Karlsruher Institut für Technologie

Website

National University of Singapore

Distributional Regression — Models and Applications

7th Joint Statistical Meeting of the Deutsche Arbeitsgemeinschaft Statistik

Nadja Klein[1] Lucas Kock[2]

[1] Scientific Computing Center, Karlsruhe Institute of Technology
[2] Department of Statistics and Data Science, National University of Singapore

Overview

- What is distributional regression and for which kind of data is it useful?
- How to perform (Bayesian) inference in distributional regression models.
- Distributional regression in practice: implementation, interpretation of estimates, model choice.
- Exercises on applications of distributional regression models.
- At the end of the tutorial you should be able to
  - understand the basic structure of distributional regression models, and
  - apply distributional regression models to your data using existing software

What are distributional regression models?

Components and examples

- The most common regression model is the linear model

$$y_i = \mathbf{x}^\top \boldsymbol{\beta} + \varepsilon_i = \beta_0 + \beta_i x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2).$$

- The parameters β_j can be 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 Gaussian linear model is not flexible enough and/or assumptions may be questionable.

What are distributional regression models?

Components and examples

- Instead of just $\mathbb{E}[y \mid x]$ the full distribution $Y \mid x$ is modeled.
- There is many different approaches to distributional regression within the statistical and machine learning literature.
- We focus on the GAMLSS framework
  - Additive predictors η can combine different effect types
  - The direct link between the distributional parameters θ and their predictors η makes the effects well interpretable
  - User friendly software packages are available
- Many alternatives are available: Quantile regression, transformation models, conditional normalizing flows, etc.

What are distributional regression models?

Components and examples

- **Example 1:** 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.

What are distributional regression models?

Components and examples

- Covariates:

Variable	Description
`vage`	Vehicle’s age.
`page`	Policyholder’s age.
`hp`	Vehicle’s horsepower.
`bm`	Bonus-malus score.
`s`	District in Belgium.
x	Vector of categorical covariates.

What are distributional regression models?

Components and examples

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

$$\log(y) \sim \mathrm{N}(\mathbf{x}^\top \boldsymbol{\beta}, \sigma^2).$$

  - Poisson model for frequencies h_i:

$$h \sim \mathrm{Po}(\exp(\mathbf{x}^\top \boldsymbol{\beta})).$$

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

$$\eta = \mathbf{x}^\top \boldsymbol{\beta} = \beta_0 + x_1\beta_1 + \ldots + x_p\beta_p.$$

What are distributional regression models?

Components and examples

- 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 $f(\texttt{page})$.
  - 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 $\beta_1\texttt{page} + \beta_2\texttt{sex} + \beta_3\texttt{page} \cdot \texttt{sex}$ consider a model of the form $f_1(\texttt{page}) + \texttt{sex} \cdot f_2(\texttt{page})$.

What are distributional regression models?

Components and examples

- 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 $f_{\text{spat}}(\mathfrak{s})$ based on the district $\mathfrak{s}$ a driver is living in.

$\Rightarrow$ Consider geoadditive regression models.

What are distributional regression models?

Components and examples

- Flexible non-linear effects:
  - Gaussian model for log-costs $\log(y)$:

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

with

$$\eta = f_1(\texttt{vage}) + f_2(\texttt{page}) + f_3(\texttt{bm}) + f_4(\texttt{hp}) + f_{spat}(\texttt{s}) + \mathbf{x}^\top \boldsymbol{\beta}.$$

  - Poisson model for frequencies h_i:

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

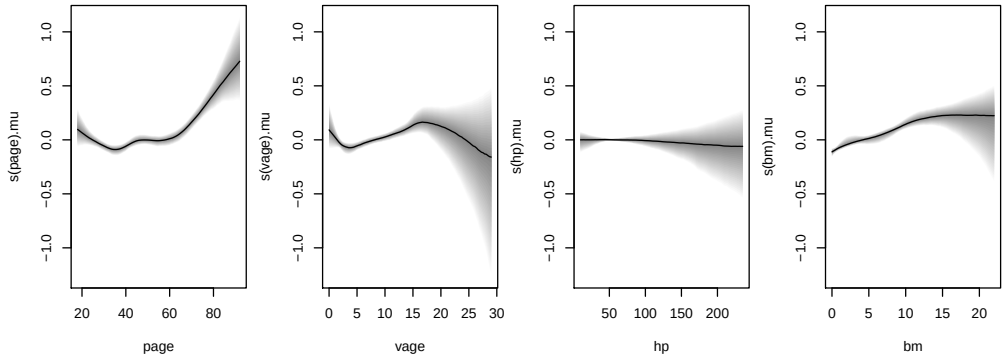
with

$$\eta = f_1(\texttt{vage}) + f_2(\texttt{page}) + f_3(\texttt{page}) \cdot \texttt{sex} + f_3(\texttt{bm}) + f_4(\texttt{hp}) + f_{spat}(\texttt{s}) + \mathbf{x}^\top \boldsymbol{\beta}.$$

What are distributional regression models?

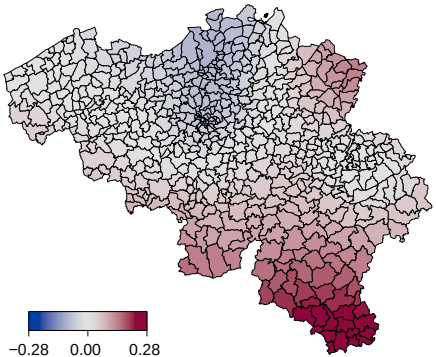
Components and examples

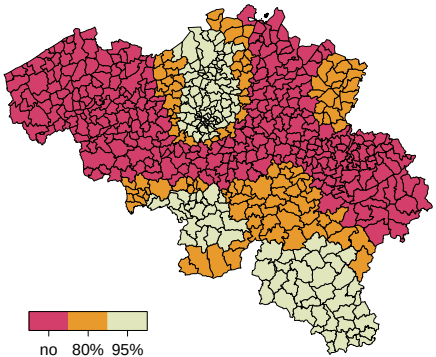
- Results for claim size:



What are distributional regression models?

Components and examples

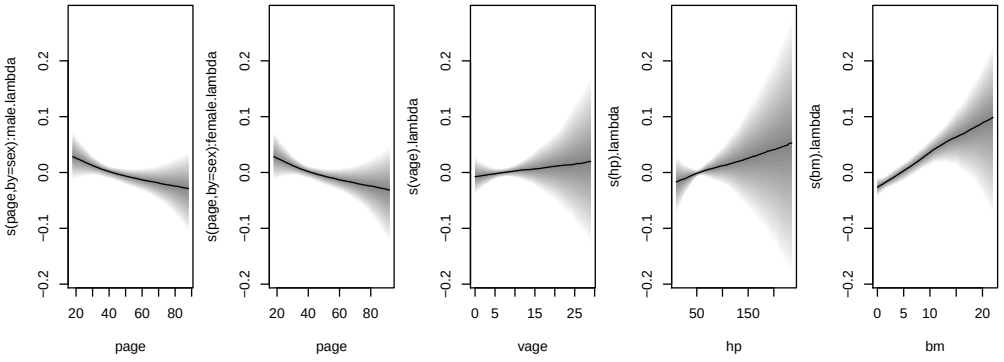




What are distributional regression models?

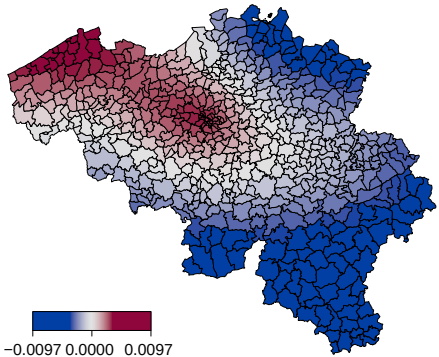
Components and examples

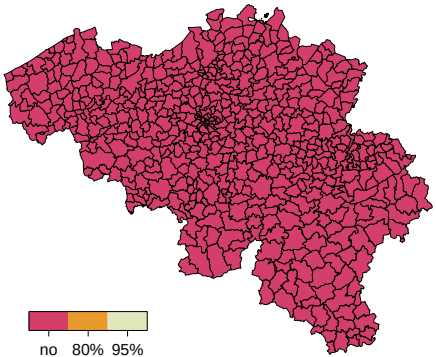
- Results for claim frequency:



What are distributional regression models?

Components and examples





What are distributional regression models?

Components and examples

- **Example 2:** Forest health data.
- 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:

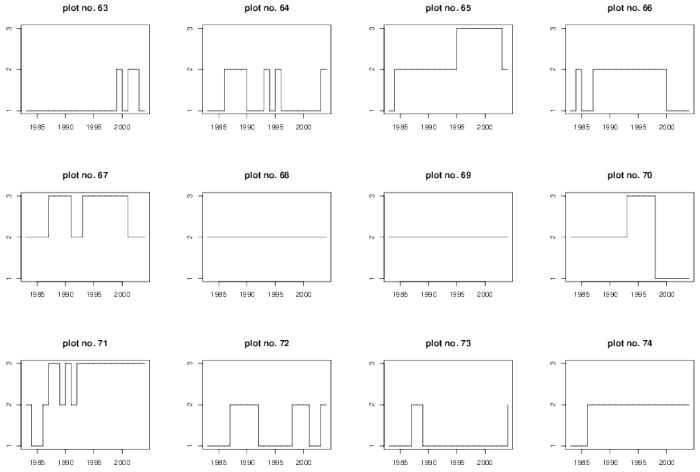
  $y_{it} = 1$ no defoliation,

  $y_{it} = 2$ defoliation 25% or less,

  $y_{it} = 3$ defoliation above 25%.

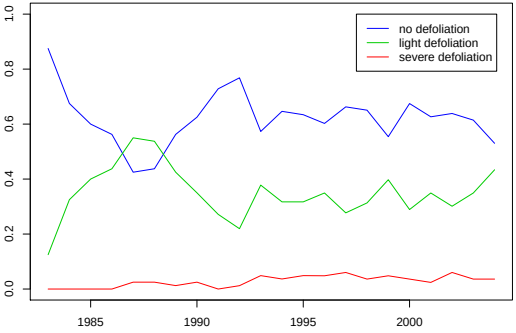
What are distributional regression models?

Components and examples



What are distributional regression models?

Components and examples

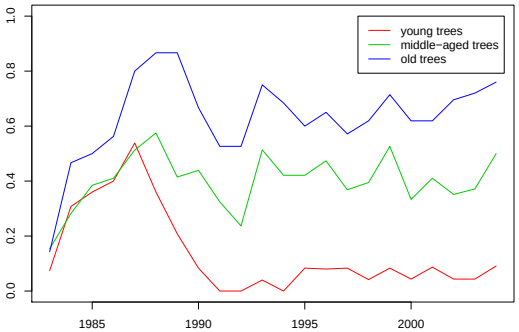


Empirical time trends.

What are distributional regression models?

Components and examples

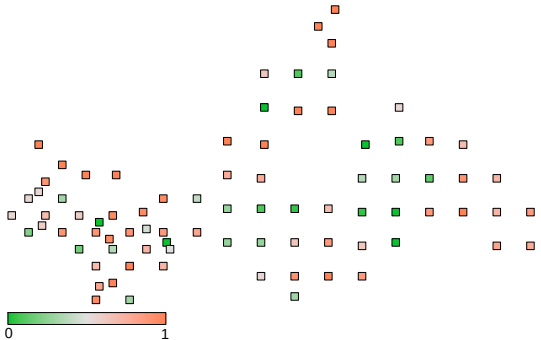
Trends for different ages.



What are distributional regression models?

Components and examples

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



What are distributional regression models?

Components and examples

- 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).

What are distributional regression models?

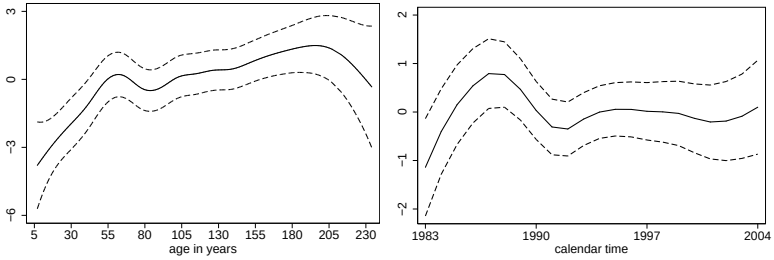
Components and examples

- Additive predictor:

$$
\begin{aligned}
\eta_{it} \quad = \quad & f_1(\texttt{age}_{it}) && \text{nonlinear effects of age,} \\
& + f_2(\texttt{inc}_i) && \text{inclination of slope, and} \\
& + f_3(\texttt{can}_{it}) && \text{canopy density.} \\
& + f_{\texttt{time}}(t) && \text{nonlinear time trend.} \\
& + f_4(t, \texttt{age}_{it}) && \text{interaction between age and} \\
& && \text{calendar time.} \\
& + f_{\texttt{spat}}(s_i) && \text{structured and} \\
& + b_i && \text{unstructured spatial effects.} \\
& + \mathbf{x}_{it}^\top \beta && \text{usual parametric effects.}
\end{aligned}
$$

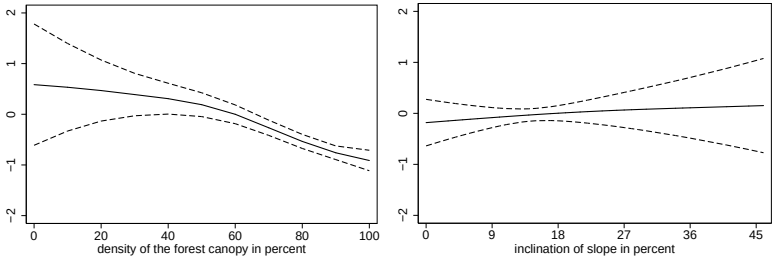
What are distributional regression models?

Components and examples



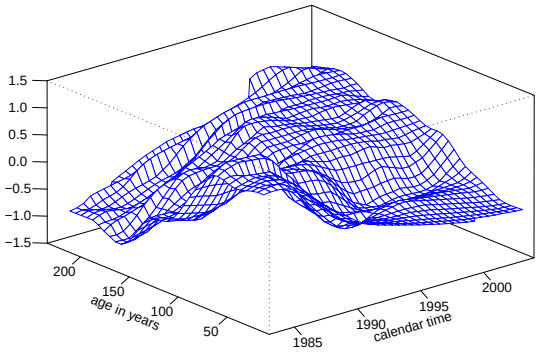
What are distributional regression models?

Components and examples



What are distributional regression models?

Components and examples



What are distributional regression models?

Components and examples

- **Example 3:** Munich rental guide.
- Most larger German cities publish rental guides to provide “average rents” as a reference point for landlords.
- These can be based on regression analyses with rent (e.g. net rent per square meter) as dependent variable.
- Available covariates:

Variable	Description
`size`	floor space in square metres,
`year`	year of construction,
`s`	subquarter in Munich the flat is located in,
`...`	further categorical covariates characterising the flat.

What are distributional regression models?

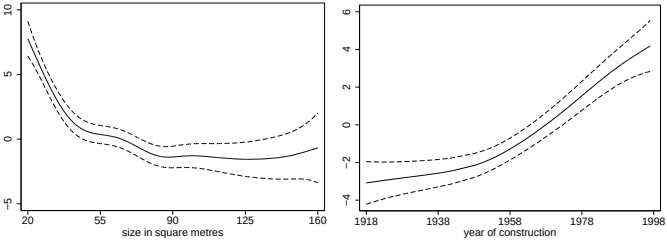
Components and examples

- Standard regression models: Include all covariates in a purely parametric, linear way.
- Geoadditive regression models extend this framework in several directions.
- Additive models include nonlinear effects of continuous covariates:

$$\texttt{netrent} = f_1(\texttt{size}) + f_2(\texttt{year}) + \ldots$$

What are distributional regression models?

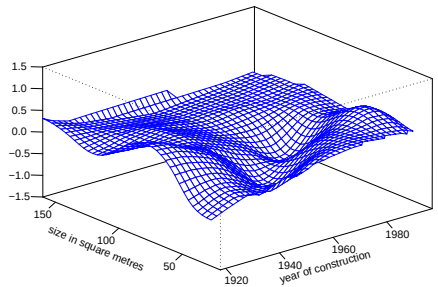
Components and examples



What are distributional regression models?

Components and examples

- Interaction surfaces: Sometimes a purely additive structure is too restrictive, i.e. there are more complex interactions between continuous covariates.



What are distributional regression models?

Components and examples

- Two modelling possibilities: Pure interaction models

$$\texttt{netrent} = f_{1|2}(\texttt{size}, \texttt{year}) + \ldots$$

or interaction models in ANOVA-style

$$\texttt{netrent} = f_1(\texttt{size}) + f_2(\texttt{year}) + f_{1|2}(\texttt{size}, \texttt{year}) + \ldots$$

What are distributional regression models?

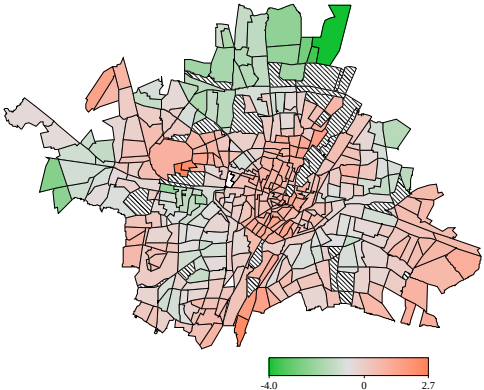
Components and examples

- The flats are spatially aligned within Munich. There may be spatial effects that are not explained by the included covariates.
- The spatial information is given in terms of subquarters $\mathtt{s}$ the flats are located in.
- Geoadditive model:

$$\mathtt{netrent} = f_1(\mathtt{size}) + f_2(\mathtt{year}) + f_{1|2}(\mathtt{size}, \mathtt{year}) + f_{spat}(\mathtt{s}) \ldots$$

What are distributional regression models?

Components and examples



What are distributional regression models?

Components and examples

- Ignoring spatial information can introduce severe bias in the remaining estimates and, in particular, may lead to underestimation of standard errors due to spatial correlations!

What are distributional regression models?

Components and examples

- Summary: Why are structured additive regression models useful?
- They offer a much larger flexibility compared to parametric models and in particular allow to
  - Detect nonlinearity of the influence of some explanatory variables,
  - Model complex interactions,
  - Account for correlations, introduced for example by spatial or temporal alignment of the data.

What are distributional regression models?

Components and examples

- Regression models may contain arbitrary combinations of the following terms:
  - Parametric effects $\mathbf{x}^\top \beta$,
  - Nonlinear effects $f(x)$ of continuous covariates x,
  - Interaction surfaces $f(x_1, x_2)$,
  - Spatial effects $f_{spat}(s)$ based on either point-referenced data $s = (s_x, s_y)$ or georeferenced data $s \in \{s_1, \ldots, s_S\}$,
  - Varying coefficient terms $uf(x)$ with continuous effect modifier or $uf(s)$ with spatial effect modifier.
- We will not work with all these possibilities in this course but will focus on nonlinear effects, varying coefficients and spatial effects.

Penalized Spline Smoothing

Scatterplot Smoothing:

- Given data (y_i, x_i) following the model

$$y_i = f(x_i) + \varepsilon_i, \quad \mathrm{E}(\varepsilon_i) = 0, \quad \mathrm{Var}(\varepsilon_i) = \sigma^2,$$

  infer the function f.
- Usual assumption: f is smooth (in a sense to be made precise later-on).
- Basis function approaches: Represent $f(x)$ as a linear combination of (suitably chosen) basis functions $B_j(x)$, i.e.

$$f(x) = \sum_{j=1}^{d} \gamma_j B_j(x).$$

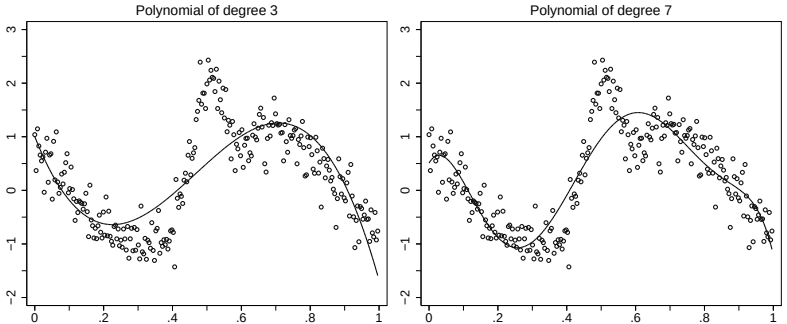
Penalized Spline Smoothing

- Main advantage: Still fits in the framework of linear models.
- The simplest basis functions are polynomials

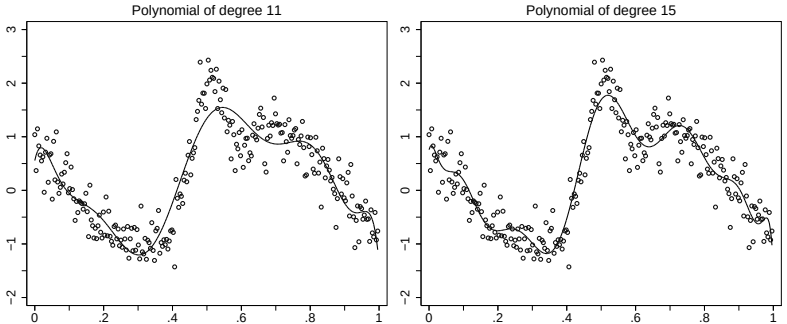
$$f(x) = \gamma_0 + \gamma_1 x + \ldots + \gamma_{d-1} x^{d-1}.$$

- However:
  - Polynomials are not very flexible.
  - Polynomials are prone to produce artificial behaviour especially at the boundaries and when extrapolating.

Penalized Spline Smoothing



Penalized Spline Smoothing

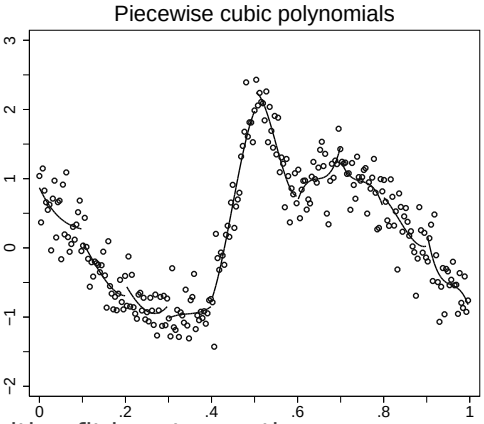


Penalized Spline Smoothing

Polynomial splines

- Aim: Construct more flexible and reliable basis functions.
- First idea: Consider piecewise polynomials, i.e. divide the codomain of the covariate into nonoverlapping intervals and fit separate polynomials.

Penalized Spline Smoothing



- Problem: The resulting fit is not smooth.

Penalized Spline Smoothing

$\Rightarrow$ Add continuity conditions at the interval boundaries.

- This naturally leads to the following class of basis functions:
  A function $f : [a, b] \to \mathbb{R}$ is called a polynomial spline of degree l, $l \in \mathbb{N}_0$, based on knots $a = \kappa_1 < \ldots < \kappa_m = b$, if it satisfies the following conditions:

  1. $f(x)$ is $(l-1)$ times continuously differentiable and
  2. $f(x)$ is a polynomial of degree l for $x \in [\kappa_j, \kappa_{j+1})$, $j = 1, \ldots, m-1$.

Penalized Spline Smoothing

- The space of polynomial splines is a $(m + l - 1)$-dimensional vector space and a subspace of the space of $(l - 1)$ times continuously differentiable functions.
- Hence every polynomial spline can be represented by a set of $d = m + l - 1$ basis functions, i.e.

$$f(x) = \sum_{j=1}^{d} \gamma_j B_j(x).$$

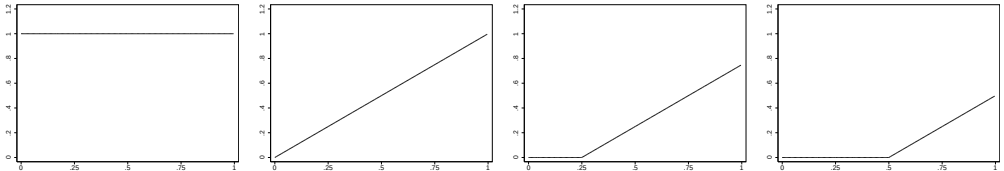
Penalized Spline Smoothing

- A simple and intuitive basis: The truncated power series

$$B_1(x) = 1, \quad B_2(x) = x, \quad \ldots, \quad B_{l+1}(x) = x^l,$$
$$B_{l+2}(x) = (x - \kappa_2)_+^l, \quad \ldots, \quad B_d(x) = (x - \kappa_{m-1})_+^l,$$

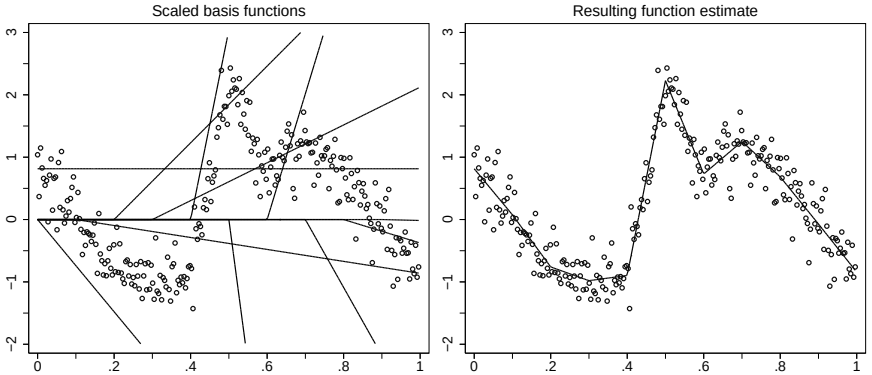
where

$$(x - \kappa_j)_+^l = \begin{cases} (x - \kappa_j)^l & x \geq \kappa_j, \\ 0 & \text{else.} \end{cases}$$



Penalized Spline Smoothing

- Fitting polynomial splines in practice:



Penalized Spline Smoothing

- Estimation of the coefficients γ_j is easy. Polynomial splines simply from large linear models

$$\mathbf{y} = \mathbf{Z}\gamma + \varepsilon$$

with

$$\begin{aligned}
\mathbf{Z} &= \begin{pmatrix} B_1(x_1) & \dots & B_d(x_1) \\ \vdots & & \vdots \\ B_1(x_n) & \dots & B_d(x_n) \end{pmatrix} \\
&= \begin{pmatrix} 1 & x_1 & \dots & x_1^l & (x_1-\kappa_2)_+^l & \dots & (x_1-\kappa_{m-1})_+^l \\ \vdots & & & & & & \vdots \\ 1 & x_n & \dots & x_n^l & (x_n-\kappa_2)_+^l & \dots & (x_n-\kappa_{m-1})_+^l \end{pmatrix}.
\end{aligned}$$

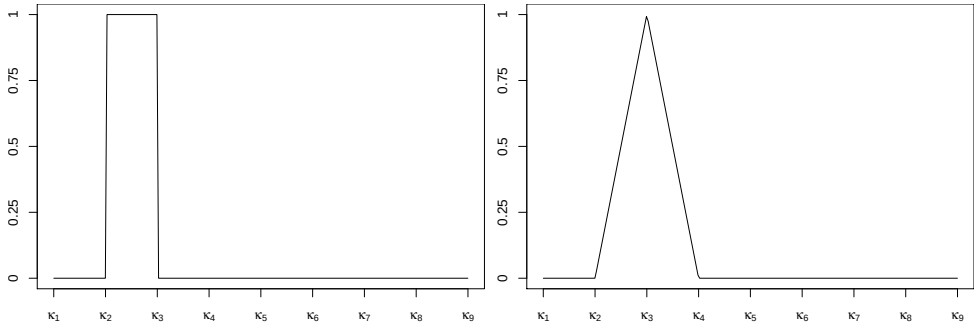
Penalized Spline Smoothing

$\Rightarrow$ Usual least-squares estimator $\hat{\gamma} = (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{y}$ and

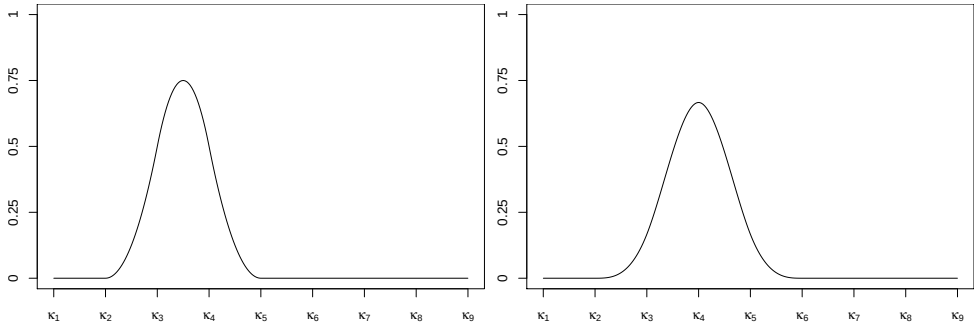
$$\hat{\mathbf{f}} = (\hat{f}(x_1), \ldots, \hat{f}(x_n))^\top = \mathbf{Z}\hat{\gamma} = \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{y}.$$

- The truncated power series basis is easy to understand, but
  1. Requires the evaluation of polynomials to the power l. This bears the risk of overflow errors (may be alleviated by standardising x).
  2. The basis functions are linearly independent from a theoretical perspective but may be numerically collinear (especially with knots close to each other).
- A more flexible and reliable basis: Basic-splines (or simply B-splines).
- B-splines of degree l are obtained by fusing $l + 1$ polynomials of degree l smoothly at $l - 1$ inner knots.

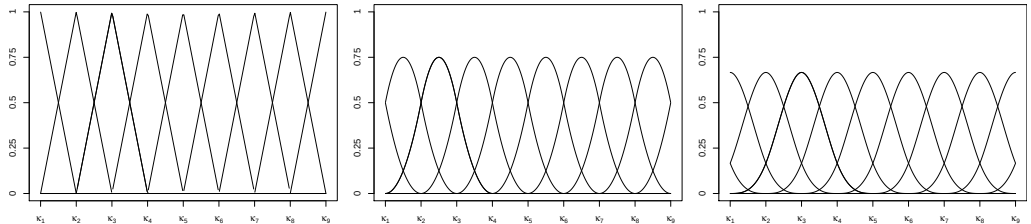
Penalized Spline Smoothing



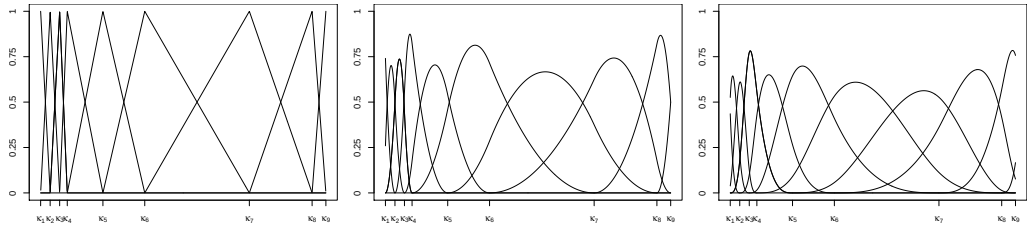
Penalized Spline Smoothing



Penalized Spline Smoothing



Penalized Spline Smoothing



Penalized Spline Smoothing

- Mathematical definition:
  B-splines of degree $l = 0$:

$$B_j^0(x) = \mathbb{1}_{[\kappa_j, \kappa_{j+1})}(x) = \begin{cases} 1 & \kappa_j \leq x < \kappa_{j+1}, \\ 0 & \text{otherwise.} \end{cases}$$

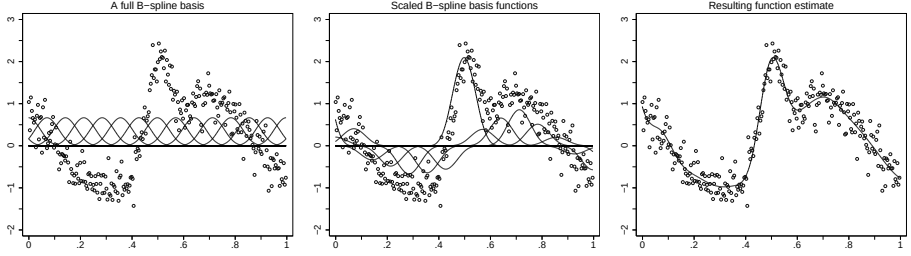
  Higher order B-splines are defined recursively as

$$B_j^l(x) = \frac{x - \kappa_j}{\kappa_{j+l} - \kappa_j} B_j^{l-1}(x) + \frac{\kappa_{j+l+1} - x}{\kappa_{j+l+1} - \kappa_{j+1}} B_{j+1}^{l-1}(x),$$

Penalized Spline Smoothing

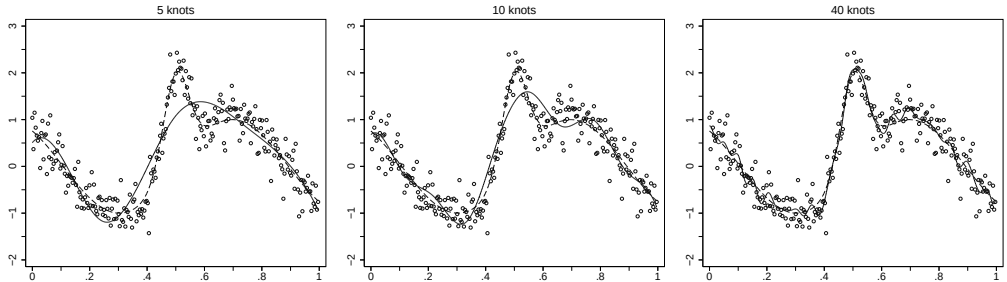
- B-splines also determine a large linear model $\mathbf{y} = \mathbf{Z}\gamma + \varepsilon$ with

$$\mathbf{Z} = \begin{pmatrix} B^l_{1-l}(x_1) & \dots & B^l_{m-1}(x_1) \\ \vdots & & \vdots \\ B^l_{1-l}(x_n) & \dots & B^l_{m-1}(x_n) \end{pmatrix}.$$



Penalized Spline Smoothing

- What determines the functional form of a polynomial spline fit?
  - The degree of the basis (can be chosen from theoretical considerations about the smoothness of f).
  - The number and position of the knots (this is the crucial question!).



Penalized Spline Smoothing

Penalisation approaches

- Basic idea of penalisation approaches:
  - Approximate the function f using a large set of knots to obtain a flexible function estimate.
  - Penalise large variation of f with an appropriate penalty term on the coefficient vector γ.
- For the TP-basis with

$$f(x) = \gamma_1 + \gamma_2 x + \ldots + \gamma_{l+1} x^l + \gamma_{l+2}(x - \kappa_2)_+^l + \ldots + \gamma_d (x - \kappa_{m-1})_+^l$$

variation in the function estimate is mostly introduced by too wiggly truncated polynomials.

$\Rightarrow$ Place a ridge penalty $\sum_{j=l+2}^{d} \gamma_j^2$

Penalized Spline Smoothing

on the corresponding coefficients and minimise the penalised least squares criterion

$$\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{d} \gamma_j B_j(x_i) \right)^2 + \lambda \sum_{j=l+2}^{d} \gamma_j^2.$$

- The smoothing parameter λ controls the trade off between fidelity to the data (λ small) and smoothness of the function estimate (λ large).
- More specifically $\hat{f}$ approaches
  - a polynomial of degree l for $\lambda \to \infty$,
  - an unpenalised polynomial spline for $\lambda \to 0$.

$\Rightarrow$ The number of free parameters varies smoothly between $l + 1$ and $d = m + l - 1$.

Penalized Spline Smoothing

- The problem of choosing the appropriate number and position of knots is reduced to the selection of an appropriate smoothing parameter.
- Penalisation for B-splines: The penalty term should represent a roughness measure for the function estimate

  $\Rightarrow$ Integrated squared derivative penalties such as

$$\text{pen}(f) = \lambda \int (f''(x))^2 dx.$$

- Derivatives of B-splines are defined in terms of differences of coefficients, e.g. $\gamma_j - \gamma_{j-1}$ for the first derivative.

  $\Rightarrow$ Construct penalty terms using differences to ensure smoothness.

Penalized Spline Smoothing

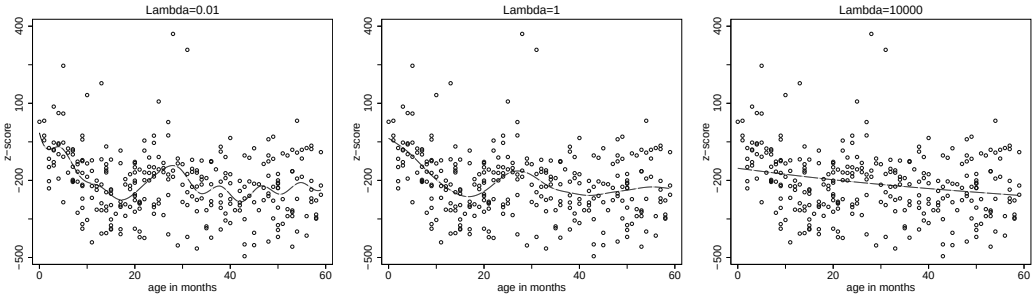
- This yields the penalised least squares objective

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{d}\gamma_j B_j(x_i)\right)^2 + \lambda \sum_{j=k+1}^{d}(\Delta^k \gamma_j)^2,$$

where

$$\begin{aligned}
\Delta^1\gamma_j &= \gamma_j - \gamma_{j-1}, \\
\Delta^2\gamma_j &= \Delta^1\Delta^1\gamma_j = \gamma_j - 2\gamma_{j-1} + \gamma_{j-2}, \\
&\vdots \\
\Delta^k\gamma_j &= \Delta^{k-1}\gamma_j - \Delta^{k-1}\gamma_{j-1}.
\end{aligned}$$

Penalized Spline Smoothing



- The estimate $\hat{f}$ approaches a polynomial of degree $k-1$ for $\lambda \to \infty$.

Penalized Spline Smoothing

Penalised Least Squares Estimation

- Write the penalty term (for B-splines) in matrix notation:

$$\begin{aligned} \lambda \sum_{j=k+1}^{d} (\Delta^k \gamma_j)^2 &= \lambda \boldsymbol{\gamma}^\top \mathbf{D}_k^\top \mathbf{D}_k \boldsymbol{\gamma} \\ &= \lambda \boldsymbol{\gamma}^\top \mathbf{K}_k \boldsymbol{\gamma} \end{aligned}$$

where $\mathbf{K}_k = \mathbf{D}_k^\top \mathbf{D}_k$ is derived from difference matrices $\mathbf{D}_k$, e.g.

$$\mathbf{D}_1 = \begin{pmatrix} -1 & 1 & & \\ & -1 & 1 & \\ & & \ddots & \ddots \\ & & & -1 \quad 1 \end{pmatrix} \qquad (d-1) \times d.$$

Penalized Spline Smoothing

- Recursive definition of higher order difference matrices:

$$\mathbf{D}_k = \mathbf{D}_1 \mathbf{D}_{k-1}.$$

- First order penalty matrix:

$$\mathbf{K}_1 = \begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{pmatrix}.$$

Penalized Spline Smoothing

- Minimisation of the penalised least squares criterion

$$(\mathbf{y} - \mathbf{Z}\gamma)^\top(\mathbf{y} - \mathbf{Z}\gamma) + \lambda\gamma^\top\mathbf{K}\gamma$$

  yields the penalised least squares estimator

$$\hat{\gamma} = (\mathbf{Z}^\top\mathbf{Z} + \lambda\mathbf{K})^{-1}\mathbf{Z}^\top\mathbf{y}.$$

- Properties of penalised least squares estimates:
  - Characterised by the hat matrix

$$\mathbf{S}_\lambda = \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z} + \lambda\mathbf{K})^{-1}\mathbf{Z}^\top.$$

  - PLS-estimates define a linear smoother $\hat{\mathbf{f}} = \mathbf{S}_\lambda\mathbf{y}$.
    $\Rightarrow$ Construct confidence intervals, confidence bands and other quantities in analogy to linear models.

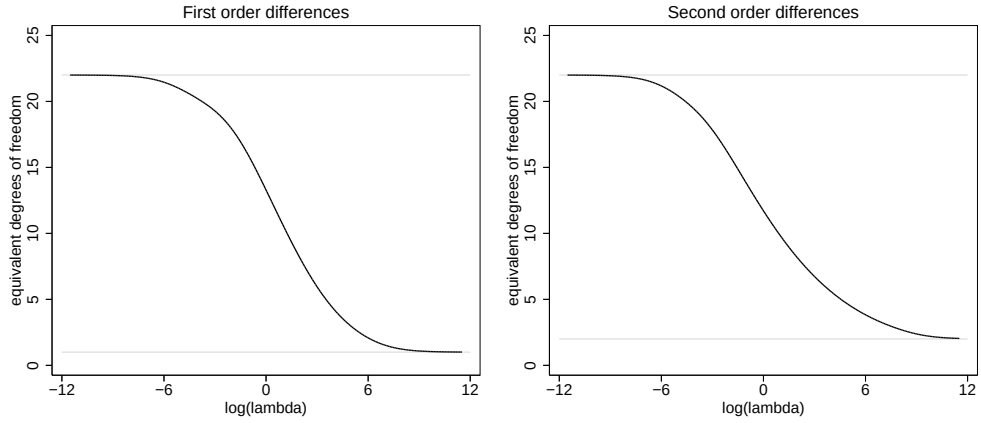
Penalized Spline Smoothing

- Degrees of freedom of a P-spline:

$$\text{df}_\lambda = \text{trace}(\mathbf{S}_\lambda) = \text{trace}((\mathbf{Z}^\top\mathbf{Z} + \lambda\mathbf{K})^{-1}\mathbf{Z}^\top\mathbf{Z}).$$

  In analogy to linear models this is also called the effective number of parameters.

Penalized Spline Smoothing



Penalized Spline Smoothing

Varying Coefficient Models:

- Penalised splines can be used to introduce interactions of the form

$$y_i = \ldots u_i f(x_i) + \ldots$$

  where the effect of the interaction variable u_i varies smoothly over the domain of the effect modifier x_i.
- Typical example: Time-varying effects $uf(t)$ where the effect of covariate u is allowed to vary nonlinearly over time.

Penalized Spline Smoothing

- The function $f(x)$ can then again be approximated by a linear spline and its coefficients can be determined via least squares in the model

$$\mathbf{y} = \mathbf{Z}\gamma + \varepsilon$$

with

$$\mathbf{Z} = \begin{pmatrix} u_1 B^l_{1-l}(x_1) & \dots & u_1 B^l_{m-1}(x_1) \\ \vdots & & \vdots \\ u_n B^l_{1-l}(x_n) & \dots & u_n B^l_{m-1}(x_n) \end{pmatrix}.$$

Penalized Spline Smoothing

Non-Gaussian Responses:

- Instead of Gaussian responses $\mathbf{y} \sim \mathrm{N}(h(\boldsymbol{\eta}), \sigma^2 \mathbf{I}_n)$ consider exponential family smoothing
$$\mathrm{E}(y|x) = h(\eta), \qquad \eta = f(x).$$
- Representing $f(x)$ as a penalised spline leads to penalised likelihood estimation
$$\ell_{\mathrm{pen}}(\boldsymbol{\gamma}) = \ell(\boldsymbol{\gamma}) - \frac{\lambda}{2}\boldsymbol{\gamma}^\top \mathbf{K}\boldsymbol{\gamma}$$
where $\ell(\boldsymbol{\gamma})$ is the log-likelihood corresponding to a GLM with predictor $\boldsymbol{eta} = \mathbf{Z}\boldsymbol{\gamma}$.
- For given smoothing parameter, estimation can be achieved by a slight variation of the usual Fisher scoring algorithm.

Penalized Spline Smoothing

- Equivalent representation: Iteratively weighted least-squares estimation

$$\hat{\gamma}^{(k+1)} = (\mathbf{Z}^\top \mathbf{W}^{(k)} \mathbf{Z} + \lambda \mathbf{K})^{-1} \mathbf{Z}^\top \mathbf{W}^{(k)} \tilde{\mathbf{y}}^{(k)}$$

  where $\mathbf{W}$ are the GLM working weights and $\tilde{\mathbf{y}}$ is the GLM working response.
- Both $\mathbf{W}$ and $\tilde{\mathbf{y}}$ depend on the current estimate.
- Repeated weighted penalised least-squares fit to a working response corresponds to the working model

$$\tilde{\mathbf{y}} \sim \mathrm{N}(\mathbf{Z}\gamma, \mathbf{W}^{-1}).$$

Penalized Spline Smoothing

Bayesian P-Splines

- The stochastic analogue to k-th order difference penalties are random walk priors of order k, e.g.

$$\gamma_j = \gamma_{j-1} + u_j, \quad u_j \sim \mathrm{N}(0, \tau^2), \quad j = 2, \ldots, d,$$

  for a first order random walk and

$$\gamma_j = 2\gamma_{j-1} - \gamma_{j-2} + u_j, \quad u_j \sim \mathrm{N}(0, \tau^2), \quad j = 3, \ldots, d,$$

  for a second order random walk.
- Noninformative priors $p(\gamma_1) \propto const$ and $p(\gamma_1, \gamma_2) \propto const$ are assigned to the starting values.

Penalized Spline Smoothing

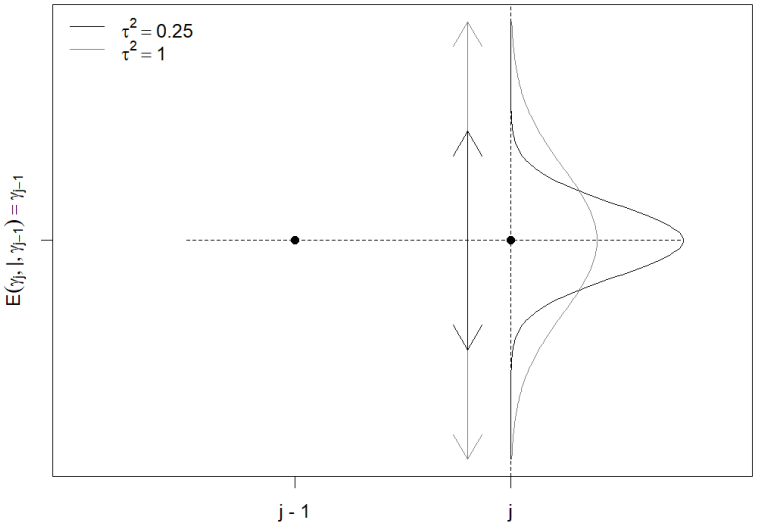
- The RW1 can be equivalently expressed as

$$\gamma_j - \gamma_{j-1} = u_j, \quad u_j \sim \mathrm{N}(0, \tau^2),$$

or

$$\gamma_j | \gamma_{j-1}, \ldots, \gamma_1 \sim \mathrm{N}(\gamma_{j-1}, \tau^2).$$

Penalized Spline Smoothing



Markov Random Fields

Motivation

- Disease Mapping: Spatial analysis of mortality due to a certain (noninfectious, rare) disease.
- We will use oral cavity cancer of males in Germany between 1985 and 1990 as an example.
- The data consist of observed and expected mortality counts in spatial regions such as the German districts (Landkreise).

Markov Random Fields

- Descriptive analysis: Compute the standardised mortality ratios

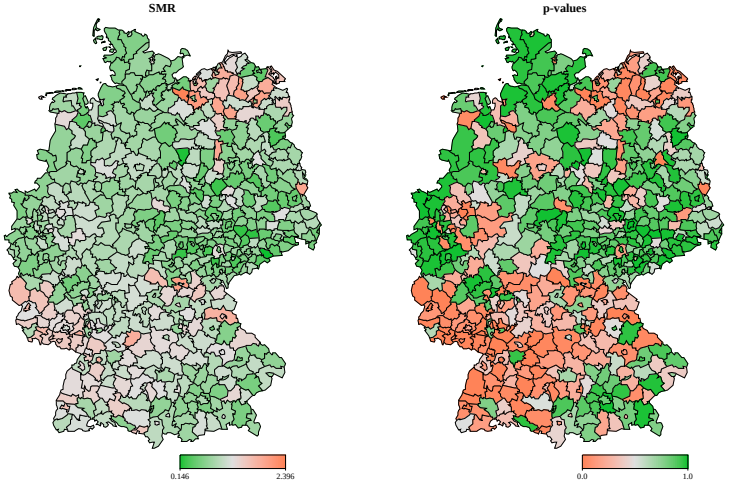
$$\mathrm{SMR}_i = \frac{y_i}{e_i} = \frac{\text{observed count in region } i}{\text{expected count in region } i}$$

or p-values

$$p_i = P(Y_i \geq y_i)$$

based on the assumption $Y_i \sim \mathrm{Po}(e_i)$ and visualise them.

Markov Random Fields



Markov Random Fields

- Both approaches have the conceptual drawback that they do not account for spatial correlations.
- Spatial smoothing would also allow to borrow strength from neighbouring regions.
- For example, we may parameterise the region-specific risk factors as

$$r_i = \exp(f_{\text{spat}}(s_i))$$

and assume

$$y_i \sim \text{Po}(e_i r_i)$$

with e_i as offset.

- The region index is a discrete spatial variable $s \in \{1, \ldots, S\}$.

Markov Random Fields

- Intuitive idea: Rates of neighboring regions should be close to each other.
- More generally, consider spatial regression models with predictors

$$\eta = \ldots + f_{\text{spat}}(s) + \ldots$$

  where $s \in \{1, \ldots, S\}$ is a discrete spatial variable.
- Parameterise $f_{\text{spat}}(s)$ in terms of region-specific regression coefficients

$$\gamma_s = f_{\text{spat}}(s), \quad s = 1, \ldots, S.$$

Markov Random Fields

- Penalise large deviations between neighbouring coefficients to obtain a spatially smooth effect:

$$\sum_{s=1}^{S} \sum_{r \in \delta_s, r<s} (\gamma_s - \gamma_r)^2 = \boldsymbol{\gamma}^\top \mathbf{K} \boldsymbol{\gamma}$$

  where δ_s contains the indices of neighbours of region s, $N_s = |\delta_s|$ denotes the number of such neighbors and the penalty matrix has elements

$$\begin{aligned} \mathbf{K}[s,s] &= N_s \\ \mathbf{K}[s,r] &= \begin{cases} -1 & \text{if } s \text{ and } r \text{ are neighbours} \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

- Extends the penalty of penalised splines where we also considered squared differences of adjacent parameters.

Markov Random Fields

- Bayesian formulation: Assume a priori that

$$\gamma_s|\boldsymbol{\gamma}_{-s} \sim \mathrm{N}\left(\frac{1}{N_s}\sum_{r\in\delta_s}\gamma_r, \frac{\tau^2}{N_s}\right)$$

  where $\boldsymbol{\gamma}_{-s}$ denotes the vector of spatial effects excluding γ_s.
- Rationale underlying this prior choice:
  - The (conditional) prior expectation for the spatial effect γ_s equals the average of the neighbouring values.
  - The precision of this prior assumption depends on the prior variance τ^2 and the number of neighbours N_s.
  - γ_s is conditionally independent of all non-neighbours.

Markov Random Fields

- It turns out that the corresponding joint prior distribution is given by

$$p(\boldsymbol{\gamma}|\tau^2) \propto \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\gamma}^\top \mathbf{K}\boldsymbol{\gamma}\right)$$

  where $\mathbf{K}$ coincides with the penalty matrix from above.
- The prior is partially improper since the rows (and columns) of $\mathbf{K}$ sum to zero.
- This is the simplest example of an (intrinsic) Gaussian Markov random field (GMRF) prior.

Modelling spatial effects with GMRFs

- The vector of spatial effects $\mathbf{f}_{\text{spat}} = (f_{\text{spat}}(s_1), \ldots, f_{\text{spat}}(s_n))^\top$ can be written as

$$\mathbf{f}_{\text{spat}} = \mathbf{Z}\boldsymbol{\gamma}$$

where

$$\mathbf{Z}[i, s] = \begin{cases} 1 & \text{if } y_i \text{ has been observed in region } s \text{ (i.e. } s_i = s) \text{ and} \\ 0 & \text{else} \end{cases}$$

and $f_{\text{spat}}(s) = \gamma_s$.

- IGMRF penalty / prior for $\boldsymbol{\gamma}$:

$$p(\boldsymbol{\gamma}) \propto \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\gamma}^\top \mathbf{K}\boldsymbol{\gamma}\right)$$

Markov Random Fields

- Neighbourhood definitions:
  - For regional data, regions are usually considered as neighbours if they share a common boundary. More general definitions may be needed if for example some regions are islands or if the map is separated.
  - We can also introduce additional neighbourhoods for example between areas connected by direct flights, etc.

Markov Random Fields

- Disease Mapping (oral cavity cancer): We consider a generalisation of the model discussed at the beginning of the lecture:

$$y_i \sim \mathrm{Po}(\lambda_i)$$

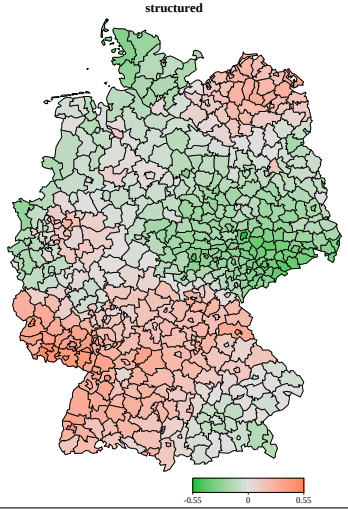
with

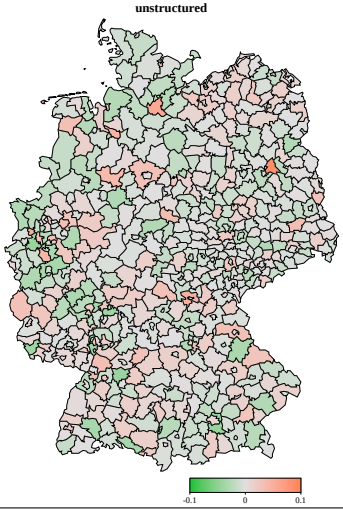
$$\lambda_i = \exp(\log(e_i) + f_{\text{spat}}(s_i) + b_i)$$

where f_{spat} is modelled by a IGMRF while b_i are i.i.d. region-specific random effects.

- The idea is, to relate the risk factors to both structured and unstructured sources of spatial variation.

Markov Random Fields





Why Distributional Regression?

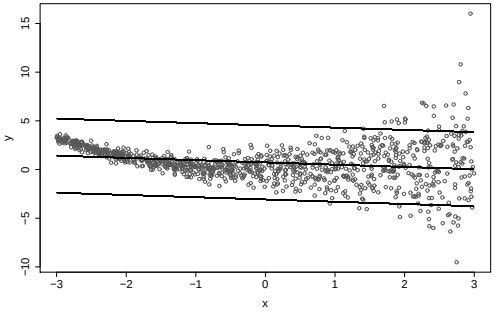
- Classical regression models within the exponential family framework focus on relating the mean of a response y_i to covariate information x_i for observations $(x_1, y_1), \ldots, (x_n, y_n)$.

- Linear model:

$$y_i = \gamma_0 + x_i\gamma + \varepsilon_i$$

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

$$\mathrm{E}(y_i) = \mu_i(x_i) = \gamma_0 + x_i\gamma,$$
$$\mathrm{Var}(y_i) = \sigma^2.$$



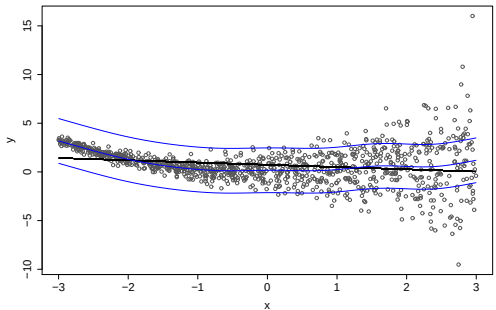
Why Distributional Regression?

- Classical regression models within the exponential family framework focus on relating the mean of a response y_i to covariate information x_i for observations $(x_1, y_1), \ldots, (x_n, y_n)$.

- Nonparametric model:

  $\mathrm{E}(y_i) = \mu_i(x_i) = \gamma_0 + f(x_i)$

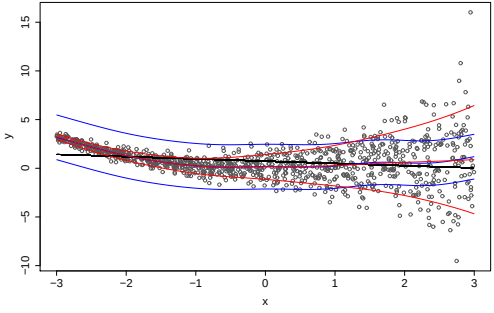
  with $y_i \sim \mathrm{N}\left(\mu_i(x_i), \sigma^2\right)$
  and σ^2 fixed.



Why Distributional Regression?

- Nonparametric model for location and scale:

$$
\begin{aligned}
\mathrm{E}(y_i) &= \mu_i(x_i) = \gamma_0^{\mu} + f^{\mu}(x_i) \\
\mathrm{Var}(y_i) &= \sigma_i^2(x_i) \\
&= \exp\left(\gamma_0^{\sigma^2} + f^{\sigma^2}(x_i)\right) \\
y_i &\sim \mathrm{N}\left(\mu_i(x_i), \sigma_i^2(x_i)\right)
\end{aligned}
$$



Why Distributional Regression?

Interest in regression models that comprise the following features:

- Embedded in a framework that is applicable for **different types of (possibly non-standard) response distributions** (e.g. continuous, discrete and mixed discrete-continuous distributions)
- **Focus on specific aspects of the data** modelled in terms of covariates such as:
  - **Zero inflation**, i.e. excess of zeros compared to standard count data distributions such as Poisson.
  - **Overdispersion**, i.e. variances of the responses exceed the expectation (unlike in Poisson regression).
  - **Heteroscedasticity or skewness** for continuous responses.
  - **Zero inflation** for continuous data.
  - **Correlation** for multivariate responses.

Why Distributional Regression?

- Specify additive predictors 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**.

$\Rightarrow$ **Structured Additive Distributional Regression**

Why Distributional Regression?

Basic idea:

- Observed data pairs $(\mathbf{y}_1, \mathbf{z}_1), \ldots, (\mathbf{y}_n, \mathbf{z}_n)$.
- **Model assumption 1** Conditional distribution of $\mathbf{y}_i$ given the covariate information $\mathbf{z}_i$, $i = 1, \ldots, n$ is from a pre-specified class of K-parametric densities

$$p(\mathbf{y}_i | \vartheta_{i1}, \ldots, \vartheta_{iK}).$$

- **Model assumption 2** Each parameter ϑ_{ik}, $k = 1, \ldots, K$ is related to a regression predictor $\eta_{ik} = \eta_k(\mathbf{z}_i)$

$$\vartheta_{ik} = h_k(\eta_{ik}) \quad \text{and} \quad \eta_{ik} = h_k^{-1}(\vartheta_{ik}).$$

Why Distributional Regression?

Examples:

- $\vartheta_{ik} = \eta_{ik}$ if no restrictions are required (e.g. for expectations),
- $\vartheta_{ik} = \exp(\eta_{ik})$ for positive parameters such as variances,
- $\vartheta_{ik} = \Phi(\eta_{ik})$ for probabilities, or
- $\vartheta_{ik} = \frac{\eta_{ik}}{\sqrt{1+\eta_{ik}^2}}$ for parameters restricted to $[-1, 1]$.

Zero-Inflated and Overdispersed Count Data

- **Count data responses** occur frequently in practice:
  - Explain the number of citations of patents based on patent characteristics.
  - Predict the number of insurance claims of a policyholder based on previous claim history.
  - Model mortality due to a specific type of disease in geographical units (disease mapping).
  - etc.
- Standard approach: **Log-linear Poisson models** embedded in the generalized linear model framework where

$$y_i \sim \mathrm{Po}(\lambda_i), \qquad \text{and} \qquad \lambda_i = \exp(\mathbf{z}_i^\top \boldsymbol{\gamma}).$$

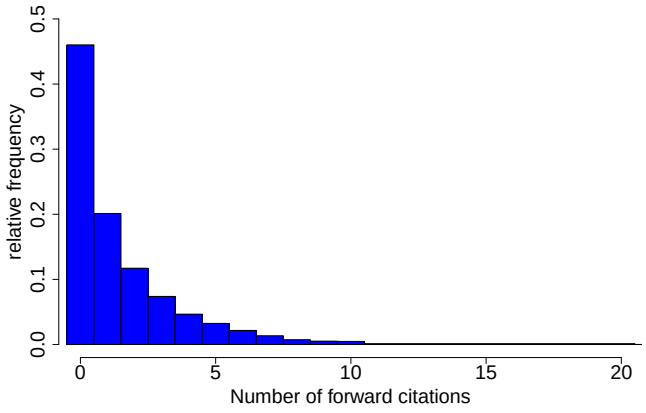
Zero-Inflated and Overdispersed Count Data

Example I: Patent Citations

- Information on 4,805 patents issued by the European Patent Office (EPO) between 1980 and 1997.
- Response variable of interest: **number of forward citations**.
- Explanatory variables include the grant year, the no. of designated states the patent applies to, and the no. of EPO claims.
- Characteristics of the response:
  - **46% zeros** (many patents are never cited).
  - Maximum no. of citations: 40.
  - Average no. of citations: 1.63
  - Variance: 7.35

Zero-Inflated and Overdispersed Count Data

- Frequency histogram:



Zero-Inflated and Overdispersed Count Data

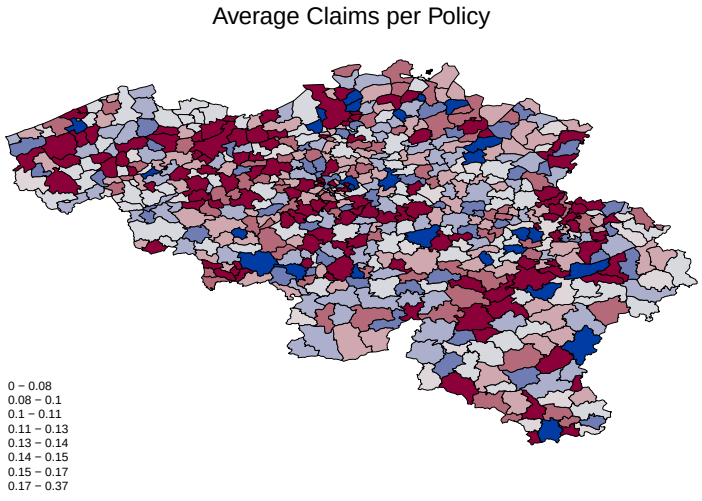
Example II: Claim Frequencies in Car Insurance

- Information on $>$160.000 observations of policyholders in Belgium.
- Response variable of interest: **claim frequency** per policyholder.
- Explanatory variables include the age of the policyholder, the age of the car, the horsepower of the car, the **geographical district** the policyholder lives in, etc.
- Characteristics of the response:
  - $>$**80% zeros** (most policyholders do not cause claims).
  - Maximum no. of claims: 5
  - Average no. of claims: 0.124
  - Variance: 0.135

Zero-Inflated and Overdispersed Count Data

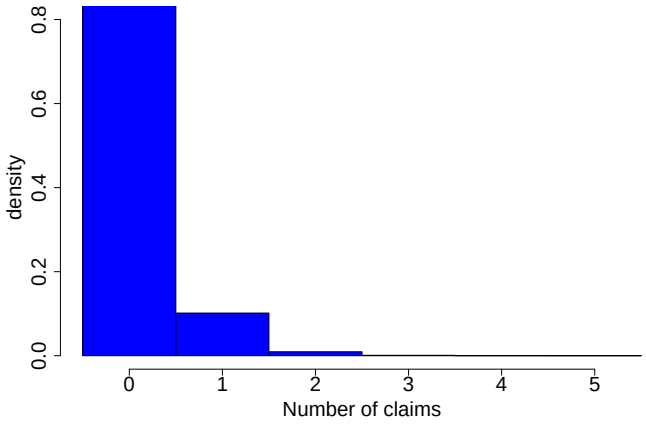
- Geographical information should be included in the regression model (**589 districts**).
- Some covariate effects (e.g. age of the policyholder) are expected to have a **nonlinear effect** on the claim frequency.

Zero-Inflated and Overdispersed Count Data



Zero-Inflated and Overdispersed Count Data

- Frequency histogram:



Zero-Inflated and Overdispersed Count Data

- We need count data regression models that comprise the following features:
  - **Zero inflation**, i.e. excess of zeros compared to standard count data distributions such as Poisson.
  - **Overdispersion**, i.e. variances of the responses exceed the expectation (unlike in Poisson regression).
  - 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.

$\Rightarrow$ Distributional Regression for Zero-Inflated Count Data.

Zero-Inflated and Overdispersed Count Data

- Basic idea of zero-inflated count data regression models: Zeros may arise from either
  - structural zeros, i.e. observations that are “always zero” and,
  - zeros arising from the count data distribution.
- If y_i is a count data response, assume that y_i is generated as

$$y_i = \kappa_i \tilde{y}_i$$

where κ_i is a **binary indicator for structural zeros**, i.e.

$$\kappa_i \sim \mathrm{B}(1 - \pi_i)$$

Zero-Inflated and Overdispersed Count Data

and $\tilde{y}_i$ follows a **standard count data distribution** such as Poisson or negative binomial, i.e.

$$\tilde{y}_i \sim \mathrm{Po}(\lambda_i) \qquad \text{or} \qquad \tilde{y}_i \sim \mathrm{NB}\left(\delta_i, \frac{\delta_i}{\delta_i + \mu_i}\right).$$

Zero-Inflated and Overdispersed Count Data

- Interpretation:
  - If the binary indicator κ_i is zero, we always obtain zero as response (structural zeros).
  - If the binary indicator κ_i is one, y_i is realized from the count data model.
  - Structural zeros occur with probability π_i.
- **Mixed density** for the responses y_i:

$$p(y_i) = \pi_i \mathbb{1}_{\{0\}}(y_i) + (1 - \pi_i)\tilde{p}(y_i)$$

where $\tilde{p}(y_i)$ is the density for the count data distribution.

Zero-Inflated and Overdispersed Count Data

- For the count data part, we will consider the
  - Poisson distribution $\tilde{y}_i \sim \mathrm{Po}(\lambda_i)$ where

$$\mathrm{E}(\tilde{y}_i) = \lambda_i \qquad \text{and} \qquad \mathrm{Var}(\tilde{y}_i) = \lambda_i.$$

  - Negative binomial distribution $\tilde{y}_i \sim \mathrm{NB}(\delta_i, \delta_i/(\delta_i + \mu_i))$ where

$$\mathrm{E}(\tilde{y}_i) = \mu_i \qquad \text{and} \qquad \mathrm{Var}(\tilde{y}_i) = \mu_i + \frac{\mu_i^2}{\delta_i}.$$

Zero-Inflated and Overdispersed Count Data

- For the zero-inflated distribution we obtain

$$\mathrm{E}(y_i) = (1 - \pi_i)\mathrm{E}(\tilde{y}_i)$$

and

$$\mathrm{Var}(y_i) = (1 - \pi_i)\mathrm{Var}(\tilde{y}_i) + \pi_i(1 - \pi_i)\left[\mathrm{E}(\tilde{y}_i)\right]^2 .$$

Zero-Inflated and Overdispersed Count Data

- Regression specification for zero-inflated count data: Relate both the probability for structural zeros and the parameters of the count data distribution to **regression predictors** using suitable link functions, e.g.

$$\eta_i^\pi = \text{logit}(\pi_i) \quad \text{and} \quad \eta_i^\lambda = \log(\lambda_i)$$

  for zero-inflated Poisson data and

$$\eta_i^\pi = \text{logit}(\pi_i), \quad \eta_i^\delta = \log(\delta_i) \quad \text{and} \quad \eta_i^\mu = \log(\mu_i)$$

  for zero-inflated negative binomial data.
- Instead of linear predictors, we will consider flexible **semiparametric specifications**.

Generic Description of Distributional Regression Models

- For any of the parameters in a distributional regression model, we assume a semiparametric predictor

$$\eta = \gamma_0 + f_1(\mathbf{z}) + \ldots + f_r(\mathbf{z}),$$

  $f_1, \ldots, f_r$ are **generic functions** of the covariate vector $\mathbf{z}$.
- Types of effects:
  - Linear effects: $f(\mathbf{z}) = \mathbf{z}^\top \gamma$.
  - Nonlinear, smooth effects of continuous covariates: $f(\mathbf{z}) = f(x)$.
  - Varying coefficients: $f(\mathbf{z}) = uf(x)$.
  - Interaction surfaces: $f(\mathbf{z}) = f(x_1, x_2)$.
  - Spatial effects: $f(\mathbf{z}) = f_{\text{spat}}(s)$.
  - Random effects: $f(\mathbf{z}) = b_c$ with cluster index c.

Generic Description of Distributional Regression Models

- Approximate each of the functions via basis functions, i.e.

$$f_j(\mathbf{z}) = \sum_{l=1}^{L} \gamma_{jl} B_{jl}(\mathbf{z})$$

  for example with B-spline basis functions for penalized splines or indicator basis functions for Markov random fields.

Generic Description of Distributional Regression Models

- This yields a generic model description based on
  - a **design matrix $\mathbf{Z}_j$**, such that the vector of function evaluations $\mathbf{f}_j = (f_j(\mathbf{z}_1), \ldots, f_j(\mathbf{z}_n))^\top$ can be written as
$$\mathbf{f}_j = \mathbf{Z}_j \gamma_j.$$
  - a quadratic **penalty term**
$$\text{pen}(f_j) = \text{pen}(\gamma_j) = \gamma_j^\top \mathbf{K}_j \gamma_j$$
    which operationalizes smoothness properties of f_j.

Generic Description of Distributional Regression Models

- Estimation based on direct optimisation of a penalised likelihood (Newton-type iterations with numerical differentiation).
- **Problems**:
  - The **selection of smoothing parameters** in particular in models with complex predictors and/or many parameters.
  - The **confidence intervals** are usually (much) too narrow.
  - **Numerical instabilities** and limited **potential complexity**.
  - Restriction to **univariate responses**.

$\Rightarrow$ Resort to **Bayesian approach** using Markov chain Monte Carlo simulations.

Generic Description of Distributional Regression Models

From a Bayesian perspective, the penalty term corresponds to **multivariate Gaussian priors** of the form $\gamma_j \sim \mathrm{N}(\mathbf{0}, \tau_j^2 \mathbf{K}_j^-)$,

$$p(\gamma_j|\tau_j^2) \propto \exp\left(-\frac{1}{2\tau_j^2}\gamma_j^\top \mathbf{K}_j \gamma_j\right)$$

where

- $\mathbf{K}_j$ is the precision of the normal distribution,
- $\mathbf{K}_j^-$ the generalised inverse of $\mathbf{K}_j$,
- τ_j^2 represents the inverse smoothing parameter.

Generic Description of Distributional Regression Models

Priors for the smoothing variances:

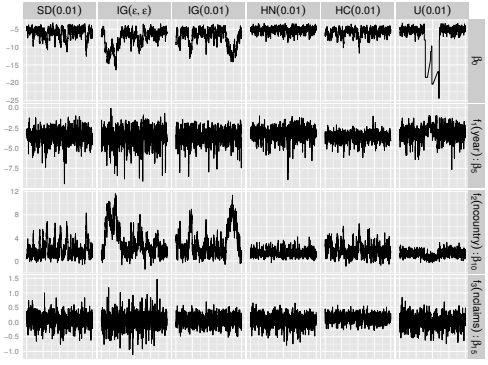
- Assign an inverse gamma prior to τ^2:

$$p(\tau^2) \propto \frac{1}{(\tau^2)^{a+1}} \exp\left(-\frac{b}{\tau^2}\right).$$

Proper for	$a > 0, b > 0$	Common choice: $a = b = \varepsilon$ small.
Improper for	$b = 0, a = -1$	Flat prior for variance τ^2,
	$b = 0, a = -\frac{1}{2}$	Flat prior for standard deviation τ.

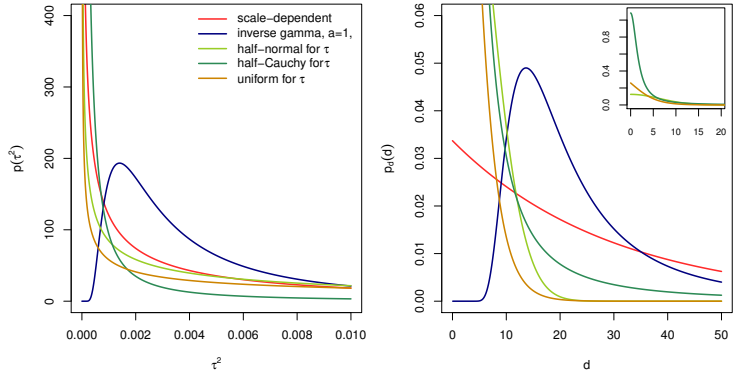
Generic Description of Distributional Regression Models

Small π, sampling paths:

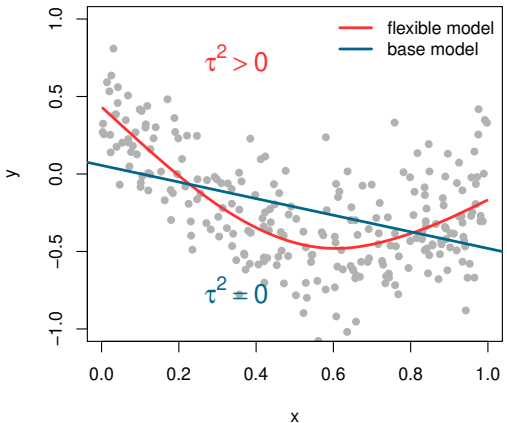


Generic Description of Distributional Regression Models

Hyperprior sensitivity:



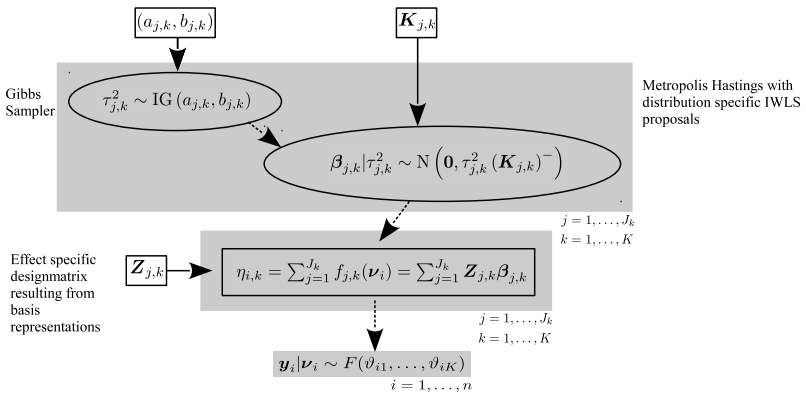
Generic Description of Distributional Regression Models



Generic Description of Distributional Regression Models

Name	Density	Information
$SD(\alpha)$	$p(\tau^2) \propto \frac{1}{2\theta}\left(\frac{\tau^2}{\theta}\right)^{-1/2} \exp\left(-\left(\frac{\tau^2}{\theta}\right)^{1/2}\right)$	Weibull prior for τ^2
$HN(\alpha)$	$p(\tau^2) \propto (\tau^2)^{1/2-1} \exp\left(-\tau^2/(2\theta^2)\right)$	gamma prior for τ^2 half-normal prior for τ
$HC(\alpha)$	$p(\tau^2) \propto \left(1+\tau^2/\theta^2\right)^{-1}\left(\tau^2/\theta^2\right)^{-1/2}$	generalised beta prime prior for τ^2/ half-Cauchy prior for τ
$U(\alpha)$	$p(\tau^2) \propto (\tau^2)^{-1/2}\left(1-\frac{\exp\left((\tau^2)^{1/2}\tilde{c}/\theta-\tilde{c}\right)}{1+\exp\left((\tau^2)^{1/2}\tilde{c}/\theta-\tilde{c}\right)}\right)$	approximate uniform prior for τ^2/ proper uniform prior for τ
$IG(\alpha)$	$p(\tau^2) \propto (\tau^2)^{-2} \exp\left(-\theta/\tau^2\right)$	flat prior for $1/\tau^2$ for $\theta \to 0$
$IG(\epsilon,\epsilon)$	$p(\tau^2) \propto (\tau^2)^{-\epsilon-1} \exp\left(-\epsilon/\tau^2\right)$	'Jeffreys'/flat prior on log-scale for $\epsilon \to 0$
$IG(-1,0)$	$p(\tau^2) \propto const$	flat prior for τ^2
$IG(-\frac{1}{2},0)$	$p(\tau^2) \propto 1/\sqrt{\tau^2}$	flat prior for τ

Generic Description of Distributional Regression Models



Generic Description of Distributional Regression Models

Full conditionals for

- Smoothing variances τ_j^2

$$\tau_j^2|\cdot \sim \text{IG}\left(a_j + \frac{1}{2}\text{rank}(\mathbf{K}_j), b_j + \frac{1}{2}\gamma_j\top\mathbf{K}_j\gamma_j\right)$$

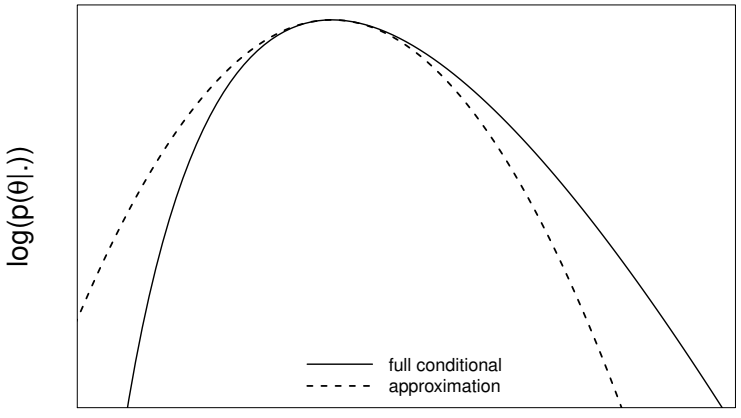
  realised with a **Gibbs sampler step**.

- Regression coefficients γ_j

$$\log(p(\gamma_j|\cdot)) \propto \log(p(\eta_k)) - \frac{1}{2\tau_j^2}\gamma_j\top\mathbf{K}_j\gamma_j$$

  usually not available in closed form.

Generic Description of Distributional Regression Models



Generic Description of Distributional Regression Models

Metropolis-Hastings updates for $\boldsymbol{\gamma}_j|\cdot$

- Propose new state from multivariate normal distribution with mean and precision matrix

$$\boldsymbol{\mu}_j = \mathbf{P}_j^{-1}\mathbf{Z}_j^\top \mathbf{W}(\tilde{\mathbf{y}} - \boldsymbol{\eta}_{-j}), \quad \mathbf{P}_j = \mathbf{Z}_j^\top \mathbf{W}\mathbf{Z}_j + \frac{1}{\tau_j^2}\mathbf{K}_j$$

with

- **Working observations**

$$\tilde{\mathbf{y}} = \boldsymbol{\eta} + \mathbf{W}^{-1}\mathbf{v}.$$

- The predictor without the *j*-th component

$$\boldsymbol{\eta}_{-j} = \boldsymbol{\eta} - \mathbf{Z}_j\boldsymbol{\gamma}_j.$$

Generic Description of Distributional Regression Models

Distribution and parameter specific quantities:

- **Working weights**

$$(\mathbf{W})_{ii} = \mathrm{E}\left(-\frac{\partial^2 \log(p)}{\partial \eta_i^2}\right), \quad (\mathbf{W})_{il} = 0 \ \text{ for } i \neq l.$$

- **Score vectors**

$$\mathbf{v} = \frac{\partial \log(p)}{\partial \boldsymbol{\eta}}.$$

Model Comparison

- To compare models with respect to
  - the response distribution and
  - the predictor specification

  we can rely on
  - **quantile residuals**,
  - **proper scoring rules**, or
  - the deviance information criterion (DIC).

Model Comparison

Quantile residuals are defined as

$$r_i = \Phi^{-1}(u_i)$$

where

- Φ^{-1} is the inverse cumulative distribution function of a standard normal distribution.
- $u_i = \hat{F}(y_i)$ for continuous responses or
- u_i a random value from the uniform distribution on the interval $\left[\hat{F}(y_i - 1), \hat{F}(y_i)\right]$ for discrete responses.
- $\hat{F}(\cdot)$ is the cumulative distribution function of the estimated model.

Model Comparison

Scoring rules:

- Let $y_1, \ldots, y_R$ be data from a hold out sample and F_r the predicted distributions with predicted parameter vectors $\hat{\theta}_r$.
- A **scoring rule** is any real-valued function $S(F_r, y_r)$ assigning a value to the event that y_r is observed under F_r. The score is

$$S = \sum_{r=1}^{R} S(F_r, y_r).$$

- Comparison of different scoring rules via expected values of the scores under the true distribution $F_{r,0}$.
- S is **proper** if $\mathrm{E}_{F_0}(S(F_0, y)) \geq \mathrm{E}_{F_0}(S(F, y))$ for any $F \neq F_{r,0}$.

Model Comparison

Example: Scores for discrete distributions.

- Brier score

$$S(p_r, y_r) = -\sum_h \left(\mathbb{1}_{\{y_r=h\}} - p_{rh}\right)^2$$

- Log-score

$$S(p_r, y_r) = \log\left(p_{ry_r}\right)$$

- Spherical score

$$S(p_r, y_r) = \frac{p_{ry_r}}{\sqrt{\sum_h p_{rh}^2}}$$

with $p_{rh} = \mathrm{P}(y_r = h)$.

Model Comparison

- Deviance information criterion: Let $\boldsymbol{\theta}^{[1]}, \ldots, \boldsymbol{\theta}^{[T]}$ be a MCMC sample from the posterior for the complete parameter vector $\boldsymbol{\theta}$. Then, the DIC is given by

$$\overline{D(\boldsymbol{\theta})} + \text{pd} = 2\overline{D(\boldsymbol{\theta})} - D(\overline{\boldsymbol{\theta}}),$$

  - $D(\boldsymbol{\theta}) = -2\log(f(\mathbf{y}|\boldsymbol{\theta}))$ is the model deviance and

$$\overline{D(\boldsymbol{\theta})} = \frac{1}{T}\sum D(\boldsymbol{\theta}^{[t]})$$

    is the posterior expected deviance.
  - an effective parameter count is given by

$$\text{pd} = \overline{D(\boldsymbol{\theta})} - D(\overline{\boldsymbol{\theta}}).$$

- The DIC resembles the structure of other information criteria such as AIC or BIC.

Patent Citations

- Variables in the data set:

Continuous covariates				
	Description	Mean	Std.	Min/Max
`year`	grant year	1991		1980/1997
`ncountry`	no. of designated states in Europe	7.77	4.12	1/17
`nclaims`	no. of EPO claims	12.33	8.13	1/50
Binary covariates				
	Description	Categories	Rel. freq.	
`biopharm`	patent from biotech/pharma sector	yes=1	43.9%	
`ustwin`	U.S. twin exists	yes=1	61.3%	
`patus`	patentholder of the patent from U.S.	yes=1	33.2%	
`patgsgr`	owner from Switzerland, Ger or GB	yes=1	23.7%	
`opp`	oppositions	yes=1	41.1%	

Patent Citations

- We compare Poisson, zero-inflated Poisson and zero-inflated negative binomial models.
- Basic predictor structure for all model parameters:

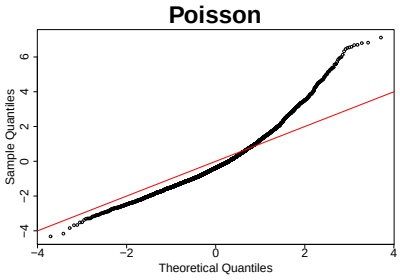
$$\eta = \mathbf{z}^\top \gamma + f_1(\texttt{year}) + f_2(\texttt{ncountry}) + f_3(\texttt{nclaims}).$$

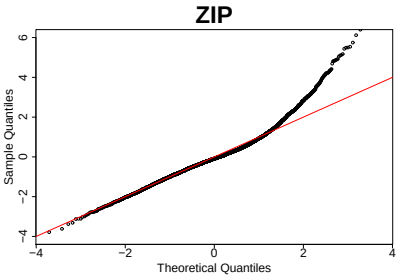
- Scores (averages from ten-fold cross-validation):

Model	Brier Score	Logarithmic Score	Spherical Score
Poisson	-0.8180	-2.3926	0.0070
ZIP	-0.7480	-2.0197	0.0077
ZINB	-0.7439	-1.7604	0.0074

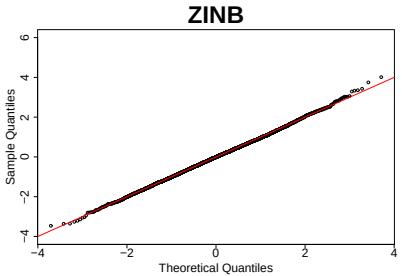
Patent Citations

- Quantile residuals:





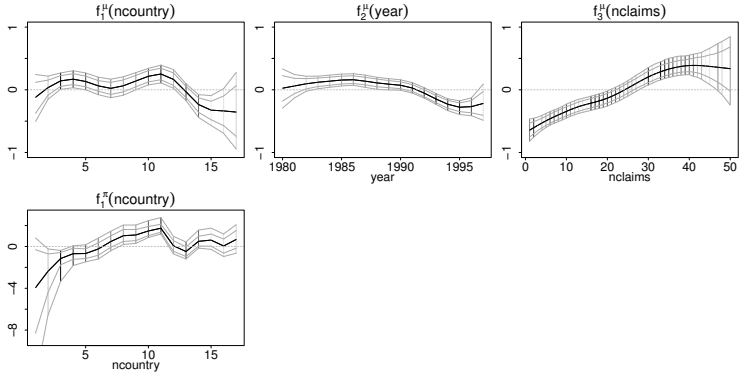
Patent Citations



Patent Citations

$$
\begin{aligned}
\eta^{\mu} &= \beta_0^{\mu} + \beta_1^{\mu} opp + \beta_2^{\mu} biopharm + \beta_3^{\mu} patus + \beta_4^{\mu} patgsgr \\
&\quad + f_1^{\mu}(ncountry) + f_2^{\mu}(year) + f_3^{\mu}(nclaims) \\
\eta^{\pi} &= \beta_0^{\pi} + \beta_1^{\pi} biopharm + f_1^{\pi}(ncountry) + \beta_2^{\pi}(year - 1991) \\
\eta^{\delta} &= \beta_0^{\delta} + \beta_1^{\delta} patus + \beta_2^{\delta} patgsgr.
\end{aligned}
$$

Patent Citations



Claim Frequencies in Car Insurance

- Continuous covariates:

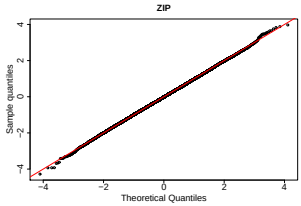
	Description	Mean	Std.	Min/Max
`ageph`	age of policyholder	47	14.82	18/78
`agec`	age of vehicle	7.31	4.04	0/22
`power`	engine power	156.02	19.00	10/243
`bm`	bonus-malus-score	3.27	4.00	0/22

- Basic predictor structure for all model parameters:

$$\eta = \mathbf{z}^\top \boldsymbol{\gamma} + f_1(\texttt{ageph}) + f_2(\texttt{agec}) + f_3(\texttt{power}) + f_4(\texttt{bm}) + f_{\text{spat}}(\texttt{dist}).$$

Claim Frequencies in Car Insurance

- Quantile residuals:

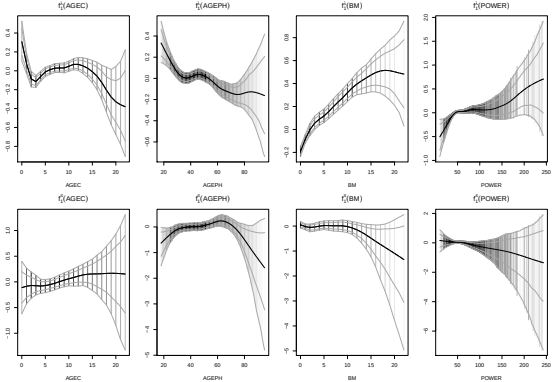


- Scores (averages from ten-fold cross-validation):

Model	Brier Score	Logarithmic Score	Spherical Score
Poisson	-0.2001	-0.4131	0.0020
ZIP	-0.1987	-0.4107	0.0021

Claim Frequencies in Car Insurance

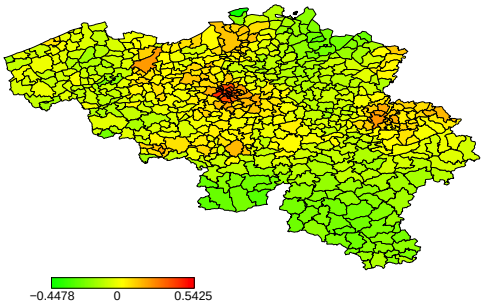
- Estimated nonlinear effects:



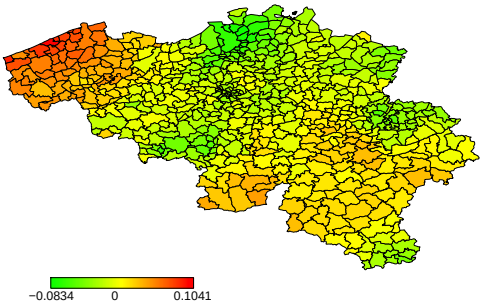
Claim Frequencies in Car Insurance

- Spatial effects:

Estimated spatial effect on λ



Estimated spatial effect on π



Nonnegative Response Models

- 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.

Nonnegative Response Models

- **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.

Nonnegative Response Models

- 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.

Nonnegative Response Models

- **Hierarchical predictor structure**:

$$\begin{aligned}\eta_i &= \beta_0 + \texttt{educ}_i\beta_1 + f_1(\texttt{age}_i) + \texttt{educ}_i f_2(\texttt{age}_i) + \\ &\quad f_3(\texttt{lmexp}_i) + f_{\texttt{spat}}(s_i) + f_{time}(\texttt{t}_i)\end{aligned}$$

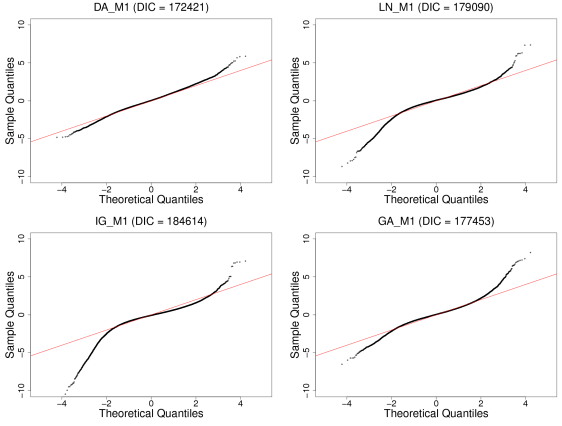
  where the spatial effects is decomposed as

$$f_{\texttt{spat}}(s) = \texttt{east}_s\gamma_1 + g_{\texttt{str}}(s) + g_{\texttt{unstr}}(s)$$

- DIC and scoring rules:

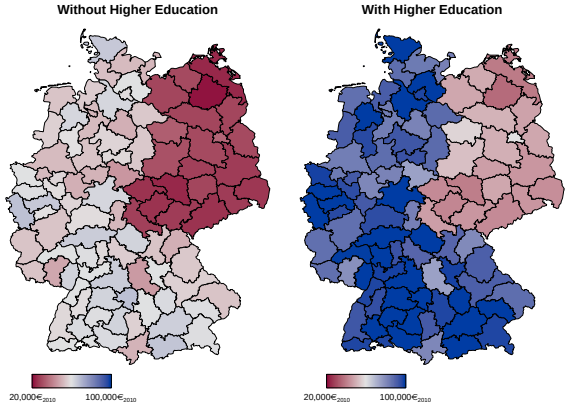
Distribution	DIC	Quadratic	Logarithmic	Spherical	CRPS
LN	179,090	0.1304	-2.4363	0.3621	-2.1581
IG	184,614	0.1464	-2.2741	0.3777	-1.6195
GA	177,453	0.1609	-2.1715	0.3963	-1.2735
DA	**172,421**	**0.1684**	**-2.1034**	**0.4053**	**-1.2662**

Nonnegative Response Models



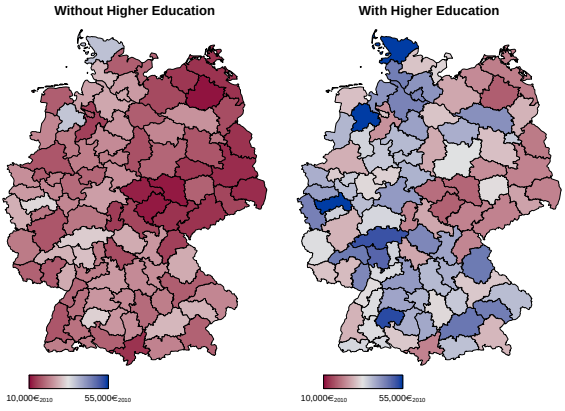
Nonnegative Response Models

- **Expected income** for an “average man” with / without higher education:



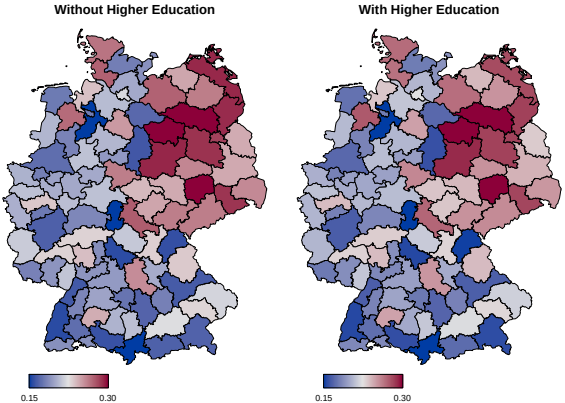
Nonnegative Response Models

- **Income standard deviation** for an “average man”:



Nonnegative Response Models

- **Income inequality** (by the Gini coefficient) for an “average man”:



Multivariate Distributional Regression

- In many empirical studies, we may not only be interested in one single response variable but rather a multivariate response vector.
- Then, typically also the dependence structure between the responses is of interest and may as well change with covariates.
- Examples:
  - Simultaneous analysis of multiple indicators for undernutrition capturing different physiological aspects.
  - Simultaneous analysis of several indicators representing different diseases.
- If a suitable multivariate distributional model can be found, the multivariate extension can easily be cast into the distributional regression framework.

Multivariate Distributional Regression

- Assumptions in **multivariate** distributional regression:
  - The conditional distribution of the **response vector** $\mathbf{y}_i$ given covariate information $\mathbf{z}_i$ is from a pre-specified class of K-parametric densities $p(\mathbf{y}_i|\vartheta_{i1}, \ldots, \vartheta_{iK})$.
  - **Each parameter** ϑ_{ik} is related to a **regression predictor** $\eta_{ik} = \eta_k(\mathbf{z}_i)$ via

$$\vartheta_{ik} = h_k(\eta_{ik}) \quad \text{and} \quad \eta_{ik} = h_k^{-1}(\vartheta_{ik})$$

    where h_k are one-to-one transformations of the predictors to ensure appropriate restrictions to the parameter space.
- The main difficulty is to construct suitable multivariate response distributions!
- In contrast to the univariate models, in multivariate models one or multiple parameters will be related to the dependence between responses.

Multivariate Normal and t Models

- **Childhood malnutrition** is one of the most urgent public health problems in developing and transition countries.
- Nutritional status is commonly assessed by scores formed from an appropriate **anthropometric measure** AI relative to a **reference population**:

$$y_i = \frac{\text{AI}_i - \mu}{\sigma}$$

  where μ and σ refer to median and standard deviation in the reference population.
- Two different indicators:
  - **Chronic** undernutrition (stunting, y_{i1}) is measured by **insufficient height for age**.
  - **Acute** undernutrition (wasting, y_{i2}) is measured by **insufficient weight for age**.

Multivariate Normal and t Models

- Common approach for a joint analysis of both indicators: **Seemingly Unrelated Regression (SUR)** with

$$y_{i1} = \eta_{i1} + \varepsilon_{i1}, \qquad y_{i2} = \eta_{i2} + \varepsilon_{i2}$$

  where $\varepsilon_{ij} \sim \mathrm{N}(0, \sigma_j^2)$, $j = 1, 2$ and $\mathrm{Cor}(\varepsilon_{i1}, \varepsilon_{i2}) = \rho$.

- Limitations of standard SUR models:
  - In most cases, only **linear predictors** where $\eta_{ij} = \mathbf{z}_{ij}^\top \gamma_j$ (no nonlinear effects, no spatial effects, etc.).
  - Joint **normality assumption** for the error terms and therefore the responses.
  - Only **effects on the expectation** of the responses (no covariate effects on the variances or the correlation coefficient).

Multivariate Normal and t Models

- We will instead consider distributional regression based on
  - the **bivariate normal distribution**

$$\begin{pmatrix} y_{i1} \\ y_{i2} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu_{i1} \\ \mu_{i2} \end{pmatrix}, \begin{pmatrix} \sigma_{i1}^2 & \sigma_{i1}\sigma_{i2}\rho_i \\ \sigma_{i1}\sigma_{i2}\rho_i & \sigma_{i2}^2 \end{pmatrix} \right)$$

    with the five parameters μ_{i1}, μ_{i2}, σ_{i1}, σ_{i2}, and ρ_i.
  - the **bivariate t distribution**

$$\begin{pmatrix} y_{i1} \\ y_{i2} \end{pmatrix} \sim \mathrm{t}\left(\mathrm{df}_i, \begin{pmatrix} \mu_{i1} \\ \mu_{i2} \end{pmatrix}, \begin{pmatrix} \sigma_{i1}^2 & \sigma_{i1}\sigma_{i2}\rho_i \\ \sigma_{i1}\sigma_{i2}\rho_i & \sigma_{i2}^2 \end{pmatrix} \right)$$

    with the six parameters df_i, μ_{i1}, μ_{i2}, σ_{i1}, σ_{i2}, and ρ_i.

Multivariate Normal and t Models

- We consider data from the 1998/99 **India Demographic and Health Survey** (`http://www.measuredhs.com`).
- Nationally representative cross-sectional study on fertility, family planning, maternal and child health, as well as child survival, HIV/AIDS, and nutrition.
- Information on 24,316 children is available (after excluding observations with missing information).

Multivariate Normal and t Models

- Possible **determinants of childhood malnutrition**:

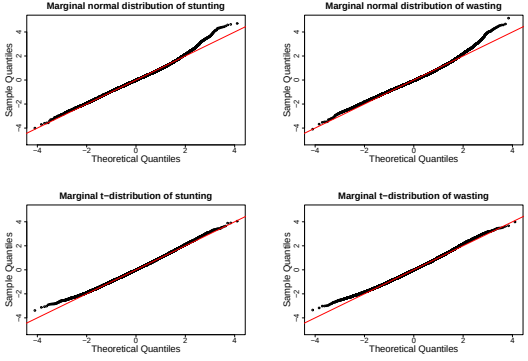
Child-specific factors:	age, gender, duration of breastfeeding, . . .
Maternal factors:	age, body mass index, years of education, employment status, . . .
Household factors:	place of residence, electricity, radio, tv, . . .
Spatial information:	district where the child lives

- Predictor structure for each parameter:

$$\begin{aligned}\eta_i \;=\; & \mathbf{z}_i^\top \boldsymbol{\gamma} + f_1(\texttt{cage}_i) + f_2(\texttt{breastfeeding}_i) + f_3(\texttt{mage}_i) + \\ & f_4(\texttt{mbmi}_i) + f_5(\texttt{medu}_i) + f_6(\texttt{edupartner}_i) + \\ & f_{\text{spat}}(\texttt{dist}_i) + \beta_{\texttt{dist}_i}\end{aligned}$$

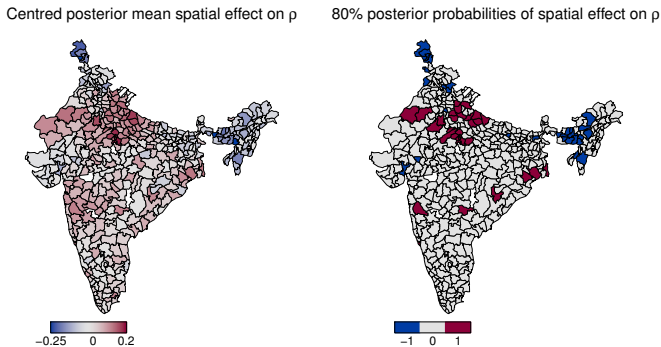
Multivariate Normal and t Models

- Model comparison based on marginal quantile residuals:



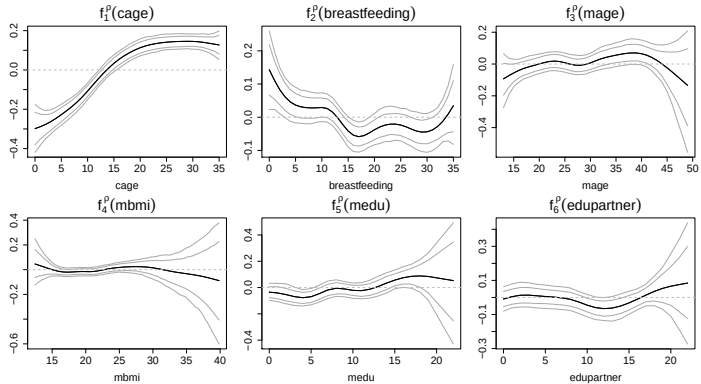
Multivariate Normal and t Models

- Estimated spatial effect (and posterior probabilities) for the correlation coefficient (t distribution):



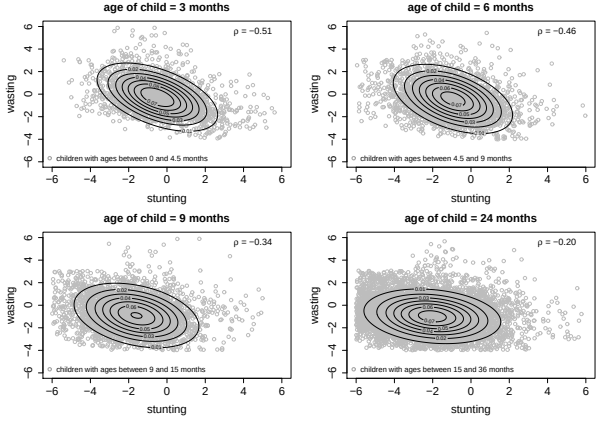
Multivariate Normal and t Models

- Estimated nonlinear effects (with 95% credible intervals) for the correlation coefficient (t distribution):



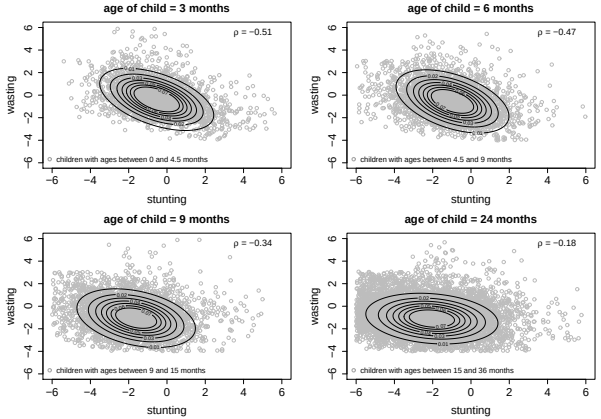
Multivariate Normal and t Models

- Contours of fitted distributions (bivariate normal):



Multivariate Normal and t Models

- Contours of fitted distributions (bivariate t):

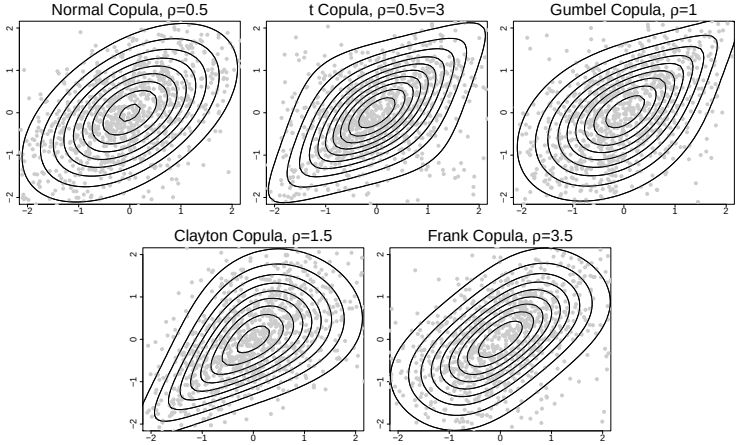


Multivariate Normal and t Models

Limitations:

- Assumption of multivariate normal/t distribution implies normal/t marginals.
- Dependence structure limited to covariances / correlation coefficients.

Multivariate Normal and t Models



Distributional Copula Regression

- A copula $C : [0,1]^D \to [0,1]$ is a cumulative distribution function with uniform margins.
- Sklar's Theorem:
  - Let $F_d(y_1|\nu)$, $d = 1, \ldots, D$ be the conditional marginal distributions of two univariate random variables $y_d|\nu$ and ν the available covariate information.
  - Let $F(y|\nu)$ be the joint conditional distribution of $y = (y_1, \ldots, y_D)|\nu$.

  Then, there exists a **conditional copula** $C(\cdot|\nu)$ such that

$$F(y|\nu) = C(F_1(y_1|\nu), \ldots, F_D(y_D|\nu)|\nu).$$

  Conversely, if $F_d(y_1|\nu)$, $d = 1, \ldots, D$ are the marginals of $y_d|\nu$ and $C(\cdot|\nu)$ is a conditional copula, then $F(y|\nu)$ defined above is a conditional bivariate distribution function with conditional marginals F_d, $d = 1, \ldots, D$.

Bivariate Copula Regression

- Structure of a copula regression model:
  1. Specify marginal distributions with densities p_1 and p_2 with distribution parameters $\vartheta_{i1}^{(1)}, \ldots \vartheta_{iK_1}^{(1)}$ and $\vartheta_{i1}^{(2)}, \ldots \vartheta_{iK_2}^{(2)}$.
  2. Specify a parametric copula with density c with parameters $\vartheta_{i1}^{(c)}, \ldots \vartheta_{iK_c}^{(c)}$.
  3. Specify structured additive predictors for all involved distribution parameters, i.e.

$$\begin{aligned}
\vartheta_{ik_1}^{(1)} &= h_{k_1}(\eta_{ik_1}^{(1)}), \quad k_1 = 1, \ldots, K_1 \\
\vartheta_{ik_2}^{(2)} &= h_{k_2}(\eta_{ik_2}^{(2)}), \quad k_2 = 1, \ldots, K_2 \\
\vartheta_{ik_c}^{(c)} &= h_{k_c}(\eta_{ik_c}^{(c)}), \quad k_c = 1, \ldots, K_c.
\end{aligned}$$

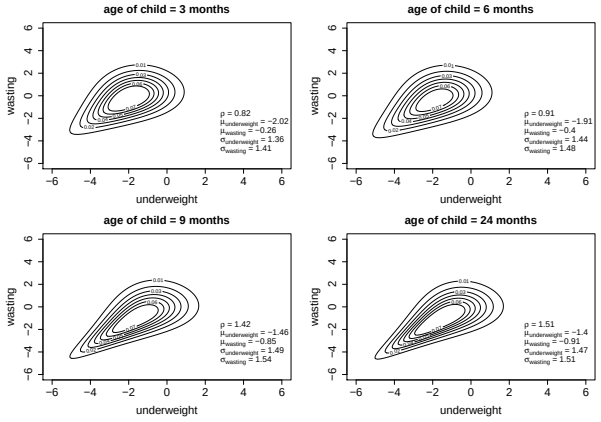
Bivariate Copula Regression

- We keep the assumption of normal margins (for convenience) and consider different copula specifications
- Focus on the relation between underweight and the two specific indicators for acute (stunting) and chronic (wasting) malnutrition.

Response **y**	Copula	DIC
(underweight,wasting)	**Clayton**	**123,157**
(underweight,wasting)	Gumbel	128,509
(underweight,wasting)	normal	145,268
(underweight,stunting)	**Clayton**	**126,417**
(underweight,stunting)	Gumbel	129,964
(underweight,stunting)	normal	135,182

Bivariate Copula Regression

- Clayton copula with response $\mathbf{y} = (\texttt{underweight}, \texttt{wasting})^\top$



Gaussian Copula GAMLSS for $D > 2$

Each marginal distribution is modeled through a univariate distributional regression model

$$y_{id} \mid x_i \sim F_d \left(y_{id} \mid \vartheta^{(d)} \right).$$

To allow for covariate-dependent associations between the components of $y_i \in \mathbb{R}^D$, we assume a Gaussian copula with covariate-dependent correlation matrix $\vartheta^{(c)} \equiv \Omega$.
Then for $i = 1, \ldots, n$,

$$u_i = (u_{i1}, \ldots, u_{iD})^\top \sim (0, \Omega)$$
$$y_{id} = F_j^{-1} \left(\Phi \left(u_{id} \right) \mid \vartheta^{(d)} \right), \quad d = 1, \ldots, D.$$

Gaussian Copula GAMLSS for $D > 2$

The parameter of the Gaussian copula

$$\Omega = \begin{pmatrix} 1 & \omega_{21} & \dots & \omega_{D1} \\ \omega_{21} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \omega_{D(D-1)} \\ \omega_{D1} & \dots & \omega_{D(D-1)} & 1 \end{pmatrix}$$

has to be a valid correlation matrix.

$D = 2$: The only condition is $\omega_{21} \in [-1, 1]$

$D = 3$: $\omega_{21}, \omega_{31}, \omega_{32} \in [-1, 1]$ and
$$\det(\Omega) = 1 + 2\omega_{21}\omega_{31}\omega_{32} - \omega_{21}^2 - \omega_{31}^2 - \omega_{32}^2 \geq 0$$

Gaussian Copula GAMLSS for $D > 2$

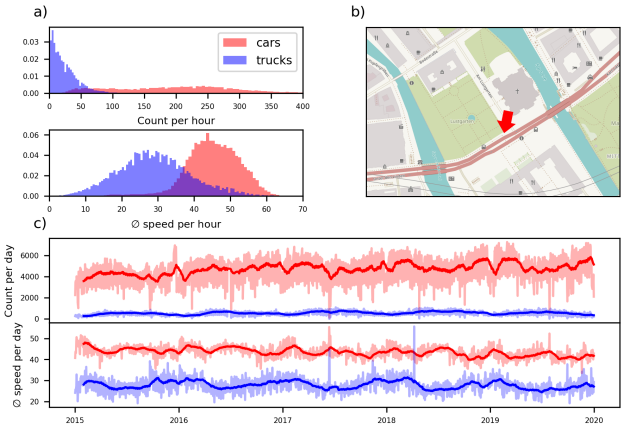
$\Omega = \text{diag}(\Sigma)^{-\frac{1}{2}} \Sigma \, \text{diag}(\Sigma)^{-\frac{1}{2}}$ is the correlation matrix corresponding to the covariance matrix $\Sigma = \left(\Lambda\Lambda^\top\right)^{-1}$ with

$$\Lambda = \begin{pmatrix} 1 & 0 & \dots & 0 \\ \lambda_{21} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ \lambda_{D1} & \dots & \lambda_{D(D-1)} & 1 \end{pmatrix}.$$

$D = 2$: $\omega_{21} = -\frac{\lambda_{21}}{\sqrt{1+\lambda_{21}^2}}$

$D = 3$: ω_{21}, ω_{31} and ω_{32} are complex non-linear combinations of the entries of Λ

Traffic Detection in Berlin



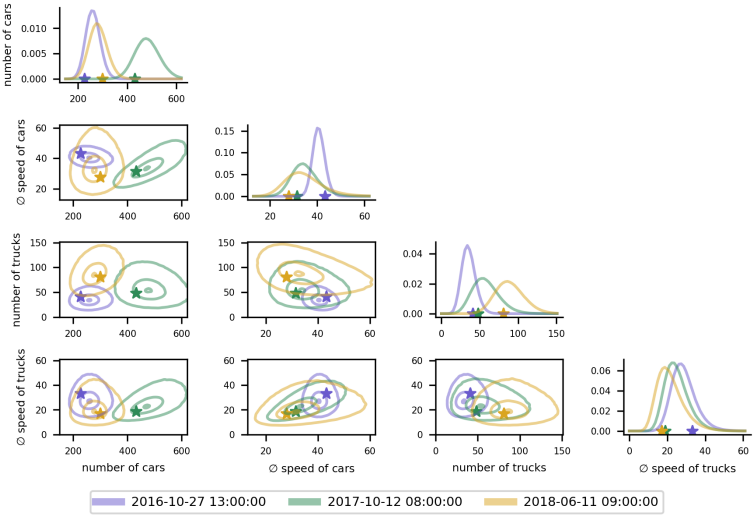
Description of data

- Hourly aggregated data over five years from 2015-–2020 ($n = 39,739$)

Challenges

- 4-dimensional response combining continuous and discrete marginals
- Complex dependence structure changing over time

Traffic Detection in Berlin



Open Research Questions

- Copula regression models often lack user friendly implementations
- Extensions to more flexible dependence structures (for $D > 2$) are challenging:
  - Find an unrestricted parameterization for the multivariate copula (that is also somewhat interpretable)
  - Inference is not straight forward due to the high dimensional parameter space and the complex interactions between different building blocks of the model

Dirichlet Regression

- Aim: Identify determinants of and spatial variation in Germany's 2009 federal election results.
- Responses: **proportions of the electorate** voting on five parties
  - Christian Democratic Union/Christian Social Union (CDU/CSU),
  - Social Democratic Party (SPD),
  - The Liberals (FDP),
  - The Left,
  - The Greens

  for each of the 413 districts (Landkreise) in Germany.
- All remaining votes are collected in category "Others".

Dirichlet Regression

- **Multivariate** response vector $\mathbf{y} = (y_1, \ldots, y_6)$ with

$$0 < y_d < 1, \qquad d = 1, \ldots, 6$$
$$\sum_{d=1}^{6} y_d = 1$$

- Suitable model: **Dirichlet distribution**, i.e. $\mathbf{y} \sim \mathrm{Dir}(\boldsymbol{\alpha})$ with parameter $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_6) \in \mathbb{R}^6_{>0}$ and density

$$p(y_1, \ldots, y_6) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{d=1}^{6} y_d^{\alpha_d - 1}$$

Dirichlet Regression

where the normalising constant is the multinomial beta function of $\boldsymbol{\alpha}$:

$$\mathrm{B}(\boldsymbol{\alpha}) = \frac{\prod_{d=1}^{D} \Gamma(\alpha_d)}{\Gamma\left(\sum_{d=1}^{D} \alpha_d\right)}.$$

- The Dirichlet distribution is a multivariate extension of the beta distribution with

$$\mathrm{E}(y_d) = \frac{\alpha_d}{\alpha_0}$$

where $\alpha_0 = \sum_{d=1}^{D} \alpha_d$ represents a precision parameter.

- Each parameter α_d is linked to a regression predictor via

$$\eta^{\alpha_d} = \log(\alpha_d).$$

Dirichlet Regression

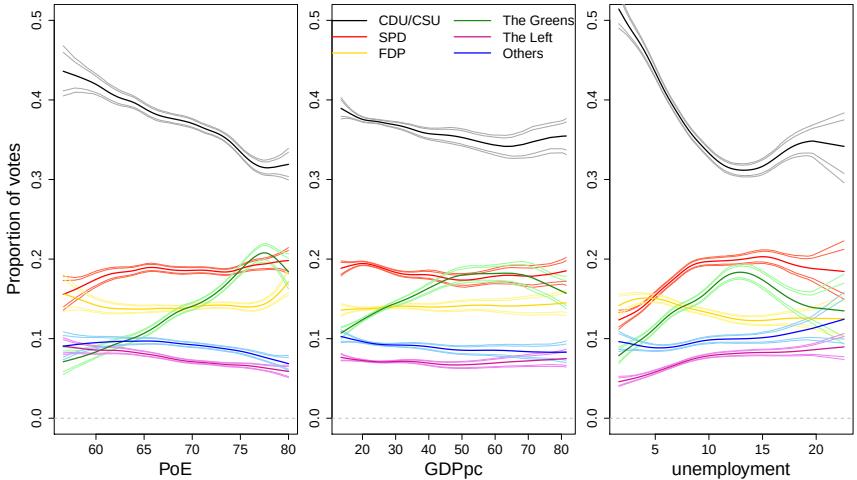
- The predictors are given as

$$\begin{aligned}\eta_i &= \beta_0 + f_1(\texttt{PoE}_i) + f_2(\texttt{GDPpc}_i) + f_3(\texttt{unemployment}_i) + \\ &\quad f_{\text{spat}}(\texttt{region}_i).\end{aligned}$$

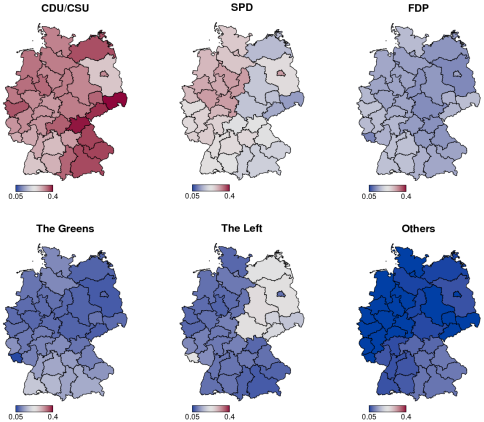
with **district-specific covariates**

- `PoE`: proportion of electorates compared to the population entitled to vote (turnout) in percent.
- `unemployment`: rate of unemployment in 2008.
- `GDPpc`: gross domestic product per capita in 2008 (measured in thousand Euros).
- `region`: one of 38 administrative regions.

Dirichlet Regression



Dirichlet Regression



Summary

- In any undergraduate course on statistics, we are teaching that means are not enough to characterise differences between distributions.
- Distributional regression is an attempt to transfer this knowledge to the regression context.
- Distributional regression allows for considerable flexibility in terms of specifying an appropriate response distribution reflecting various features of the response distribution, e.g.
  - zero-inflation and overdispersion for count data.
  - nonnegativity for continuous responses.
  - boundedness to a prespecified interval for continuous responses.
  - mixtures of discrete and continuous parts.
  - Etc.

Thank you!

References

- Klein, N. (2024). Distributional regression for data analysis. *Annual Review of Statistics and Its Application*, 11:321-346
- N. Klein, M. Carlan, T. Kneib, S. Lang and H. Wagner (2021). Bayesian effect selection in structured additive distributional regression models. *Bayesian Analysis*, 6(2):545–573.
- N. Umlauf, N. Klein and A. Zeileis (2018). BAMLSS: Bayesian additive models for location, scale, and shape (and beyond) *Journal of Computational and Graphical Statistics.*
- L. Kock, N. Klein (2025). Truly multivariate structured additive distributional regression. *Journal of Computational and Graphical Statistics.*
- Michaelis, P., Klein, N. and Kneib, T. (2018): Bayesian Multivariate Distributional Regression with Skewed Responses and Skewed Random Effects. *To appear in Journal of Computational and Graphical Statistics.*
- Klein, N. and Kneib, T. (2016): Scale-Dependent Priors for Variance Parameters in Structured Additive Distributional Regression. *Bayesian Analysis*, 11, 1071–1106, URL: `https://projecteuclid.org/euclid.ba/1448323525`.

References

- Klein, N. and Kneib, T. (2016): Simultaneous Inference in Structured Additive Conditional Copula Regression Models: A Unifying Bayesian Approach. *Statistics and Computing*, 26, 841–860, doi:10.1007/s11222-015-9573-6.
- Klein, N., Kneib, T. and Lang, S. (2015): Bayesian Generalized Additive Models for Location, Scale and Shape for Zero-Inflated and Overdispersed Count Data. *Journal of the American Statistical Association*, 110, 405–419, URL: `http://amstat.tandfonline.com/doi/abs/10.1080/01621459.2014.912955`.
- Klein, N., Kneib, T., Lang, S. and Sohn, A. (2015): Bayesian Structured Additive Distributional Regression with an Application to Regional Income Inequality in Germany. *Annals of Applied Statistics*, 9, 1024–1052, URL: `http://www.imstat.org/aoas/next_issue.html`.
- Herwartz, H., Klein, N., and Strumann, C. (2015): Modelling Hospital Admission and Length of Stay by Means of Generalised Count Data Models. *Journal of Applied Econometrics*, 6, 1159-1182, URL: `https://doi.org/10.1002/jae.2454`.

References

Klein, N., Kneib, T., Klasen, S. and Lang, S. (2015): Bayesian Structured Additive Distributional Regression for Multivariate Responses. *Journal of the Royal Statistical Society, Series C*, 64, 569–591, URL: `http://onlinelibrary.wiley.com/doi/10.1111/rssc.12090/abstract`.

Klein, N., Denuit, M., Lang, S. and Kneib, T. (2014): Nonlife Ratemaking and Risk Management with Bayesian Generalized Additive Models for Location, Scale, and Shape. *Insurance: Mathematics and Economics*, 55(1):225–249