posterior of the transformed parameter.
- No restrictions on the transformation $g(\cdot)$ and no need to deal with asymptotic considerations

The Bayes factor for model comparison:

- If there are L competing models $M_1, \ldots, M_L$ with associated parameters $\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_L$, the posterior for $\boldsymbol{\vartheta}_l$ given the model M_l is

$$f(\boldsymbol{\vartheta}_l|\boldsymbol{y}, M_l) = \frac{f(\boldsymbol{y}|\boldsymbol{\vartheta}_l, M_l) f(\boldsymbol{\vartheta}_l|M_l)}{f(\boldsymbol{y}|M_l)},$$

  where $f(\boldsymbol{y}|\boldsymbol{\vartheta}_l, M_l)$ and $f(\boldsymbol{\vartheta}_l|M_l)$ are the likelihood and the prior of $\boldsymbol{\vartheta}_l$ under model M_l, respectively, and

$$f(\boldsymbol{y}|M_l) = \int f(\boldsymbol{y}|\boldsymbol{\vartheta}_l, M_l) f(\boldsymbol{\vartheta}_l|M_l) d\boldsymbol{\vartheta}_l$$

  is the marginal likelihood of model M_l.

- For model selection, we assign prior distributions to the competing models, $f(M_l)$, and compare models through their marginal posteriors

$$f(M_l|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|M_l)f(M_l)}{f(\boldsymbol{y})}, \qquad l = 1, \ldots, L,$$

  where $f(\boldsymbol{y}) = \sum_{l=1}^{L} f(\boldsymbol{y}|M_l)f(M_l)$.

- Prefer model M_l against model M_s, $s \neq l$ if

$$f(M_l|\boldsymbol{y}) > f(M_s|\boldsymbol{y})$$

  or in other words if

$$\frac{f(M_l|\boldsymbol{y})}{f(M_s|\boldsymbol{y})} = \frac{f(M_l)f(\boldsymbol{y}|M_l)}{f(M_s)f(\boldsymbol{y}|M_s)} > 1.$$

- The ratio of marginal likelihoods is referred to as the Bayes factor

$$\mathrm{BF}_{ls} = \frac{f(\boldsymbol{y}|M_l)}{f(\boldsymbol{y}|M_s)} \ .$$

which reflects model preference under equal prior probabilities

$$f(M_1) = \ldots = f(M_L)$$

- The marginal likelihoods can be estimated as arithmetic mean

$$\hat{f}(\boldsymbol{y}) = \frac{1}{T}\sum_{t=1}^{T} f(\boldsymbol{y}|\boldsymbol{\vartheta}^{[t]})$$

or harmonic mean

$$\hat{f}(\boldsymbol{y}) = \left(\frac{1}{T}\sum_{t=1}^{T} \frac{1}{f(\boldsymbol{y}|\boldsymbol{\vartheta}^{[t]})}\right)^{-1}$$

with samples $\boldsymbol{\vartheta}^{[t]}$, $t = 1, \ldots, T$ from the posterior distribution.

Bayesian information criteria:

- Bayesian information criterion (BIC)

$$\text{BIC}(M_l) = -2l(\hat{\boldsymbol{\vartheta}}_l) + \log(n)p_l$$

where p_l is the number of parameters in model l.

- Deviance information criterion (DIC)

$$\mathrm{DIC} = \overline{D(\boldsymbol{\vartheta})} + p_{\mathrm{DIC}}$$

where

$$D(\boldsymbol{\vartheta}) = -2\log(f(\boldsymbol{y}|\boldsymbol{\vartheta})) = \frac{1}{T}\sum_{t=1}^{T} D(\boldsymbol{\vartheta}^{[t]})$$

denotes the model deviance and

$$p_{\mathrm{DIC}} = \overline{D(\boldsymbol{\vartheta})} - D(\overline{\boldsymbol{\vartheta}}) = \frac{1}{T}\sum_{t=1}^{T} D(\boldsymbol{\vartheta}^{[t]}) - D\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\vartheta}^{[t]}\right)$$

provides an estimate for the effective parameter count.

- Widely applicable information criterion (WAIC)

$$\text{WAIC} = 2\left(D_{\text{WAIC}} + p_{\text{WAIC}}\right)$$

with

$$D_{\text{WAIC}} = -\sum_{i=1}^{n} \log\left(\frac{1}{T}\sum_{t=1}^{T} p\left(y_i|\boldsymbol{\vartheta}^{[t]}\right)\right)$$

as the measure of model fit,

$$p_{\text{WAIC}} = \sum_{i=1}^{n} \widehat{\text{Var}}\left(\log\left(p\left(y_i|\boldsymbol{\vartheta}\right)\right)\right)$$

as the measure of model complexity, and the empirical variance

$$\widehat{\text{Var}}\left(a\right) = \frac{1}{T-1}\sum_{t=1}^{T}\left(a_t - \bar{a}\right)^2.$$

3 Bayesian Additive Regression

3.1 Introduction

Linear models:

- The work horse of statistical modelling and analysis is the linear model where

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2).$$

- The parameters β_j can be related to the expected change in the response associated with differences in x_j.

  $\Rightarrow$ Parameters have a specific meaning and purpose.

- Statistical inference is facilitated by the distributional assumptions on the error terms.
- However, in many practical situations the linear model is not flexible enough and/or assumptions may be questionable.

Nonlinear effects:

- Common practice if the linearity of the effect of x_j is questionable: Include low-order polynomials, e.g. replace $x_j\beta_j$ by

$$x_j\beta_j + x_j^2\beta_{j+1} + x_j^3\beta_{j+2}.$$

- Imposes strong assumptions on the form of the effect and is not very flexible.
- Ideally, the form of an effect should be left unspecified and should be determined by the data (under mild, qualitative assumptions).
- Additive model:

$$y_i = \beta_0 + s_1(x_{i1}) + \ldots + s_k(x_{ik}) + \varepsilon_i, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2).$$

- We will use penalized splines to represent the effects $s_j(x_{ij})$.

Clustered data:

- For longitudinal data $(y_{it}, \boldsymbol{x}_{it})$, $i = 1 \ldots, n$, $t = 1, \ldots, T$, a classical model of the form

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}$$

  may be questionable for a number of reasons:

  - Unobserved heterogeneity due to individual-specific, unobserved confounders that have not been included in the model,
  - Dependence between observations on one individual, or
  - Individual-specific regression coefficients.

- Similarly applies to other grouping structures (families, geographical regions, school classes, . . .)

- Random effects models are then often considered, e.g. random intercepts

$$y_{it} = \gamma_{0i} + \boldsymbol{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}$$

  with $\gamma_{i0} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \tau^2)$.

- More complex models can also have individual-specific random slopes or other additional structures.

Spatial dependence:

- For spatial regression data $(y(s), \boldsymbol{x}(s))$, one may similarly question whether linear models take unobserved spatial heterogeneity and/or dependence into account.
- Include spatially correlated random effects, leading to

$$y(s) = \gamma(s) + \boldsymbol{x}(s)'\boldsymbol{\beta} + \varepsilon(s)$$

with $\gamma(s)$ being an appropriately specified spatial stochastic process.

Bayesian additive regression:

- Bayesian additive regression provides a unifying framework for dealing with the challenges discussed so far.
- The model also supports other effect types, e.g. varying coefficients or interaction surfaces.
- The models can be conveniently represented in a hierarchical fashion that enables us to benefit from the flexibility of Bayesian inference.
- Tomorrow, we will discuss Bayesian distributional regression that allows us to overcome the normality assumption for the error terms.

Example: Car insurance data from two insurance companies in Belgium

- Sample of approximately 160.000 policyholders.
- Aims: Separate risk analyses for claim size and claim frequency to predict risk premium from covariates.
- Variables of primary interest: Claim size y_i or claim frequency h_i of policyholders.
- Covariates:

$vage$	vehicle's age
$page$	policyholder's age
hp	vehicle's horsepower
bm	bonus-malus score
d	district in Belgium
$\boldsymbol{x}$	vector of categorical covariates

- Generalised linear models:
  - Gaussian model for log-costs $\log(y)$:

$$\log(y) \sim \mathrm{N}(\boldsymbol{z}'\boldsymbol{\gamma}, \sigma^2).$$

  - Poisson model for frequencies h_i:

$$h \sim \mathrm{Po}(\exp(\boldsymbol{z}'\boldsymbol{\gamma})).$$

  - Linear predictors formed as a linear combination of (possibly transformed) covariates:

$$\eta = \boldsymbol{z}'\boldsymbol{\gamma} = \gamma_0 + x_1\gamma_1 + \ldots + x_p\gamma_p.$$

- Subject-matter knowledge:
  - Young and old drivers have a higher claims expenditure. This hints at a quadratic instead of a linear age effect, but the precise form is unknown.

    $\Rightarrow$ Replace the parametric effect with a nonparametric effect $s(page)$.
  - Male and female drivers have a different claims expenditure. This hints at an interaction between age and gender, but the effect should be allowed to vary with age.

    $\Rightarrow$ Instead of a parametric model of the form $\gamma_1 page + \gamma_2 sex + \gamma_3 page \cdot sex$ consider a model of the form $s_1(page) + sex\, s_2(page)$.
  - Drivers in rural areas cause less accidents with a higher average claim amount while drivers in urban areas cause more but smaller claims. The effect may change smoothly between rural and urban areas such that modeling based on a rural vs. urban dummy is too simplistic.

    $\Rightarrow$ Include a spatial function $s_{\text{spat}}(d)$ based on the district d a driver is living in.

- Model specifications:
  - Gaussian model for log-costs $\log(y)$:

$$\log(y) \sim \mathrm{N}(\eta, \sigma^2)$$

with

$$\eta = s_1(vage) + s_2(page) + s_3(bm) + s_4(hp) + s_{spat}(d) + \boldsymbol{z}'\boldsymbol{\gamma}.$$

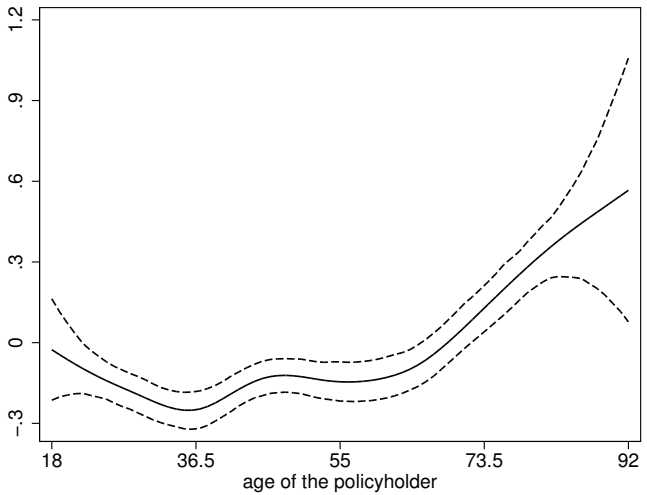
  - Poisson model for frequencies h_i:

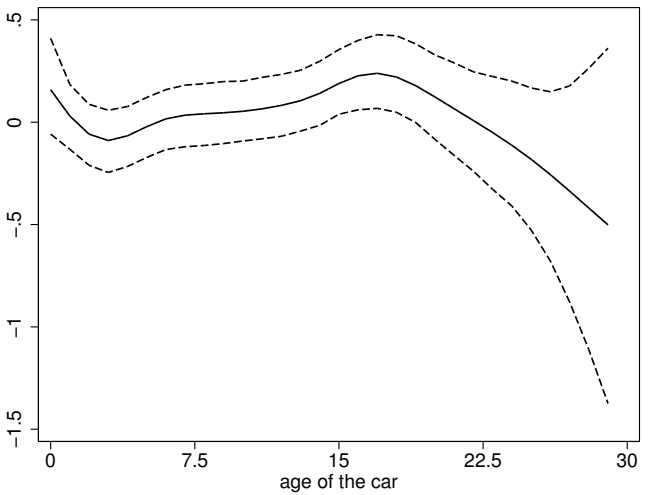
$$h \sim \mathrm{Po}(\exp(\eta))$$

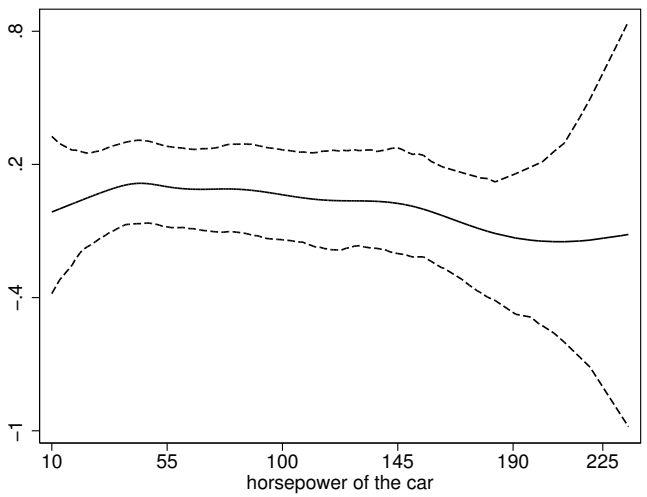
with

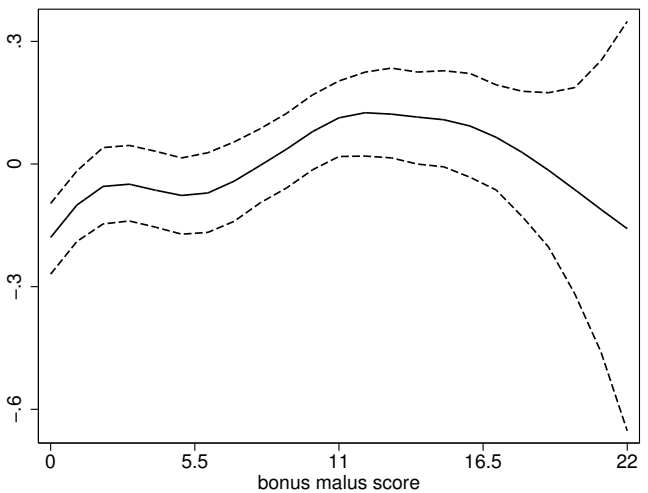
$$\eta = s_1(vage) + s_2(page) + s_3(page)sex + s_3(bm) + s_4(hp) + s_{spat}(d) + \boldsymbol{z}'\boldsymbol{\gamma}.$$

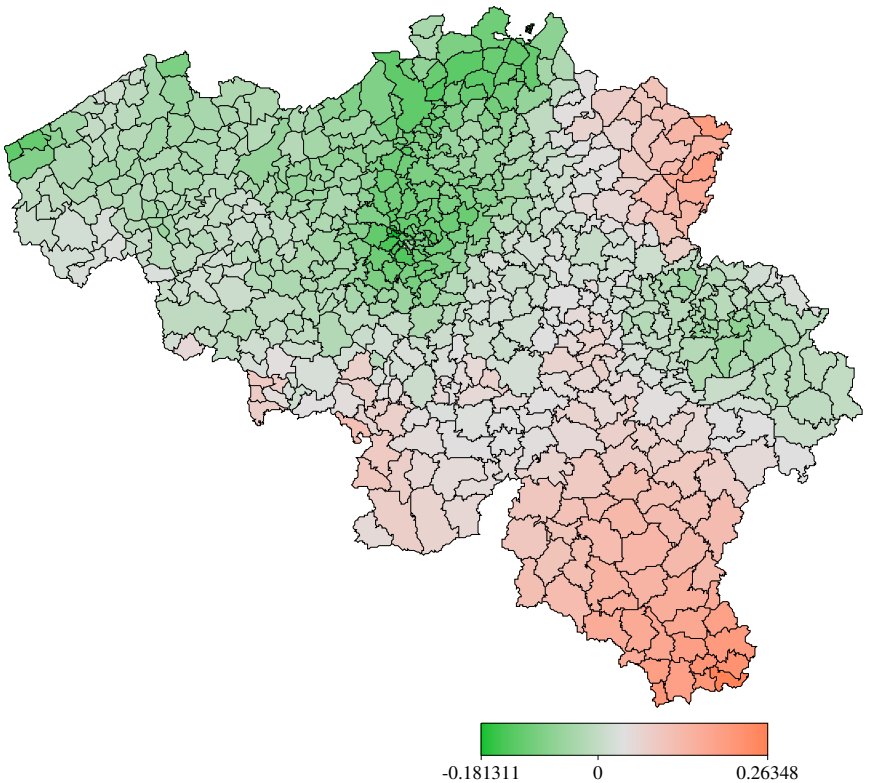
- Results for claim size:

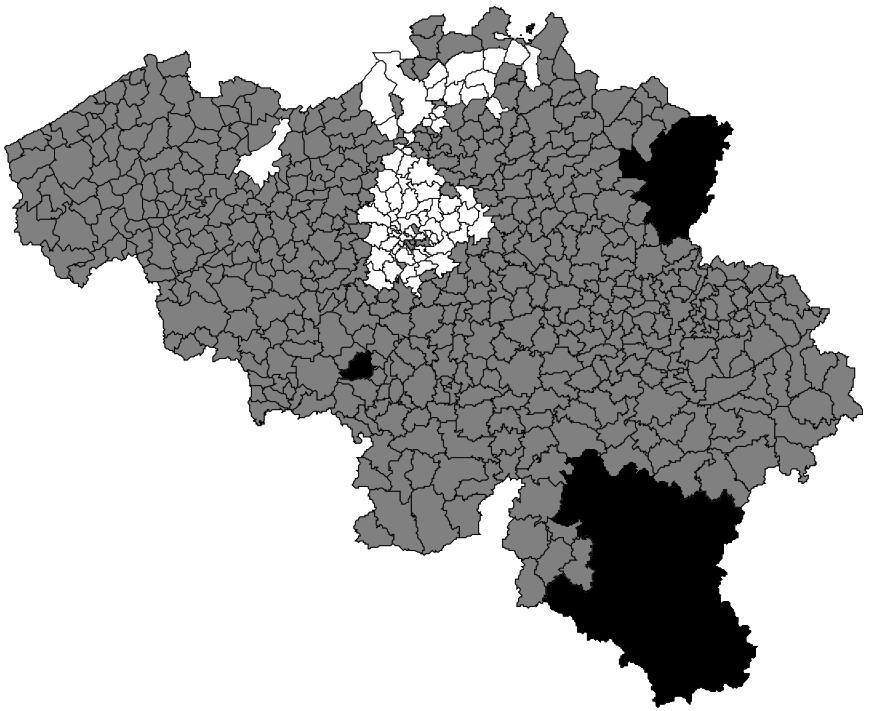




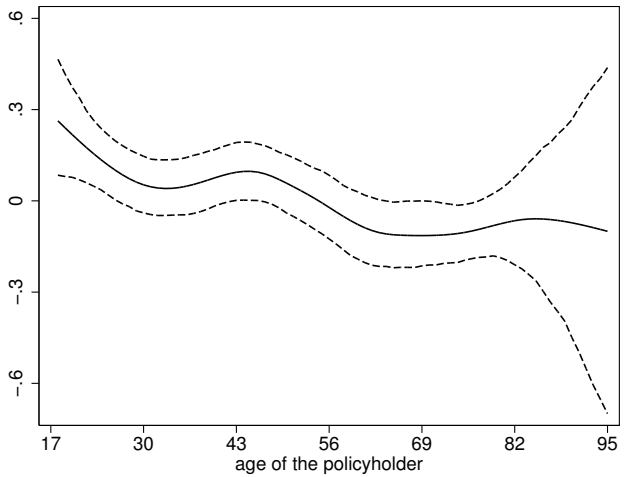


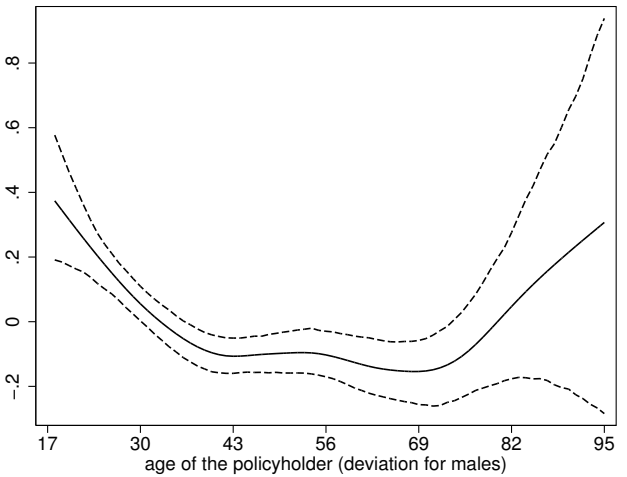


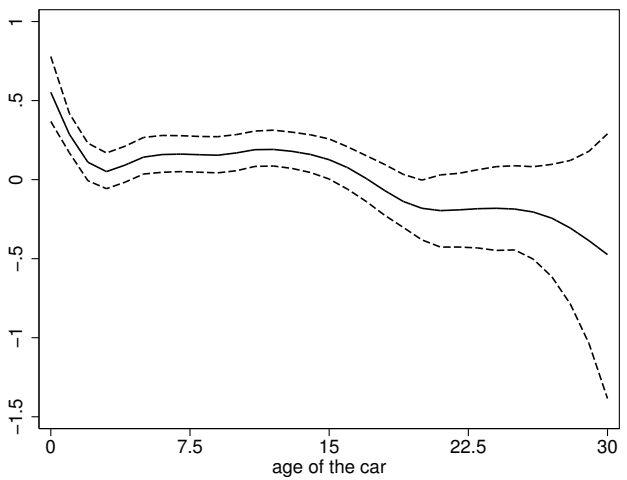

-0.181311
0
0.26348

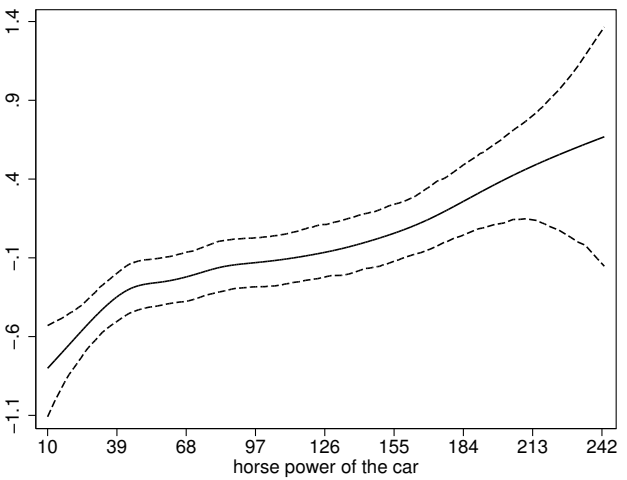


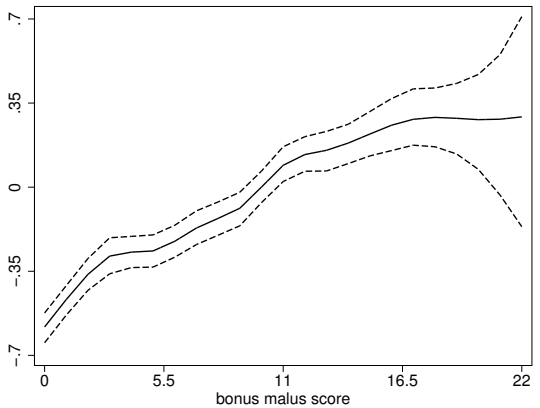
- Results for claim frequency:

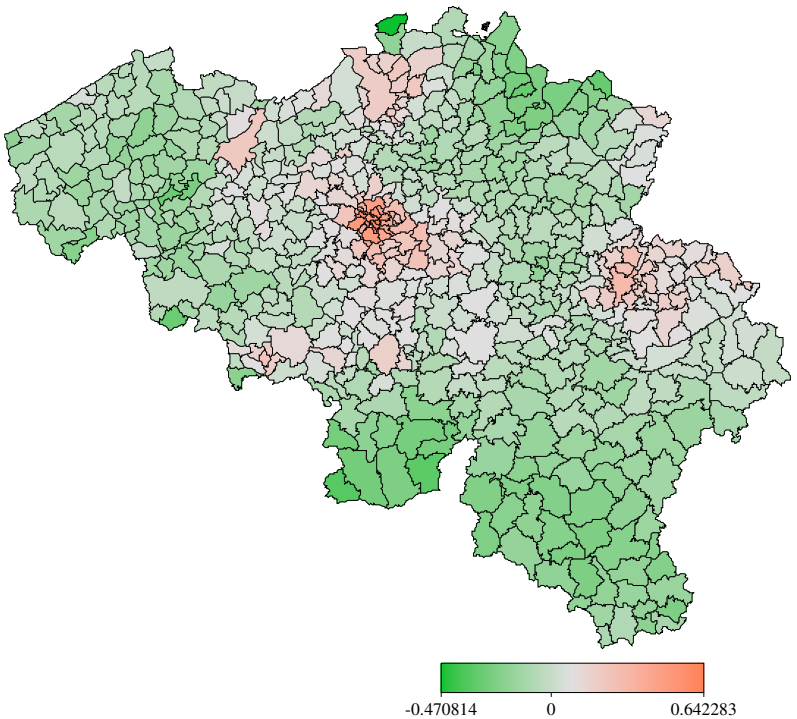


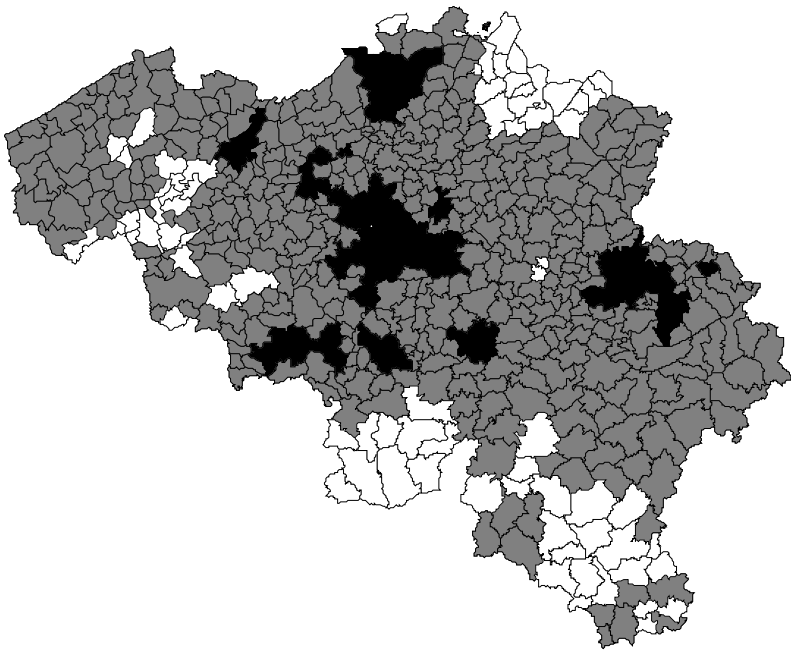







.7
.35
0
−.35
−.7
0
5.5
11
16.5
22
bonus malus score


-0.470814
0
0.642283



3.2 Penalized Spline Smoothing

Scatterplot smoothing:

- Start from scatterplot smoothing

$$y_i = s(z_i) + \varepsilon_i, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2)$$

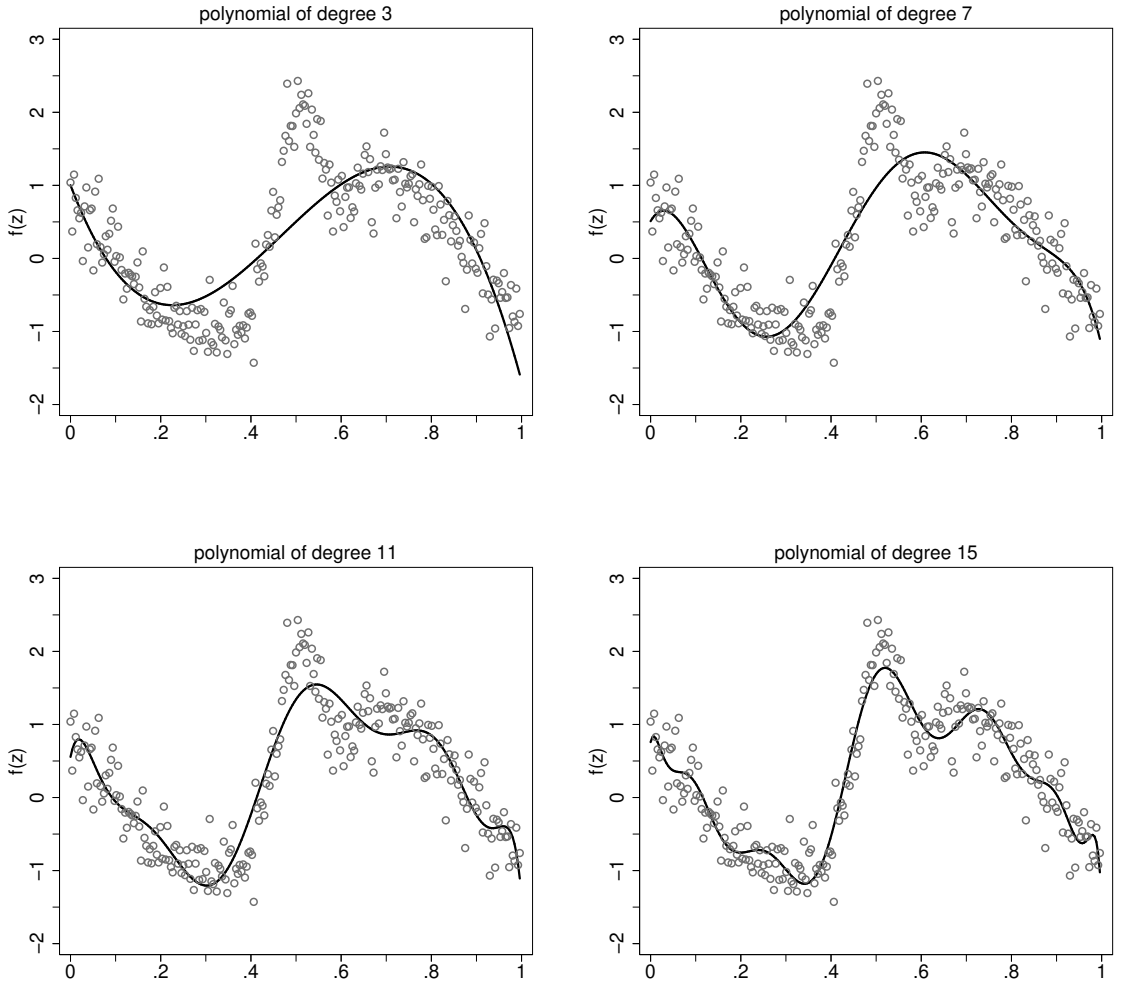
  where $s(z)$ should be inferred based on observations (z_i, y_i), $i = 1, \ldots, n$, for a continuous covariate z and response y.

- Common approach: Approximate $s(z)$ by a low-order polynomial

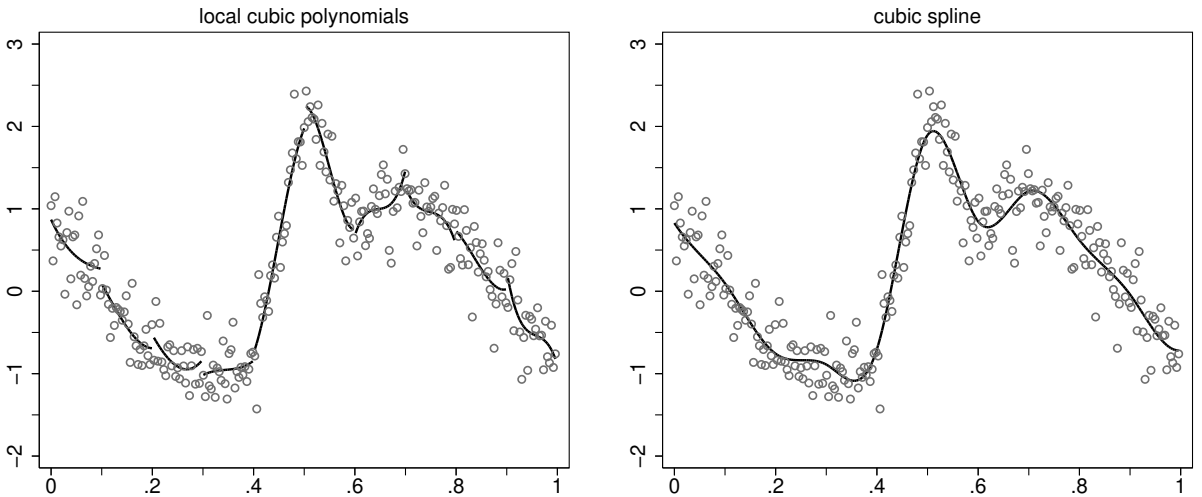
$$s(z_i) = \gamma_0 + \gamma_1 z_i + \ldots + \gamma_l z_i^l$$

  since any smooth function $s(\cdot)$ can be approximated arbitrarily accurately if the degree l is chosen large enough.

- In statistics, the problem of estimating the coefficients $\gamma_0, \ldots, \gamma_l$ limits the applicability of high polynomial degrees:



- Moreover, polynomials have the following disadvantages:
  - A polynomial assumes a global amount of smoothness for the function s.
  - Polynomial estimates tend to be unstable at the boundaries of the covariate space.
  - Low order polynomials induce very specific types of functional forms.
- A possibility to overcome some of these problems is to define polynomials piecewise on partial intervals of the covariate domain.



- Advantages:
  - Localized fits instead of global smoothness, and
  - high flexibility.
- Disadvantages:
  - Potentially large number of regression coefficients.
  - The resulting function is no longer smooth since the function pieces on the intervals are fitted separately.

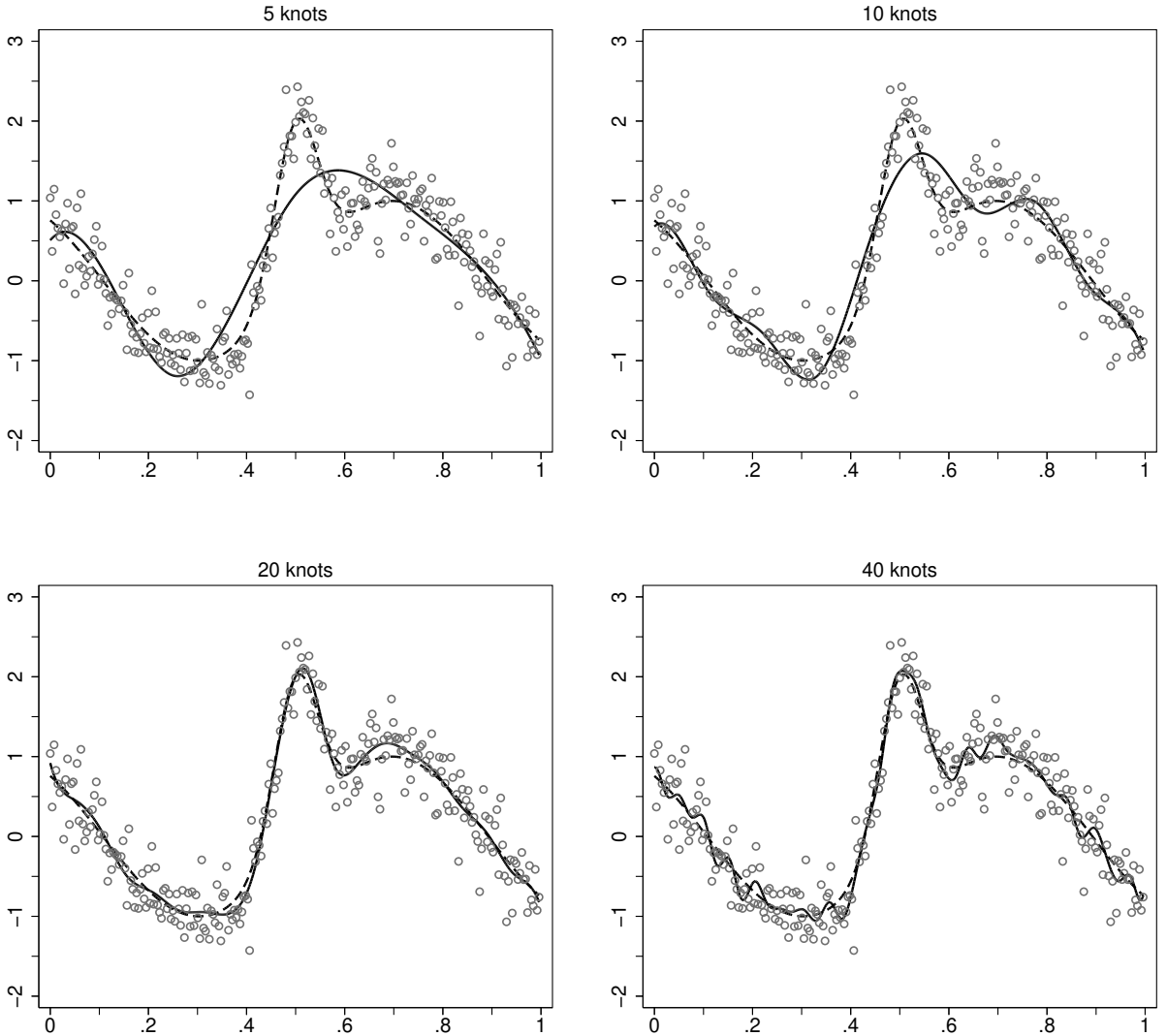
$\Rightarrow$ Combine global polynomials and piecewise polynomials to obtain polynomial splines.

Polynomial splines:

- A function s is a polynomial spline of degree l with knots $a = \kappa_1 < \ldots < \kappa_m = b$ if
  - $s(z)$ is $(l-1)$ times continuously differentiable,
  - $s(z)$ is polynomial of degree l on each of the intervals $[\kappa_j, \kappa_{j+1})$.

  $\Rightarrow$ Piecewise polynomial with global smoothness.

- The flexibility of a polynomial spline is determined by the number of knots.



- Polynomial splines form a vector space such that they can be represented in terms of $L = m + l - 1$ basis functions, i.e.

$$s(z) = \sum_{l=1}^{L} \gamma_l B_l(z).$$

- Different basis representations exist, but we focus on B-splines due to their advantageous numerical properties.

- The model can now be represented in matrix notation as

$$\boldsymbol{y} = \boldsymbol{B}\boldsymbol{\gamma} + \boldsymbol{\varepsilon},$$

where

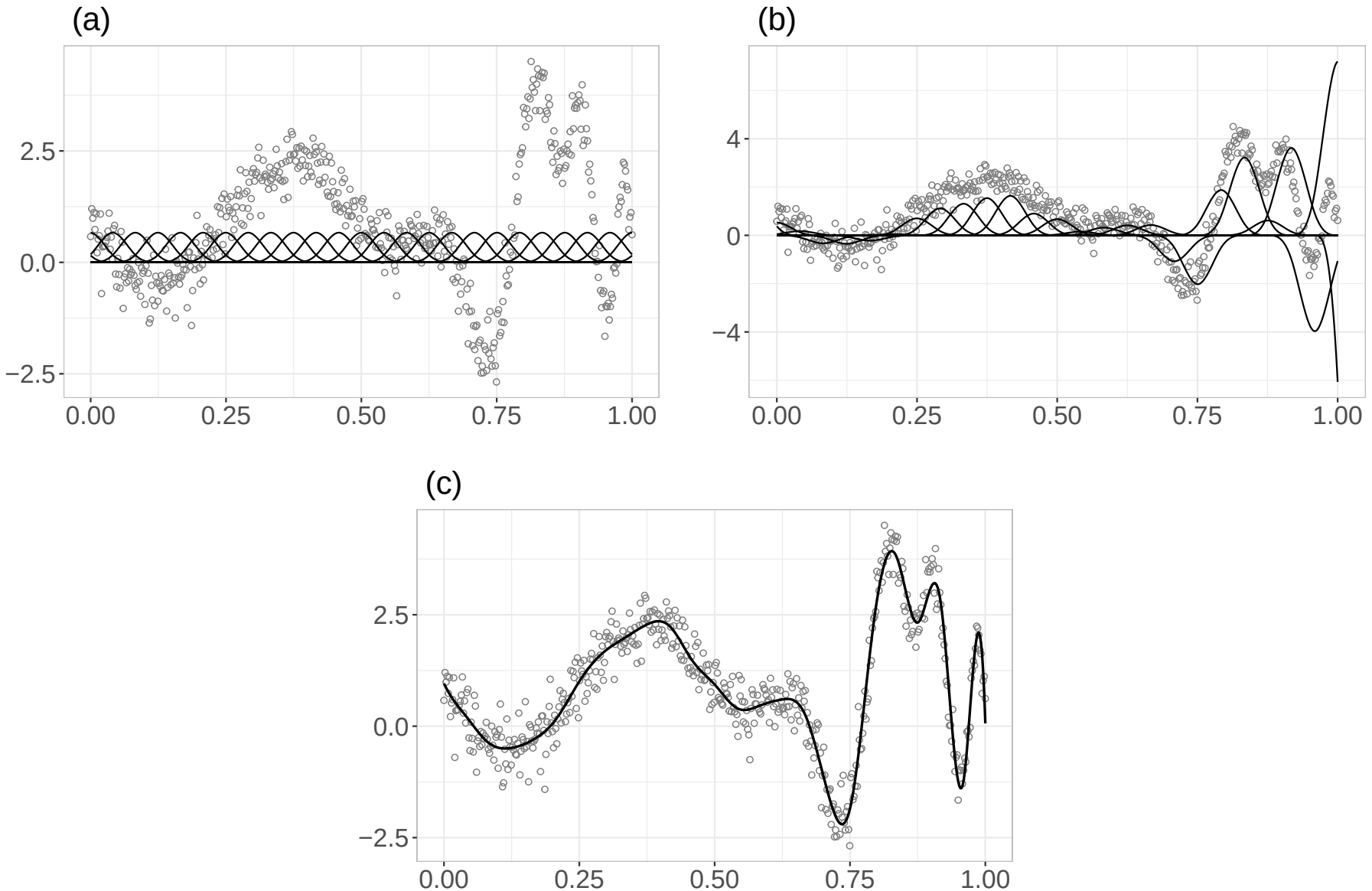
$$\boldsymbol{B} = \begin{pmatrix} B_1(z_1) & \dots & B_L(z_1) \\ \vdots & & \vdots \\ B_1(z_n) & \dots & B_L(z_n) \end{pmatrix}.$$

- Estimate the basis coefficients via least squares as

$$\hat{\boldsymbol{\gamma}} = (\boldsymbol{B}'\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{y}$$

and the function evaluations as

$$\hat{\boldsymbol{s}} = \boldsymbol{B}\hat{\boldsymbol{\gamma}}.$$


(a)
2.5
0.0
−2.5
0.00
0.25
0.50
0.75
1.00
(b)
4
0
−4
0.00
0.25
0.50
0.75
1.00
(c)
2.5
0.0
−2.5
0.00
0.25
0.50
0.75
1.00

Penalized splines:

- To avoid the need to optimize the number and location of the knots for polynomial splines, we
  - approximate $s(z)$ based on a rich spline basis (usually about 20 to 40 basis functions) to ensure enough flexibility of the estimate, and
  - regularize estimation by adding a penalty term to the least squares fit criterion to ensure smoothness of the estimate.
- From a Bayesian perspective, regularisation is achieved by assigning an informative prior that encourages smoothness.

- More precisely, use random walk priors of order k, e.g.

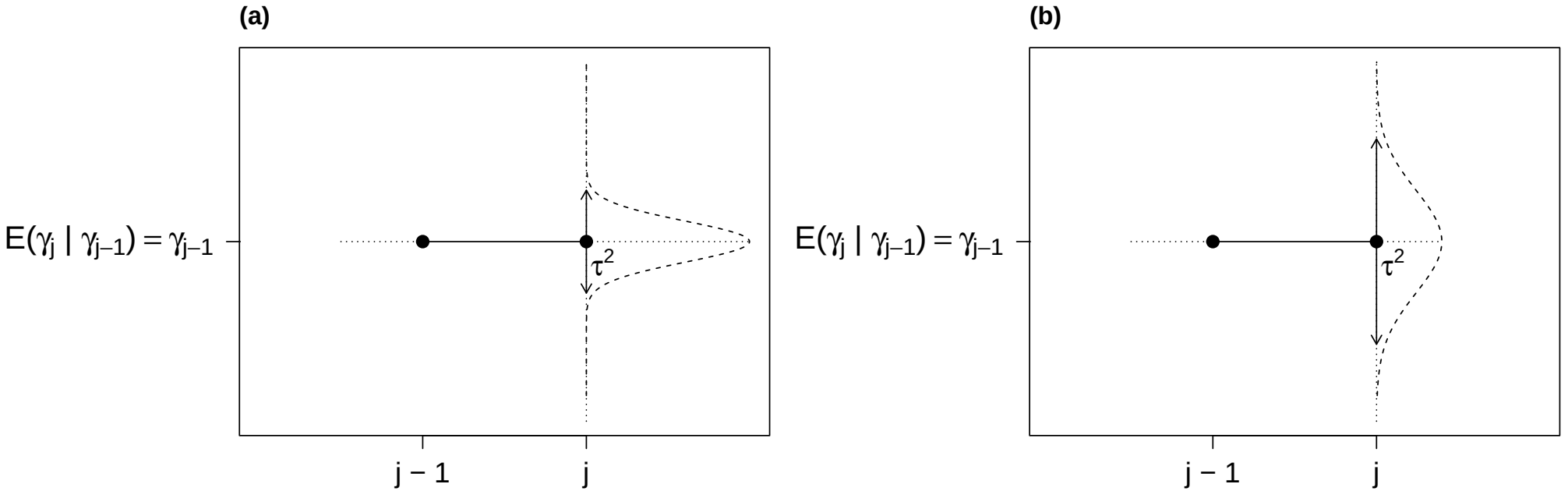
$$\gamma_l = \gamma_{l-1} + u_l, \quad u_l \sim \mathrm{N}(0, \tau^2), \quad l = 2, \ldots, L,$$

  for a first order random walk and

$$\gamma_l = 2\gamma_{l-1} - \gamma_{l-2} + u_l, \quad u_l \sim \mathrm{N}(0, \tau^2), \quad l = 3, \ldots, L,$$

  for a second order random walk with noninformative priors $f(\gamma_1) \propto const$ and $f(\gamma_1, \gamma_2) \propto const$ for initial values.

- The random walk variance τ^2 determines the impact of the prior on the posterior



- The random walk assumptions imply a (partially improper) multivariate Gaussian prior

$$f(\boldsymbol{\gamma}|\tau^2) \propto \left(\frac{1}{\tau^2}\right)^{(L-k)/2} \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\gamma}'\boldsymbol{K}\boldsymbol{\gamma}\right),$$

  where k is the order of the random walk and $\boldsymbol{K}$ is the prior precision matrix.

- Penalized splines can be modified, for example
  - to obtain cyclic effects when smoothing over temporal domains,
  - to add shape constraints such as monotonicity or concavity, or
  - to make the amount of smoothness adaptive to the covariate space.
- They can also be extended to bivariate (or higher order) smoothing.

3.3 Generic Basis Function Framework

- Penalized splines are one representative for a variety of effects that can be cast into a generic basis function framework.
- Let $\boldsymbol{\nu}_i$ denote some generic type of covariate information and assume

$$s(\boldsymbol{\nu}_i) = \sum_{l=1}^{L} \gamma_l B_l(\boldsymbol{\nu}_i)$$

  with L basis functions $B_l(\boldsymbol{\nu}_i)$.
- The vector of function evaluations $\boldsymbol{s} = (s(\boldsymbol{x}_1), \ldots, s(\boldsymbol{x}_n))'$ at the observed covariate values is then given by

$$\boldsymbol{s} = \boldsymbol{B}\boldsymbol{\gamma},$$

  where $\boldsymbol{B}$ is the design matrix obtained from the basis function evaluations and $\boldsymbol{\gamma}$ is the corresponding vector of basis coefficients.

- To regularize estimation, assign the prior

$$f(\boldsymbol{\gamma}|\tau^2) \propto \left(\frac{1}{\tau^2}\right)^{0.5\,\mathrm{rg}(\boldsymbol{K})} \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\gamma}'\boldsymbol{K}\boldsymbol{\gamma}\right)$$

  with positive semidefinite prior precision matrix $\boldsymbol{K}$ and prior variance parameter τ^2.

- An effect is then characterized by the chosen basis functions $B_l(\boldsymbol{\nu})$, the structure of the precision matrix $\boldsymbol{K}$ and the hyperprior assumed for τ^2

- For the latter, it is common to use conjugate inverse gamma hyperpriors.

3.4 Special Cases

Spatial effects for regional data:

- Each observation i is assumed to belong to one of the spatial regions represented by a spatial indicator $r_i \in \{1, \ldots, L\}$.
- Assign separate regression coefficients γ_l, $l = 1, \ldots, L$, to each of the L regions.
- The spatial effect $\gamma_{r_i} = s(r_i)$ of an individual observation i collected in region r_i can then be expressed as

$$s(r_i) = \sum_{l=1}^{L} \gamma_l B_l(r_i),$$

where

$$B_l(r_i) = \begin{cases} 1 & \text{if } r_i = l \\ 0 & \text{otherwise.} \end{cases}$$

- In matrix notation this yields the $(n \times L)$ design matrix $\boldsymbol{B}$ with entries

$$\boldsymbol{B}[i, l] = \begin{cases} 1 & \text{if } r_i = l \\ 0 & \text{otherwise} \end{cases}$$

  and the complete vector of spatial effects is given by $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_L)'$.

- Assume a Gaussian Markov random field prior, where the conditional distribution of γ_l given all the neighboring effects is specified as

$$\gamma_l \,|\, \gamma_r, r \neq l \sim \mathrm{N}\left(\frac{1}{|N(l)|} \sum_{r:r\sim l} \gamma_r, \frac{\tau^2}{|N(l)|}\right),$$

  where $l \sim r$ indicates that l and r are neighbors and $|N(l)|$ is the number of neighbors of region l.

- This implies that
  - the prior expectation for the spatial effect in region l is given by the average of all spatial effects of neighboring regions,
  - the effect in region l is conditionally independent of all non-neighbors, and
  - the variance of the conditional prior distribution in region l is inversely proportional to the number of neighbors.

- The conditional distributions yield a multivariate Gaussian joint distribution for $\boldsymbol{\gamma}$ given by

$$f(\boldsymbol{\gamma} \mid \tau^2) \propto \left(\frac{1}{\tau^2}\right)^{(\text{rg}(\boldsymbol{K}))/2} \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\gamma}'\boldsymbol{K}\boldsymbol{\gamma}\right),$$

  with prior precision matrix

$$\boldsymbol{K}[l, r] = \begin{cases} -1 & l \neq r, l \sim r, \\ 0 & l \neq r, l \nsim r, \\ |N(l)| & l = r, \end{cases}$$

- If each region has at least one neighbor and the map is fully connected, the rank of the spatial adjacency matrix is given by $\text{rg}(\boldsymbol{K}) = L - 1$.

Random effects:

- Assume that the data are grouped into L disjoint sets of observations and define group-specific regression coefficients γ_l, $l = 1, \ldots, L$.
- If group membership is represented by the indicator $g_i \in \{1, \ldots, L\}$, the group-specific effects can be represented as

$$s(g_i) = \sum_{l=1}^{L} \gamma_l B_l(g_i)$$

with indicator basis functions

$$B_l(g_i) = \begin{cases} 1 & \text{if } g_i = l \\ 0 & \text{otherwise.} \end{cases}$$

- In matrix notation, the indicator basis functions imply a dummy-coded design matrix $\boldsymbol{B}$ with elements

$$\boldsymbol{B}[i, l] = \begin{cases} 1 & \text{if } s_i = l \\ 0 & \text{otherwise.} \end{cases}$$

- The assumption of i.i.d. random intercepts translates to $\boldsymbol{\gamma} \sim \mathrm{N}(\mathbf{0}, \tau^2 \boldsymbol{I}_L)$ for the complete vector of random effects with prior precision matrix $\boldsymbol{K} = \boldsymbol{I}_L$.

Other effect types:

- The framework supports a number of further effect types such as
  - Spatial effects $s(x_1, x_2)$ for spatial domains that are continuously indexed by coordinates (x_1, x_2),
  - varying coefficients $\nu_1 s(\nu_2)$ with effect modifier ν_2 and interaction variable ν_1 (e.g. random slopes), or
  - interaction surfaces $s(z_1, z_2)$ with two continuous covariates z_1 and z_2.

3.5 Hyperprior Specifications

Inverse gamma prior:

- The smoothing variance τ^2 determines how strongly estimates $\hat{s}(\boldsymbol{\nu})$ are affected by the smoothness properties induced by the precision matrix $\boldsymbol{K}$.
- Inference for τ^2 is therefore extremely important (similar as determining the regularisation parameter for, e.g., the LASSO).
- Bayesian approach: Assign a hyperprior $f(\tau^2)$ and include τ^2 as a hyperparameter.
- The de facto standard are inverse gamma priors $\tau^2 \sim \text{IG}(a, b)$, $a > 0$, $b > 0$ since this is conjugate to the multivariate normal prior $f(\boldsymbol{\gamma}|\tau^2)$.
- Allows updating τ^2 in a simple Gibbs step without requiring a proposal density and/or acceptance step.

- How should the parameters a and b of the inverse gamma distribution $\tau^2 \sim \text{IG}(a, b)$ be chosen?
- Limiting cases:
  - $a \to 0$, $b \to 0$ leads to a flat prior for $\log(\tau^2)$ (this is also Jeffreys' prior).
  - $a = 1$, $b \to 0$ leads to a flat prior for the precision $1/\tau^2$.
  - $a = -1$, $b = 0$ leads to a flat prior for τ^2.
  - $a = -0.5$, $b = 0$ leads to a flat prior for the standard deviation τ.

- In practice, the limit of $a \to 0$ / $b \to 0$ is often approximated by a small constant ϵ and different values are tried out to study prior sensitivity.
- $a < 0$ leads to improper priors (prior does not integrate to one), such that propriety of the posterior has to be ensured.
- Especially in situations with little information per parameter (e.g. random effects models with small groups), there has been considerable debate about prior sensitivity and the suitability of the inverse gamma distribution.

Alternatives to the conjugate inverse gamma:

- Alternative prior distributions that have been considered:
  - Half-normal $\tau^2 \sim \mathrm{HN}(0, \theta^2)$
  - Half-Cauchy $\tau^2 \sim \mathrm{HC}(0, \theta^2)$
  - Uniform $\tau^2 \sim \mathrm{U}(0, \theta)$
- The hyperparameter θ has to be chosen with respect to the prior beliefs.

Scale-dependent hyperpriors

- Goal: Determine a prior based on a simple set of principles and derive an intuitive way of eliciting hyperparameters.
- Principle 1: Occam's Razor.
  - The hyperprior should invoke the principle of parsimony.
  - Simple base model for each effect is preferred unless the data provide convincing evidence for more complex modelling.
  - For structured additive regression terms, $\tau^2 \to 0$ results in the base model $f_b(\boldsymbol{\gamma}|\tau^2 = 0)$ determined by the nullspace of $\boldsymbol{K}$.

- Principle 2: Measure of Complexity.
  - The increased complexity is measured by the Kullback-Leibler divergence

$$\text{KLD}(f||f_b) = 2\int f(u)\log\left(\frac{f(u)}{f_b(u)}\right)\mathrm{d}u$$

    for the base model f_b and an alternative flexible model f.
  - Gives a measure of the information loss when the base model is used to approximate the more flexible models.
  - Define

$$d(f||f_b) = \sqrt{2\text{KLD}(f||f_b)}$$

    as the unidirectional 'distance' from the flexible model to the base model.

- Principle 3: Constant Rate Penalisation.
  - Constant rate penalisation implies an exponential prior on the distance scale, i.e.

$$f_d(d) = \lambda \exp(-\lambda d), \quad \lambda > 0$$

    with mode at $d = 0$.
  - Constant rate of decay in the distance prior from f_b to stronger deviations from f_b.
  - λ determines the rate of penalisation.

- The change of variable theorem gives

$$f(\tau^2) = \lambda \exp(-\lambda d(\tau^2)) \left| \frac{\partial d(\tau^2)}{\partial \tau^2} \right| \text{ with } d(\tau^2) = \sqrt{2\text{KLD}}.$$

- For structured additive regression terms, this induces a Weibull prior $\tau^2 \sim \text{We}(0.5, \theta)$, i.e.

$$f(\tau^2|\theta) = \frac{1}{2\theta} \left(\frac{\tau^2}{\theta} \right)^{-1/2} \exp\left(- \left(\frac{\tau^2}{\theta} \right)^{1/2} \right).$$

- By construction, the scale-dependent prior is invariant under transformations, i.e. we obtain equivalent priors for τ, τ^2, $1/\tau^2$, etc.

Hyperprior elicitation

- The decay rate $\exp(-\lambda)$ can be controlled by the condition

$$\mathbb{P}(q(\tau^2) \leq c) = 1 - \alpha$$

  for a suitable transformation $q(\cdot)$ of τ^2 and user-defined values c and α.

- Alternative: Prior knowledge about the scale of functional effects $s(\boldsymbol{\nu})$ allows to specify a certain interval with high marginal probability:

$$\mathbb{P}(|s(\boldsymbol{\nu})| \leq c;\ \forall \boldsymbol{\nu} \in \mathcal{D}) \geq 1 - \alpha.$$

- The marginal density of $s(\boldsymbol{\nu}) = \boldsymbol{b}'\boldsymbol{\gamma}$ is

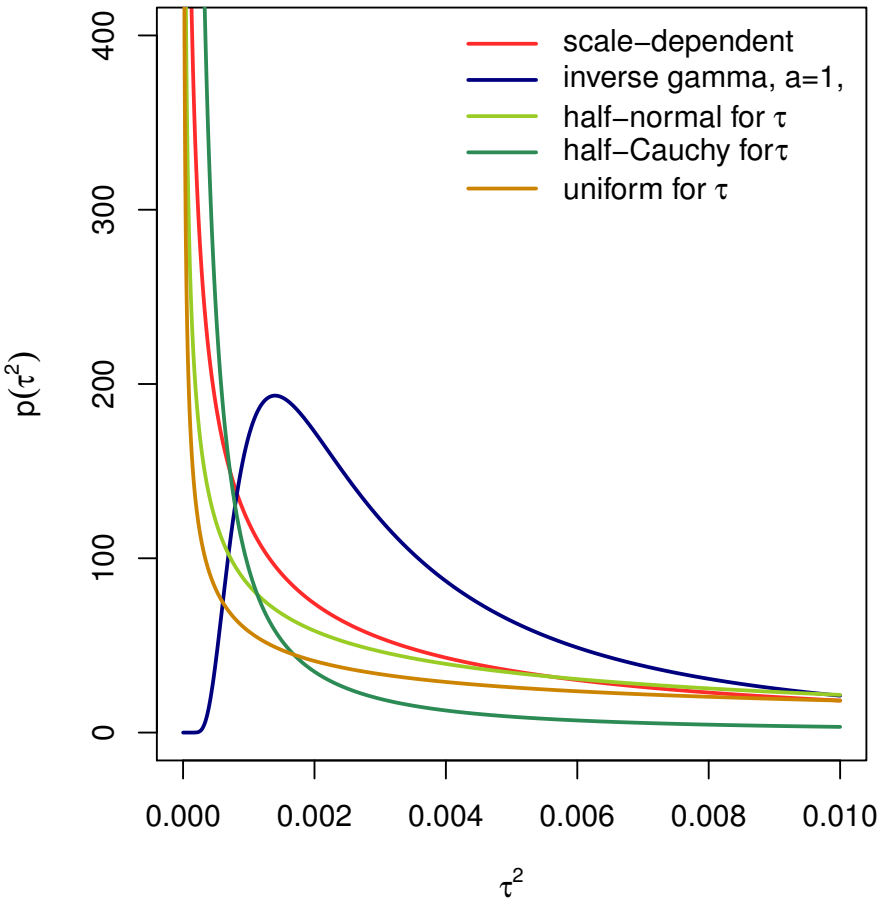
$$f(\boldsymbol{b}'\boldsymbol{\gamma}) = \int_0^\infty f(\boldsymbol{b}'\boldsymbol{\gamma}, \tau^2)\mathrm{d}\tau^2 = \int_0^\infty f(\boldsymbol{b}'\boldsymbol{\gamma}|\tau^2) f(\tau^2|\theta)\mathrm{d}\tau^2$$

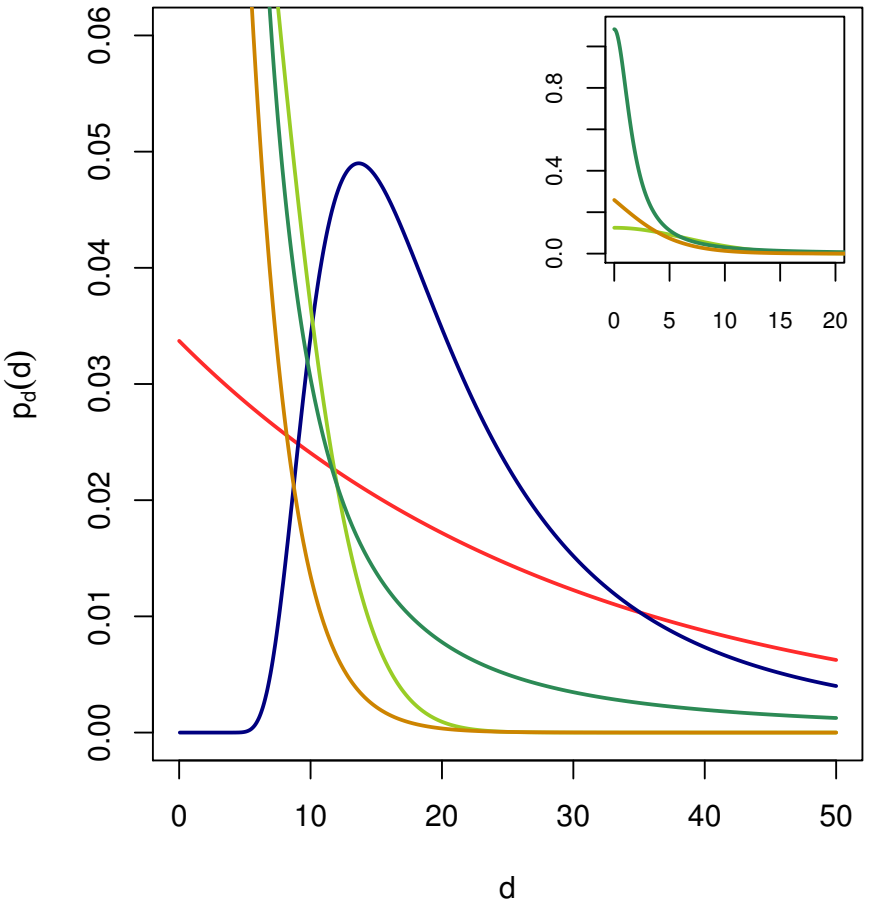
- θ can be chosen such that

$$\left(1 - \int_{-c}^{c} \int_0^\infty f_{\boldsymbol{b}'\boldsymbol{\gamma}}(u|\tau^2) f(\tau^2|\theta)\mathrm{d}\tau^2 \mathrm{d}u\right) = \alpha.$$

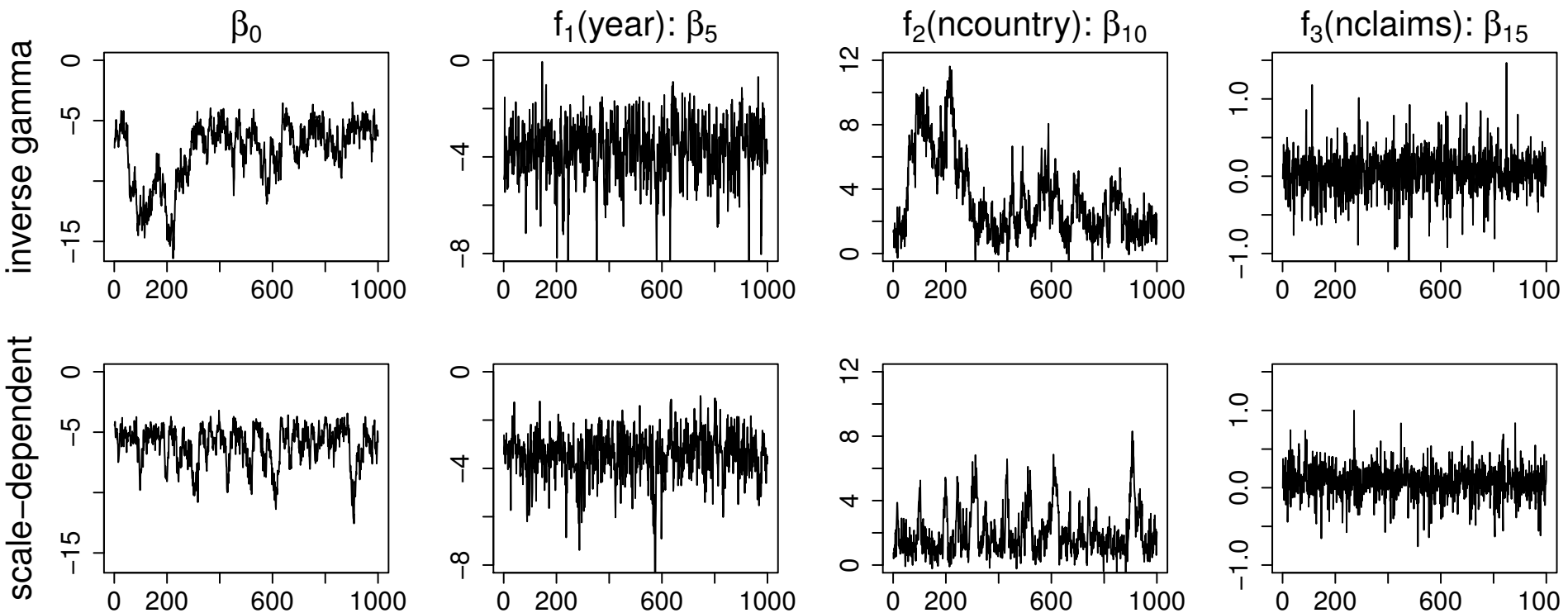
- The integral can be approximated by Monte Carlo sampling from the prior.

- The scaling criterion can also be employed for other prior structures.
- Comparison of resulting priors:





- The prior can make a difference:



3.6 Posterior Inference

- For Gaussian responses and inverse gamma hyperpriors, a Gibbs sampler can be derived.
- For other response types or non-conjugate priors, more general steps such as Metropolis-Hastings have to be included.
- Bayesian additive models can be conveniently expressed in a hierarchical fashion that can be exploited in the implementation of MCMC iterations.

- Stylized model hierarchy:

$$
\begin{aligned}
y \mid \boldsymbol{\nu}, \sigma^2 &\sim \mathrm{N}(\mu(\boldsymbol{\nu}), \sigma^2), \\
\mu(\boldsymbol{\nu}) &= s_1(\boldsymbol{\nu}) + \cdots + s_J(\boldsymbol{\nu}), \\
s_j(\boldsymbol{\nu}) &= \sum_{l=1}^{L} B_{lj}(\boldsymbol{\nu}) \gamma_{lj} \\
\boldsymbol{\gamma}_j \mid \tau_j^2 &\sim \mathrm{N}(0, \tau_j^2 \boldsymbol{K}_j^-), \\
\tau_j^2 &\sim \mathrm{IG}(a, b).
\end{aligned}
$$

Updates for the regression coefficients:

- The full conditional for $\boldsymbol{\gamma}_j$ is a multivariate normal:

$$\boldsymbol{\gamma}_j|\cdot \sim \mathrm{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$

- The parameters of this are given by

$$\boldsymbol{\mu}_j = \left(\boldsymbol{B}_j'\boldsymbol{B}_j + \frac{\sigma^2}{\tau_j^2}\boldsymbol{K}_j\right)^{-1} \boldsymbol{B}_j'\left(\boldsymbol{y} - \boldsymbol{\eta}_{-j}\right)$$

and

$$\boldsymbol{\Sigma}_j = \sigma^2\left(\boldsymbol{B}_j'\boldsymbol{B}_j + \frac{\sigma^2}{\tau_j^2}\boldsymbol{K}_j\right)^{-1}$$

where $\boldsymbol{\eta}_{-j}$ is the predictor without $\boldsymbol{\gamma}_j$.

Efficient sampling from multivariate normals:

- Naively working with multivariate normal distributions can lead to large computation times.
- Particular bottleneck: The covariance matrix $\boldsymbol{\Sigma}_j$ (and its inverse).
- Efficient simulation from $\mathrm{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j^{-1})$ avoiding the explicit computation of $\boldsymbol{\Sigma}_j$:
  - Compute the Cholesky factorisation $\boldsymbol{\Sigma}_j^{-1} = \boldsymbol{L}\boldsymbol{L}'$.
  - Sample $\boldsymbol{z} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I})$.
  - Solve $\boldsymbol{L}'\boldsymbol{v} = \boldsymbol{z}$.
  - Return $\boldsymbol{\mu}_j + \boldsymbol{v}$.
- Can be extended to handle also the determination of $\boldsymbol{\mu}_j$ without computing $\boldsymbol{\Sigma}_j^{-1}$
- In many cases, $\boldsymbol{\Sigma}_j^{-1}$ has a sparse matrix structure that can be exploited for further efficiency gains.

Updates for the variance parameters:

- Full conditional for the smoothing variances:

$$\tau_j^2|\cdot \sim \text{IG}(\tilde{a}_j, \tilde{b}_j)$$

with

$$\tilde{a}_j = a_j + \frac{1}{2}\text{rg}(\boldsymbol{K}_j), \qquad \tilde{b}_j = b_j + \frac{1}{2}\boldsymbol{\gamma}_j'\boldsymbol{K}_j\boldsymbol{\gamma}_j.$$

- Full conditional for the residual variance:

$$\sigma^2|\cdot \sim \text{IG}(\tilde{a}, \tilde{b})$$

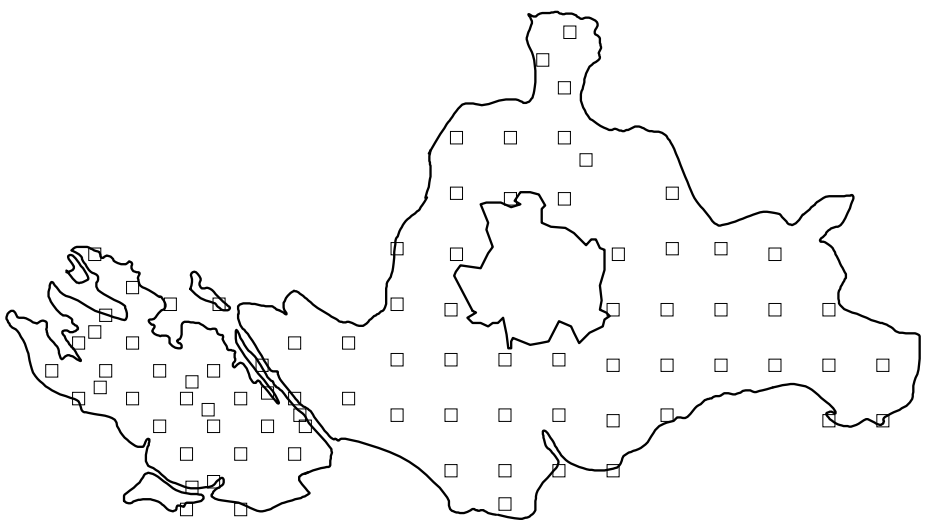
with

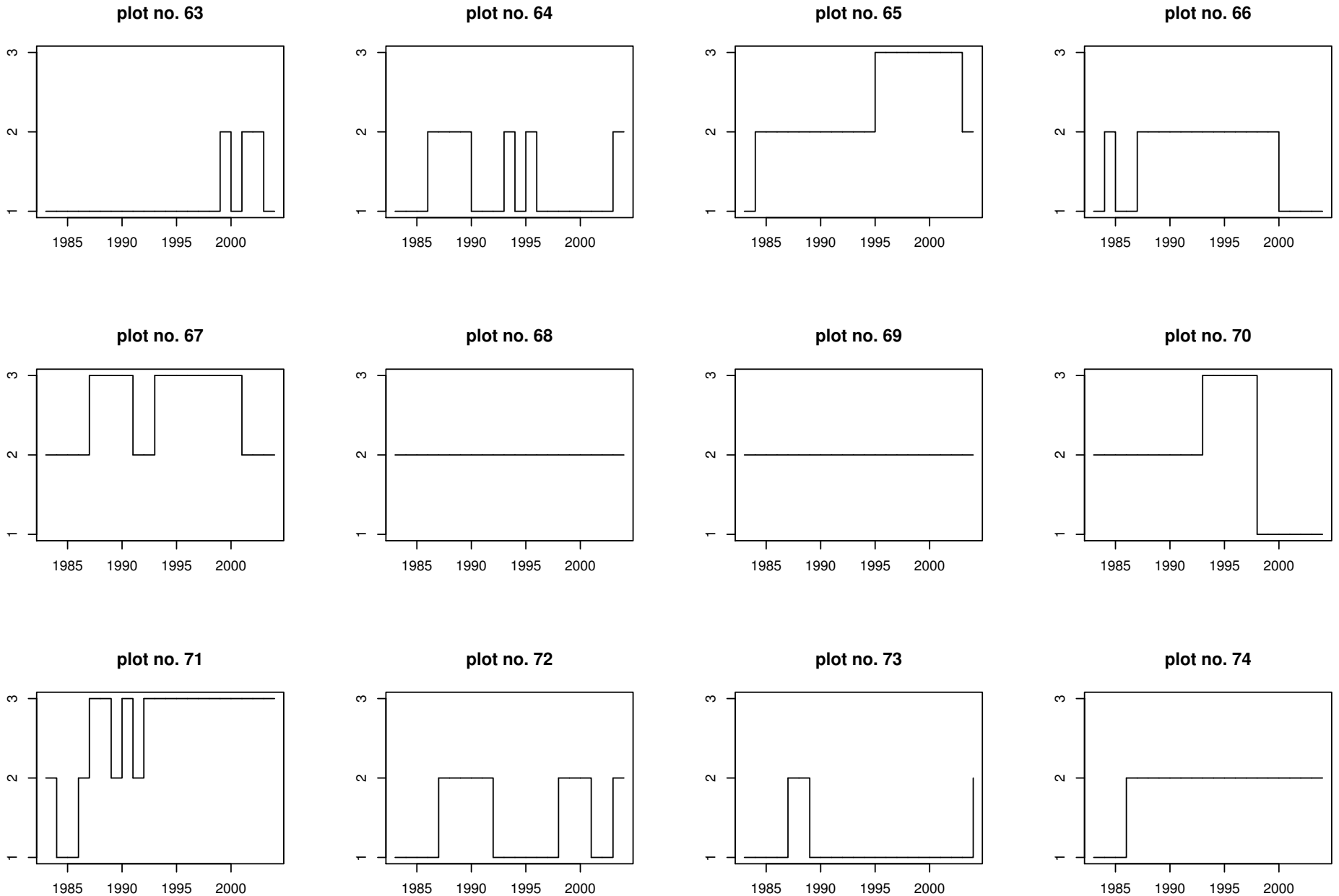
$$\tilde{a} = a + \frac{n}{2}, \qquad \tilde{b} = b + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{\eta})'(\boldsymbol{y} - \boldsymbol{\eta}).$$

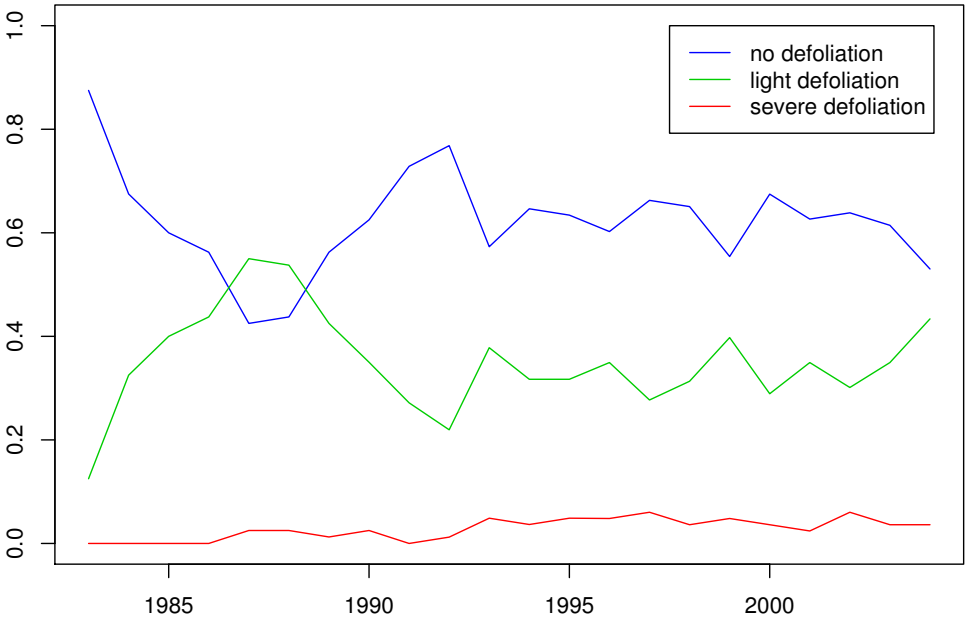
3.7 Case Study on Forest Health

- Aim of the study: Identify factors influencing the health status of trees.
- Database: Yearly visual forest health inventories carried out from 1983 to 2004 in a northern Bavarian forest district.
- 83 observation plots of beeches within a 15 km times 10 km area.
- Response: defoliation degree at plot i in year t, measured in three ordered categories:

$y_{it} = 1$ no defoliation,
$y_{it} = 2$ defoliation 25% or less,
$y_{it} = 3$ defoliation above 25%.

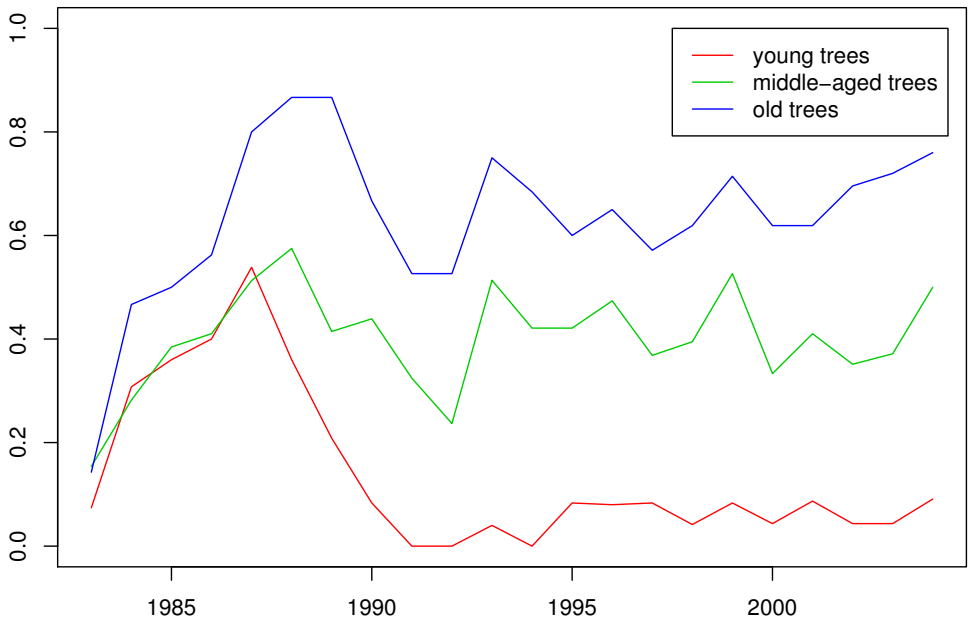


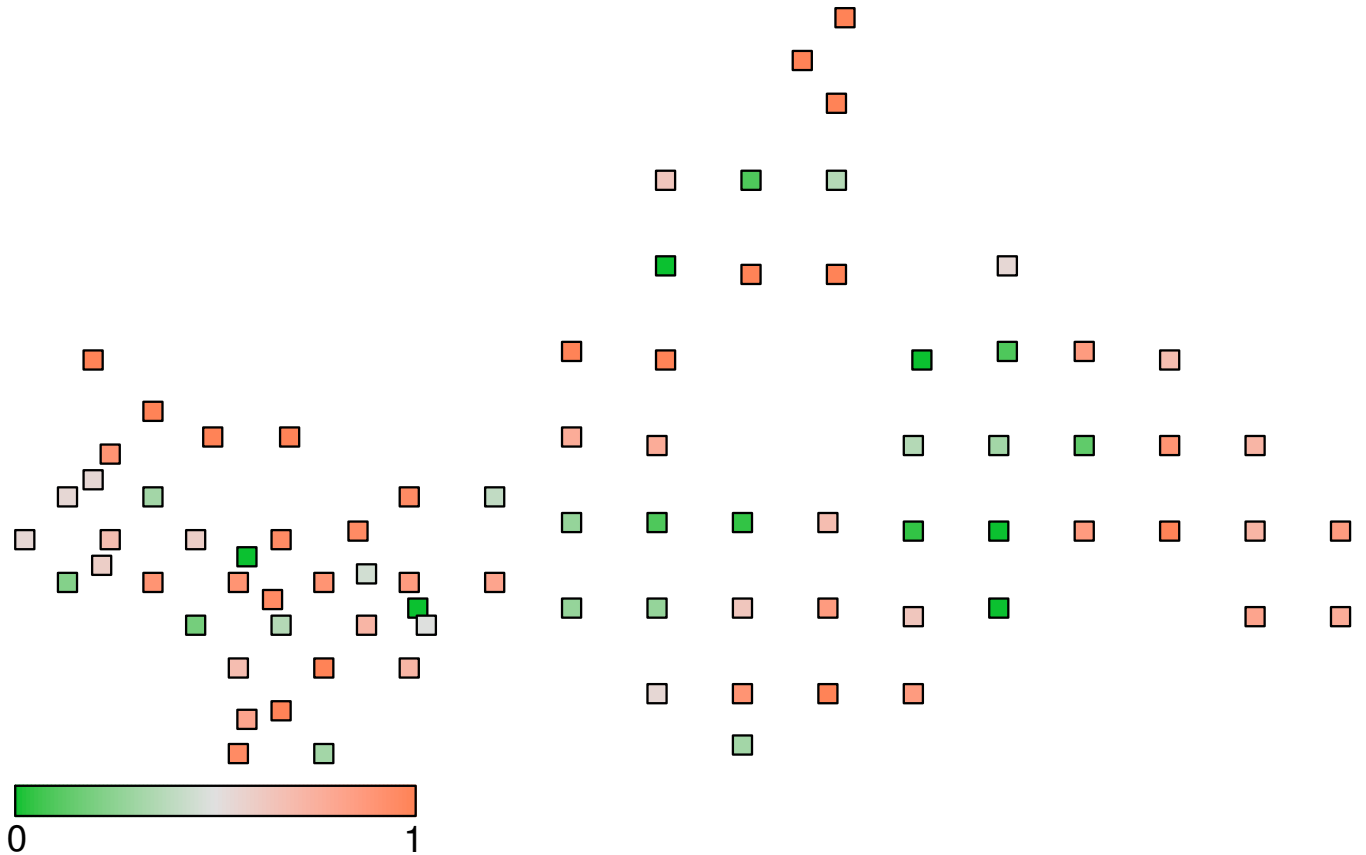

plot no. 63
plot no. 64
plot no. 65
plot no. 66
plot no. 67
plot no. 68
plot no. 69
plot no. 70
plot no. 71
plot no. 72
plot no. 73
plot no. 74
1
2
3
1985
1990
1995
2000



Empirical time trends.

Trends for different ages.





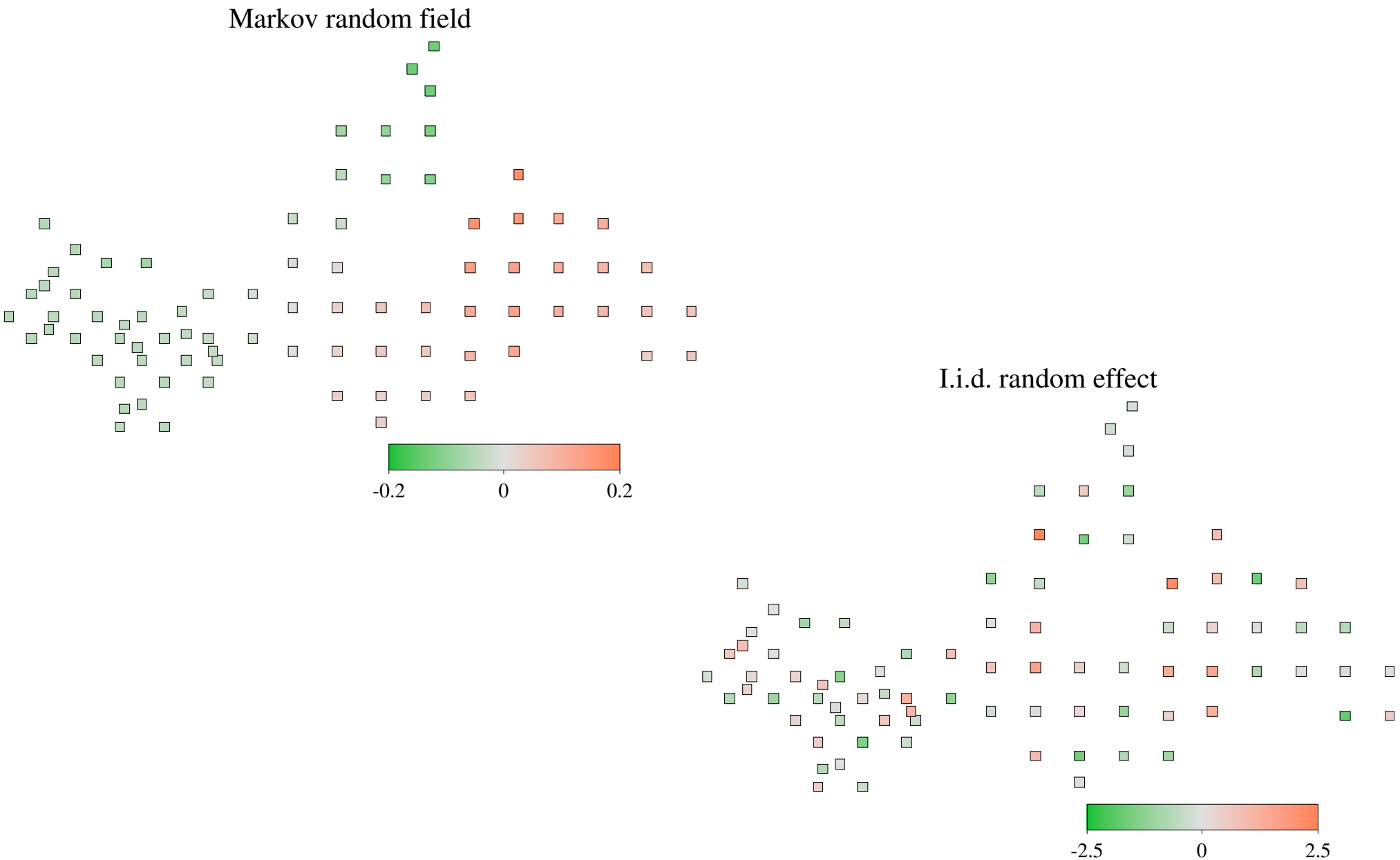
Percentage of time points for which an observation plot was classified to be defoliated.

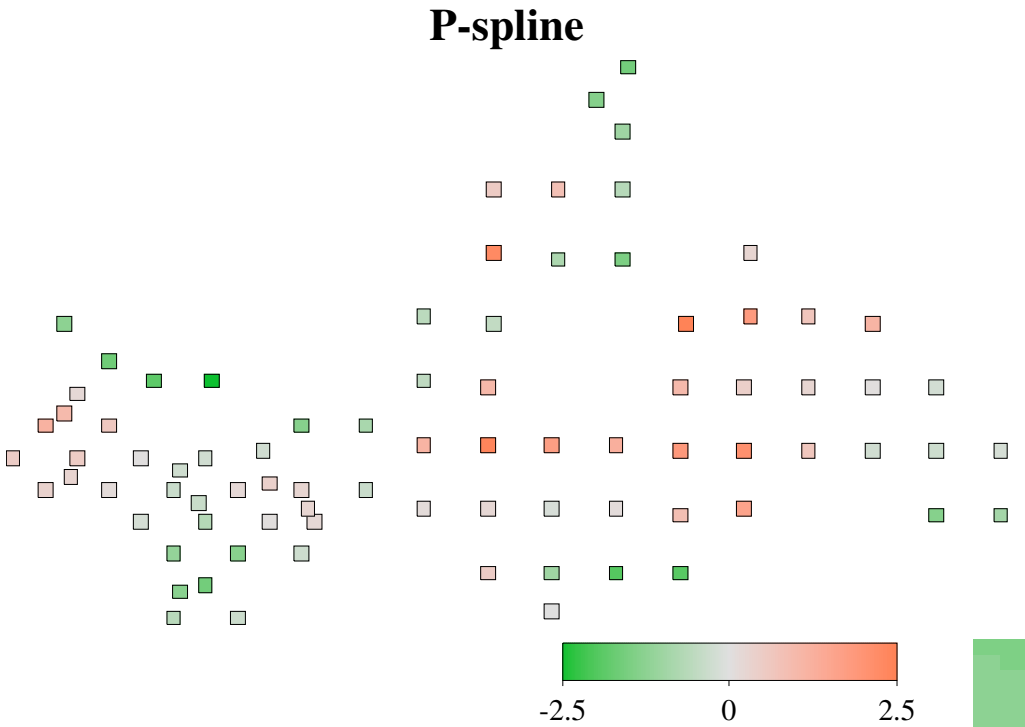
- We need a model that can simultaneously deal with the following issues:
  - A spatially aligned set of time series.

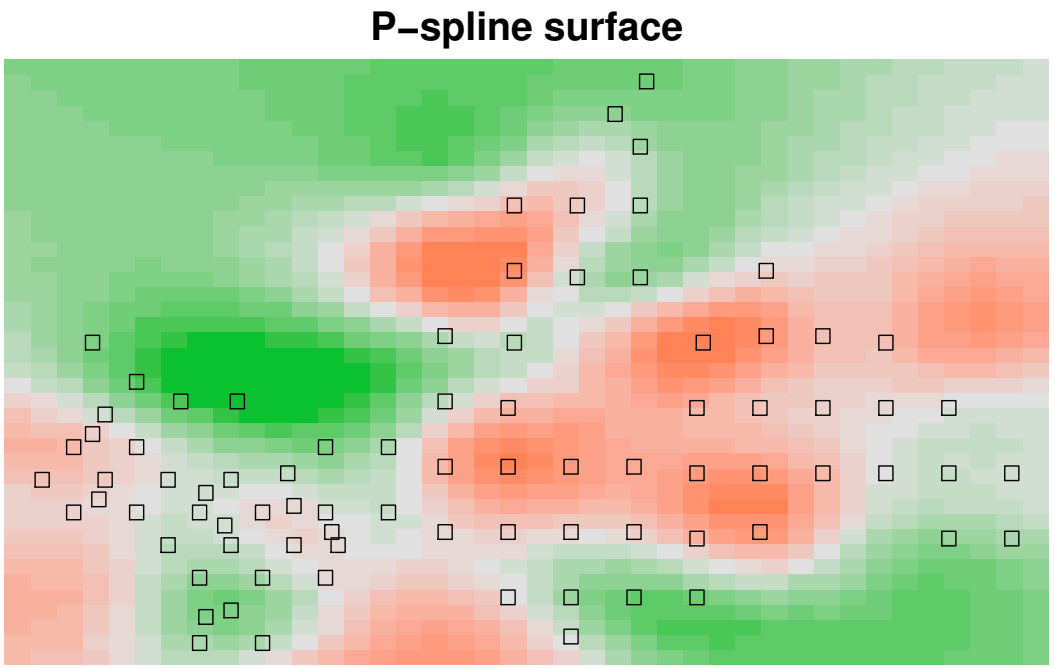
    $\Rightarrow$ Both spatial and temporal correlations have to be considered.
  - Decide whether unobserved heterogeneity is spatially structured or not.
  - Nonlinear effects of continuous covariates (e.g. age).
  - A possibly time-varying effect of age (i.e. an interaction between age and calendar time).

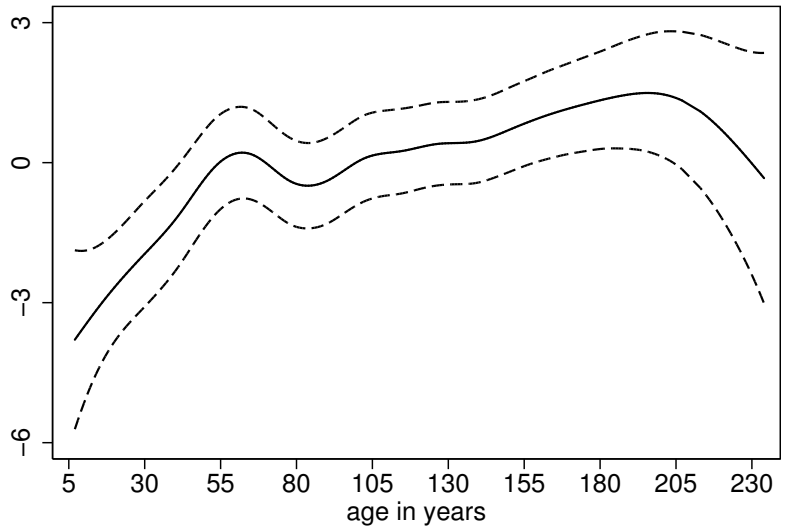
- Predictor specification:

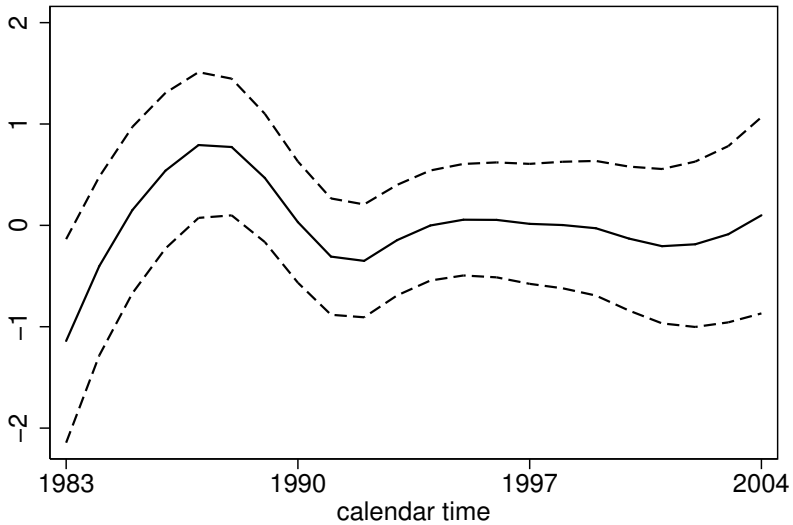
$$
\begin{aligned}
\eta_{it} = \quad & s_1(age_{it}) && \text{nonlinear effects of age,} \\
& + s_2(inc_i) && \text{inclination of slope, and} \\
& + s_3(can_{it}) && \text{canopy density.} \\
& + s_{time}(t) && \text{nonlinear time trend.} \\
& + s_4(t, age_{it}) && \text{interaction between age and calendar time.} \\
& + s_{spat}(p_i) && \text{structured and} \\
& + b_i && \text{unstructured spatial effects.} \\
& + \boldsymbol{z}_{it}'\boldsymbol{\gamma} && \text{usual parametric effects.}
\end{aligned}
$$

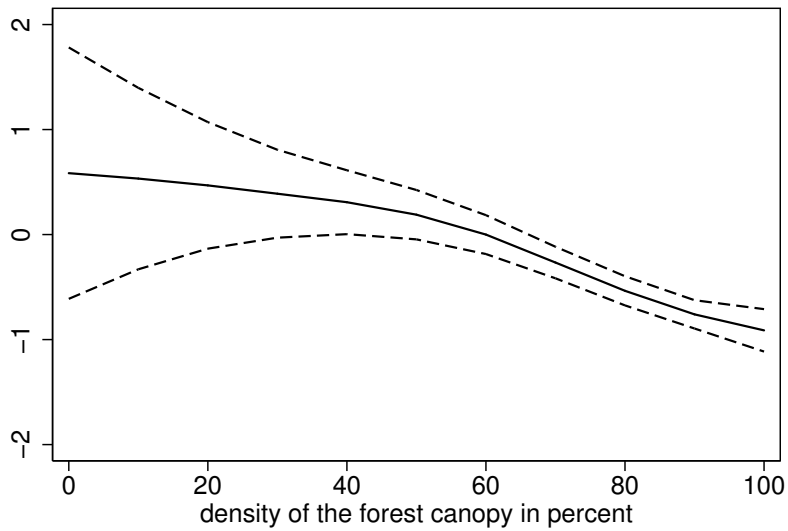

Markov random field
-0.2
0
0.2
I.i.d. random effect
-2.5
0
2.5

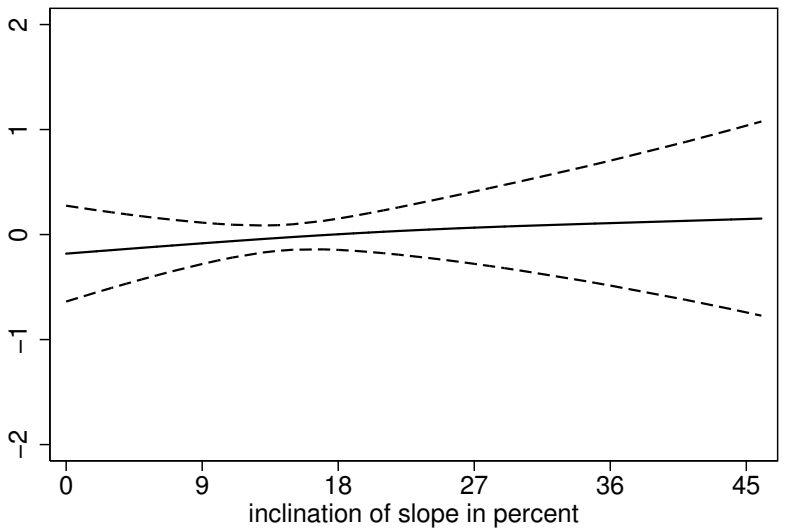

P-spline
-2.5
0
2.5

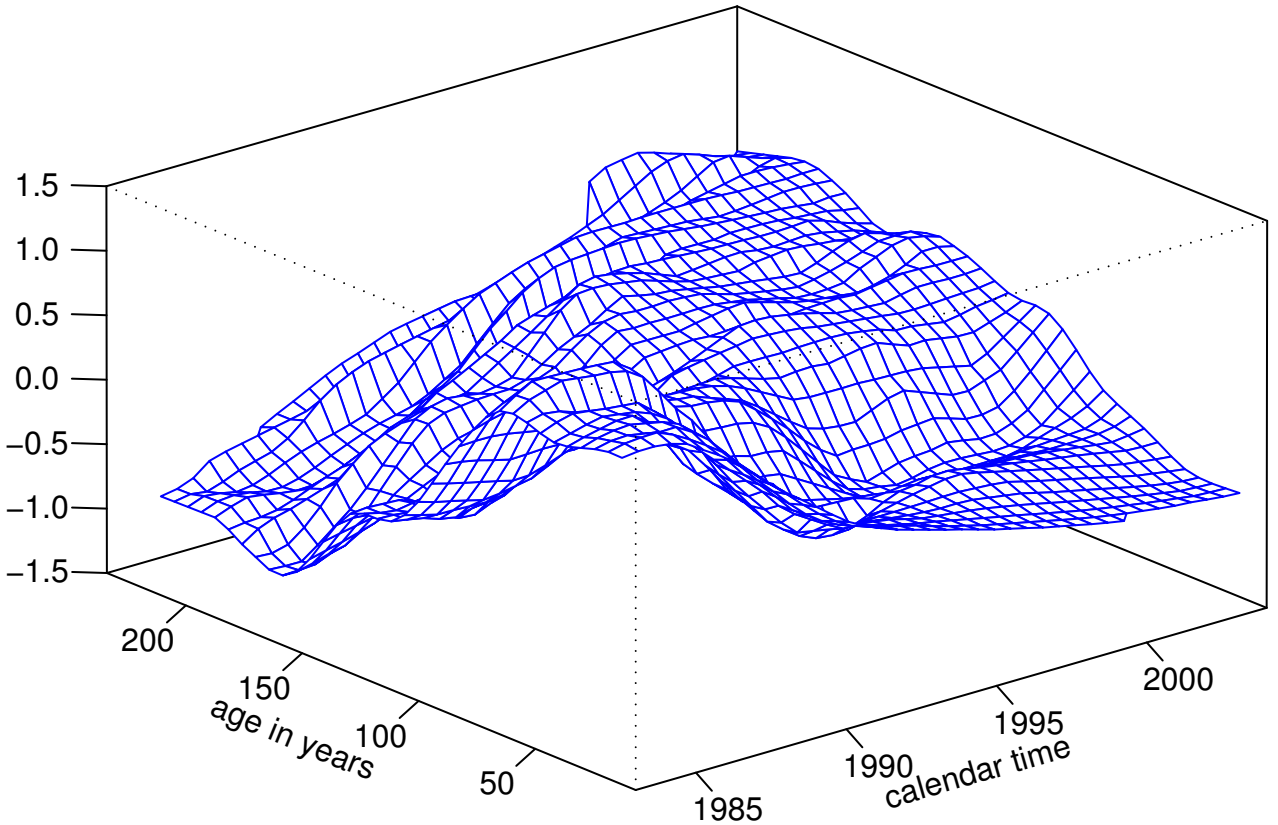

P−spline surface


3
0
−3
−6
5
30
55
80
105
130
155
180
205
230
age in years


2
1
0
−1
−2
1983
1990
1997
2004
calendar time


2
1
0
−1
−2
0
20
40
60
80
100
density of the forest canopy in percent


2
1
0
−1
−2
0
9
18
27
36
45
inclination of slope in percent


1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
200
150
100
50
age in years
1985
1990
1995
2000
calendar time

4 Bayesian Distributional Regression

4.1 Introduction

- Classical regression has focused on relating the conditional mean of a response y_i to covariate information x_i for observations $(x_1, y_1), \ldots, (x_n, y_n)$.

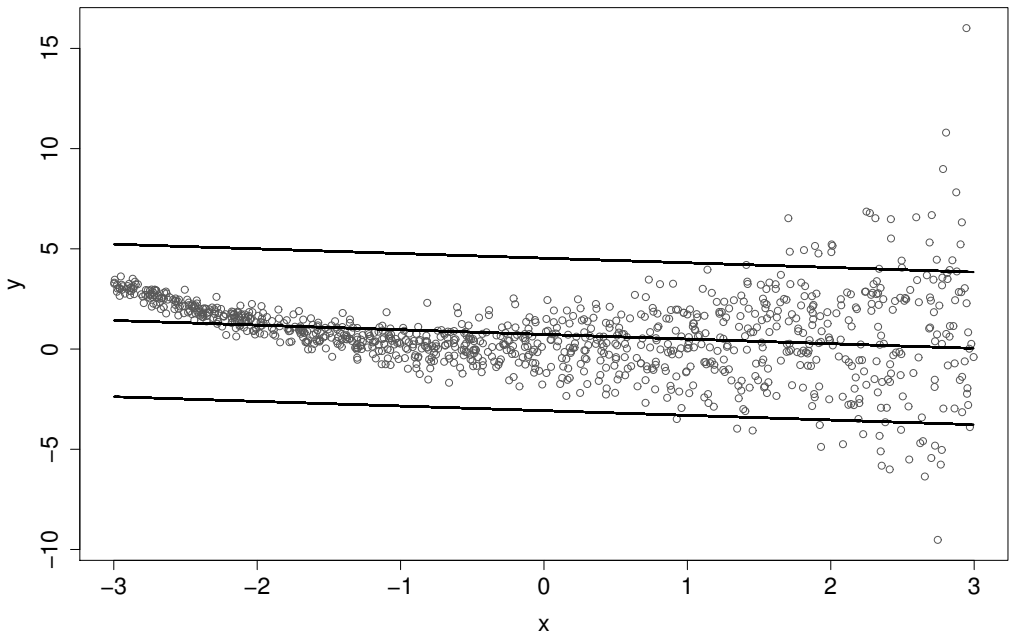
- Linear model:

$$y_i = \beta_0 + \beta x_i + \varepsilon_i$$

with ε_i i.i.d. $\mathrm{N}\left(0, \sigma^2\right)$

$$\Rightarrow \quad \mathrm{E}(y_i|x_i) = \mu_i(x_i) = \beta_0 + \beta x_i$$

$$\Rightarrow \quad \mathrm{Var}(y_i|x_i) = \sigma^2$$



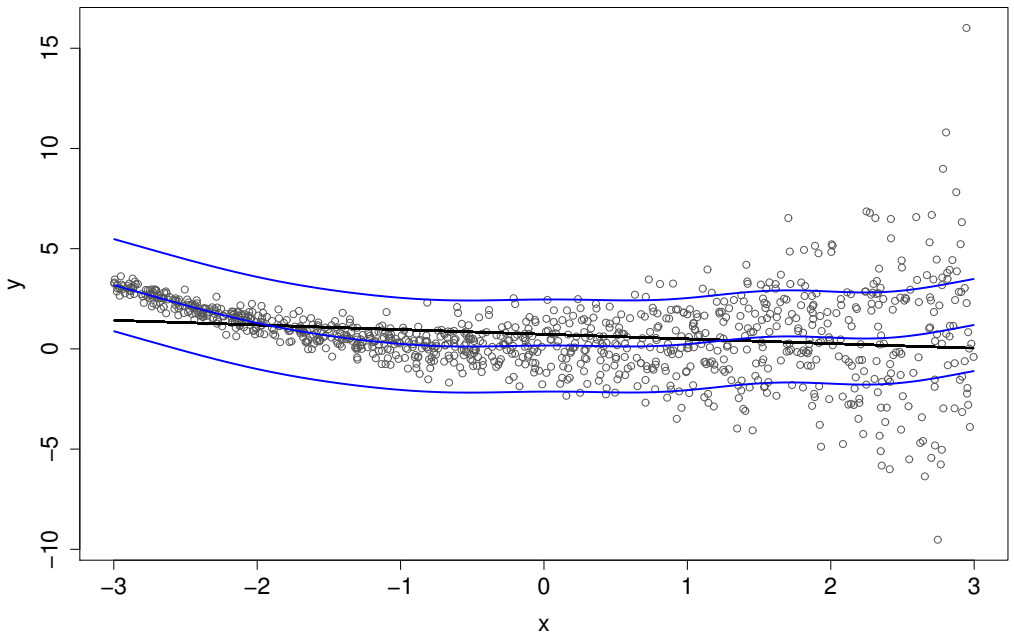
- Classical regression has focused on relating the conditional mean of a response y_i to covariate information x_i for observations $(x_1, y_1), \ldots, (x_n, y_n)$.

- Nonparametric model:

$$\mathrm{E}(y_i|x_i) = \mu_i(x_i) = \beta_0 + f(x_i)$$

with $y_i|x_i \sim \mathrm{N}\left(\mu_i(x_i), \sigma^2\right)$

σ^2 fixed

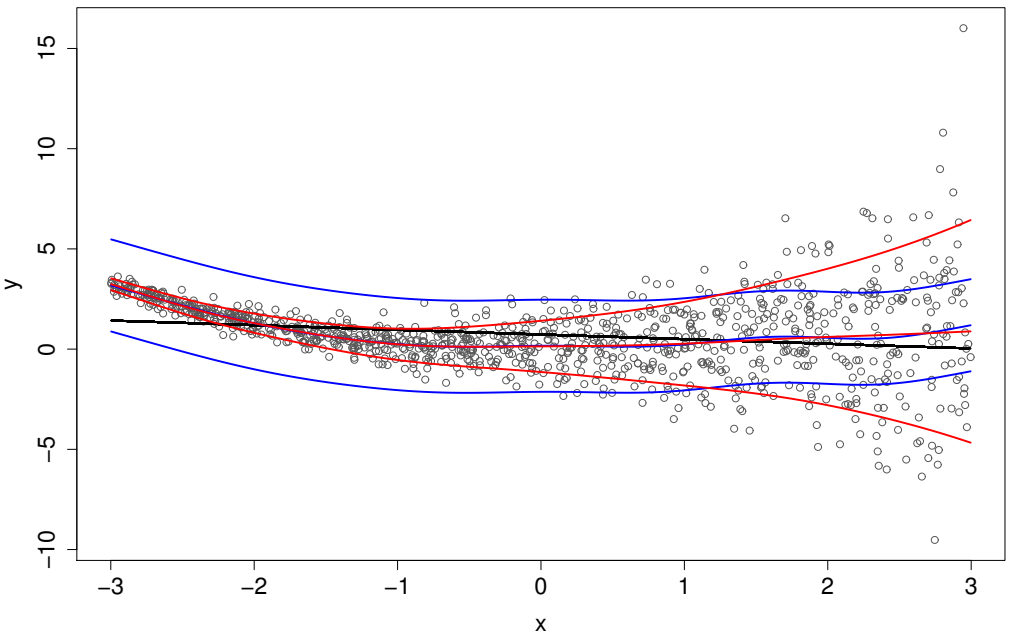


- Nonparametric model for location and scale:

$$\mathrm{E}(y_i|x_i) = \mu_i(x_i) = \beta_0^\mu + f^\mu(x_i)$$

$$\mathrm{Var}(y_i|x_i) = \sigma_i^2(x_i)$$
$$= \exp\left(\beta_0^{\sigma^2} + f^{\sigma^2}(x_i)\right)$$

with $y_i|x_i \sim \mathrm{N}\left(\mu_i(x_i), \sigma_i^2(x_i)\right)$



- So why should we focus on the mean alone if this gives only such an incomplete picture about the (conditional) distribution of the response y_i?

4.2 Bayesian Distributional Regression

Model Setup:

- Assume a parametric specification for the conditional distribution of the responses y_i given covariates $\boldsymbol{\nu}_i$ such that

$$f(y_i|\boldsymbol{\nu}_i) = f(y_i|\boldsymbol{\vartheta}(\boldsymbol{\nu}_i)),$$

  where $\boldsymbol{\vartheta}(\boldsymbol{\nu}_i) = (\vartheta_1(\boldsymbol{\nu}_i), \ldots, \vartheta_K(\boldsymbol{\nu}_i))^\top$ is a K-dimensional vector of distributional parameters.

- Each parameter $\vartheta_k(\boldsymbol{\nu}_i)$ is linked to a regression predictor $\eta_k(\boldsymbol{\nu}_i)$ based on a response function $h_k(\cdot)$:

$$\vartheta_k(\boldsymbol{\nu}_i) = h_k(\eta_k(\boldsymbol{\nu}_i)) \quad \text{and} \quad \eta_k(\boldsymbol{\nu}_i) = h_k^{-1}(\vartheta_k(\boldsymbol{\nu}_i)).$$

Additive predictors:

- Flexibility attained by specifying additive predictors

$$\eta_k(\boldsymbol{\nu}_i) = s_{k1}(\boldsymbol{\nu}_i) + \ldots + s_{kJ_k}(\boldsymbol{\nu}_i)$$

  for each parameter of interest, comprising

  – flexible nonlinear effects of continuous covariates where the amount of smoothness is determined based on the data.

  – spatial effects to capture unobserved spatial heterogeneity and spatial correlations.

  – interaction terms such as varying coefficients or interaction surfaces.

  – cluster-specific random effects.

Examples:

- The normal location-scale model

$$y_i|\boldsymbol{\nu}_i \sim \mathrm{N}(\mu(\boldsymbol{\nu}_i), \sigma^2(\boldsymbol{\nu}_i))$$

with

$$\begin{aligned} \mu(\boldsymbol{\nu}_i) &= \eta_\mu(\boldsymbol{\nu}_i) \\ \sigma^2(\boldsymbol{\nu}_i) &= \exp(\eta_{\sigma^2}(\boldsymbol{\nu}_i)) \end{aligned}$$

is the most commonly known distributional regression model.

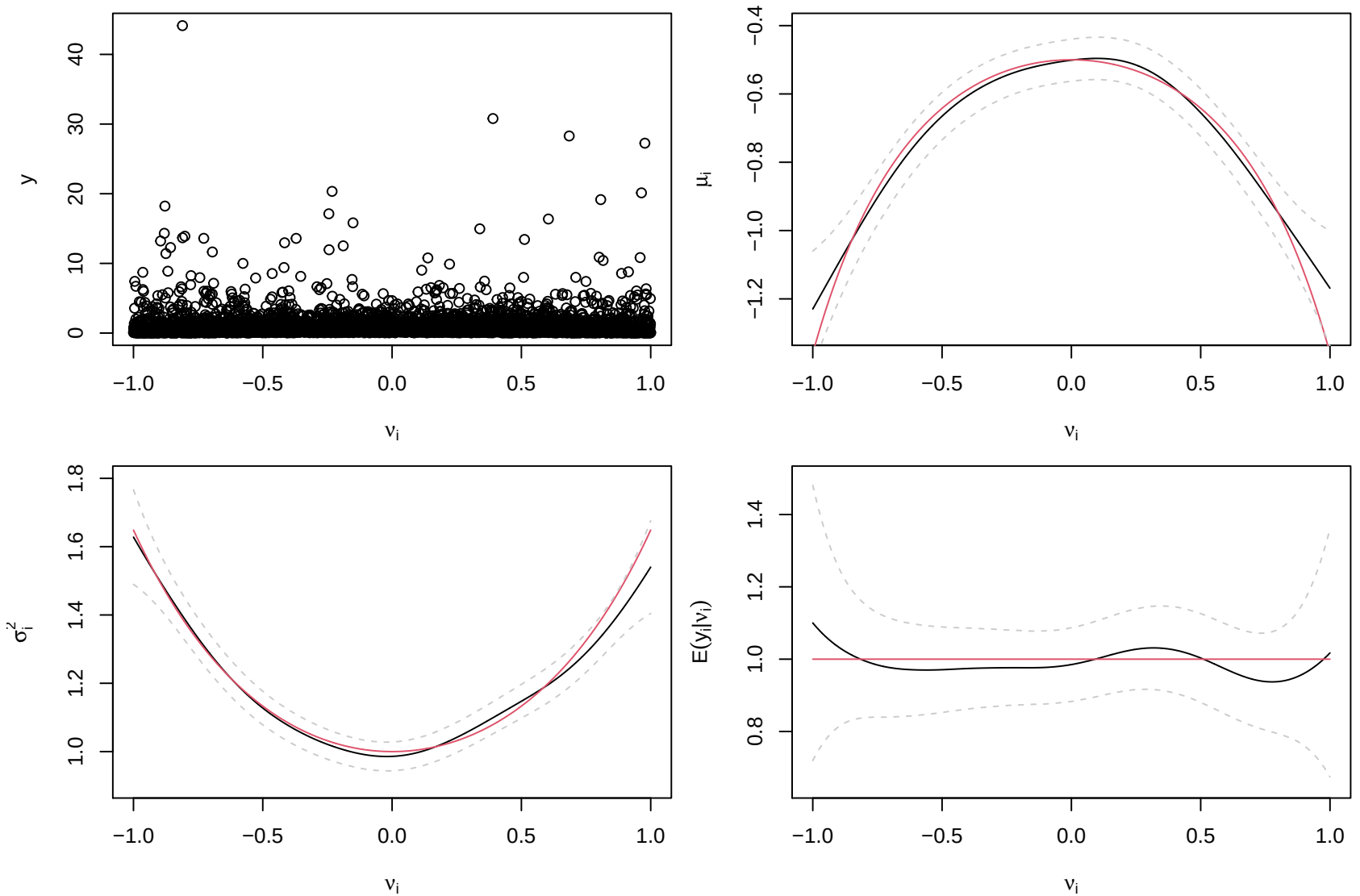
- However, there is a huge variety of models to pick from!

- Examples for interesting response structures:
  - Zero-inflated and / or overdispersed count data, i.e. responses with an excess of zeros and / or variances exceeding the expectation.
  - Responses with heteroscedastic or skewed distribution.
  - Continuous data with a spike in zero.
  - Fractional responses restricted to [0,1] (possibly with inflation in 0 and 1).
  - Multivariate responses with regression effects on the dependency parameters for example based on copula specifications.

Challenges:

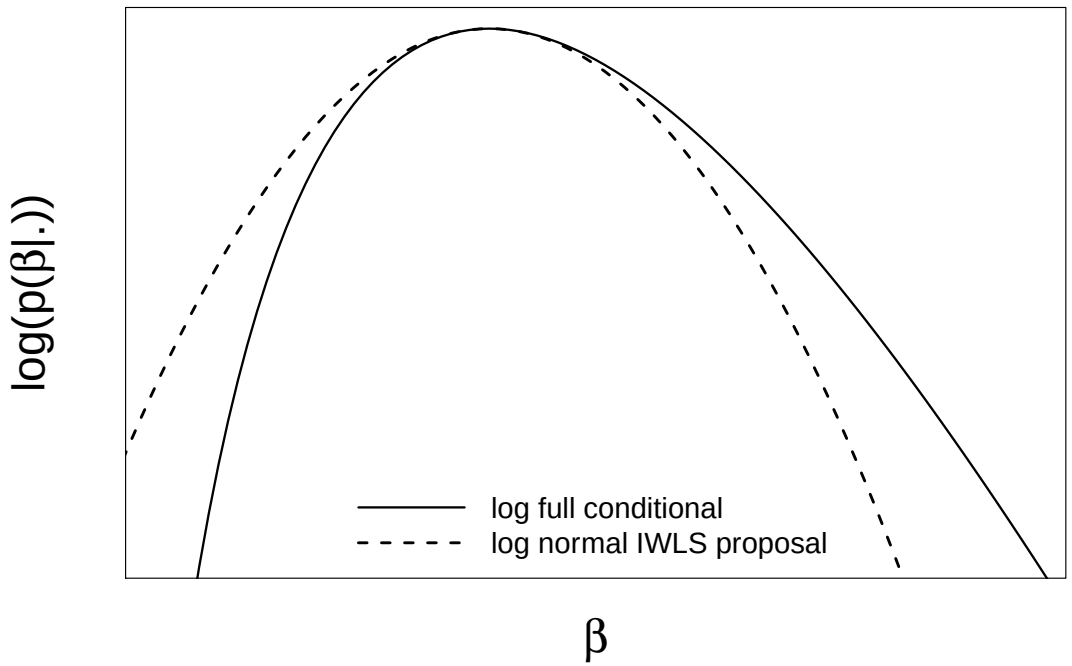
- Conceptually, distributional regression is very appealing and intuitive, but it comes with a number of challenges:
  - No conjugate priors for the observation models such that MCMC requires more care.
  - Interpretation of the estimated effects more difficult due to link functions and multi-parameter setup.
  - Model choice and checking to avoid model miss-specification.

Illustration for simulated log-normal data:



IWSL proposals for MCMC:

- Basic idea: Construct a quadratic approximation to the log-full conditional.



- Can also be motivated from a normal approximation of the working observations

$$\tilde{\boldsymbol{y}}_k = \boldsymbol{\eta}_k + (\boldsymbol{W}_k)^{-1} \boldsymbol{v}_k$$

where

$$\boldsymbol{v}_k = \frac{\partial \ell(\boldsymbol{\eta}_k)}{\partial \boldsymbol{\eta}_k}$$

is the vector of first derivatives of the log-likelihood with respect to the predictor (i.e. the score vector) and

$$\begin{aligned} \boldsymbol{W}_k &= \text{diag}(\{w_{ik}, i = 1, \ldots, n\}) \\ &= \text{diag}\left(\left\{\mathbb{E}\left(-\frac{\partial^2 \ell_i(\eta_{ik})}{\partial \eta_{ik}^2}\right), i = 1, \ldots, n\right\}\right) \end{aligned}$$

is the diagonal matrix of working weights corresponding to the individual contributions to the expected Fisher information.

- The resulting approximate normal distribution for the working observations then implies

$$\tilde{\boldsymbol{y}}_k \sim \mathrm{N}\left(\boldsymbol{\eta}_k, \boldsymbol{W}_k^{-1}\right).$$

- The IWLS proposal density for γ_{jk} is then constructed from the resulting working model for $\tilde{\boldsymbol{y}}_k$ as $\gamma_{jk} \sim \mathrm{N}(\boldsymbol{\mu}_{jk}, \boldsymbol{P}_{jk}^{-1})$ with expectation

$$\boldsymbol{\mu}_{jk} = \boldsymbol{P}_{jk}^{-1} \boldsymbol{B}_{jk}' \boldsymbol{W}_k (\tilde{\boldsymbol{y}}_k - \boldsymbol{\eta}_{-jk})$$

and precision matrix

$$\boldsymbol{P}_{jk} = \boldsymbol{B}_{jk}' \boldsymbol{W}_k \boldsymbol{B}_{jk} + \frac{1}{\tau_{jk}^2} \boldsymbol{K}_{jk}$$

where $\boldsymbol{\eta}_{-jk} = \boldsymbol{\eta}_k - \boldsymbol{B}_{jk}\gamma_{jk}$ is the predictor without the jth component.

- More precisely, a proposal γ_j^* is determined from the density $q(\gamma_{jk}^*|\gamma_{jk}^c)$ of the normal distribution

$$\mathrm{N}\left(\boldsymbol{\mu}_{jk}^c, (\boldsymbol{P}_{jk}^c)^{-1}\right)$$

where $\boldsymbol{\gamma}_{jk}^c$ denotes the current value of the parameter vector $\boldsymbol{\gamma}_{jk}$ and

$$\boldsymbol{\mu}_{jk}^c = \boldsymbol{\mu}_{jk}(\boldsymbol{\gamma}_{jk}^c) = \boldsymbol{P}_{jk}^{-1}(\boldsymbol{\gamma}_{jk}^c)\boldsymbol{B}_{jk}'\boldsymbol{W}_k(\boldsymbol{\gamma}_{jk}^c)(\tilde{\boldsymbol{y}}_k(\boldsymbol{\gamma}_{jk}^c) - \boldsymbol{\eta}_{-jk})$$

and

$$\boldsymbol{P}_{jk}^c = \boldsymbol{P}_{jk}(\boldsymbol{\gamma}_{jk}^c) = \boldsymbol{B}_{jk}'\boldsymbol{W}_k(\boldsymbol{\gamma}_{jk}^c)\boldsymbol{B}_{jk} + \frac{1}{\tau_{jk}^2}\boldsymbol{K}_{jk}$$

correspond to the mean and precision matrix with the matrix of working weights $\boldsymbol{W}_k(\boldsymbol{\gamma}_{jk}^c)$ and the vector of working observations $\tilde{\boldsymbol{y}}_k(\boldsymbol{\gamma}_{jk}^c)$ evaluated at this current value.

- The proposal $\boldsymbol{\gamma}_{jk}^*$ is then accepted with probability

$$\alpha\left(\boldsymbol{\gamma}_{jk}^*|\boldsymbol{\gamma}_{jk}^c\right) = \min\left\{\frac{f(\boldsymbol{\gamma}_{jk}^*|\cdot)q(\boldsymbol{\gamma}_{jk}^c|\boldsymbol{\gamma}_{jk}^*)}{f(\boldsymbol{\gamma}_{jk}^c|\cdot)q(\boldsymbol{\gamma}_{jk}^*|\boldsymbol{\gamma}_{jk}^c)}, 1\right\}.$$

Model Choice and Model Checking:

- Quantile residuals:
  - For a continuous random variable $Y \sim F$ with cumulative distribution function F, we have $F(Y) \sim \mathrm{U}(0,1)$ (probabilitiy integral transform) or

$$\Phi^{-1}(F(Y)) \sim \mathrm{N}(0,1).$$

  - For a correctly specified distributional regression model, we should therefore have

$$u_i = \Phi^{-1}(F(y_i|\hat{\boldsymbol{\vartheta}}(\nu_i))) \overset{a}{\sim} \mathrm{N}(0,1)$$

    and the quantile residuals u_i can, e.g., be visualized in a quantile-quantile plot.
  - For discrete or multivariate data, appropriate generalisations are needed.

- Information criteria:
  - DIC or WAIC can straightforwardly be used in the context of distributional regression.
- Proper scoring rules:
  - In a distributional setting, typical predictive measures such as the mean squared error of prediction or the mean absolute error of prediction are not adequate.
  - Proper scoring rules provide a framework for evaluating predictive distributions rather than point predictions.
  - Underlying theory ensures that proper scores encourage the analyst to honestly report their uncertainty in terms of the predictive distribution.
  - The cross-validated log-likelihood is the most commonly used proper score.

- Scoring rules for real-valued outcomes with predictive density $f(y)$ and observed outcome y_0:
  - Spherical score
$$\text{SPS}(f(y), y_0) = -\frac{f(y_0)}{\left(\int f^2(t)dt\right)^{1/2}}.$$
  - Logarithmic score
$$\text{LS}(f(y), y_0) = -\log(f(y_0)).$$
  - Continuously ranked probability score
$$\text{CRPS}(f(y), y_0) = \int \left[F(t) - \mathbb{1}_{[y_0,\infty)}(t)\right]^2 dt.$$
  - Note: All scores are negatively oriented, i.e. smaller values indicate a better agreement between the predictive distribution and the observed values.

Case Study: Conditional Income Distributions

- Utilise information from the German Socio-Economic Panel to study real gross annual personal labour income in Germany for the years 2001 to 2010.
- Specific focus on changes in spatial differences in income inequality.
- Response: income of males in full time employment in the age range 20–60.
- Information available on $7,216$ individuals with a total of $n = 40,965$ observations.
- Potential response distributions:
  - Log-normal $\mathrm{LN}(\mu, \sigma^2)$.
  - Gamma $\mathrm{Ga}(\mu, \sigma)$.
  - Inverse Gaussian $\mathrm{IG}(\mu, \sigma^2)$.
  - Dagum $\mathrm{Da}(a, b, c)$.

  with covariate effects on potentially all distributional parameters.

- Covariates:
  - $educ$: Educational level measured as a binary indicator for completed higher education (according to the UNESCO International Standard Classification of Education 1997).
  - age: age in years.
  - $lmexp$: previous labour market experience in years.
  - t calendar time.
  - s: area of residence in terms of geographical district ($Raumordnungsregion$).
  - $east$: indicator in effect coding for districts belonging to the eastern part of Germany.

- Hierarchical predictor structure:

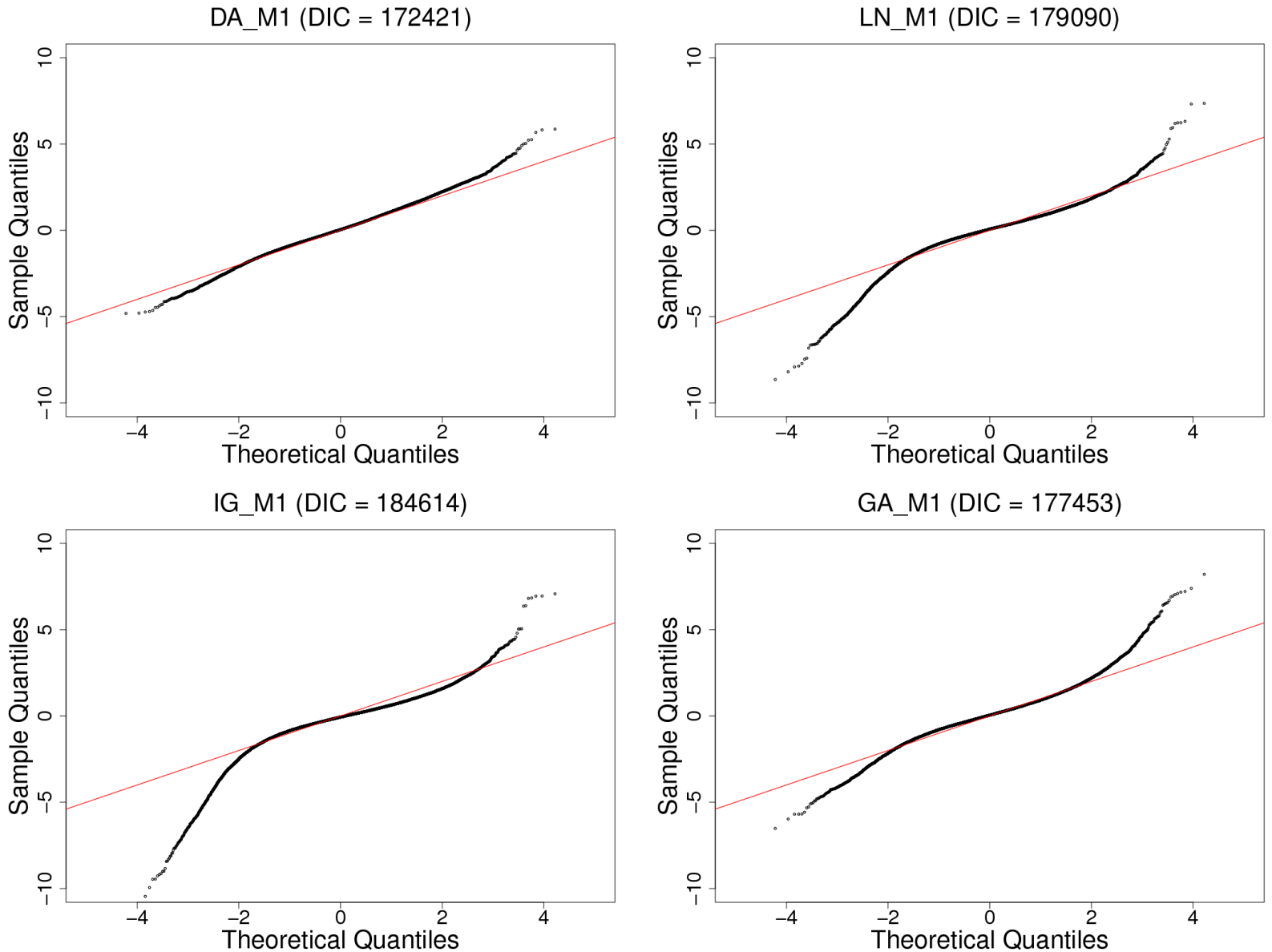
$$\eta_i = \beta_0 + educ_i\beta_1 + f_1(age_i) + educ_i f_2(age_i) + f_3(lmexp_i) + f_{spat}(s_i) + f_{time}(t_i)$$

where the spatial effects is decomposed as

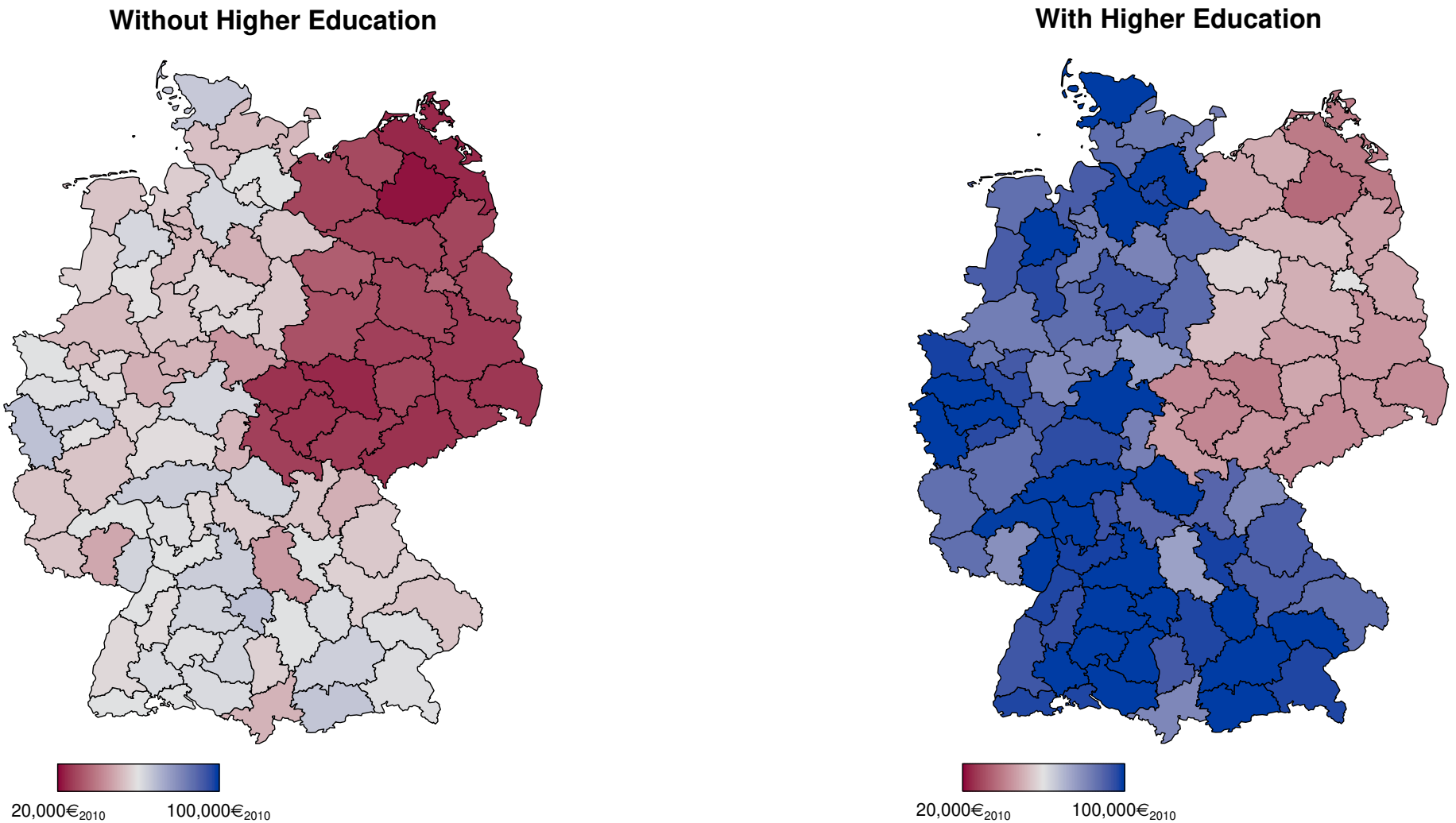
$$f_{spat}(s) = east_s\gamma_1 + g_{str}(s) + g_{unstr}(s)$$

- DIC and scoring rules:

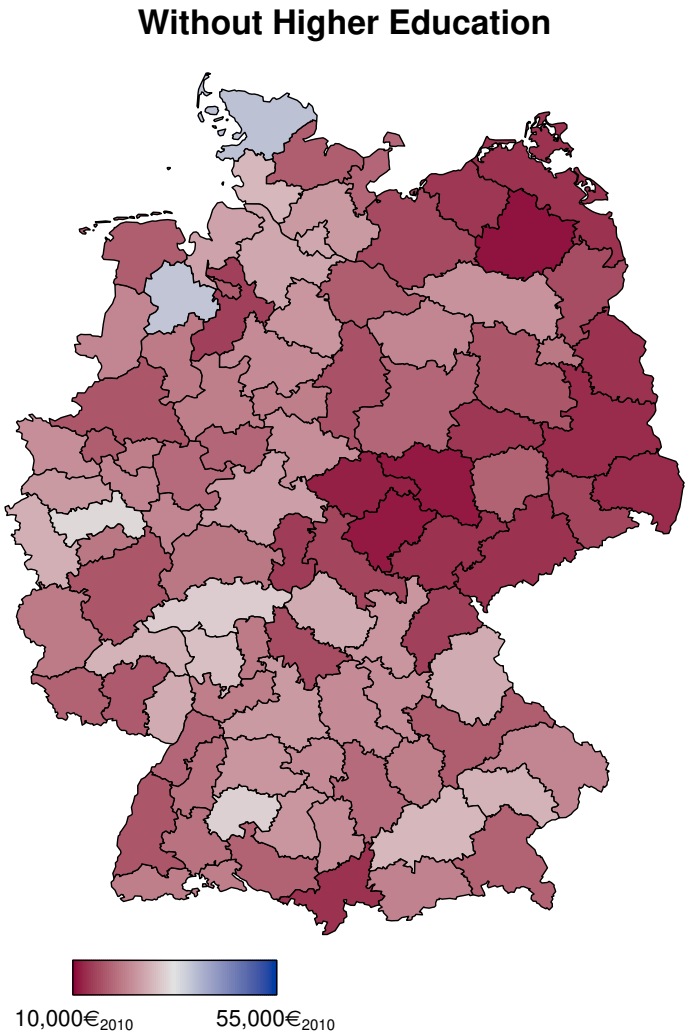
Distribution	DIC	Quadratic	Logarithmic	Spherical	CRPS
LN	179,090	0.1304	-2.4363	0.3621	-2.1581
IG	184,614	0.1464	-2.2741	0.3777	-1.6195
GA	177,453	0.1609	-2.1715	0.3963	-1.2735
DA	**172,421**	**0.1684**	**-2.1034**	**0.4053**	**-1.2662**

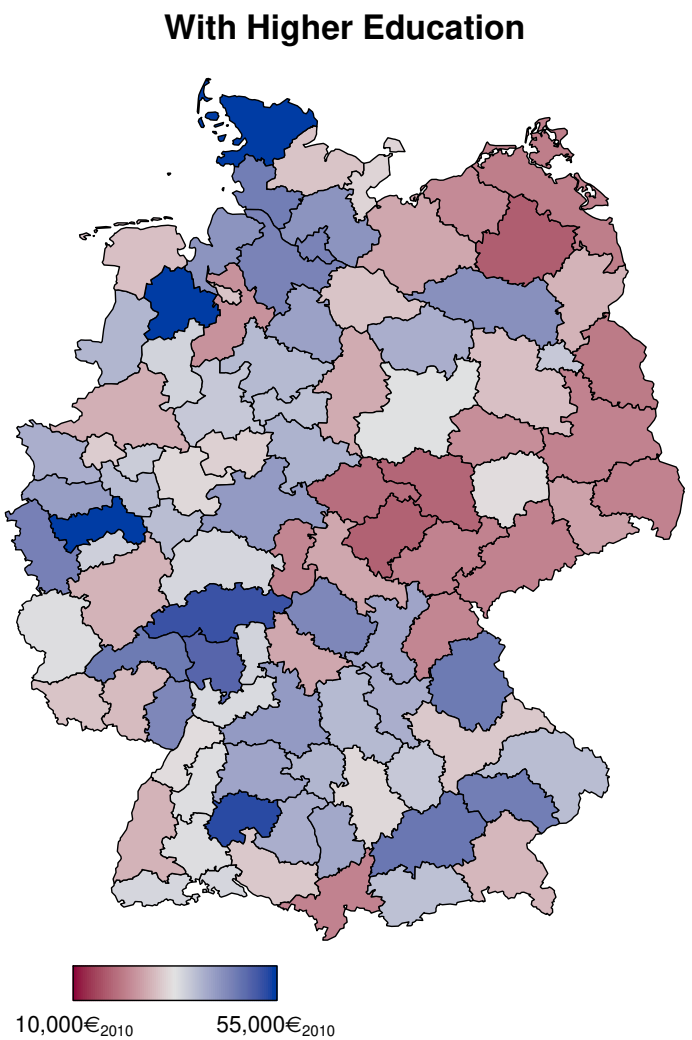

DA_M1 (DIC = 172421)
LN_M1 (DIC = 179090)
IG_M1 (DIC = 184614)
GA_M1 (DIC = 177453)
Sample Quantiles
Theoretical Quantiles
10
5
0
−5
−10
−4
−2
0
2
4

- Expected income for an "average man" with / without higher education:

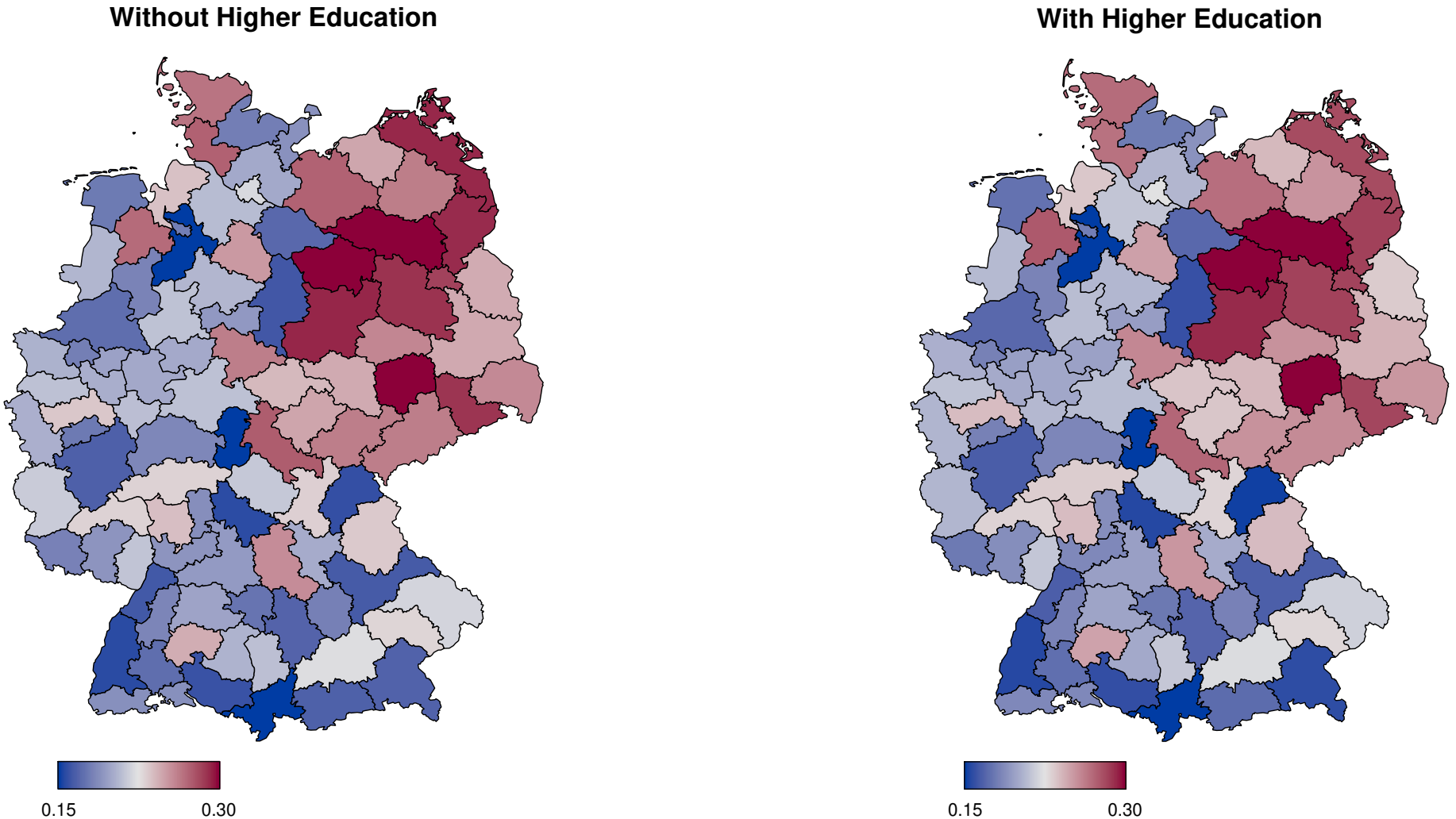


- Income standard deviation for an "average man":





- Income inequality (measured by the Gini coefficient) for an "average man":



Modelling Complex Interactions

Identifiability in additive regression models:

- In additive regression models, the additive effects have to be constrained to ensure identifiability since

$$s_1(\boldsymbol{x}) + s_2(\boldsymbol{x}) = s_1(\boldsymbol{x}) + c + s_2(\boldsymbol{x}) - c.$$

- This is often operationalized as

$$\sum_{i=1}^{n} s(\boldsymbol{x}_i) = 0$$

or

$$\int s(\boldsymbol{x}) d\boldsymbol{x} = 0.$$

- In practice, both constraints can be implemented by imposing

$$\boldsymbol{A}\boldsymbol{\gamma} = \mathbf{0}$$

  with an appropriate $(A \times L)$ constraint matrix $\boldsymbol{A}$.

- This can easily be integrated into sampling $\boldsymbol{\gamma}$ in MCMC:
  - Compute the $L \times A$ matrix $\boldsymbol{V} = \boldsymbol{P}^{-1}\boldsymbol{A}'$ by solving the equation systems $\boldsymbol{P}\boldsymbol{V} = \boldsymbol{A}'$ for each of the columns of $\boldsymbol{V}$.
  - Compute the $A \times A$ matrix $\boldsymbol{W} = \boldsymbol{A}\boldsymbol{V}$.
  - Compute the $A \times L$ matrix $\boldsymbol{U} = \boldsymbol{W}^{-1}\boldsymbol{V}'$ by solving the equation systems $\boldsymbol{W}\boldsymbol{U} = \boldsymbol{V}'$ for each of the columns of $\boldsymbol{U}$.
  - Compute the constrained sample $\boldsymbol{\gamma}^* = \boldsymbol{\gamma} - \boldsymbol{U}'\boldsymbol{A}\boldsymbol{\gamma}$ where $\boldsymbol{\gamma}$ is an unconstrained sample from $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{P}^{-1})$.

Construction of general constraint matrices:

- For an effect $s(\boldsymbol{x}) \in \mathcal{F}$, we are interested in removing all effects from the function space $\mathcal{H} \subset \mathcal{F}$ generated as

$$\mathcal{H} = \left\{ h(\boldsymbol{x}) : h(\boldsymbol{x}) = \sum_{a=1}^{A} \delta_a H_a(\boldsymbol{x}) \right\}.$$

- This implies a set of A linear constraints on the vector of regression coefficients $\boldsymbol{\gamma}$ that represents $s(\boldsymbol{x})$.
- Construct the constraints from orthogonality of the basis functions of $s(\boldsymbol{x})$ to the basis functions of $\mathcal{H}$, i.e. assume

$$\int B_l(\boldsymbol{x}) H_a(\boldsymbol{x}) d\boldsymbol{x} = 0, \quad a = 1, \ldots, A, l = 1, \ldots, L$$

- This implies a constraint $\boldsymbol{A}\boldsymbol{\gamma} = \mathbf{0}$ where

$$\boldsymbol{A}[a, l] = \int B_l(\boldsymbol{x}) H_a(\boldsymbol{x}) d\boldsymbol{x}, \quad a = 1, \ldots, A, l = 1, \ldots, L$$

- The prior for $\boldsymbol{\gamma}$ has to be changed to

$$p(\boldsymbol{\gamma}|\tau^2) \propto \left(\frac{1}{\tau^2}\right)^{\frac{\text{effdim}(\boldsymbol{\gamma})}{2}} \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\gamma}'\boldsymbol{K}\boldsymbol{\gamma}\right) \mathbb{1}(\boldsymbol{A}\boldsymbol{\gamma} = \mathbf{0})$$

where

$$\text{effdim}(\boldsymbol{\gamma}) = L - \tilde{L}$$

and

$$\tilde{L} = \dim\left(\text{span}(\ker(\boldsymbol{K})) \cup \text{span}(\boldsymbol{A}')\right).$$

- An alternative is to construct constraints such that $\boldsymbol{A}$ corresponds to the null space of the precision matrix $\boldsymbol{K}$.

- Explicitly reparameterize the original function to obtain a basis of the constraint effect:
  - Can be applied in virtually any inferential approach to estimating structured additive regression models.
  - The constraints really remove parameters and therefore reduce the dimensionality of the problem.
  - Typically destroys sparse matrix structures that could be used for efficient computations.
- Determine estimates under the constraint but in the original parameterisation:
  - Allows to maintain sparse matrix structures and therefore efficient MCMC samples.
  - Requires constrained sampling / optimisation under linear constraints.
  - Keeps the original dimensionality and therefore increases the number of parameters when decomposing effects.

Application to tensor product interactions:

- Construct the interaction of two effects

$$s_1(x_1) = \sum_{l_1=1}^{L_1} \gamma_{l_1} B_{l_1}(x_1), \qquad s_2(x_2) = \sum_{l_2=1}^{L_2} \gamma_{l_2} B_{l_2}(x_2)$$

with priors

$$p(\boldsymbol{\gamma}_d|\tau_d^2) \propto \left(\frac{1}{\tau_d^2}\right)^{\frac{\text{rg}(\boldsymbol{K}_d)}{2}} \exp\left(-\frac{1}{2\tau_d^2}\boldsymbol{\gamma}_d'\boldsymbol{K}_d\boldsymbol{\gamma}_d\right), \quad d = 1, 2.$$

- The tensor product interaction of these two effects is then given by

$$s(x_1, x_2) = \sum_{l_1=1}^{L_1} \sum_{l_2=1}^{L_2} \gamma_{l_1 l_2} B_{l_1 l_2}(x_1, x_2)$$

with tensor product basis functions

$$B_{l_1 l_2}(x_1, x_2) = B_{l_1}(x_1) B_{l_2}(x_2).$$

- The regression coefficients of the interaction are given by

$$\boldsymbol{\gamma} = \begin{pmatrix} \gamma_{11} & \cdots & \gamma_{1L_2} \\ \vdots & & \vdots \\ \gamma_{L_1,1} & \cdots & \gamma_{L_1 L_2} \end{pmatrix}'$$

- The precision matrix for the tensor product can be defined as

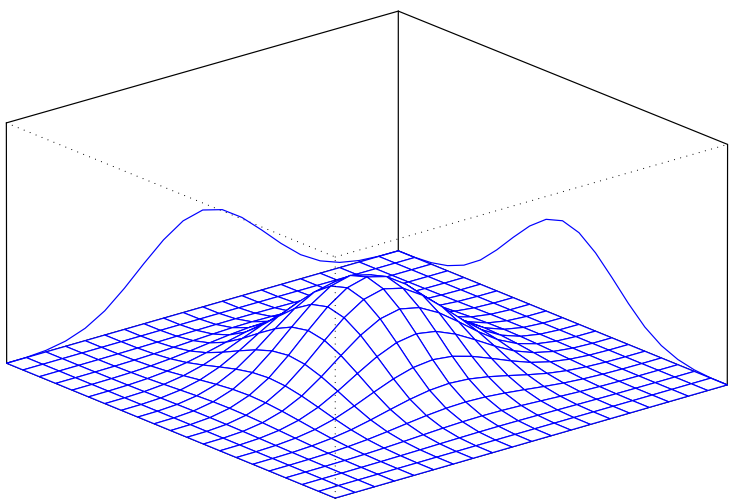
$$\frac{1}{\tau^2}\boldsymbol{K} = \frac{1}{\tau^2}\left[\omega \boldsymbol{K}_1 \otimes \boldsymbol{I}_{L_2} + (1-\omega)\boldsymbol{I}_{L_1} \otimes \boldsymbol{K}_2\right]$$

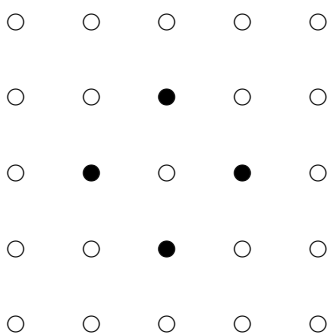
  where

  – $\omega \in [0, 1]$ relates to the prior weight of the first effect relative to the second while

  – τ^2 relates to the overall importance of the interaction effect.

- Illustration for a bivariate tensor product spline:





- We can use any two components from structured additive regression to formulate a tensor product interaction.
- Constraints can now be used to
  - improve interpretation by decomposing the interaction into main effects and (maybe several) interaction effects.
  - make models involving multiple tensor product interactions identifiable.

- For example, a tensor product spline can be decomposed as

$$s(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + s_1^c(x_1) + s_2^c(x_2) + s_{1,2}^c(x_1, x_2).$$

or

$$\begin{aligned} s(x_1, x_2) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \\ &\quad + s_1^{\text{main}}(x_1) + x_2 s_1^{\text{vcm}}(x_1) \\ &\quad + s_2^{\text{main}}(x_2) + x_1 s_2^{\text{vcm}}(x_2) \\ &\quad + s_{1,2}^{(c)}(x_1, x_2). \end{aligned}$$

- Similarly, a spatio-temporal interaction can be decomposed as

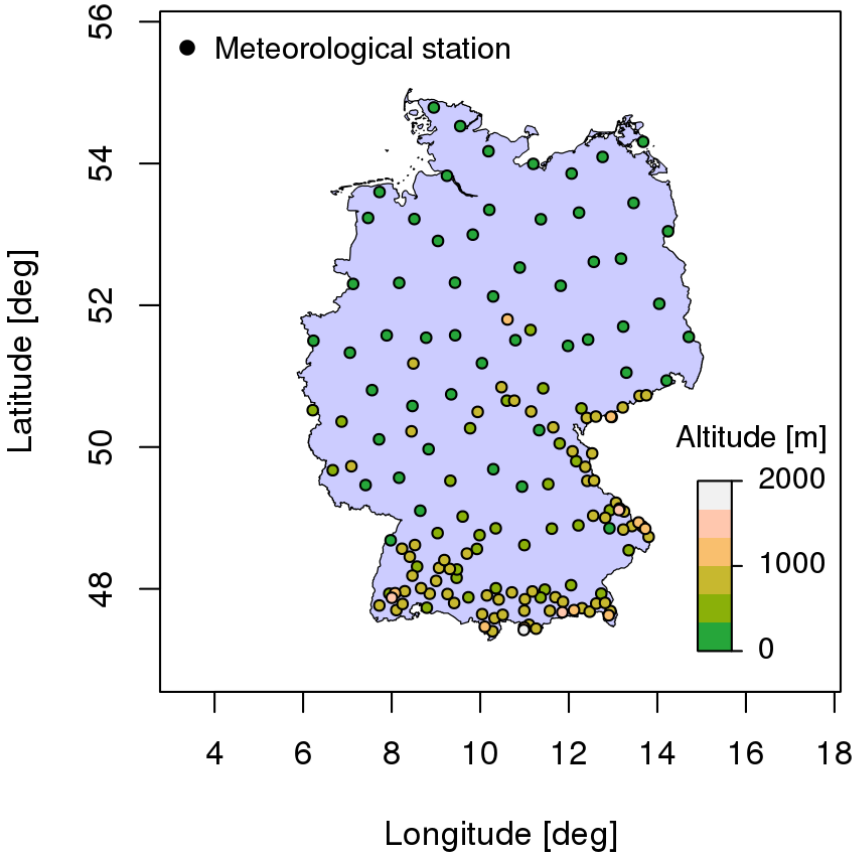
$$s(\boldsymbol{s}, t) = \beta_0 + \beta_1 t + s_{\text{space}}^{\text{main}}(\boldsymbol{s}) + t s_{\text{space}}^{\text{vcm}}(\boldsymbol{s}) + s_{\text{time}}^{\text{main}}(\boldsymbol{s}) + s_{1,2}^{(c)}(\boldsymbol{s}, t).$$

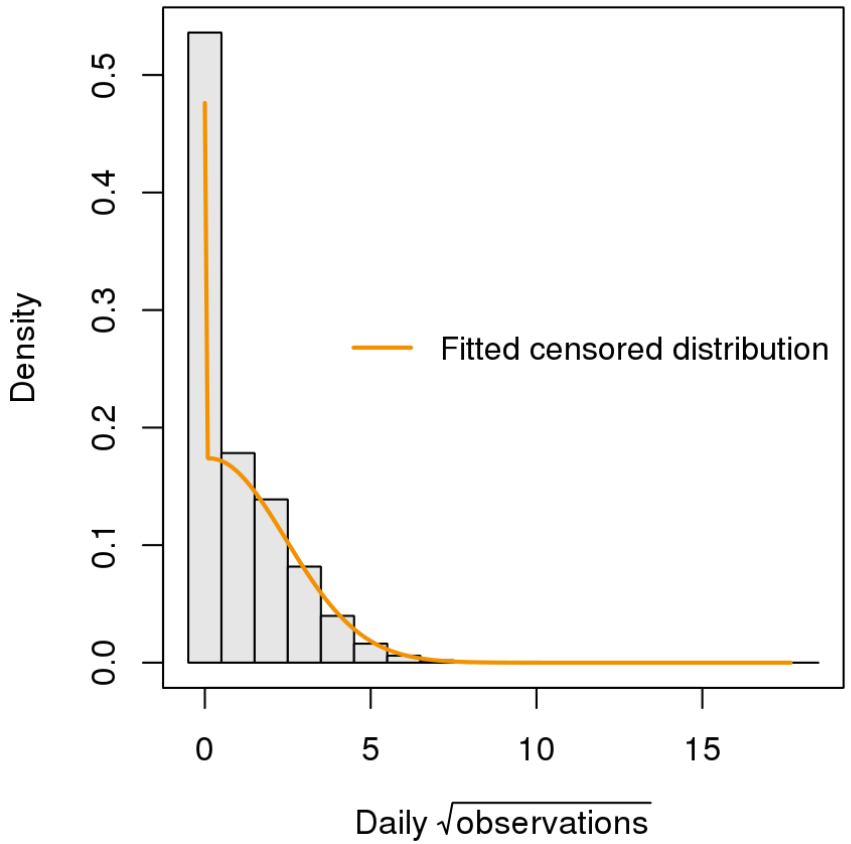
Space-Time Precipation Analysis

Daily precipitation sums

- Model daily precipitation sums in the period 1986 to 2015.
- Include all stations from *Deutscher Wetterdienst* above 900m sea level, while for stations below 900m we selected a representative subset with good coverage all over Germany.
- Results in a total of 164 meteorological stations and over 1.6 million spatio-temporally aligned observations.

- Stylized view on the data:





- Square-root transformation of the precipitation sums to improve the fit of a censored normal model where

$$y_{st} = \max(0, y_{st}^\star)$$

and

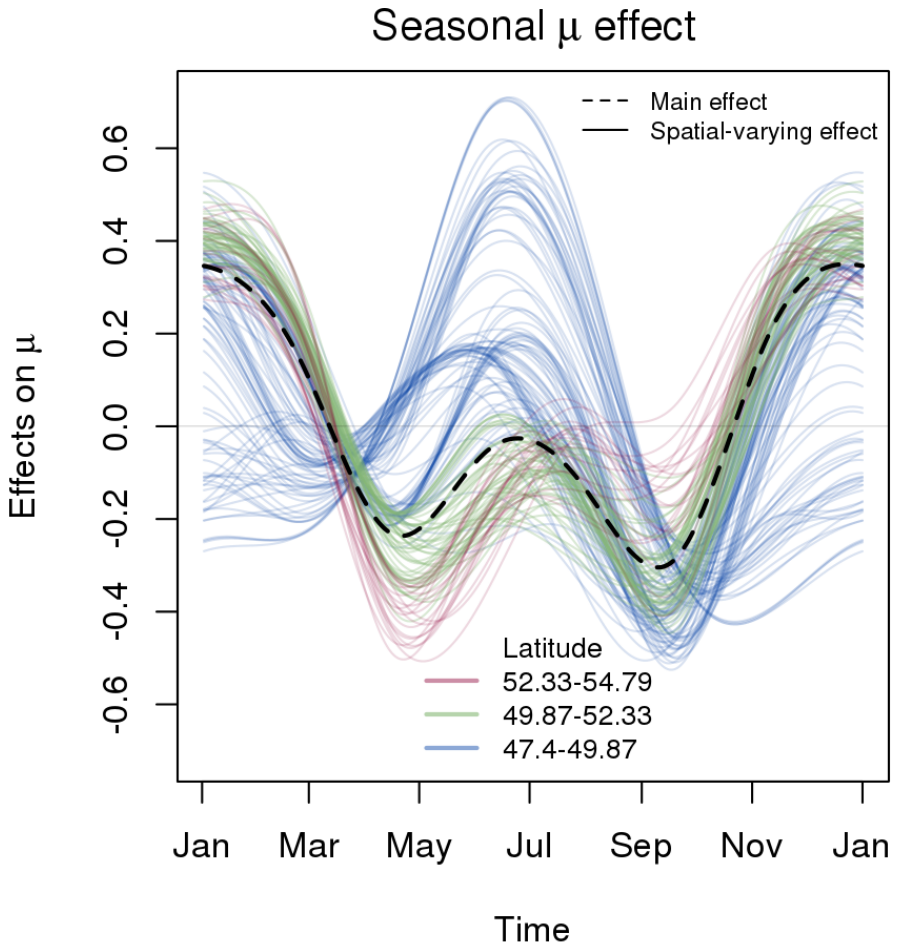
$$y_{st}^\star \sim \mathcal{N}(\mu_{st}, \sigma_{st}^2)$$

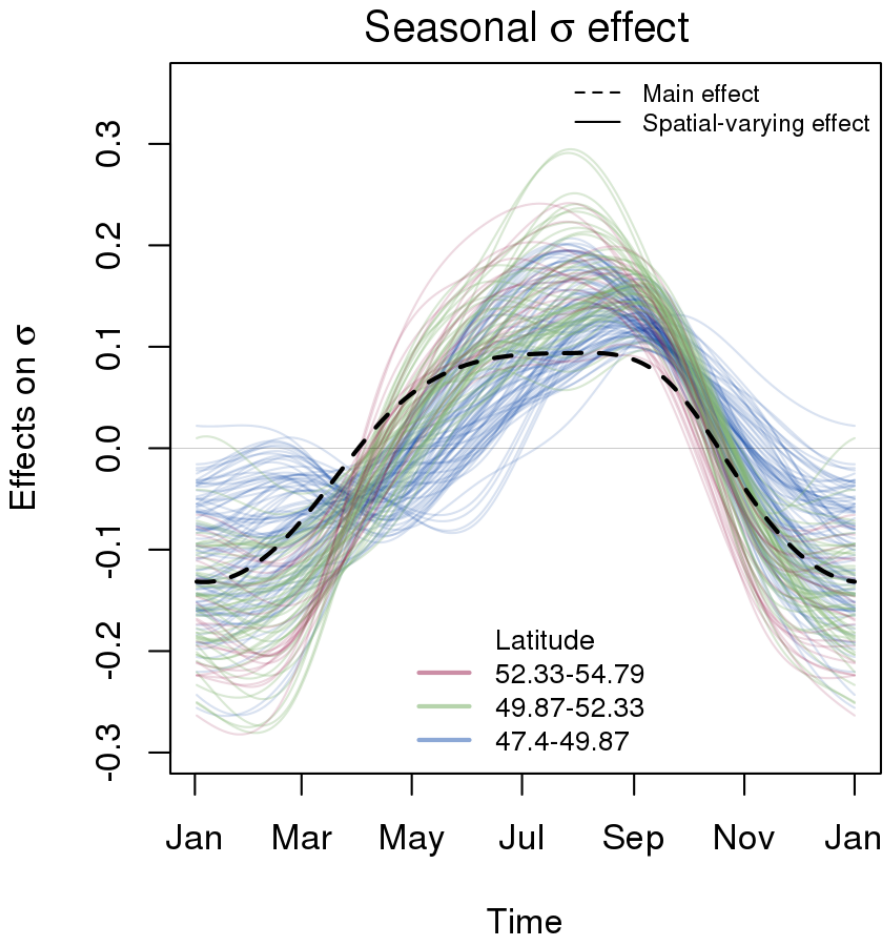
with s indexing the spatial locations of the meteorological stations and t indexing the daily measurement time points.

- Predictor structure for both the "location" and the "scale" parameter:

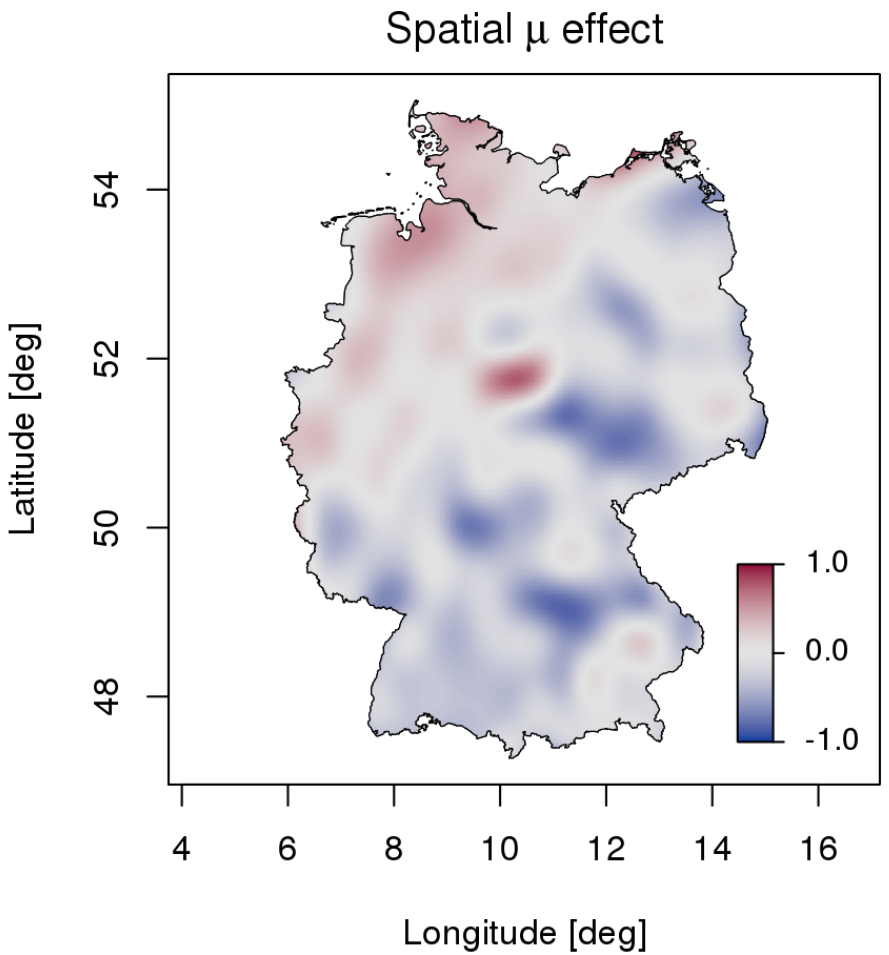
$$\eta = \beta_0 + f_1(\texttt{alt}) + f_2(\texttt{day}) + f_3(\texttt{lon}, \texttt{lat}) + f_4(\texttt{day}, \texttt{lon}, \texttt{lat}),$$

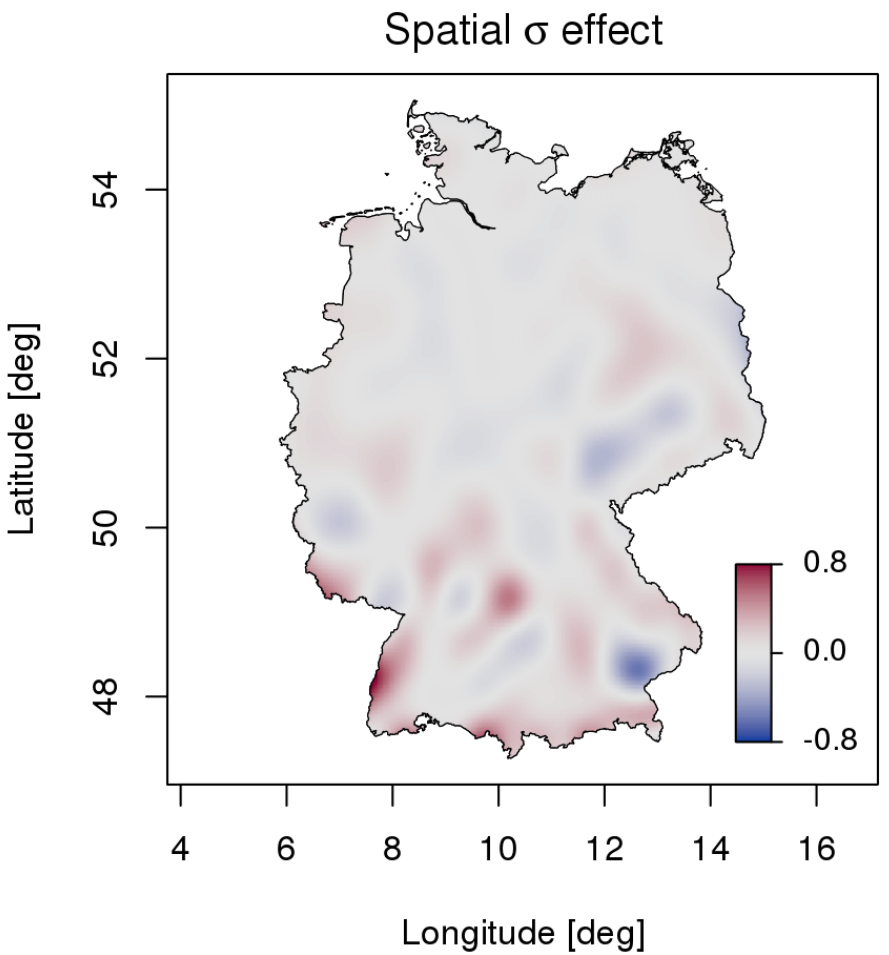
- Seasonal main effects and interactions:



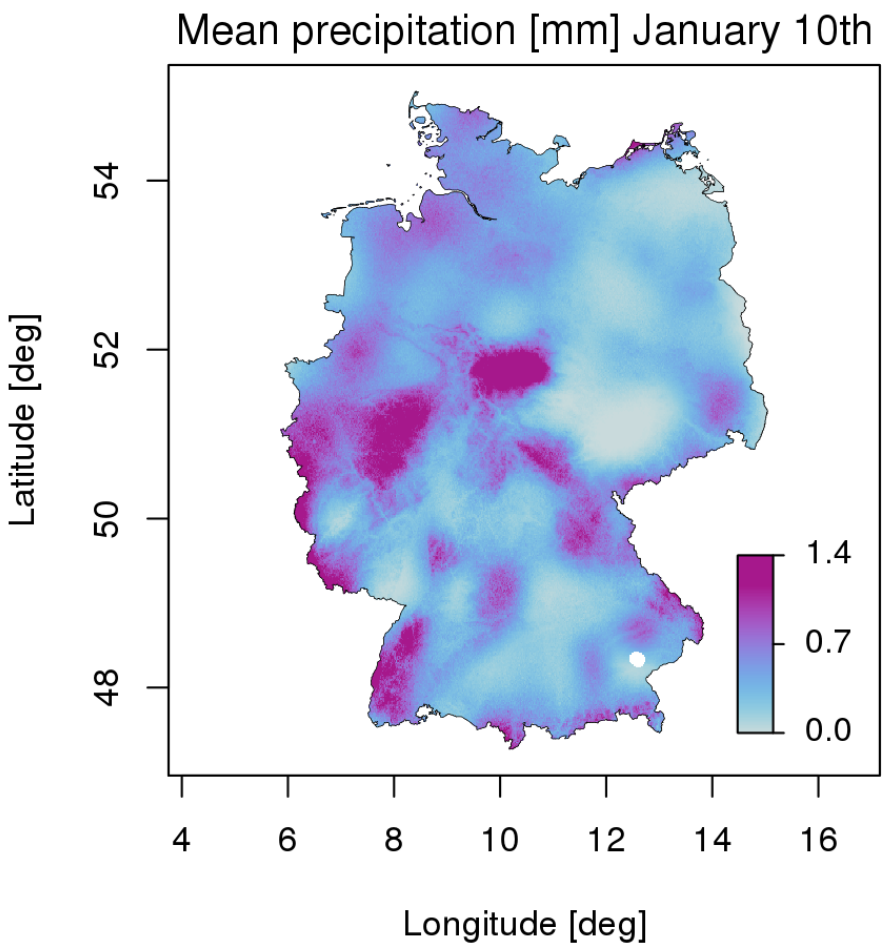


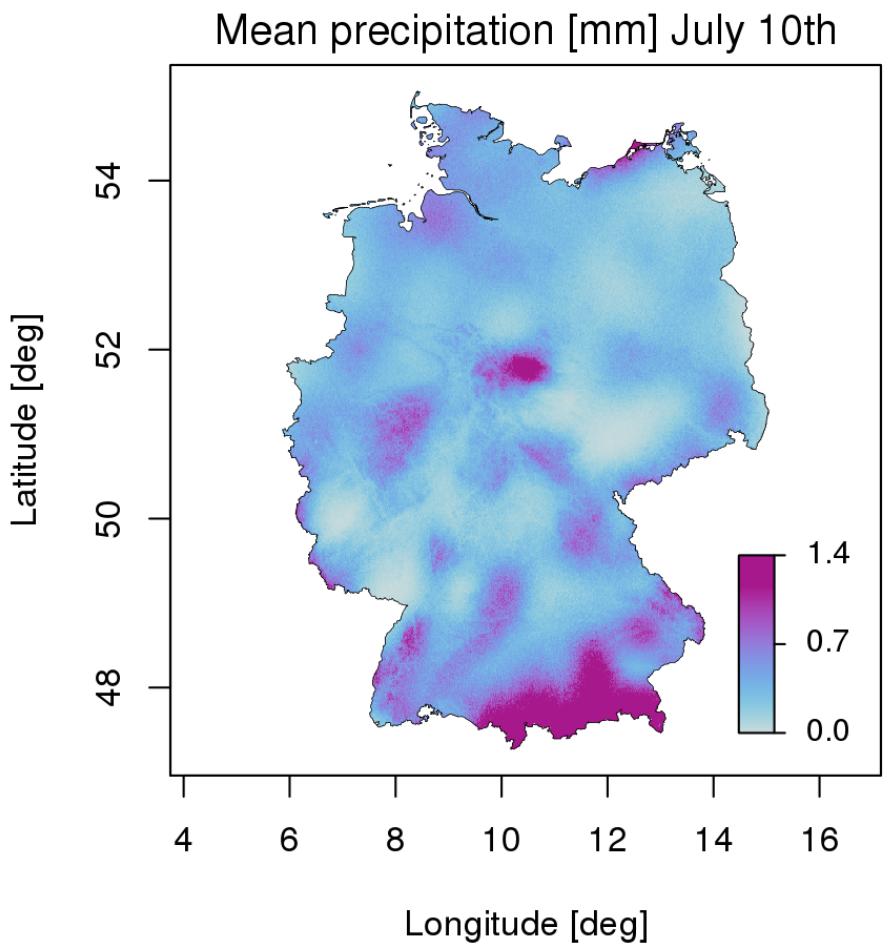
- Spatial main effects:





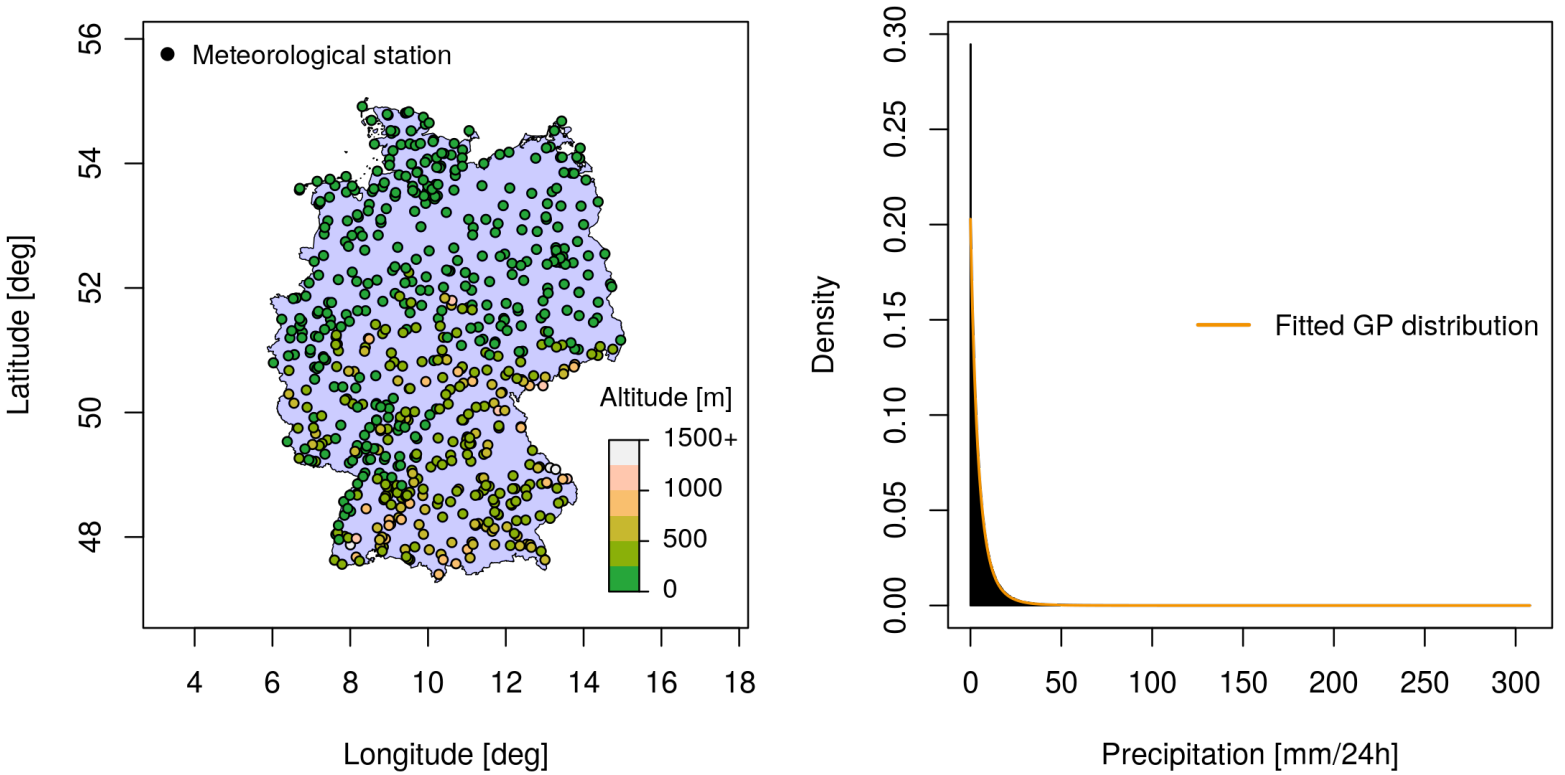
- Predicted precipitation climatology for January and July 10th:





Extreme precipitation events

- Determine spatio-temporal variation in 100 year return levels of precipitation in Germany.
- $1.138.868$ observations of 569 meteorological stations.

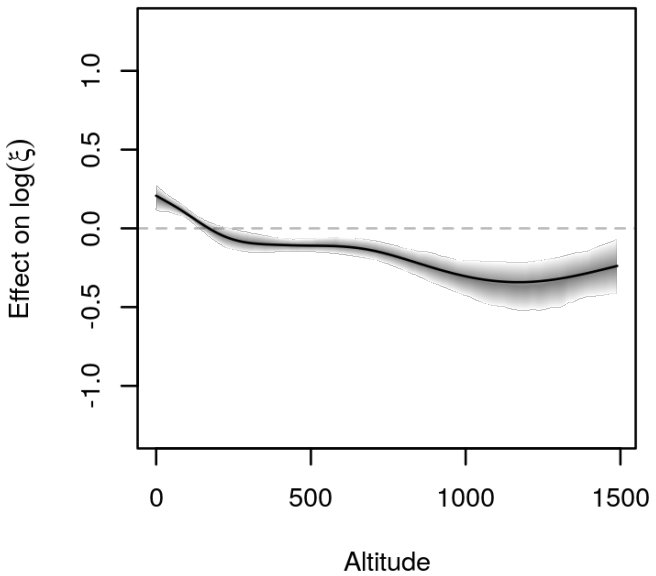


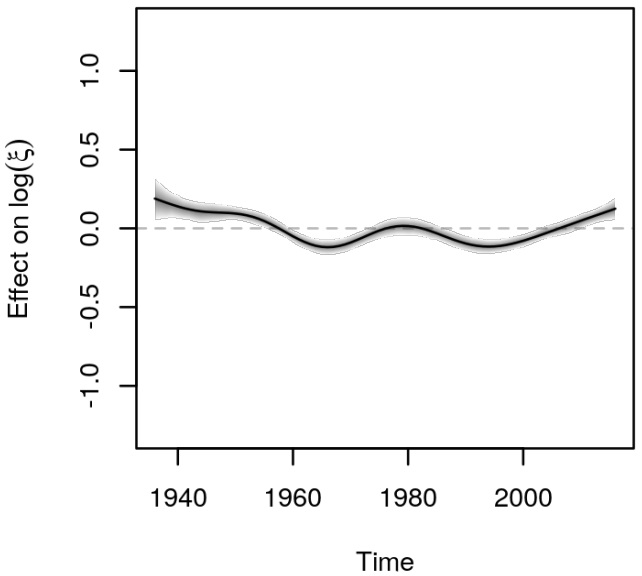
- Assume a generalized Pareto model

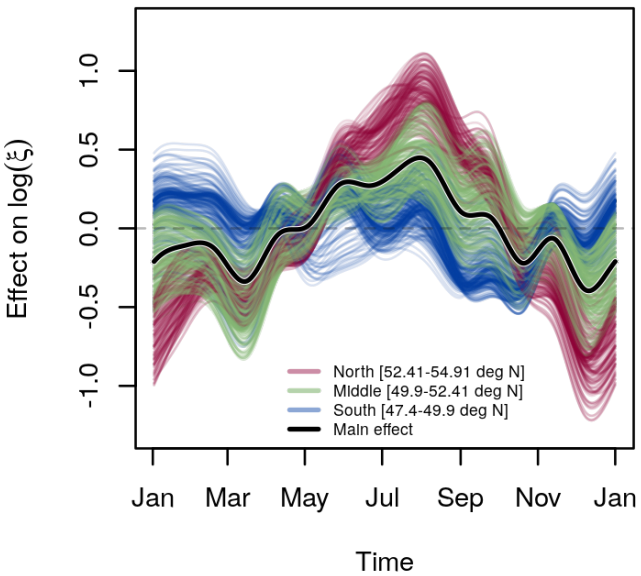
$$\texttt{P}_i \sim \text{GP}(\xi(\boldsymbol{x}_i), \sigma(\boldsymbol{x}_i))$$

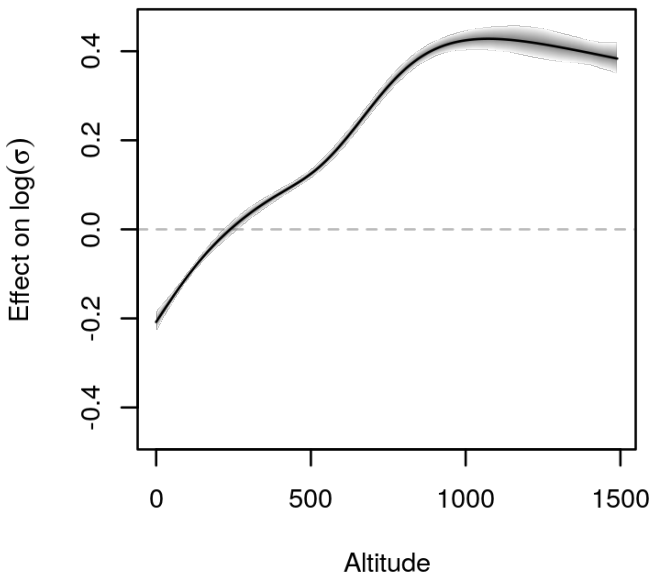
with the following predictor structure for both parameters:

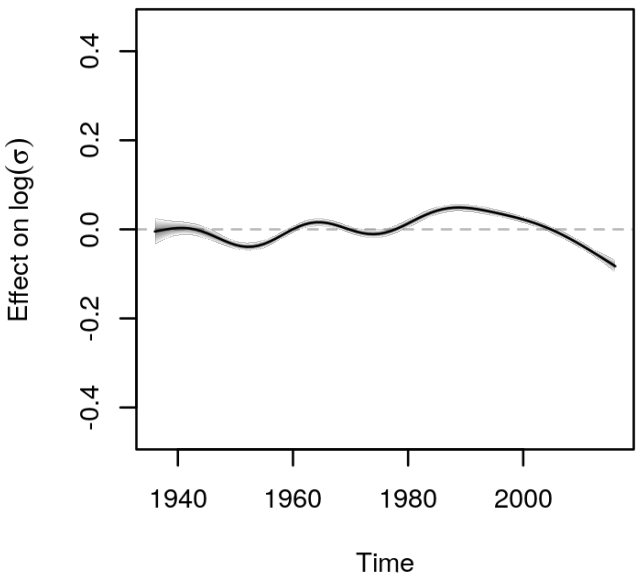
$$\begin{aligned} \eta_i \quad = \quad & \beta_0 + f_1(\texttt{alt}) + f_2(\texttt{year}_i) + f_3(\texttt{yday}_i) + \\ & f_4(\texttt{lon}_i, \texttt{lat}_i) + f_5(\texttt{yday}_i, \texttt{lon}_i, \texttt{lat}_i), \quad i = 1, \ldots, 1138868, \end{aligned}$$

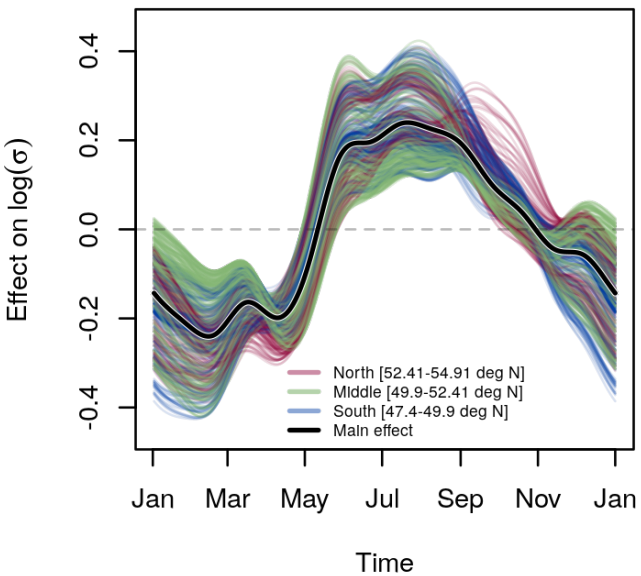

Effect on log(ξ)
1.0
0.5
0.0
-0.5
-1.0
0
500
1000
1500
Altitude

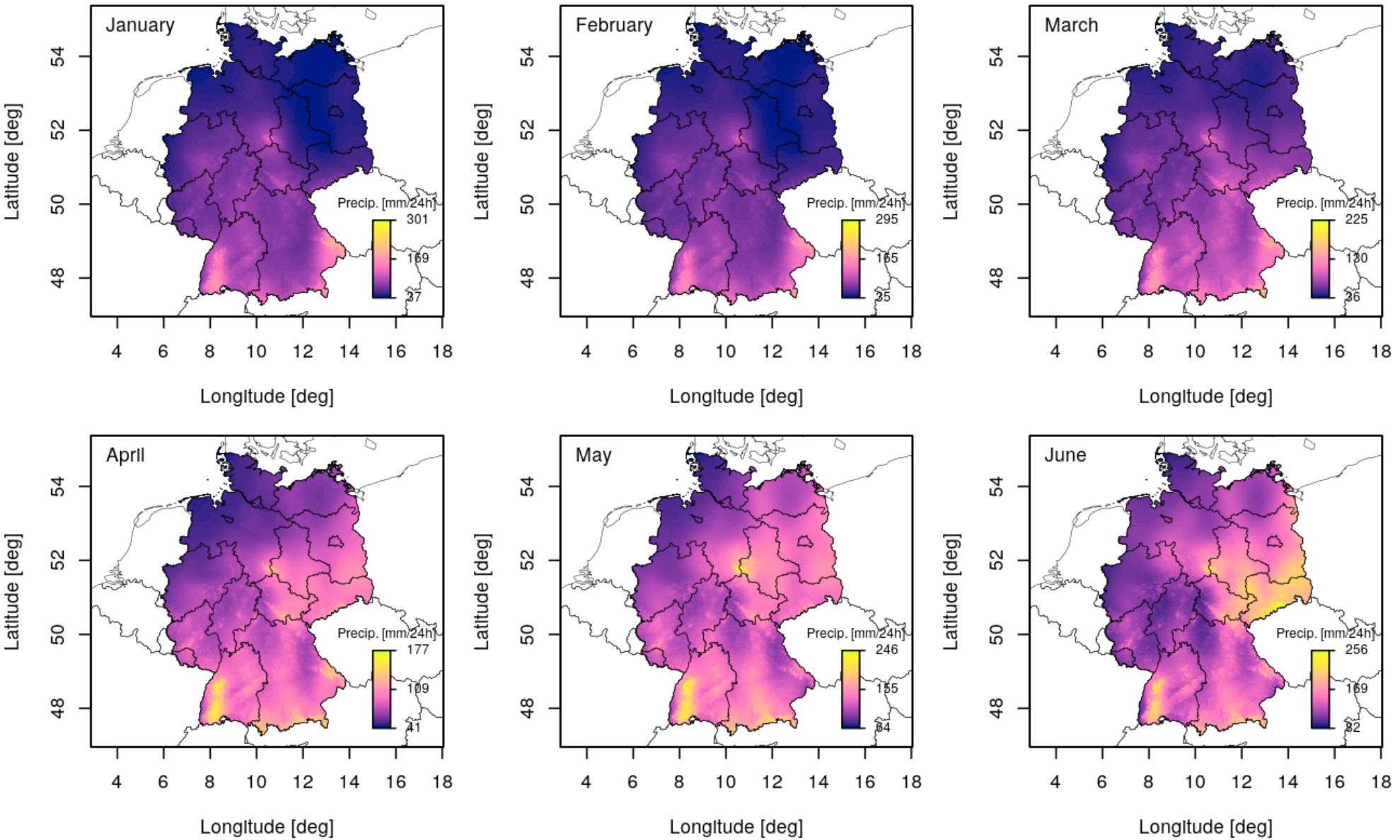

Effect on log(ξ)
1.0
0.5
0.0
-0.5
-1.0
1940
1960
1980
2000
Time

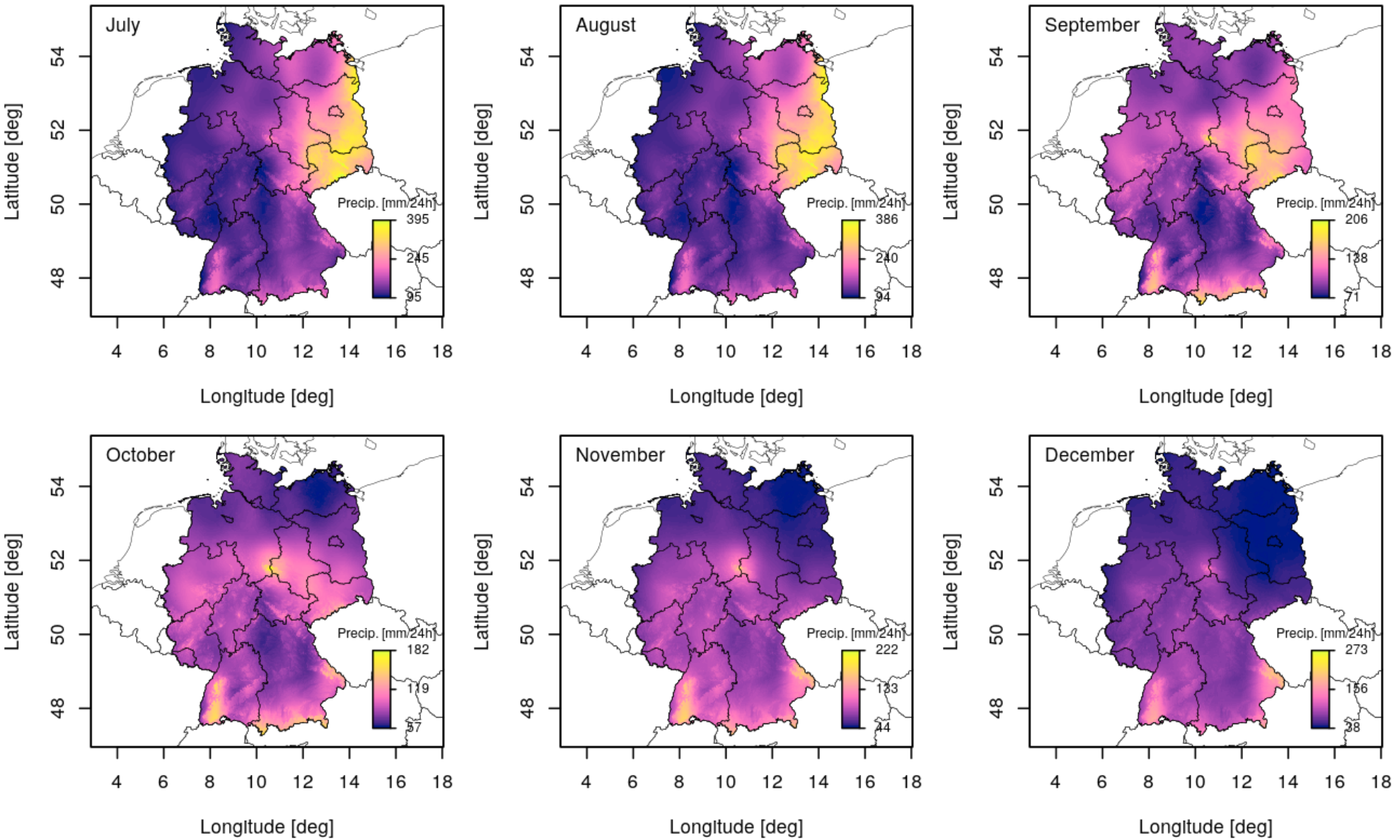

Effect on log(ξ)
1.0
0.5
0.0
-0.5
-1.0
North [52.41-54.91 deg N]
Middle [49.9-52.41 deg N]
South [47.4-49.9 deg N]
Main effect
Jan
Mar
May
Jul
Sep
Nov
Jan
Time


Effect on log(σ)
0.4
0.2
0.0
-0.2
-0.4
0
500
1000
1500
Altitude


Effect on log(σ)
0.4
0.2
0.0
-0.2
-0.4
1940
1960
1980
2000
Time


Effect on log(σ)
0.4
0.2
0.0
-0.2
-0.4
North [52.41-54.91 deg N]
Middle [49.9-52.41 deg N]
South [47.4-49.9 deg N]
Main effect
Jan
Mar
May
Jul
Sep
Nov
Jan
Time


January
Precip. [mm/24h]
301
169
37
February
Precip. [mm/24h]
295
165
35
March
Precip. [mm/24h]
225
130
36
April
Precip. [mm/24h]
177
109
41
May
Precip. [mm/24h]
246
155
84
June
Precip. [mm/24h]
256
169
82
Latitude [deg]
Longitude [deg]
48
50
52
54
4
6
8
10
12
14
16
18


July
Precip. [mm/24h]
395
245
95
August
Precip. [mm/24h]
386
240
94
September
Precip. [mm/24h]
206
138
71
October
Precip. [mm/24h]
182
119
57
November
Precip. [mm/24h]
222
133
44
December
Precip. [mm/24h]
273
156
38
Latitude [deg]
48
50
52
54
Longitude [deg]
4
6
8
10
12
14
16
18

Conditional Transformation Models

- Transform the responses towards a fully specified reference distribution, i.e.

$$h_{\boldsymbol{x}_i}(y_i) \overset{\mathcal{D}}{=} z_i, \quad z_i \overset{\text{i.i.d.}}{\sim} p_{\text{ref}},$$

  where $h_{\boldsymbol{x}_i}(\cdot)$ denotes the transformation function.

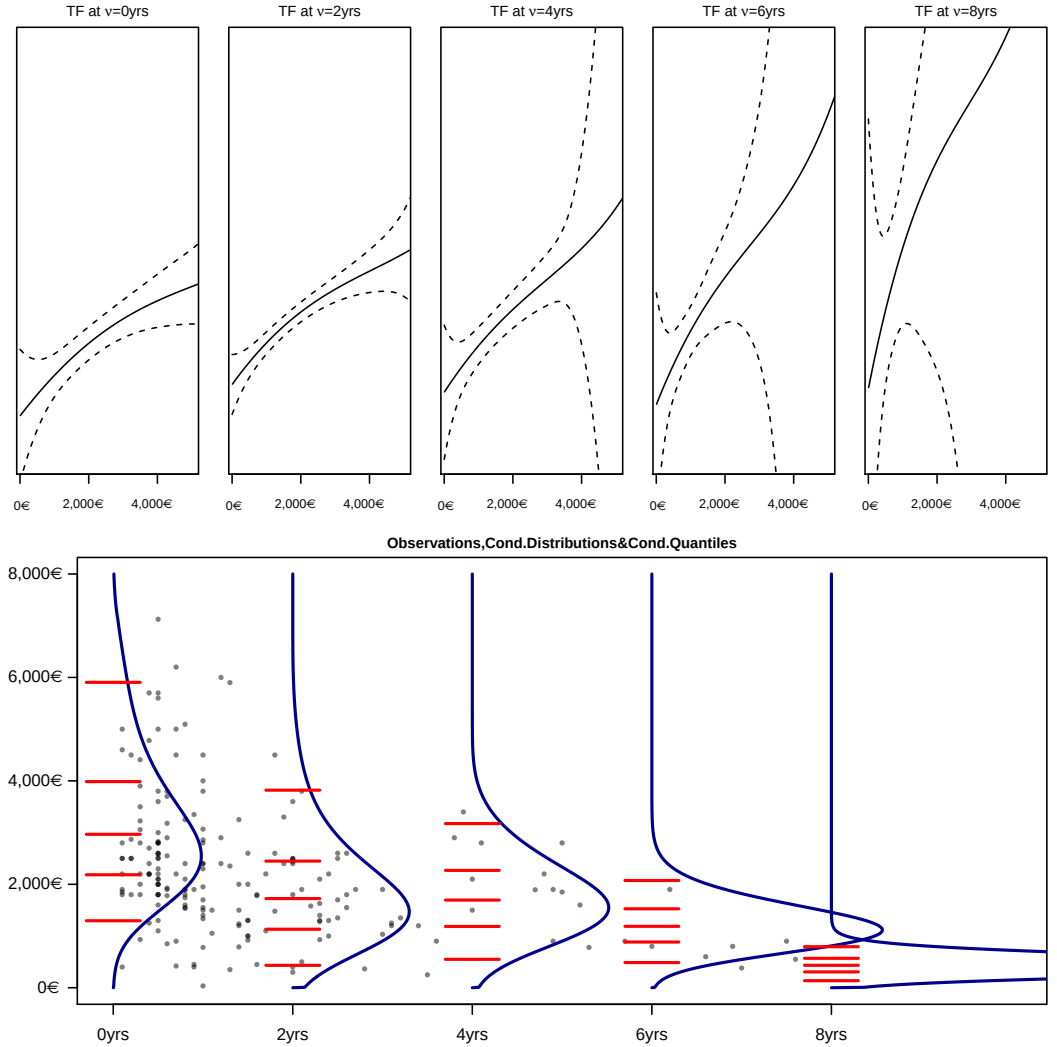
- Implies the density

$$p(y_i|\boldsymbol{x}_i) = p_{\text{ref}}(h_{\boldsymbol{x}_i}(y_i)) \left| \frac{\partial h_{\boldsymbol{x}_i}(y_i)}{\partial y_i} \right|$$

  and cumulative distribution function

$$\mathbb{P}(y_i \leq c|\boldsymbol{x}_i) = F_{y_i}(c|\boldsymbol{x}_i) = F_{\text{ref}}(h_{\boldsymbol{x}_i}(c)) = \mathbb{P}(z_i \leq h_{\boldsymbol{x}_i}(c)).$$

- Illustration with an analysis of income as a function of unemployment duration:



- Major advantages:
  - Encapsulates various statistical models as special cases.
  - Extensions to discrete, censored, or multivariate responses are available and can be treated in a unifying framework.
  - Also likelihood-based, such that standard ways of fitting and checking models are available.
- Main challenges:
  - Setting up the transformation function in a way that combines flexibility with feasibility.
  - The transformation function is specified on the "wrong" scale making interpretation challenging.
  - User-friendly software still under development.

Quantile and Expectile Regression

- Focus on local properties of the response distribution indexed by an extremeness parameter $\tau \in (0, 1)$.
- Formulate a local model specification where

$$y_i = \eta_\tau(\boldsymbol{x}_i) + \varepsilon_{i\tau}, \quad i = 1, \ldots, n,$$

  and augment this with suitable assumptions for the error terms.
- For example,

$$Q_{\varepsilon_{i\tau}}(\tau) = F^{-1}_{\varepsilon_{i\tau}}(\tau) = 0$$

  leads to quantile regression where the τ-quantile of the conditional response distribution is

$$Q_{y_i}(\tau) = \eta_\tau(\boldsymbol{x}_i),$$

- Nonparametric estimation based on weighted error criterion

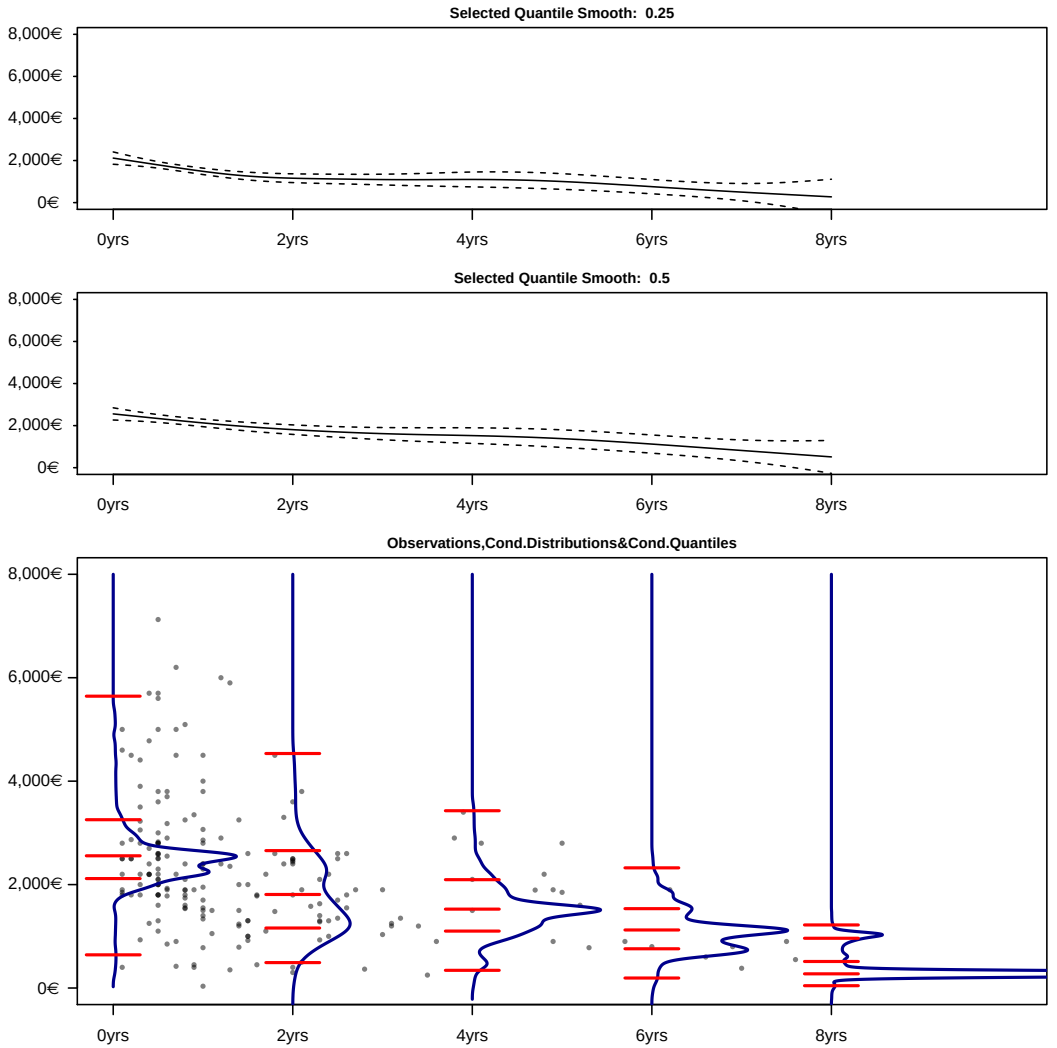
$$\sum_{i=1}^{n} w_\tau(y_i, \eta_{i\tau}(\boldsymbol{x}_i))|y_i - \eta_{i\tau}(\boldsymbol{x}_i)|^p$$

with weights

$$w_\tau(y_i, \eta_{i\tau}(\boldsymbol{x}_i)) = \begin{cases} \tau & y_i > \eta_{i\tau}(\boldsymbol{x}_i) \\ 1-\tau & \text{otherwise} \end{cases}$$

leading to quantile ($p = 1$) and expectile ($p = 2$) regression.

- Illustration with the income data:



- Major advantages:
  - Nonparametric approach that avoids the necessity of a global probabilistic model.
  - Robustness of quantile regression.
  - Link to risk measures for expectile regression.
- Main challenges:
  - Local estimates that are not necessarily consistent with each other (quantile crossing).
  - Inferential approaches more complex due to the absence of likelihoods.
  - Non-differentiable optimisation criterion for quantile regression.
  - Bayesian approaches only based on auxiliary distributional assumptions.

Summary

- Distributional regression modeling as a way to go "beyond the mean" in regression analyses.
- Bayesian inference convenient, but not necessary.
- The basic methods are there (including software), but there is still room for improvements as well as relevant and novel applications.