University of Innsbruck

Bayesian structured additive distributional regression

Nadja Klein, Thomas Kneib, Stefan Lang

Working Papers in Economics and Statistics

2013-23

Research Platform eeecon Empirical and Experimental Economics
University of Innsbruck
http://eeecon.uibk.ac.at/

University of Innsbruck
Working Papers in Economics and Statistics

The series is jointly edited and published by

- Department of Economics
- Department of Public Finance
- Department of Statistics

Contact Address:
University of Innsbruck
Department of Public Finance
Universitaetsstrasse 15
A-6020 Innsbruck
Austria
Tel: + 43 512 507 7171
Fax: + 43 512 507 2970
E-mail: eeecon@uibk.ac.at

The most recent version of all working papers can be downloaded at
http://eeecon.uibk.ac.at/wopec/

For a list of recent papers see the backpages of this paper.

Bayesian Structured Additive Distributional Regression

Nadja Klein, Thomas Kneib
Chair of Statistics
Georg-August-University Göttingen

Stefan Lang
Department of Statistics
University of Innsbruck

Abstract

In this paper, we propose a generic Bayesian framework for inference in distributional regression models in which each parameter of a potentially complex response distribution and not only the mean is related to a structured additive predictor. The latter is composed additively of a variety of different functional effect types such as nonlinear effects, spatial effects, random coefficients, interaction surfaces or other (possibly non-standard) basis function representations. To enforce specific properties of the functional effects such as smoothness, informative multivariate Gaussian priors are assigned to the basis function coefficients. Inference is then based on efficient Markov chain Monte Carlo simulation techniques where a generic procedure makes use of distribution-specific iteratively weighted least squares approximations to the full conditionals. We study properties of the resulting model class and provide detailed guidance on practical aspects of model choice including selecting an appropriate response distribution and predictor specification. The importance and flexibility of Bayesian structured additive distributional regression to estimate all parameters as functions of explanatory variables and therefore to obtain more realistic models, is exemplified in two applications with complex response distributions.

Key words: generalised additive models for location scale and shape; iteratively weighted least squares proposal; Markov chain Monte Carlo simulation; penalised splines; semiparametric regression.

1 Introduction

Classical regression models within the exponential family framework, such as generalised linear models [McCullagh and Nelder, 1989] or generalised additive models

(GAMs, [Hastie and Tibshirani, 1990, Ruppert et al., 2003, Wood, 2006]), focus exclusively on relating the mean of a response variable to covariates but neglect the potential dependence of higher order moments or other features of the response distribution on covariates. As a consequence, the advantage of obtaining covariate effects that are straightforward to estimate and easy to interpret is at least partly offset by the likely misspecification of the model that may render inferential conclusions invalid. A completely distribution-free alternative to mean regression is provided by quantile or expectile regression where the assumptions on the error term are generalised such that the regression predictor is related to a local feature of the response distribution, indexed by a pre-specified asymmetry parameter (the quantile or expectile level), see Koenker and Bassett [1978], Newey and Powell [1987] for the original references and Koenker [2005], Yu and Moyeed [2001], Schnabel and Eilers [2009], Sobotka and Kneib [2012] for more recent overviews. Both approaches have the distinct advantage that basically no assumptions on the specific type of the response distribution or homogeneity of certain parameters such as the variance are required. However, this flexibility also comes at a price since properties of the determined estimates are more difficult to obtain, the flexibility of the predictor specification is somewhat limited and estimates for a set of asymmetries may cross leading to incoherent distributions for the response. Moreover, model choice and model comparison tend to be difficult since the models only relate to local properties of the response. Finally, if prior knowledge on specific aspects of the response distribution is available, quantile and expectile regression may be less efficient and are also less appropriate for discrete distributions or mixed discrete continuous distributions.

As a consequence, it is of considerable interest to derive models that are in between the simplistic framework of exponential family mean regression and distribution-free approaches. Such an approach is given by the class of generalised additive models for location, scale and shape (GAMLSS, [Rigby and Stasinopoulos, 2005]) in which all parameters of a potentially complex response distribution are related to additive regression predictors in the spirit of GAMs [Ruppert et al., 2003, Wood, 2004, 2008, Fahrmeir et al., 2004, 2013]. Estimation is then based on Newton-Raphson-type or Fisher scoring approaches derived from a penalised likelihood where in many cases the score function and observed Fisher information are determined by numerical dif-

ferentiation. In this paper, we build upon GAMLSS to develop a generic Bayesian treatment of distributional regression relying on Markov chain Monte Carlo simulation algorithms. To construct suitable proposal densities, we follow the idea of iteratively weighted least squares proposals [Gamerman, 1997, Brezger and Lang, 2006] and construct local quadratic approximations to the full conditionals. To approximate the mode, we explicitly derive expressions for the score function and expected Fisher information. In our experience, this considerably enhances numerical stability as compared to using numerical derivatives and the observed Fisher information.

We use the notion of distributional regression for our approach since in most cases, the parameters of the response distribution are in fact not related to location, scale and shape but are general parameters of the response distribution and only indirectly determine location, scale and shape. As an example, consider the application on the proportion of farms' outputs achieved by cereal that we will analyse in more detail in Section 6. The response variable in this study is given by the percentage of the total output that farms in England and Wales obtain from the cultivation of cereal. Hence, the response variable is restricted to the interval [0,1] and represents the amount of cereal products by the farms relative to their total output. A natural candidate for analysing such ratios is the beta distribution parametrised such that one parameter represents the expectation while the other one relates to a general shape parameter [see for example Ferrari and Cribari-Neto, 2004]. However, a further complication arises in our data set from the fact that there is a considerable fraction of observations with either no production of cereal at all or complete specialisation on cereal (cereal output equal to 100%) such that the beta distribution has to be inflated with zeros and ones similar as in zero-inflated count data regression. As a consequence, we end up with a mixed discrete-continuous distribution with four distributional parameters relating the probability for excess zeros, the probability for excess ones and the two parameters of the beta distribution to regression predictors (we provide appropriate link functions for such a model in the next section). Moreover, the two probabilities have to be linked such that their sum is always smaller or equal than one. Since none of the parameters involved in the response distribution is actually related to location, scale and shape, we prefer the term distributional regression as compared to GAMLSS. Note also that the structure of the response variable in this example

renders both quantile regression and transformations of the response distribution inappropriate due to the mixture of a discrete and a continuous part.

The full potential of distributional regression is only exploited when the regression predictor is also broadened beyond the scope of simple linear or additive specifications. We will consider structured additive predictors [Fahrmeir et al., 2013, Brezger and Lang, 2006] where each predictor is determined as an additive combination of various types of functional effects, such as nonlinear effects of continuous covariates, seasonal effects of time trends, spatial effects, random intercepts and slopes, varying coefficient terms or interaction surfaces. All of these approaches can be represented in terms of possibly non-standard basis functions in combination with a multivariate Gaussian prior to enforce desired properties of the estimates, such as smoothness or sparsity.

The main advantages of Bayesian structured additive distributional regression can be summarized as follows:

- It provides a broad and generic framework for distributional regression, including continuous, discrete and mixed discrete-continuous distributions and therefore considerably expands the common exponential family framework. We will provide more details on examples in the next section.

- Compared to penalised maximum likelihood inference developed in the GAMLSS framework, the Bayesian approach provides valid confidence intervals even without relying on asymptotic arguments, automatically yields estimates for the smoothing parameters determining the impact of the prior and allows for a very modular inferential approach. Iteratively weighted least squares proposals based on the expected Fisher information yield numerically stable and adaptive proposal densities that do not require manual tuning.

- General guidelines for important model choice issues such as choosing an appropriate response distribution and determining a suitable predictor specification can be obtained based on quantile residuals, the deviance information criterion and proper scoring rules [Dunn and Smyth, 1996, Spiegelhalter et al., 2002, Gneiting and Raftery, 2007].

- Theoretical results on the positive definiteness of the precision matrix in the

proposal densities and the propriety of the posterior can be provided.

- The Bayesian approach allows to borrow extensions developed for Bayesian mean regression such as multilevel structures, monotonicity constraint estimates, variable selection and regularisation priors without the necessity to redevelop the complete inferential machinery.

The remainder of this paper is structured as follows: Section 2 provides a more detailed introduction to distributional regression comprising several important choices for the response distribution and predictor components. Section 3 develops a Markov chain Monte Carlo simulation algorithm for Bayesian inference, discusses theoretical results and provides some information on the efficient implementation. Model choice issues are treated in Section 4. Sections 5 and 6 illustrate distributional regression based on two complex applications utilising non-negative skewed distributions, such as log-normal, gamma or Dagum for income distributions on the one hand and the zero-one-inflated beta distribution for the proportion of cereal output produced by farms on the other hand. Finally, Section 7 provides a summary and comments on directions of future research.

2 Structured Additive Distributional Regression

We assume that observations on a scalar response variable $y_1, \ldots, y_n$ as well as covariate information $\boldsymbol{\nu}_i$, $i = 1, \ldots, n$, have been collected for n individuals. The conditional distribution of observation y_i given the covariate information $\boldsymbol{\nu}_i$ is assumed to be from a pre-specified class of K-parametric distributions $f_i(y_i|\vartheta_{i1}, \ldots, \vartheta_{iK})$ indexed by the (in general covariate-dependent) parameters $\vartheta_{i1}, \ldots, \vartheta_{iK}$. Note that f_i is considered a general density, i.e. we use the same notation for continuous responses, discrete responses and also mixed discrete-continuous responses. Each parameter ϑ_{ik} is linked to a semiparametric regression predictor η_{ik} formed of the covariates via a suitable (one-to-one) response function such that $\vartheta_{ik} = h_k(\eta_{ik})$ and $\eta_{ik} = h_k^{-1}(\vartheta_{ik})$. The response function is usually chosen to ensure appropriate restrictions on the parameter space such as the exponential function $\vartheta_{ik} = \exp(\eta_{ik})$, to ensure positivity, the logit link $\vartheta_{ik} = \exp(\eta_{ik})/(1 + \exp(\eta_{ik}))$ for parameters representing probabilities or the identity function if the parameter space is unrestricted.

An overview of some interesting examples of response distributions is provided in Table 1, listed together with density or probability mass function and restrictions on the parameters. Note that this is of course not an exhaustive list but simply reflects a useful subset of distributions we have already some experience with. Other distributions may be added following the inferential procedure outlined in Section 3. In the following, we will discuss some important response distributions and the specification of structured additive predictors in more detail.

2.1 Response Distributions

Real-Valued Responses Examples in which the response is real-valued are the well known normal distribution, where not only the expectation $\mu_i \in \mathbb{R}$ but also the variance $\sigma_i^2 > 0$ can explicitly be modelled in terms of covariates or the t-distribution, where in addition to a location parameter (corresponding to the expectation μ_i of the response, if existing) and a scale parameter σ_i^2, the degrees of freedom $n_{d,i} > 0$ may be linked to an additive predictor. Although the t-distribution is symmetric and bell-shaped like the normal distribution, it has heavier tails and may therefore be considered a robust alternative to the normal which is less affected by extreme values. Effects on the degrees of freedom also allow to determine the deviation from normality since the t-distribution leads to a normal distribution for $n_{d,i} \to \infty$.

Non-Negative Responses The inverse Gaussian distribution is a two-parameter distribution with expectation μ_i and variance $\text{Var}(y_i) = \sigma_i^2 \mu_i^3$ which is proportional to the second parameter σ_i^2. This distribution has for example been used by Heller et al. [2006] for modelling the extreme right skewness of claim size distributions arising in car insurances. There are many more distributions with positive support and if one is interested in describing the conditional distribution of for instance incomes, costs or other quantities with positive values, the generalised beta family provides a huge number of distributions with up to five parameters allowing for skewness, kurtosis or parameters that affect the shape of the distribution in general. Special cases include the gamma distribution with parameters $\mu_i > 0$, $\sigma_i > 0$, which is often used in insurances to model small claim sizes. With the density as parametrised in Table 1, the expectation of $y_i > 0$ is directly given by the location parameter μ_i,

this is $E(y_i) = \mu_i$. The variance is $\text{Var}(y_i) = \frac{\mu_i^2}{\sigma_i}$ such that the second parameter σ_i is inversely proportional to the variance. A generalised gamma distribution can be defined by including an additional shape parameter τ_i that allows an extended flexibility compared to the gamma distribution. Furthermore, for special parameter values, the generalised gamma distribution includes distributions like the log-normal, exponential, gamma or Weibull distribution. The latter one is often used for instances in survival analysis to represent failure times and is determined by a shape parameter $\alpha_i > 0$ and a scale parameter $\lambda_i > 0$. While a value of $\alpha_i < 1$ indicates decreasing failure rates over time, $\alpha_i = 1$ and $\alpha_i > 1$ stand for constant and increasing rates with time, respectively. For $\alpha_i = 1$ one gets the exponential distribution whereas $\alpha_i = 2$ leads to the Rayleigh distribution. In economic research, the Dagum distribution [Chotikapanich, 2008] has become quite famous for modelling income distributions. Containing one scale parameter $b_i > 0$ and two shape parameters $a_i > 0$, $p_i > 0$, this distribution becomes very flexible. The expectation of y_i is well defined for $a_i > 0$ and in this case given by

$$E(y_i) = -\frac{b_i}{a_i} \frac{\Gamma\left(-\frac{1}{a_i}\right) \Gamma\left(p_i + \frac{1}{a_i}\right)}{\Gamma(p_i)} \tag{1}$$

Hence, the expected value of y_i is proportional to the parameter b_i. Although the other parameters are not that easily interpretable at first sight, the Dagum distribution has the great advantage that both the conditional mode and conditional quantiles can be expressed in closed form. More specifically, for $a_i p_i > 1$ there exists an interior mode of the form

$$\text{Mode}(y_i) = b_i \left(\frac{a_i p_i - 1}{a_i + 1}\right)^{1/a_i} \tag{2}$$

and for $a_i p_i < 1$ the density has a pole at zero. The former case is suited to model unimodal income distributions. In this way, both the factors $a_i p_i$ and a_i represent the probability mass in the tails since they can be interpreted as the rate of increase (decrease) from (to) zero for y_i tending to zero (infinity). For $\alpha \in (0, 1)$, we furthermore get an explicit form to compute quantiles of the distribution as

$$F_i^{-1}(\alpha) = b_i \left(\alpha^{-1/p_i} - 1\right)^{-1/a_i}. \tag{3}$$

If the probability of measuring a particular value varies inversely as a power of that value, e.g. the density of y_i can be written in terms of $c y_i^{-\alpha}$ for some constant c, the

Pareto distribution is one famous distribution widely used in physics, social science, biology, finance and earth sciences to model different quantities of interest like city populations, sizes of earthquakes, wars, sales of books, compare Newman [2005] for further examples and references. With the parametrisation given in Table 1 the expectation is given by $p_i/(b_i - 1)$, where $p_i > 0$ is a shape parameter which is known as the tail or Pareto index and $b_i > 0$ is a scale parameter that corresponds to the mode of the distribution.

Discrete Responses Discrete responses occur frequently in practice, for example in explaining the number of citations of patents based on patent characteristics, in predicting the number of insurance claims of policyholders on the basis of previous claim histories [Denuit and Lang, 2004], or in modelling mortality due to a specific type of diseases (disease mapping). Compared to standard Poisson regression one often faces the problems of excess of zeros (zero-inflation), when the number of zeros is larger than expected from a Poisson distribution, and overdispersion, where the variance exceeds the expectation of the true distribution. Appropriate approaches to overcome the limitations of Poisson regression are the negative binomial distribution with an additional overdispersion parameter $\delta_i > 0$ or zero-inflated distributions, in which structural zeros are introduced with probability $\pi_i \in (0, 1)$. The resulting density can be written in mixed form as

$$f_i(y_i) = \pi_i \mathbb{1}_{\{0\}}(y_i) + (1 - \pi_i) g_i(y_i),$$

with density g_i corresponding to a count data distribution. A detailed description of zero inflated and overdispersed count data regression with real data analysis is described in Klein et al. [2013b].

Mixed Discrete-Continuous Distributions In many applications, e.g. insurances [Klein et al., 2013a, Heller et al., 2006] or weather forecasts [Gneiting and Ranjan, 2011], distributions with point masses at zero are of great interest. The flexibility of distributional regression allows to consider such mixed distributions where the (conditional) density has the general form

$$f_i(y_i) = (1 - \pi_i) \mathbb{1}_{\{0\}}(y_i) + \pi_i g_i(y_i)(1 - \mathbb{1}_{\{0\}}(y_i))$$

with $g(y_i)$ any parametric density of a positive real random variable and π_i is the probability of observing a value of y_i greater than zero. For example in case of claim sizes arising in car insurances such models give information about the probability of observing a positive claim and the distribution of the corresponding conditional claim size in one single model. The expectation of y_i is then decreased by the factor π_i compared to the expectation of the continuous part.

Distributions with Compact Support Like in our application of cereal products, beta regression is a useful tool to describe the conditional distribution of responses that take values in a pre-specified interval such as $(0, 1)$ for proportions or relative amounts of quantities of interest. With the parametrisation given in Table 1, $E(y_i) = \mu_i \in (0, 1)$ holds and the second parameter $\sigma_i^2 \in (0, 1)$ is proportional to the variance $\text{Var}(y_i) = \sigma_i^2 \mu_i (1 - \mu_i)$. An extension of beta regression and a special case of mixed-discrete continuous distributions described before is the zero-one-inflated beta distribution where $y_i = 0$ or $y_i = 1$ are assigned positive probabilities and the probabilities for these boundary values can be estimated in dependency of covariates. The additional parameters $\nu_i > 0$ and $\tau_i > 0$ control the probabilities $p_1 = f_i(y_i = 0) = \nu_i/(1+\nu_i+\tau_i)$ and $p_2 = f_i(y_i = 1) = \tau_i/(1+\nu_i+\tau_i)$. Note that this definition ensures a complete probability smaller or equal to one. The expectation of y_i, compared to a beta distributed variable, is reduced by the factor $1 - p_1 - p_2$ but with additional term p_2:

$$E(y_i) = \left(1 - \frac{\nu_i + \tau_i}{1 + \nu_i + \tau_i}\right) \mu_i + \frac{\tau_i}{1 + \nu_i + \tau_i}.$$

The special cases in which either $p_1 = 0$ or $p_2 = 0$ result in a one-inflated and zero-inflated beta distribution with expectations $\frac{\mu_i+\tau_i}{1+\tau_i}$ and $\frac{\mu_i}{1+\nu_i}$, respectively.

2.2 Structured Additive Predictors

2.2.1 Generic Structure

In many recent applications, linear regression models are too restrictive to capture the underlying true, complex structure of real life problems. We therefore consider structured additive distributional regression models, a generic framework in which each of the K model parameters $\boldsymbol{\vartheta}_k = (\vartheta_{1k}, \ldots, \vartheta_{nk})'$, $k = 1, \ldots, K$ is related to a

1. Continuous distributions on $\mathbb{R}$	Density	Parameters
Normal	$f(y\|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$	$\mu \in \mathbb{R}, \sigma^2 > 0$
t	$f(y\|\mu,\sigma^2,n_d) = \frac{\Gamma((n_d+1)/2)}{\Gamma(1/2)\Gamma(n_d/2)\sqrt{n_d\sigma^2}} \left(1+\frac{(y-\mu)^2}{n_d\sigma^2}\right)^{-\frac{n_d+1}{2}}$	$\mu \in \mathbb{R}, n_d, \sigma^2 > 0$
2. Continuous distributions on $\mathbb{R}^+$		
Log-normal	$f(y\|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}y} \exp\left(-\frac{(\log(y)-\mu)^2}{2\sigma^2}\right)$	$\mu \in \mathbb{R}, \sigma^2 > 0$
Inverse Gaussian	$f(y\|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}y^{3/2}} \exp\left(-\frac{(y-\mu)^2}{2y\mu^2\sigma^2}\right)$	$\mu, \sigma^2 > 0$
Gamma	$f(y\|\mu,\sigma) = \left(\frac{\sigma}{\mu}\right)^{\sigma} \frac{y^{\sigma-1}}{\Gamma(\sigma)} \exp\left(-\frac{\sigma}{\mu}y\right)$	$\mu, \sigma > 0$
Weibull	$f(y\|\lambda,\alpha) = \frac{\alpha y^{\alpha-1}\exp(-(y/\lambda)^{\alpha})}{\lambda^{\alpha}}$	$\alpha, \lambda > 0$
Pareto	$f(y\|b,p) = pb^p(y+b)^{-p-1}$	$b, p > 0$
Generalized gamma	$f(y\|\mu,\sigma,\tau) = \left(\frac{\sigma}{\mu}\right)^{\sigma\tau} \frac{\tau y^{\sigma\tau-1}}{\Gamma(\sigma)} \exp\left(-\left(\frac{\sigma}{\mu}y\right)^{\tau}\right)$	$\mu, \sigma, \tau > 0$
Dagum	$f(y\|a,b,p) = \frac{apy^{ap-1}}{b^{ap}(1+(y/b)^a)^{p+1}}$	$a, b, p > 0$
3. Discrete distributions		
Poisson	$f_1(y\|\lambda) = \frac{\lambda^y \exp(-\lambda)}{y!}$	$\lambda > 0$
Negative binomial	$f_2(y\|\mu,\delta) = \frac{\Gamma(y+\delta)}{\Gamma(y+1)\Gamma(\delta)} \left(\frac{\delta}{\delta+\mu}\right)^{\delta} \left(\frac{\mu}{\delta+\mu}\right)^{y}$	$\mu, \delta > 0$
Zero-inflated Poisson	$f(y\|\pi,\mu,\delta) = \pi\mathbb{1}_{\{0\}}(y) + (1-\pi)f_1$	$\pi \in (0,1)$
Zero-inflated negative binomial	$f(y\|\pi,\mu,\delta) = \pi\mathbb{1}_{\{0\}}(y) + (1-\pi)f_2$	$\pi \in (0,1)$
4. Mixed discrete-continuous distributions		
Zero-adjusted	$f(y\|\pi,g(y)) = (1-\pi)\mathbb{1}_{\{0\}}(y) + \pi g(y)\mathbb{1}_{(0,\infty)}(y)$ $g(y)$ a distribution from 2.	$\pi \in (0,1)$
5. Distributions with compact support		
Beta	$f(y\|\mu,\sigma^2) = \frac{y^{p-1}(1-y)^{q-1}}{\mathrm{B}(p,q)}$ $\mu = \frac{p}{p+q}$, $\sigma^2 = \frac{1}{p+q+1}$	$\mu, \sigma^2 \in (0,1)$
Zero-One-inflated Beta	$f(y\|\mu,\sigma^2,\nu,\tau) = \begin{cases} \frac{\nu}{1+\nu+\tau} & y = 0 \\ \left(1-\frac{\nu+\tau}{1+\nu+\tau}\right)\frac{y^{p-1}(1-y)^{q-1}}{\mathrm{B}(p,q)} & y \in (0,1) \\ \frac{\tau}{1+\nu+\tau} & y = 1 \end{cases}$	$\nu, \tau > 0$
Zero-inflated Beta	$f(y\|\mu,\sigma^2,\nu) = \begin{cases} \frac{\nu}{1+\nu} & y = 0 \\ \left(1-\frac{\nu}{1+\nu}\right)\frac{y^{p-1}(1-y)^{q-1}}{\mathrm{B}(p,q)} & y \in (0,1) \end{cases}$	$\nu > 0$
One-inflated Beta	$f(y\|\mu,\sigma^2,\tau) = \begin{cases} \left(1-\frac{\tau}{1+\tau}\right)\frac{y^{p-1}(1-y)^{q-1}}{\mathrm{B}(p,q)} & y \in (0,1) \\ \frac{\tau}{1+\tau} & y = 1 \end{cases}$	$\tau > 0$

Table 1: List of important response distributions in distributional regression.

semiparametric predictor with the general form

$$\eta_i^{\vartheta_k} = \beta_0^{\vartheta_k} + f_1^{\vartheta_k}(\boldsymbol{\nu}_i) + \ldots + f_{J_k}^{\vartheta_k}(\boldsymbol{\nu}_i)$$

where β_0 represents the overall level of the predictor and the functions $f_j(\boldsymbol{\nu}_i)$, $j = 1, \ldots, J_k$, relate to different covariate effects required in the applications. Note that of course each parameter vector $\boldsymbol{\vartheta}_k$ may depend on different covariates and especially a different number of effects J_k. To simplify notation, we suppress this possibility and also drop the parameter index in the following.

In structured additive regression, each function f_j is approximated by a linear combination of D_j appropriate basis functions, i.e. $f_j(\boldsymbol{\nu}_i) = \sum_{d_j=1}^{D_j} \beta_{j,d_j} B_{j,d_j}(\boldsymbol{\nu}_i)$ such that in matrix notation we can write $\boldsymbol{f}_j = (f_j(\boldsymbol{\nu}_1), \ldots, f_j(\boldsymbol{\nu}_n))' = \boldsymbol{Z}_j\boldsymbol{\beta}_j$ where $\boldsymbol{Z}_j[i, d_j] = B_{j,d_j}(\boldsymbol{\nu}_i)$ is a design matrix and $\boldsymbol{\beta}_j$ is the vector of coefficients to be estimated. In the next section, we will give some examples to elucidate on the potential of structured additive regression.

For regularisation reasons it is common to add a penalty term $\text{pen}(\boldsymbol{f}_j) = \text{pen}(\boldsymbol{\beta}_j) = \boldsymbol{\beta}_j'\boldsymbol{K}_j\boldsymbol{\beta}_j$ that controls specific smoothness or sparseness properties. The Bayesian equivalent to this frequentist formulation is to put multivariate Gaussian priors

$$p(\boldsymbol{\beta}_j) \propto \left(\frac{1}{\tau_j^2}\right)^{\frac{\text{rk}(\boldsymbol{K}_j)}{2}} \exp\left(-\frac{1}{2\tau_j^2}\boldsymbol{\beta}_j'\boldsymbol{K}_j\boldsymbol{\beta}_j\right) \tag{4}$$

on the regression coefficients $\boldsymbol{\beta}_j$ with prior precision matrix $\boldsymbol{K}_j$ which corresponds to the penalty matrix in a frequentist formulation. The hyperparameters τ_j^2 are assigned inverse gamma hyperpriors $\tau_j^2 \sim \text{IG}(a_j, b_j)$ (with $a_j = b_j = 0.001$ as a default option) in order to obtain a data-driven amount of smoothness.

2.2.2 Special Cases

Linear Effects For linear effects we write $f_j(\boldsymbol{\nu}_i) = \boldsymbol{x}_i'\boldsymbol{\beta}_j$ with $\boldsymbol{x}_i$ a vector of original binary or categorical covariates. The design matrix $\boldsymbol{Z}_j$ then consists of the column vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. In general, for linear effects a noninformative prior is chosen such that $\boldsymbol{K}_j = \boldsymbol{0}$. An alternative for cases where the dimension of $\boldsymbol{\beta}_j$ is large is the ridge prior with $\boldsymbol{K}_j = \boldsymbol{I}$.

Continuous Covariates For potentially nonlinear effects $f_j(\boldsymbol{\nu}_i) = f_j(x_i)$ of a single continuous covariate x_i, P(enalised)-splines [Eilers and Marx, 1996] are a convenient

and parsimonious modelling framework where $f_j(x_i)$ is approximated by a linear combination of $D_j = m_j + l_j - 1$ B-spline basis functions that are constructed from piecewise polynomials of a certain degree l_j upon an equidistant grid of knots and under certain regularity assumptions to achieve the desired smoothness constraints. To be more specific, assume an equidistant grid of inner knots $\kappa_0, \ldots, \kappa_{m_j-1}$ within the range of x_i. Each B-spline basis function consists of (l_j+1) polynomials of degree l_j which are joint in an $(l_j - 1)$-times continuous differentiable way. Advantages of B-splines are that they are a local basis with positive values on $l_j + 2$ knots each and that they are bounded from above (and therefore cause less numerical problems compared to the truncated power series). Choices about the number of knots and the degree of B-splines are discussed in Lang and Brezger [2004]. We will usually stick to the recommended values of twenty inner knots and cubic B-splines.

Regularisation of the function estimates is realised by the introduction of roughness penalties or by imposing an appropriate prior assumption. The stochastic analogues for the first or second order difference penalty suggested by Eilers and Marx [1996] are a first or second order random walk. They are defined by

$$\begin{aligned} \beta_{j,d_j} &= \beta_{j,d_j-1} + \epsilon_{j,d_j}, & d_j &= 2, \ldots, D_j \\ \beta_{j,d_j} &= 2\beta_{j,d_j-1} - \beta_{j,d_j-2} + \epsilon_{j,d_j}, & d_j &= 3, \ldots, D_j \end{aligned}$$

with Gaussian errors $\epsilon_{j,d_j} \sim \mathrm{N}(0, \tau_j^2)$ and noninformative priors for β_{j1} or β_{j1} and β_{j2}. The joint distribution of $\boldsymbol{\beta}_j$ is then given as the product of the conditional densities and follows the form of equation (4) with $\boldsymbol{K}_j = \boldsymbol{D}'\boldsymbol{D}$, where $\boldsymbol{D}$ is a difference matrix of appropriate order.

Spatial Effects For a discrete spatial variable observed on an irregular grid or regions, Markov random fields are a common approach for spatial effects, see Rue and Held [2005] for a general introduction. The generic effect $f_j(\boldsymbol{\nu}_i)$ is then given by $f_j(s_i)$ where $s_i \in \{1, \ldots, S\}$ denotes the spatial index or region of the i-th observation. Usually we estimate one separate coefficient β_{js} for each region and collect them in the vector $(\beta_{j1}, \ldots, \beta_{jS})' = \boldsymbol{\beta}_j \in \mathbb{R}^S$ such that the design matrix is an indicator matrix connecting individual observations to the corresponding region, i.e. $\boldsymbol{Z}[i, s] = 1$ if observation i belongs to region s and zero otherwise. To enforce spatial smoothness, we assume a neighbourhood structure ∂_s, i.e. two regions are

neighbours if they share common borders, and establish an adjacency matrix $\boldsymbol{K}_j$ indicating which regions are neighbours. If two regions are neighbours we write $r \sim s$ and $r \not\sim s$ otherwise. The simplest spatial smoothness prior is then given by

$$\beta_{j,s}|\beta_{j,r},\ r \neq s,\ \tau_j^2 \sim \mathrm{N}\left(\sum_{r \in \partial_s} \frac{1}{N_s}\boldsymbol{\beta}_{j,r}, \frac{\tau_j^2}{N_s}\right),$$

where N_s is the number of neighbours of region s. In consequence, the conditional mean of $\beta_{j,s}$ given all other coefficients is the average of the neighbourhood regions. This implies the penalty matrix $\boldsymbol{K}_j$

$$\boldsymbol{K}_j[s,r] \quad = \quad \begin{cases} -1 & s \neq r, \quad s \sim r \\ 0 & s \neq r, \quad s \not\sim r \\ N_s & s = r. \end{cases}$$

Random Effects Correlations between repeated measurements of the same individual or cluster can be taken into account to some extend by modelling additional individual- or cluster-specific random effects $f_j(\boldsymbol{\nu}_i) = \beta_{j,g_i}$ where $g_i \in \{1, \ldots, G\}$ indicates one of the G different groups observation y_i belongs to. In this case, the design matrix $\boldsymbol{Z}_j$ is an incidence matrix with zeros and ones, linking individual observations or clusters, e.g $\boldsymbol{Z}_j[i,g] = 1$ if observation i is in group g and zero otherwise. We assume coefficients of different groups to be i.i.d. distributed such that the penalty matrix $\boldsymbol{K}_j$ reduces to the identity matrix, i.e. $\boldsymbol{K}_j = \boldsymbol{I}$.

Further Basis Function Representations The so far presented types of effects are just a selection of the most important function estimates related to the applications in Sections 5 and 6. A more detailed exposition of further regression specifications comprising also bivariate surfaces based on 2D-basis function evaluations and penalty matrices arising from tensor products $\boldsymbol{K}_j = \boldsymbol{I} \otimes \boldsymbol{K}_{j,1} + \boldsymbol{K}_{j,2} \otimes \boldsymbol{I}$ with penalty matrices $\boldsymbol{K}_{j,1}$ and $\boldsymbol{K}_{j,2}$ as for univariate P-splines, kriging based on correlation functions or varying coefficient terms where an interaction variable interacts with a smooth term, can be found in Fahrmeir et al. [2013]. Multiplicative random effects or random scaling factors of the form $(1 + \beta_{j,g_i})f_j(x_i)$ have been discussed in Lang et al. [2013a]. Here, f_j is a smooth function of the continuous covariate x_i modelled by P-splines and β_{j,g_i} is a random scaling factor, accounting for unobserved

heterogeneity by allowing the covariate curves f_j to vary between different groups g_i. Monotonicity constraints are very helpful in cases where smooth functions of continuous covariates are presumed to have an monotonic relation to the response variable, see Brezger and Steiner [2008]. All these extensions are readily available in a Bayesian treatment of structured additive regression.

3 Bayesian Inference

The complex likelihood structures of non-standard distributions utilised in distributional regression result in most cases in full conditionals for the unknown regression coefficients that are not analytically accessible. In this section, we will therefore present a procedure that leads to a generic Metropolis-Hastings algorithm with distribution-specific working weights and working responses. The proposal densities are then based on iteratively weighted least squares (IWLS) approximations to the full conditionals. If for specific distribution parameters the Gaussian priors yield a conjugate model structure, the Metropolis-Hastings update is replaced by a Gibbs sampling step. Note that updating the smoothing variances τ_j^2 is always realised by a Gibbs update since the full conditionals follow inverse gamma distributions

$$\tau_j^2|\cdot \sim \text{IG}(a_j', b_j'), \quad a_j' = \frac{\text{rk}(\boldsymbol{K}_j)}{2} + a_j, \quad b_j' = \frac{1}{2}\boldsymbol{\beta}_j'\boldsymbol{K}_j\boldsymbol{\beta}_j + b_j \tag{5}$$

with updated parameters a_j', b_j'.

3.1 Approximations to the Full Conditionals

Since the full conditionals ($p(\boldsymbol{\beta}_j|\cdot)$) (conditional distribution of $\boldsymbol{\beta}_j$ given all other parameters and the data) cannot be written in closed form, it is obvious to find suitable approximations for them. In this section, we assume a generic predictor $\boldsymbol{\eta}$ and drop the parameter index for simplicity. For a typical parameter block $\boldsymbol{\beta}_j$, the full conditional is proportional to $l(\boldsymbol{\eta}) - \frac{1}{2\tau_j^2}\boldsymbol{\beta}_j'\boldsymbol{K}_j\boldsymbol{\beta}_j$, where $l(\boldsymbol{\eta})$ denotes the log-likelihood part depending on $\boldsymbol{\eta}$. In a frequentist setting, this can be interpreted as a penalised log-likelihood. Finding iteratively the roots of the derivative of the Taylor expansion to the full conditional of degree two around the mode leads to a Newton method type

approximation of the form

$$\frac{\partial l^{[t]}}{\partial \eta_i} - \frac{\partial^2 l^{[t]}}{\partial \eta_i^2} \cdot \left(\eta_i^{[t+1]} - \eta_i^{[t]}\right) = 0$$

where t indexes the iteration of the Newton algorithm. Since the maximum likelihood estimate is asymptotically normal distributed with zero mean and expected Fisher information as covariance matrix, we interpret the working observations $\boldsymbol{z}^{[t]} = \boldsymbol{\eta}^{[t]} + \left(\boldsymbol{W}^{[t]}\right)^{-1} \boldsymbol{v}^{[t]}$ as random variables with distribution

$$\boldsymbol{z}^{[t]} \sim \mathrm{N}\left(\boldsymbol{\eta}^{[t]}, \left(\boldsymbol{W}^{[t]}\right)^{-1}\right)$$

where $\boldsymbol{v}^{[t]} = \partial l^{[t]}/\partial \boldsymbol{\eta}$ is the score vector and $\boldsymbol{W}^{[t]}$ are working weight matrices with $w_i^{[t]} = E(-\partial^2 l^{[t]}/\partial \eta_i^2)$ on the diagonals and zero otherwise. The proposal density for $\boldsymbol{\beta}_j$ is then constructed from the resulting working model for $\boldsymbol{z}$ as $\boldsymbol{\beta}_j \sim \mathrm{N}(\boldsymbol{\mu}_j, \boldsymbol{P}_j^{-1})$ with expectation and precision matrix

$$\boldsymbol{\mu}_j = \boldsymbol{P}_j^{-1} \boldsymbol{Z}_j' \boldsymbol{W} (\boldsymbol{z} - \boldsymbol{\eta}_{-j}) \qquad \boldsymbol{P}_j = \boldsymbol{Z}_j' \boldsymbol{W} \boldsymbol{Z}_j + \frac{1}{\tau_j^2} \boldsymbol{K}_j \tag{6}$$

where $\boldsymbol{\eta}_{-j} = \boldsymbol{\eta} - \boldsymbol{Z}_j \boldsymbol{\beta}_j$ is the predictor without the j-th component.

Algorithm As shown in Section 2.2, every single effect $\boldsymbol{f}_j$ contributing to one of the predictors $\boldsymbol{\eta}_k$ is determined by an appropriate design matrix $\boldsymbol{Z}_j^{\vartheta_k}$ and a penalty structure $\boldsymbol{K}_j^{\vartheta_k}$ with smoothing variances $\left(\tau_j^{\vartheta_k}\right)^2$ sampled from (5). For the MCMC sampler, the working observations $\boldsymbol{z}^{\vartheta_k}$ and the working weights $\boldsymbol{W}^{\vartheta_k}$ are required. Given these quantities, the resulting Metropolis Hastings algorithm can be summarised as follows: Fix the number of MCMC iterations T. While $t < T$ loop over all distribution parameters $\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_K$ and for $j = 1, \ldots, J_k$, $k = 1, \ldots, K$, draw a proposal $\boldsymbol{\beta}_j^p$ from the density

$$q\left(\left(\boldsymbol{\beta}_j^{\vartheta_k}\right)^{[t]}, \boldsymbol{\beta}_j^p\right) = \mathrm{N}\left(\left(\boldsymbol{\mu}_j^{\vartheta_k}\right)^{[t]}, \left(\left(\boldsymbol{P}_j^{\vartheta_k}\right)^{[t]}\right)^{-1}\right)$$

with expectation $\boldsymbol{\mu}_j^{\vartheta_k}$ and precision matrix $\boldsymbol{P}_j^{\vartheta_k}$ given in (6). Accept $\boldsymbol{\beta}_j^p$ as a new state of $\left(\boldsymbol{\beta}_j^{\vartheta_k}\right)^{[t]}$ with acceptance probability

$$\alpha\left(\left(\boldsymbol{\beta}_j^{\vartheta_k}\right)^{[t]}, \boldsymbol{\beta}_j^p\right) = \min\left\{\frac{p\left(\boldsymbol{\beta}_j^p|\cdot\right) q\left(\boldsymbol{\beta}_j^p, \left(\boldsymbol{\beta}_j^{\vartheta_k}\right)^{[t]}\right)}{p\left(\left(\boldsymbol{\beta}_j^{\vartheta_k}\right)^{[t]}\middle|\cdot\right) q\left(\left(\boldsymbol{\beta}_j^{\vartheta_k}\right)^{[t]}, \boldsymbol{\beta}_j^p\right)}, 1\right\}.$$

To solve the identifiability problem inherent to additive models, correct the sampled effect according to Algorithm 2.6 in Rue and Held [2005] such that $\boldsymbol{A}\boldsymbol{\beta}_j = \mathbf{0}$ holds with an appropriate matrix $\boldsymbol{A}$. For updating $\left(\tau_j^{\vartheta_k}\right)^2$, generate a random number from the inverse Gamma distribution IG $\left(a_j', (b_j')^{[t]}\right)$ with a_j' and $(b_j')^{[t]}$ given in (5).

Working Weights In principle, this algorithm is applicable to all distributions where first and second derivative of the log-likelihood exists. However, in the previous section it became obvious that the IWLS proposal densities are distribution-specific, depending on the score vectors $\boldsymbol{v}$, the first derivatives of the log-likelihood with respect to the different predictors and the weights $\boldsymbol{W}$. For $\boldsymbol{W}$ we usually choose the diagonal matrix with expectations of the negative second derivatives of the log-likelihood similar as in a Fisher-scoring algorithm. Alternatives would be to simply take the negative second derivative (Newton-Raphson type) or the quadratic score (quasi-Newton-Raphson), since the computations of the required expectations can be quite challenging in cases of complex likelihood structures. Nevertheless, we observed that it is quite worth computing the expectations because of positive mixing behaviours in combination with better acceptance rates, as well as numerical stability. Furthermore, in many cases it is possible to show that the working weights as defined in the previous section are positive definite when utilising the expected negative second derivatives. This ensures that the precision matrix $\boldsymbol{P}_j$ of the proposal density is invertible if the design matrices $\boldsymbol{Z}_j$ have full column rank. Explicit derivations of score vectors and working weights for distributions of Table 1, as well as the consideration of their positiveness can be found in Section C of the supplement.

Propriety of the Posterior The question whether the posterior distribution is proper is justified since the model specifications include several partially improper priors. Sun et al. [2001] or Fahrmeir and Kneib [2009] treated this question in exponential family regression and Klein et al. [2013b] generalised these results to the distributional framework of structured additive regression in the very special case of count data regression. However, the sufficient conditions derived in this paper do not require count data distributions and hold for all distributions presented in Section 2.1 such that the findings directly carry over to structured additive distributional regression.

Software The presented regression models are implemented in the free, open source software BayesX [Belitz et al., 2012]. As described in Lang et al. [2013b] the implementation makes use of efficient storing even for large data sets and sparse matrix algorithms for sampling from multivariate Gaussian distributions. An additional feature is the access to hierarchical models which have been proposed by Lang et al. [2013b]. The idea of the multilevel regression is to allow for hierarchical prior specifications for regression effects where each parameter vector may again be assigned an additive predictor.

4 Choice of the Response Distribution and the Predictor Specifications

The framework of distributional regression has the advantage to provide several flexible candidates of distributions for discrete, continuous as well as mixed discrete-continuous distributions and comprises regression for model parameters referring to location, scale, shape or further parameters of the response distribution. The approach allows therefore to focus on special aspects of the data that go beyond the mean. However, it is required that users are aware of the characteristics of distributions they use and tools are needed to facilitate both the choice of suitable conditional distributions and variable selection in all parameters of the distribution. We propose normalized quantile residuals [Dunn and Smyth, 1996], the deviance information criterion (DIC) [Spiegelhalter et al., 2002], as well as proper scoring rules [Gneiting and Raftery, 2007] and combine these tools to determine final models under the aspects of the fit to the data (quality of estimation) and the predictive ability in terms of probabilistic forecasts (quality of prediction).

4.1 DIC

The DIC is a commonly used criterion for model choice in Bayesian inference that has become quite popular due to the fact that it can easily be computed from the MCMC output. If $\boldsymbol{\vartheta}^{[1]}, \ldots, \boldsymbol{\vartheta}^{[T]}$ is a MCMC sample from the posterior for the complete parameter vector, then the DIC is based on the deviance $D(\boldsymbol{\vartheta}) = -2\log(f(\boldsymbol{y}|\boldsymbol{\vartheta}))$ of

the model and the effective number of parameters p_D [Spiegelhalter et al., 2002]. The latter can be shown to equal the difference between the posterior mean of the deviance and deviance of the posterior means for the parameters, i.e. $p_D = \overline{D(\boldsymbol{\vartheta})} - D(\overline{\boldsymbol{\vartheta}})$ where

$$\overline{D(\boldsymbol{\vartheta})} = \frac{1}{T}\sum_{t=1}^{T} D\left(\boldsymbol{\vartheta}^{[t]}\right) \quad \text{and} \quad \overline{\boldsymbol{\vartheta}} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\vartheta}^{[t]}.$$

The DIC is then defined as

$$\text{DIC} = \overline{D(\boldsymbol{\vartheta})} + 2p_D = 2\overline{D(\boldsymbol{\vartheta})} - D(\overline{\boldsymbol{\vartheta}})$$

indicating a close relationship to the frequentist Akaike information criterion that provides a similar compromise between fidelity to the data and model complexity. A rough rule of thumb says that DIC differences of 10 and more between two competing models indicate the model with the lower DIC to be superior. For a fixed response distribution, we suggest to use the DIC for variable selection in all predictors. If for one distribution with different parameter specifications several models have similar DIC with differences smaller than 10, we usually decide for the more parsimonious models in the sense that additional non-significant effects are excluded from the predictors. It has turned out as a helpful procedure to focus in a first step on specifying the location parameter and to determine the remaining predictors by a stepwise search from either very simple models just including constants or rather complex specifications depending on the data and number of covariates. For count data models, the performance of the DIC was positively evaluated in Klein et al. [2013b] where several misspecified models have been compared to the true model in terms of the DIC.

4.2 Quantile Residuals

If $F_i(\cdot|\hat{\boldsymbol{\vartheta}}_i)$ is the assumed distribution with plugged in estimates, quantile residuals are given by $\hat{r}_i = \Phi^{-1}(u_i)$ with inverse cumulative distribution function of a standard normal distribution Φ^{-1} and $u_i = F_i(y_i|\hat{\boldsymbol{\vartheta}}_i)$ if y_i is a realisation of a continuous response variable, while u_i is a random number from the uniform distribution on the interval $[F_i(y_i - 1|\hat{\boldsymbol{\vartheta}}_i), F_i(y_i|\hat{\boldsymbol{\vartheta}}_i)]$ for discrete responses. If the estimated model is close to the true model, the quantile residuals approximately follow a standard normal distribution, even if the model distribution itself is not a normal distribution, compare Dunn and Smyth [1996]. In practice, the residuals can be assessed graphically in

terms of quantile-quantile-plots: the closer the residuals to the bisecting line, the better the fit to the data. We suggest to use quantile residuals as an effective tool for deciding between different distributional options where strong deviations from the bisecting line allow us to sort out distributions that do not fit the data well.

4.3 Proper Scoring Rules

Gneiting and Raftery [2007] propose proper scoring rules as summary measures for the evaluation of probabilistic forecasts based on the predictive distribution and the observed realisations. Let $y_1 \ldots, y_R$, be data in a hold-out sample and F_r the predictive distributions with predicted parameter vectors $\hat{\boldsymbol{\vartheta}}_r = (\hat{\vartheta}_{r1}, \ldots, \hat{\vartheta}_{rK})'$. The score is obtained by summing over individual contributions, $S = \frac{1}{R}\sum_{r=1}^{R} S(F_r, y_r)$. If $F_{r,0}$ is the true distribution Gneiting and Raftery [2007] suggest to take the expected value of the score under $F_{r,0}$ in order to compare different scoring rules. A scoring rule is proper if the expectation of the score $S(F_{r,0}, F_{r,0})$ fulfils $S(F_{r,0}, F_{r,0}) \geq S(F_r, F_{r,0})$ for any predictive distribution $F_{r,0}$ and it is strictly proper if equality holds if and only if $F_r = F_{r,0}$. This means that a scoring rule is proper if the expected score for an observation drawn from F_r is maximised if F_r is issued rather than $\tilde{F}_r \neq F_r$.

In practice, we obtain the predictive distributions F_r for observations y_r by cross validation, i.e. the data set is divided into subsets of approximately equal size and predictions for one of the subsets are obtained from estimates based on all the remaining subsets. Since we will only consider proper scores in the following, higher scores deliver better probabilistic forecasts when comparing two different models. We will now discuss specific scores for different types of distributions.

Discrete Distributions In regression models with discrete responses, we consider three common scores, namely the Brier score or quadratic score, $S(f_r, y_r) = -\sum_h (\mathbb{1}(y_r = h) - f_{rh})^2$, the logarithmic score, $S(f_r, y_r) = \log(f_{ry_r})$, and the spherical score $S(f_r, y_r) = \frac{f_{ry_r}}{\sqrt{\sum_h f_{rh}}}$, with $f_{rh} = \mathbb{P}(y_r = h)$. All these scoring rules are strictly proper but the logarithmic scoring rule has the drawback that it only takes into account one single probability of the predictive distribution and is therefore susceptible to extreme observations.

Continuous Distributions For continuous responses, the three mentioned scoring rules can be formulated in terms of density forecasts as the quadratic score $S(f_r, y_r) = 2f_r(y_r) - \|f_r\|_2^2$, the spherical score $S(f_r, y_r) = f_r(y_r)/\|f_r\|_2$ and the logarithmic score $S(f_r, y_r) = \log(f_r(y_r))$, with $\|f_r\|_2^2 = \int f_r(\omega)^2 d\omega$, dominated by the Lebesgue measure. Gneiting and Raftery [2007] give further theoretical details for more general σ-finite measures μ_r on the measurable space $(\Omega, \mathcal{A})$.

Mixed Discrete-Continuous Distributions A greater challenge is to define proper scoring rules in case of mixed distributions as for instances in zero-adjusted models or the zero-one-inflated beta regression model. In order to make the continuous versions of Brier, logarithmic and spherical score accessible for discrete-continuous distributions, consider in case of zero-adjusted distributions the sample space $\Omega = \mathbb{R}_{\geq 0} \cup \{0\}$ with σ-algebra $\mathcal{A} = \mathcal{B}(\mathbb{R}_{\geq 0}) \cup \sigma(\{0\}) = \sigma\left(\{(a, b] | a > 0, b \geq a\}\right) \cup \sigma\left(\{0\}, \emptyset\right)$. It is easy to verify that $\mathcal{A}$ is a σ-algebra. We furthermore define the probability measure μ_r on the measurable space $(\Omega, \mathcal{A})$,

$$\mu_r(A) = \begin{cases} (1 - \pi_r) + \pi_r \int_A g_r(y) dy & 0 \in A \\ \pi_r \int_A g_r(y) dy & 0 \notin A, \end{cases}$$

with $A \in \mathcal{A}$, $\pi_r \in (0, 1)$ and g_r a density of a continuous nonnegative real random variable. Let f_r be a predictive density of the form $f_r = (1 - \pi_r)\mathbb{1}_{\{0\}}(y_r) + \pi_r g_r(1 - \mathbb{1}_{\{0\}}(y_r))$. Scoring rules corresponding to the ones for discrete or continuous variables are the quadratic score $S(f_r, y_r) = 2f_r(y_r) - \|f_r\|_2^2$, the spherical score $S(f_r, y_r) = f_r(y_r)/\|f_r\|_2$ and the logarithmic score $S(f_r, y_r) = \log(f_r(y_r))$, with $\|f_r\|_2^2 = \int f_r(\omega)^2 \mu_r(\mathrm{d}\omega) = (1 - \pi_r)^2 + \pi_r^2 \int g_r(\omega)^2 \mathrm{d}\omega$. Note that the construction of scores for zero-one-inflated beta regression models would work in complete analogy.

Since S is only a summary measure for the complete predictive distribution, it is difficult to assess parts of the the true distribution that are reflected well with the model and which aspects differ from the truth. There may be, for example, a distribution that captures the central part of the distribution well but has difficulties in the tails. Then, it might be helpful to follow an alternative given by Gneiting and Raftery [2007] and Gneiting and Ranjan [2011] who suggest to define scoring rules directly in terms of predictive cumulative distribution functions if the forecasts involve distributions with a point mass at zero. This leads to the continuous ranked probability score

(CRPS), which is defined as $S(F_r, y_r) = -\int_{-\infty}^{\infty} \left(F_r(x) - \mathbb{1}_{\{x \geq y_r\}}\right)^2 \mathrm{d}x$, with predictive cumulative distribution function F_r and threshold x. Laio and Tamea [2007] showed that the CRPS score can also be written in terms of $F_r^{-1}(\alpha)$ for the quantile at level $\alpha \in (0,1)$ as $-2\int_0^1 \left(\mathbb{1}_{\{y_r \leq F_r^{-1}(\alpha)\}} - \alpha\right)\left(F_r^{-1}(\alpha) - y_r\right) \mathrm{d}\alpha$. This formulation allows us not only to look at the sum of all score contributions (whole integral) but also to perform a quantile decomposition [Gneiting and Ranjan, 2011] and to plot the mean quantile scores versus α in order to compare fits of specific quantiles. This decomposition is especially helpful in situations where the quantile score can be interpreted as an economically relevant loss function [Gneiting, 2011].

5 Spatio-Temporal Dynamics of Labour Income in Germany

As a first application, we consider the monthly income of male full-time workers in Germany obtained from the German Socio-Economic Panel (SOEP, [Frick et al., 2007]). For illustration purposes, we use a relatively small subsample of $i = 1, \ldots, 526$ male full time workers with complete longitudinal information available in the calendar time $t = 1996, \ldots, 2011$. To adjust for the sampling design and the drop-out probability, all results will be weighted according to inverse sample probabilities and the probability of staying in the sample. The labour incomes have been inflation-adjusted with 2010 as reference period. The average labour income over all observations is 3,640 Euro with a minimum observed income of 202 Euro and 26,400 Euro as maximum. Because of the nonnegative nature of incomes and the possible right skewness of income distributions, we consider the log-normal, inverse Gaussian, gamma and Dagum distribution as suitable candidates to describe the conditional behaviour of labour incomes. Especially the log-normal and Dagum distribution are used regularly in economic research for analysing income distributions, compare e.g. Chotikapanich [2008] for further references.

To study the spatio-temporal dynamics of labour income, we consider the generic predictor

$$\eta_{it} = f_1(age_{it}) + f_2(t) + f_{spat}(region_{it}) + \beta_i,$$

where f_1 adjusts income for age (one of the most important income determinants), f_2 represents a general time trend, f_{spat} captures spatial heterogeneity and the random effects β_i are included to account for the longitudinal nature of the data. The non-linear effects f_1 and f_2 are modelled as cubic penalised splines with 20 inner knots and second order random walk prior, the spatial effect is based on 96 administrative regions (Raumordnungsregionen) in Germany and decomposed in a smooth part modelled by a Markov random field prior to capture smoothly varying effects as well as an unstructured spatial random effect that represents localised spatial pattern. The latter is assigned an i.i.d. Gaussian prior just like the individual specific random effects β_i. For all parameters of the four distribution candidates, we have compared different predictor specifications based on the DIC in combination with significances of effects. In the Dagum model, the optimal predictors for the parameters a, b, p are

$$\begin{aligned}
\eta_{it}^a &= f_1^a(age_{it}) + f_2^a(t) + f_{spat}^a(region_{it}) + \beta_i^a \\
\eta_{it}^b &= f_1^b(age_{it}) + f_{spat}^b(region_{it}) + \beta_i^b \\
\eta_{it}^p &= f_1^p(age_{it}) + f_2^p(t).
\end{aligned}$$

Assuming an inverse Gaussian distribution yields

$$\begin{aligned}
\eta_{it}^\mu &= f_1^\mu(age_{it}) + f_{spat}^\mu(region_{it}) + \beta_i^\mu \\
\eta_{it}^{\sigma^2} &= f_1^{\sigma^2}(age_{it}) + f_2^{\sigma^2}(t) + f_{spat}^{\sigma^2}(region_{it}) + \beta_i^{\sigma^2}.
\end{aligned}$$

Model specifications of the log-normal and gamma distribution can be found in the supplementary material Section A.1.

For the four models, we first checked the ability to fit the data based on quantile residuals depicted in Figure 1. While none of the distributions provides a perfect fit for the data, the Dagum distribution turns out to be most appropriate for residuals in the range between -3 and 3 but deviates from the diagonal line for extreme residuals. In contrast, the log-normal and inverse Gaussian distribution seem to have problems in capturing the overall shape of the income distribution resulting in sigmoid deviations from the diagonal line but fits better for extreme residuals. While the gamma distribution also seems to fit reasonably well in general, deviations for extreme observations already start at smaller residual values as compared to the Dagum distribution. The DIC resulting from estimations based on the whole data set in Table 2 is the smallest for the Dagum distribution.

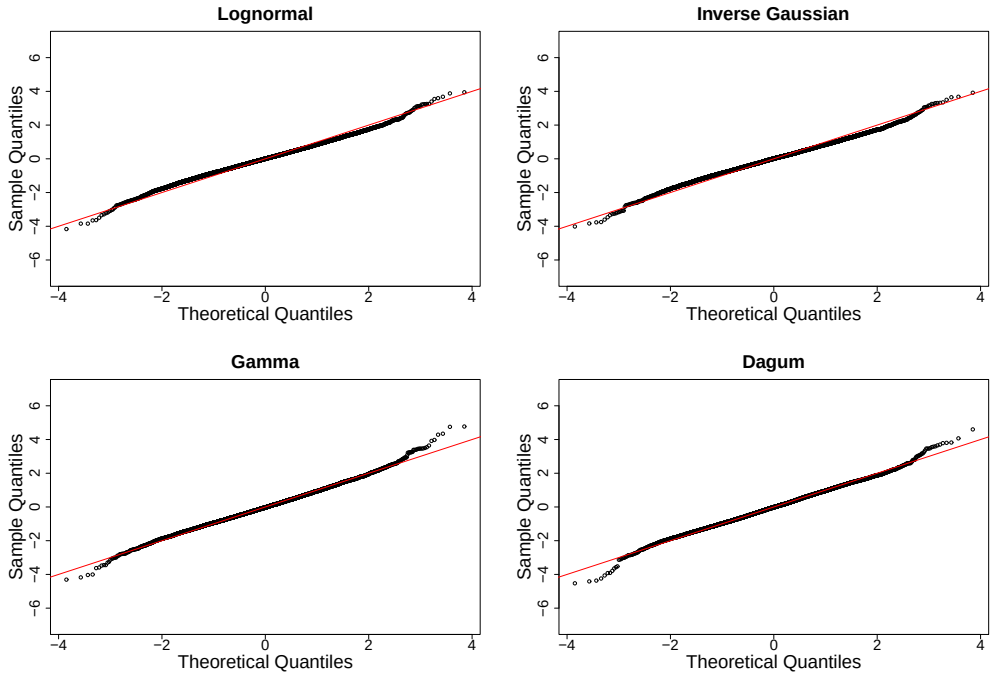

Figure 1: SOEP data. Comparison of quantile residuals for log-normal (topleft), inverse Gaussian (topright), gamma (bottomleft), Dagum (bottomright) distribution.

Distribution	DIC	Quadratic Score	Logarithmic Score	Spherical Score	CRPS
Log-normal	10,854	0.063	-4.324	0.428	-0.998
Inverse Gaussian	11,007	**0.075**	-4.358	**0.431**	**-0.99**
Gamma	10,841	0.019	-5.793	0.412	-1.006
Dagum	**10,666**	-0.006	**-3.053**	0.403	-1.004

Table 2: SOEP data. Comparison of DIC achieved of estimates based on the whole data set and average score contributions obtained from ten-fold cross validations.

We also performed a ten-fold cross validation to compute proper scoring rules for evaluating the predictive ability of our models. For the construction of the folds, we left out one tenth of the individuals at a time, i.e. complete longitudinal curves are dropped from the data set and predictions for these curves are obtained from the remaining nine tenth of the data. As scoring rules we used the quadratic, logarithmic, spherical and CRPS score for continuous random variables, see again Table 2. A general result from considering the summarised scores over all observations is that it is difficult to select one specific distribution since the logarithmic score is in favour of the Dagum distribution while the spherical score and CRPS are very close for all four distributions (with a slight tendency towards the inverse Gaussian distribution). Since the logarithmic score reacts susceptible to outliers, it is a good indicator that the Dagum distribution gives a better prediction to extreme observations or outliers. In addition to the sums over the ten folds, the proper scoring rules can also be used to assess the predictive distributions in more detail. We illustrate this along a decomposition of the CRPS over quantile levels of the predictive distribution and a decomposition of the scores over the cross validation folds. For the former, we do no longer look at the complete integral defining the score but at partial contributions. This allows us to study the performance of a specific response distribution with respect to the lower or upper tails of the distribution. Figure 2 illustrates this for the Dagum,

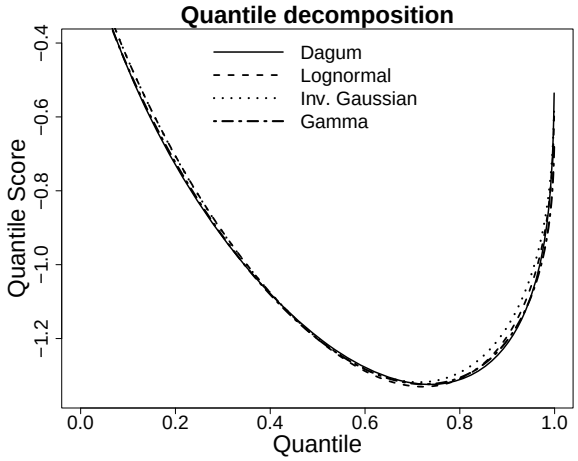

Figure 2: SOEP Data. Quantile decomposition of CRPS.

the log-normal, the inverse Gaussian and gamma distribution. Here we find that the central parts of the distribution are generally described better than the tails and that the increase towards the upper tail is much faster but to a smaller maximum

absolute value than towards the lower tail. Figure 3 displays the summarised scores separately for each of the ten cross-validation folds and for each of the four distribution candidates. Although the quadratic and spherical score are higher for the log-normal and inverse Gaussian distribution, the logarithmic score is conform with the DIC and in favour of the Dagum distribution.

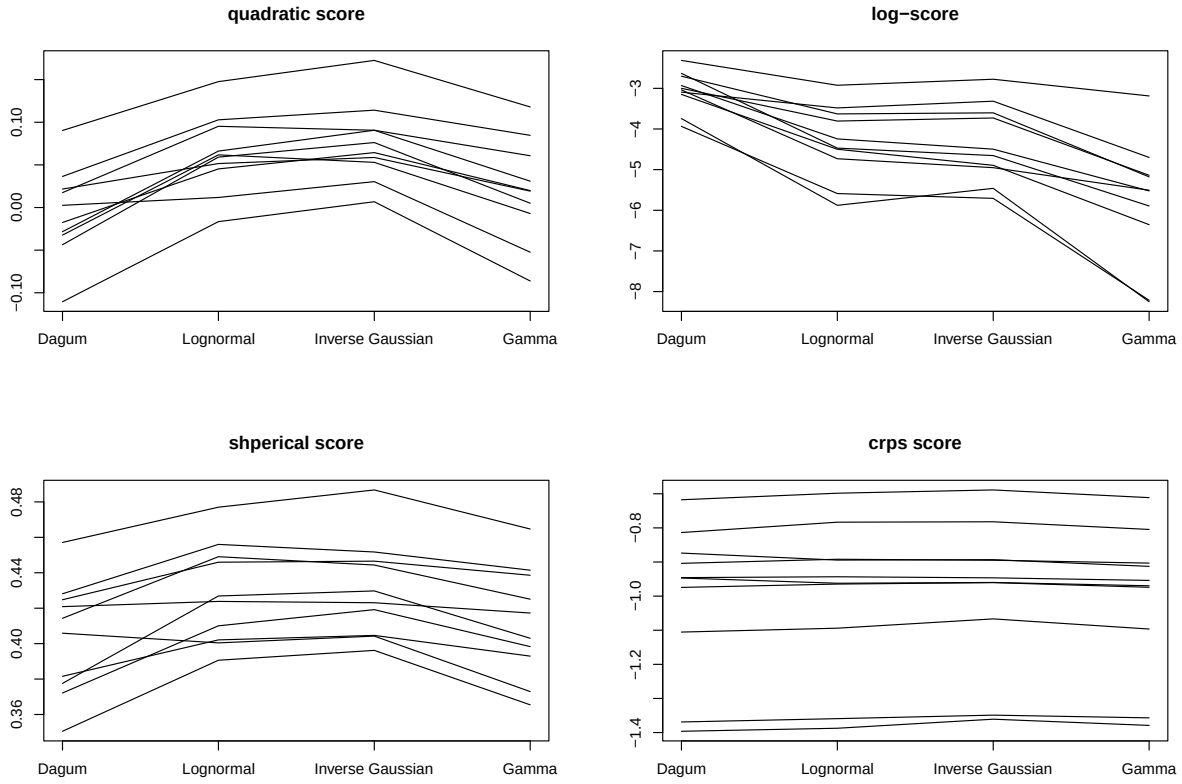

Figure 3: SOEP data. Vertical lines connect the average score contributions of the ten cross validation folds in the Dagum, lognormal, inverse Gaussain and gamma model for the quadratic score (topleft), logarithmic score (topright), spherical score (bottomleft) and the CRPS (bottomright).

Based on the model fit evaluated by quantile residuals as well as the DIC and logarithmic score, we illustrate the estimation results for our data set with the Dagum distribution. Since based on the remaining scores, the inverse Gaussian distribution would also be a reasonable candidate, the corresponding results of the inverse Gaussian distribution are shown in the supplement Section A.3. In addition, we give direct comparisons of selected figures in the following.

Equations (1), (2) and (3) of Section 2.1 indicate already that it is not straight forward to interpret the estimated effects directly on the different parameters of the Dagum distribution, but the samples of MCMC allow us to use these equations in

order to compute several quantities of the income distribution when one effect varies and all other effects are kept constant. We therefore depict in Figure 4 (first row) the estimated posterior expectation, 5% and 95% quantiles, as well as the posterior mode of the income distribution over the two nonlinear effects *age* and *year* while the other effects are set to their posterior mean at average covariate values. In the second row, the same quantities are shown for the inverse Gaussian model where the mode is computed as $\mu\left(\sqrt{1+2.25\mu^2\sigma^4}-1.5\mu\sigma^2\right)$. The corresponding estimates of expectation, mode as well as quantiles of the income distribution varying over regions are shown in Figure 5 for the Dagum distribution and in Figure A6 of the supplement for the inverse Gaussian distribution. The raw effects (centred around zero) of *age* on b and μ are shown in Figure 6 while figures of all effects and parameters for both distributions can be found in the supplement Section A.2.

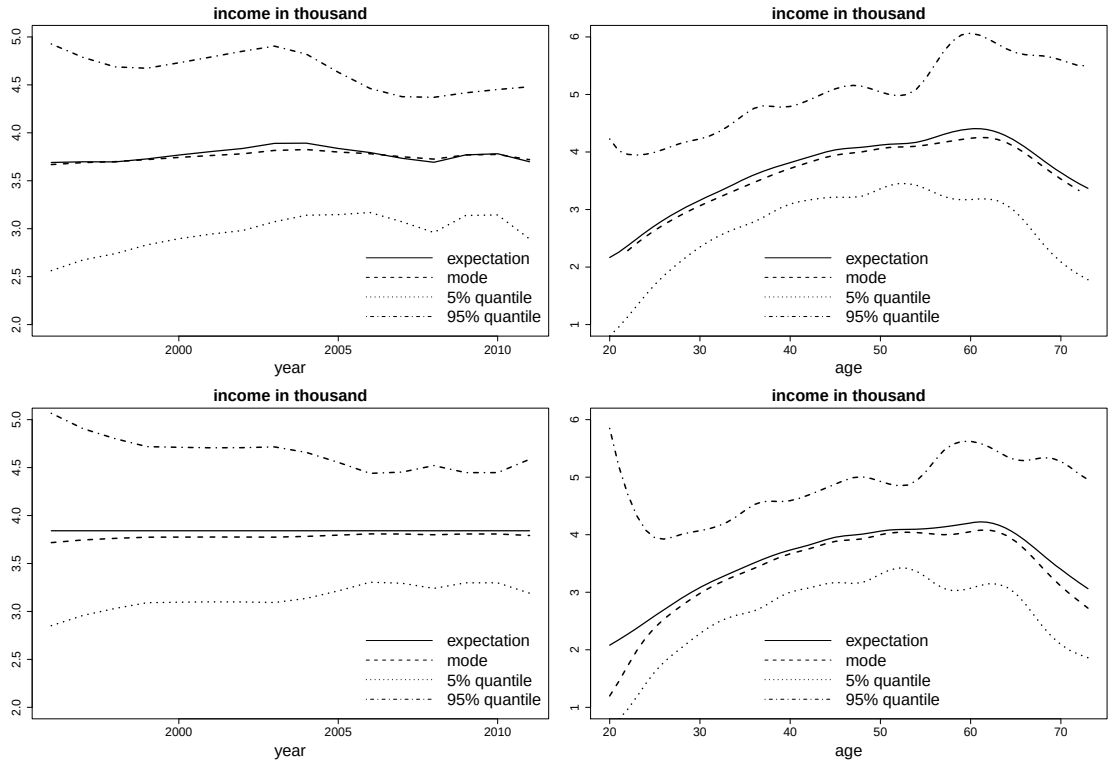

Figure 4: SOEP data. Estimated posterior mode, expectation, 5% and 95% quantile of the income distributions over year and age in the Dagum modell (first row) and inverse Gaussian model (second row), the remaining effects are kept constant.

In a nutshell, it can be said that a gradient of smaller incomes exists between several regions of the eastern part of Germany (especially in Thuringia and Saxony), compared to the rest of Germany, which is notable not only in the expectation but also

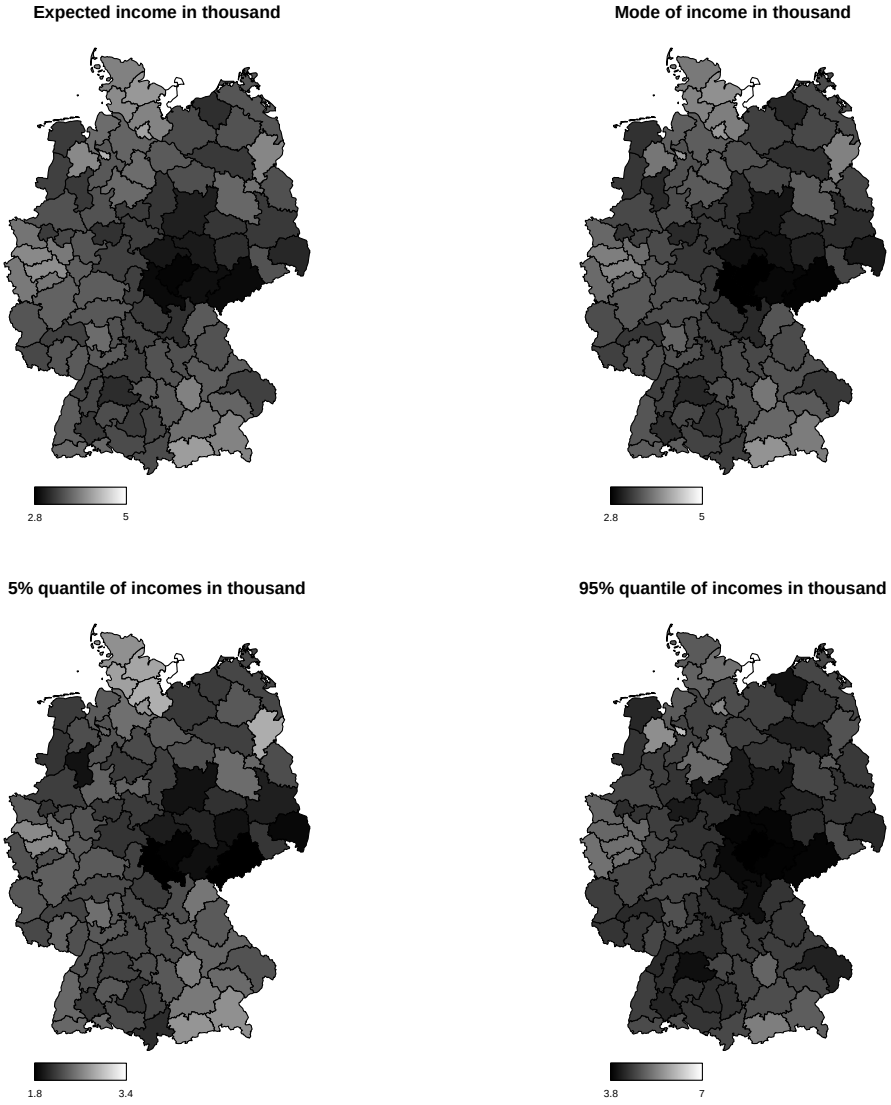

Figure 5: SOEP data. Estimated posterior mode, expectation, 5% and 95% quantile of the income distributiosn over region in the Dagum model, the remaining effects are kept constant. Note that axes of sub figures have different ranges to enhance visibility of estimated effects.

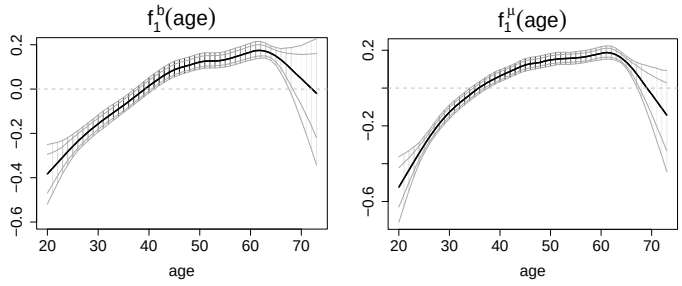

Figure 6: SOEP data. Posterior mean estimates of effect *age* (centred around zero) on b (Dagum model, left) and μ (inverse Gaussian model, right) together with pointwise 80% and 95% credible intervals.

in the mode and 10%, 90% quantiles of the estimated distribution. We furthermore estimate a steady increase of the incomes of full-time working males of ages smaller than 60 years whereas the time trend of inflation-adjusted incomes is only positive for low-income earners and even negative for top earners. The decline of incomes for males older than 60 can be explained by lower retirement pensions. Trends of the inverse Gaussian distribution behave similar compared to Dagum but with smaller estimated lower quantiles of time and spatial trend and with higher estimated upper quantiles of the income distribution over age. For the person-specific random effects (shown in Figures A3 and A7 of the supplement), the kernel density estimate of the effect on the parameter b (Dagum model) and μ (inverse Gaussian model) is more shallow and with lighter tails than the kernel density estimate of the random effect on a respectively σ^2. Since b is proportional to the expectation, the effects of *age* in Figure 6 resemble each other.

6 Production of Cereals in Farms of England and Wales

As a second illustration of Bayesian distributional regression, we consider the output proportions of 1232 different farms produced by the cultivation of cereal (including e.g. wheat, rice, maize) in England and Wales in the year 2007. The data have been collected by the Department of Environment, Food and Rural Affairs and National Assembly for Wales, Farm Business Survey, 2006-2007, and are provided by the UK Data Service (Colchester, Essex: UK Data Archive, 2008, http://dx.doi.org/10.5255/UKDA-SN-5838-1). The total output of the farms is subdivided in cereals and several animal products and if a farm reports positive output of cereals this can either be used for internal purposes (e.g. feeding of animals) or for selling.

Around 55% of the farms do not produce any cereal, 5% are completely specialised on the cultivation of cereal (100% of output are cereal) and the remaining farms have an output that is based on cereal and animal production. Therefore, the zero-one inflated beta distribution with parameters μ, σ^2, ν and τ as described in Section 2.1 is an appropriate candidate for analysing output shares on the cultivation of cereal.

From the list of potential explanatory variables related to farm business, we focus on the capital of the farms measured in pounds by their machinery, buildings and land maintenance, as well as running costs and the sum of the annual family and the annual hired hours of labour. Furthermore, we consider the utilised agricultural area in hectare and to capture spatial variations, we also take into account the geographic information on the location of the farms in one of the counties in England and Wales. Therefore, a generic predictor for farm i can be written as

$$\eta_i = \beta_0 + f_1(lland_i) + f_2(llabour_i) + f_3(lcapital_i) + f_{spat}(county_i)$$

where β_0 represents the overall level of the predictor, f_1 to f_3 are nonlinear functions of logarithmic land size *lland*, the logarithm of capital *lcapital*, as well as the logarithmic labour *llabour* modelled by B-splines, and the spatial effect f_{spat} is based on a Markov random field prior.

After comparing different models in terms of DIC, we found that several effects have linear influences on σ^2, ν and τ and ended up with the following predictor structures for the four distribution parameters:

$$\begin{aligned}
\eta_i^{\mu} &= \beta_0^{\mu} + f_1^{\mu}(lland_i) + f_2^{\mu}(llabour_i) + f_3^{\mu}(lcapital_i) + f_{spat}^{\mu}(county_i) \\
\eta_i^{\sigma^2} &= \beta_0^{\sigma^2} + lland_i\beta_1^{\sigma^2} + f_2^{\sigma^2}(llabour_i) + lcapital_i\beta_3^{\sigma^2} + f_{spat}^{\sigma^2}(county_i) \\
\eta_i^{\nu} &= \beta_0^{\nu} + f_1^{\nu}(lland_i) + f_2^{\nu}(llabour_i) + f_3^{\nu}(lcapital_i) + f_{spat}^{\nu}(county_i) \\
\eta_i^{\tau} &= \beta_0^{\tau} + lland_i\beta_1^{\tau} + llabour_i\beta_2^{\tau} + f_{spat}^{\tau}(county_i).
\end{aligned}$$

Figure 7 shows posterior mean estimates of the probabilities $\frac{\nu}{1+\nu+\tau}$ and $\frac{\tau}{1+\nu+\tau}$ as well as posterior mean expectation $\frac{\mu+\tau}{1+\nu+\tau}$ of the response which correspond to the probability of exclusively and no cereal output respectively to the expected proportion of the total output produced by cereal, varying over the counties. All other covariates are kept constant at the estimates obtained for average covariate values. The influence of the nonlinear effects *lland*, *llabour* and *lcapital* on the expectation of the response is given in Figure 8, where posterior mean estimates of $\frac{\mu+\tau}{1+\nu+\tau}$ are plotted with pointwise confidence intervals, bars are indicating the distribution of the observed covariate values, and again all other effects are kept constant. The nonlinear effects (centred around zero) on μ, ν and σ^2 are depicted in Figure B8 of the supplement together with pointwise credible intervals and bars again indicating the distribution of the

observed covariate values. The raw centred posterior mean spatial effects on the four distribution parameters can be found in Figure B9 of the supplement and estimates of linear effects on σ^2 and τ are given in Table 3.

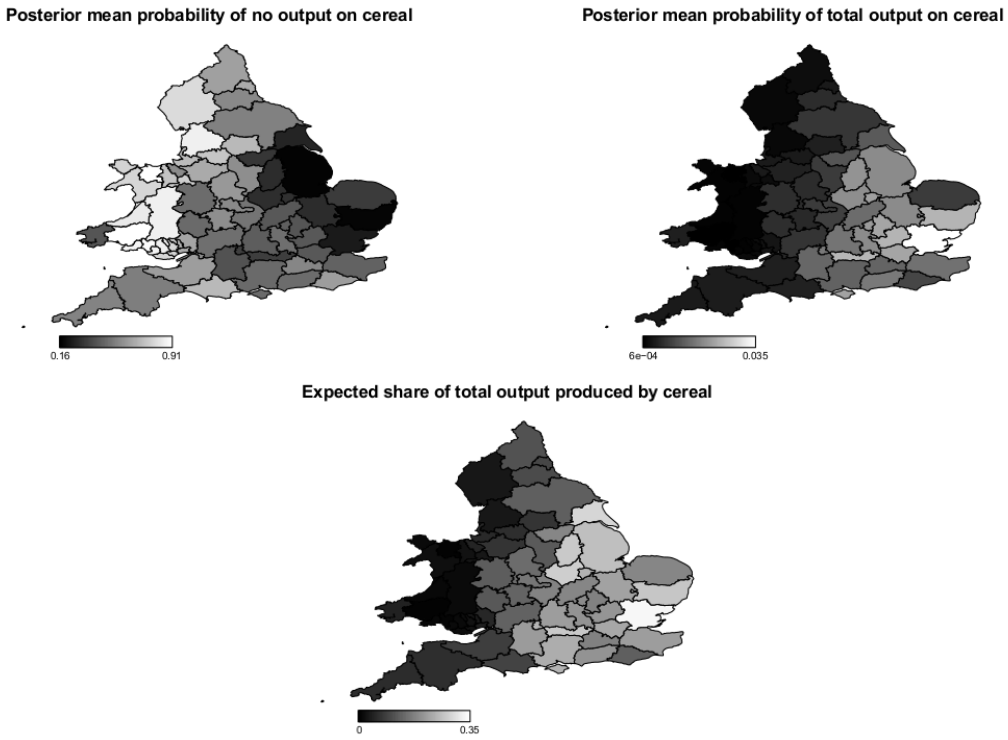

Figure 7: Cereal data. Estimated posterior mean probabilities $\frac{\nu}{1+\nu+\tau}$ (topleft), $\frac{\tau}{1+\nu+\tau}$ (topright) and posterior mean expected proportion $\frac{\mu+\tau}{1+\nu+\tau}$ of the total output on cereal (bottom) in the zero-one inflated beta model. The other effects are kept constant. Note that axes of sub figures have different ranges to enhance visibility of estimated effects.

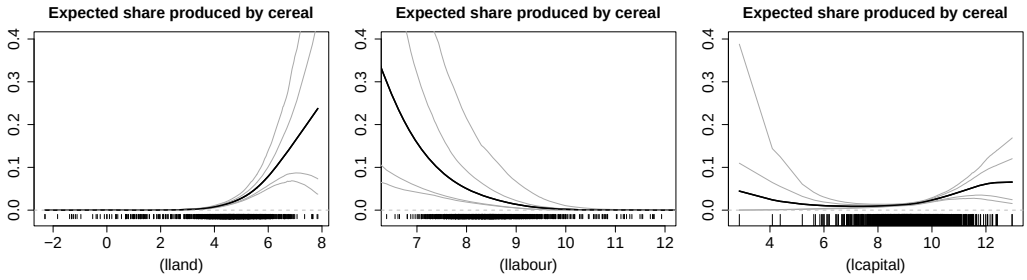

Figure 8: Cereal data. Posterior mean expected proportions $\frac{\mu+\tau}{1+\nu+\tau}$ on the production of cereal varying over *lland* (left), *llabour* (middle) and *lcapital* (right) together with pointwise 80% and 95% credible intervals in the zero-one inflated beta model. The other effects are kept constant.

Findings on selected effects can be summarised as follows: Figure 7 indicates a gradient between the East of England and Wales or the West of England with higher

Parameter	mean	2.5% quantile	median	97.5% quantile
β_0^{μ} (intercept)	-1.76	-1.98	-1.75	-1.54
$\beta_0^{\sigma^2}$ (intercept)	-5.52	-7.36	-5.3	-3.61
$\beta_1^{\sigma^2}$ ($lland$)	0.41	0.14	0.41	0.68
$\beta_3^{\sigma^2}$ ($lcaptial$)	0.19	-0.06	0.19	0.42
β_0^{ν} (intercept)	0.75	0.478	0.75	1.05
β_0^{τ} (intercept)	5.26	2.30	5.27	8.43
β_1^{τ} ($lland$)	1.02	0.58	1.02	1.45
β_2^{τ} ($llabour$)	-1.53	-2.01	-1.53	-1.08

Table 3: Cereal data. Summary of posterior distribution of linear effects in the zero-one inflated beta model.

production of cereal in the former one. This can be explained by the stony and hilly landscape of Wales and the coast in the East of England which is suited for the cultivation of cereal (compare also Figure B9). Figure 7 furthermore reveals that the probability of no cereal output is in general higher than the probability of obtaining a farm the produces only cereal. The total agricultural area of a farm is estimated to be a crucial factor for or against a large production of cereals, compare Figure B8 and β_1^{τ} in Table 3. For farms that do not have extreme capital compared to the remaining farms, the effect of capital is more or less insignificant. One possible explanation is the general problem of measuring capital in one quantity. Looking at the effects of labour it is important to be aware of the fact that e.g. a decreasing effect of f_2^{μ} in Figure B8 is not directly linked to the expected proportion of cereal products since human labour is also needed for the remaining output. However, the effect can be an indicator for the rising use of machineries and the aim to reduce costs in order to increase the efficiency of the farm. As in Figure B8, one percent increase of the nonlinear effects in Figure 8 have to be interpreted as one percent point increase of corresponding predictors respectively the expected proportion of cereal output.

7 Conclusion

Distributional regression and the closely related class of generalised additive models for location, scale and shape provide a flexible, comprehensive toolbox for solving

complex regression problems with potentially complex, non-standard response types. They are therefore extremely useful to overcome the limitations of common mean regression models and to enable a proper, realistic assessment of regression relationships. In this paper, we provided a Bayesian approach to distributional regression and described solutions for the most important applied problems including the selection of a suitable predictor specification and the most appropriate response distribution. Based on efficient MCMC simulation techniques, we developed a generic framework for inference in Bayesian structured additive distributional regression relying on distribution specific iteratively weighted least squares proposals as a core feature of the algorithms.

Despite the practical solutions outlined in this paper, model choice and variable selection remain relatively tedious and more automatic procedures would be highly desirable. Suitable approaches may be in the spirit of Belitz and Lang [2008] in a frequentist setting or based on spike and slab priors for Bayesian inference as developed in [Scheipl et al., 2012] for mean regression.

It will also be of interest to extend the distributional regression approach to the multivariate setting. For example, in case of multivariate Gaussian responses, covariate effects on the correlation parameter may be very interesting in specific applications. Similarly, multivariate extensions of beta regression lead to Dirichlet distributed responses representing multiple percentages that sum up to one.

References

C. Belitz and S. Lang. Simultaneous selection of variables and smoothing parameters in structured additive regression models. Computational Statistics and Data Analysis, 53:61–81, 2008.

C. Belitz, A. Brezger, T. Kneib, S. Lang, and N. Umlauf. Bayesx, 2012. – Software for Bayesian inference in structured additive regression models. Version 2.1. Available from http://www.bayesx.org.

A. Brezger and S. Lang. Generalized structured additive regression based on Bayesian P-splines. Computational Statistics & Data Analysis, 50:967–991, 2006.

A. Brezger and W. Steiner. Monotonic regression based on Bayesian P–splines. Journal of Business & Economic Statistics, 26(1):90–104, 2008.

D. Chotikapanich. Modeling Income Distributions and Lorenz Curves. Springer, London, 2008. Series: Economic Studies in Inequality, Social Exclusion and Well-Being, Vol. 5.

M. Denuit and S. Lang. Non-life rate-making with Bayesian gams. Insurance: Mathematics and Economics, 35:627–647, 2004.

P. K. Dunn and G. K. Smyth. Randomized quantile residuals. Computational and Graphical Statistics, 5:236–245, 1996.

P. H. Eilers and B. D. Marx. Flexible smoothing using B-splines and penalized likelihood. Statistical Science, 11:89–121, 1996.

L. Fahrmeir and T. Kneib. Propriety of posteriors in structured additive regression models: Theory and empirical evidence. Journal of Statistical Planning and Inference, 39:843–859, 2009.

L. Fahrmeir, T. Kneib, and S. Lang. Penalized structured additive regression for space-time data: a Bayesian perspective. Statistica Sinica, 14:731–761, 2004.

L. Fahrmeir, T. Kneib, S. Lang, and B. Marx. Regression - Models, Methods and Applications. Springer, 2013.

S. L. P. Ferrari and F. Cribari-Neto. Beta regression for modelling rates and proportions. Journal of the Royal Statistical Society, Series C, 31:799–815, 2004.

J. Frick, S. P. Jenkins, D. R. Lillard, O. Lipps, and M. Wooden. The cross-national equivalent file (cnef) and its member country household panel studies. Schmollers Jahrbuch, 4:626–654, 2007.

D. Gamerman. Sampling from the posterior distribution in generalized linear mixed models. Statistics and Computing, 7:57–68, 1997.

T. Gneiting. Quantiles as optimal point forecasts. International Journal of Forecasting, 27:197–207, 2011.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 102(477):359–378, 2007.

T. Gneiting and R. Ranjan. Comparing density forecasts using threshold and quantile-weighted scoring rules. Journal of Business & Economic Statistics, 29(3):411–421, 2011.

T. J. Hastie and R. J. Tibshirani. Generalized Additive Models. Chapman & Hall, 1990.

G. Heller, Stasinopoulos D. M., and Rigby R. A. The zero-adjusted inverse gaussian distribution as a model for insurance data. In J.Newell J. Hinde, J.Einbeck, editor, Proceedings of the 21th International Workshop on Statistical Modelling, 2006.

N. Klein, M. Denuit, T. Kneib, and S. Lang. Nonlife ratemaking and risk management with Bayesian additive model for location scale and shape. Technical report, 2013a. URL `http://eeecon.uibk.ac.at/wopec2/repec/inn/wpaper/2013-24.pdf`.

N. Klein, T. Kneib, and S. Lang. Bayesian generalized additive models for location, scale and shape for zero-inflated and overdispersed count data. Technical report, 2013b. URL http://eeecon.uibk.ac.at/wopec2/repec/inn/wpaper/2013-12.pdf.

R. Koenker. Quantile Regression. Cambrigde University Press, New York, 2005. Economic Society Monographs.

R. Koenker and G. Bassett. Regression quantiles. Econometrica, 46:33–50, 1978.

F. Laio and S. Tamea. Verification tools for probabilistic forecasts of continuous hydrological variables. Hydrology and Earth System Sciences, 11:1267–1277, 2007.

S. Lang and A. Brezger. Bayesian P-splines. Journal of Computational and Graphical Statistics, 13:183–212, 2004.

S. Lang, W. Steiner, A. Weber, and P. Wechselberger. Accommodating heterogeneity and functional flexibility in store sales models: A Bayesian semiparametric approach. Technical report, 2013a.

S. Lang, N. Umlauf, P. Wechselberger, K. Harttgen, and T. Kneib. Multilevel structured additive regression. Statistics and Computing, 23, 2013b.

P McCullagh and J. A. Nelder. Generalized Linear Models. Chapman & Hall, 1989.

W. K. Newey and J. L. Powell. Asymmetric least squares estimation. Econometrica, 55:819–847, 1987.

M. E. J. Newman. Power laws, pareto distributions and zipf's law. Contemporary Physics, 46(5), 2005.

R. A. Rigby and D. M. Stasinopoulos. Generalized additive models for location, scale and shape (with discussion). Applied Statistics, 54:507–554, 2005.

H. Rue and L. Held. Gaussian Markov Random Fields. Chapman & Hall / CRC, 2005.

D. Ruppert, M. P. Wand, and R. J. Carroll. Semiparametric Regression. Cambridge University Press, 2003.

F. Scheipl, L. Fahrmeir, and T. Kneib. Spike-and-slab priors for function selection in structured additive regression models. Journal of the American Statistical Association, 107:1518–1532, 2012.

S. K. Schnabel and P. Eilers. Optimal expectile smoothing. Computational Statistics & Data Analysis, 53:4168–4177, 2009.

F. Sobotka and T. Kneib. Geoadditive expectile regression. Computational Statistics & Data Analysis, 56:755–767, 2012.

D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society, 65(B):583–639, 2002.

D. Sun, R. K. Tsutakawa, and H. Zhuoqiong. Propriety of posteriors with improper priors in hierarchical linear mixed models. Statistica Sinica, 11:77–95, 2001.

S. N. Wood. Stable and efficient multiple smoothing parameter estimation for generalized additive models. Journal of the American Statistical Association, 99:673–686, 2004.

S. N. Wood. Generalized Additive Models : An Introduction with R. Chapmann & Hall, 2006.

S. N. Wood. Fast stable direct fitting and smoothness selection for generalized additive models. Journal of the Royal Statistical Society, Series B, 70:495–518, 2008.

K. Yu and R. A. Moyeed. Bayesian quantile regression. Statistics & Probability Letters, 54:437–447, 2001.

Bayesian Structured Additive Distributional Regression

Supplement

Nadja Klein, Thomas Kneib
Chair of Statistics
Georg-August-University Göttingen

Stefan Lang
Department of Statistics
University of Innsbruck

A Supplemenatry Material to Section 5

A.1 Predictor Specifications of Final Models

In the main paper we documented predictors for the Dagum and inverse Gaussian distribution. This section also provides the the final predictor specifications of the log-normal and gamma distribution which have been selected by the DIC. A description of the variables included is given in the main paper. For the LN distribution we used

$$\begin{aligned}\eta_{it}^{\mu} &= f_1^{\mu}(age_{it}) + f_2^{\mu}(t) + f_{spat}^{\mu}(region_{it}) + \beta_i^{\mu} \\ \eta_{it}^{\sigma^2} &= f_1^{\sigma^2}(age_{it}) + f_2^{\sigma^2}(t) + f_{spat}^{\sigma^2}(region_{it}) + \beta_i^{\sigma^2}.\end{aligned}$$

and for the gamma distribution

$$\begin{aligned}\eta_{it}^{\mu} &= f_1^{\mu}(age_{it}) + f_2^{\mu}(t) + f_{spat}^{\mu}(region_{it}) + \beta_i^{\mu} \\ \eta_{it}^{\sigma} &= f_1^{\sigma}(age_{it}) + f_2^{\sigma}(t) + f_{spat}^{\sigma}(region_{it}) + \beta_i^{\sigma}.\end{aligned}$$

A.2 Plots of Effects of the Final Dagum Model

In Figures A1 and A2 we depict the centred nonlinear effects of parameters a, b and p as well as the estimated spatial effects on a, b in the final Dagum model described in the main paper.

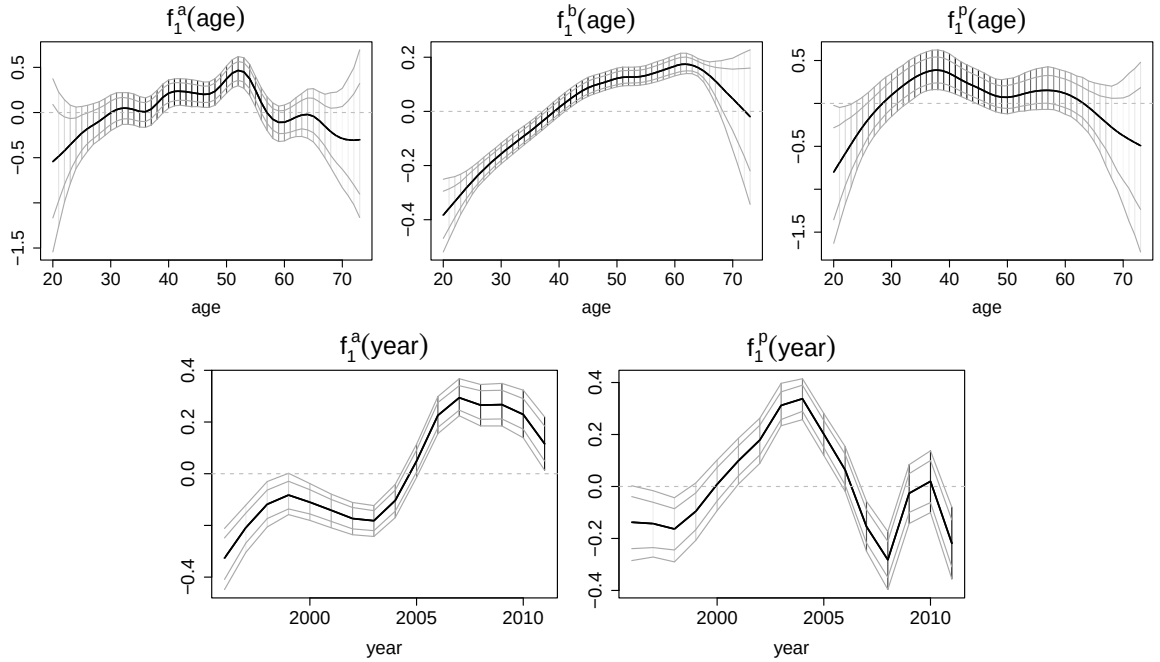

Figure A1: SOEP data. Posterior mean estimates of centred nonlinear effects together with pointwise 80% and 95% credible intervals in the Dagum model. Note that axes of sub figures have different ranges to enhance visibility of estimated effects.

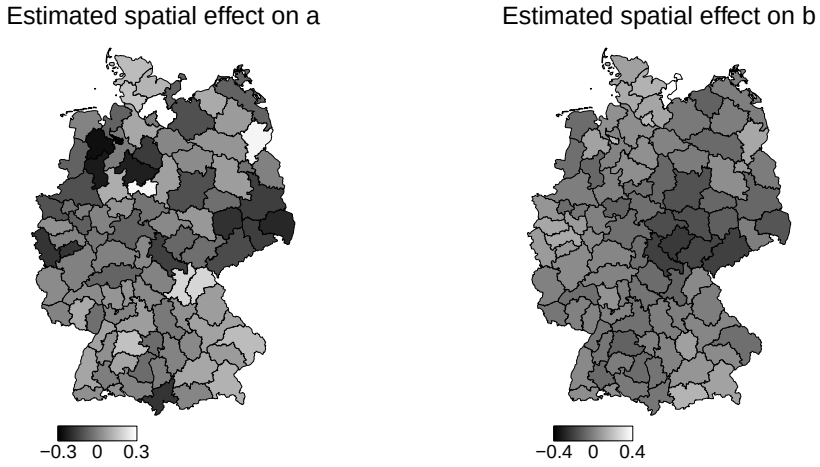

Figure A2: SOEP data. Posterior mean estimates of complete spatial effects in the Dagum model, centred around zero. Note that axes of sub figures have different ranges to enhance visibility of estimated effects.

Figure A3 additionally depicts the kernel density estimates of the individual specific random effects ob a and b of 526 persons resulting from an Epanechnikov kernel with bandwidth of around 0.1.

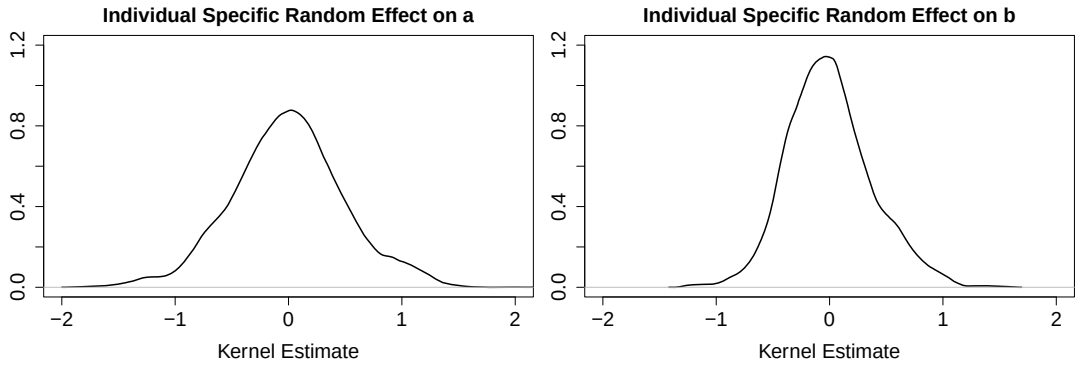

Figure A3: SOEP data. Kernel density estimates of $n = 526$ persons for parameters a and b in the Dagum model.

A.3 Plots of the Final Inverse Gaussian Model

In this section, we summarize the analysis of spatio-temporal dynamics of labour income in Germany (compare Section 5 of the main paper) for the inverse Gaussian distribution. Figure A4 shows spatial variations while the other effects are kept constant. Time and age trends can be found in Figure 4 of the main paper. In Figures A5 and A6 the raw centred posterior mean estimates of age, year and region on μ and σ^2 are depicted. Additionally, Figure A7 depicts the kernel density estimates of the individual specific random effects on μ and σ^2 of 526 persons resulting from an Epanechnikov kernel with bandwidth of around 0.1. For further descriptions and explanations of the contents of the plots, compare Section 5 of the main paper.

B Supplemtary Plots to Section 6

In Figure B9 the posterior mean spatial effects (centred around zero) of the analysis on output shares on cereal in England and Wales (compare Section 6 of the main paper) are shown for each of the four distribution parameters μ, σ^2, ν and τ in the zero-one inflated beta model. The centred nonlinear effects on μ, σ^2 and ν together with pointwise confidence intervals are given in Figure B8.

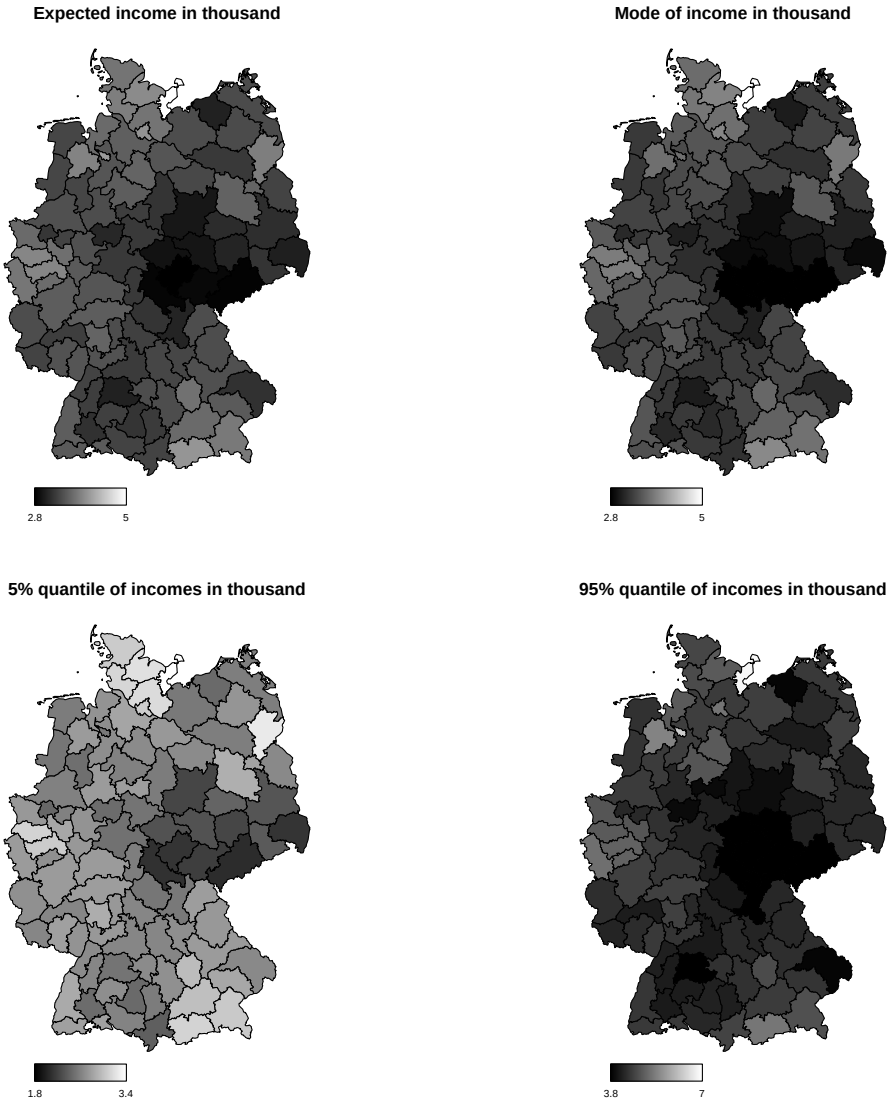

Figure A4: SOEP data. Estimated mode, expectation, 5% and 95% quantile of the income distributions over region in the inverse Gaussian model, the remaining effects are kept constant. Note that axes of sub figures have different ranges to enhance visibility of estimated effects.

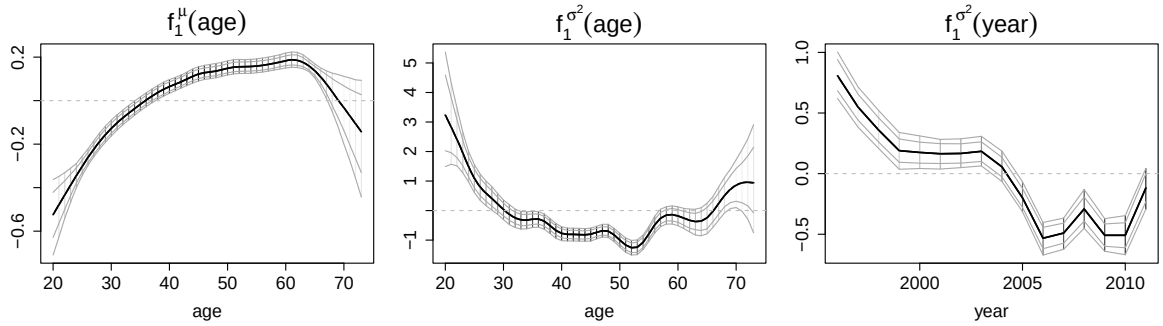

Figure A5: SOEP data. Posterior mean estimates of centred nonlinear effects together with pointwise 80% and 95% credible intervals in the inverse Gaussian model. Note that axes of sub figures have different ranges to enhance visibility of estimated effects.

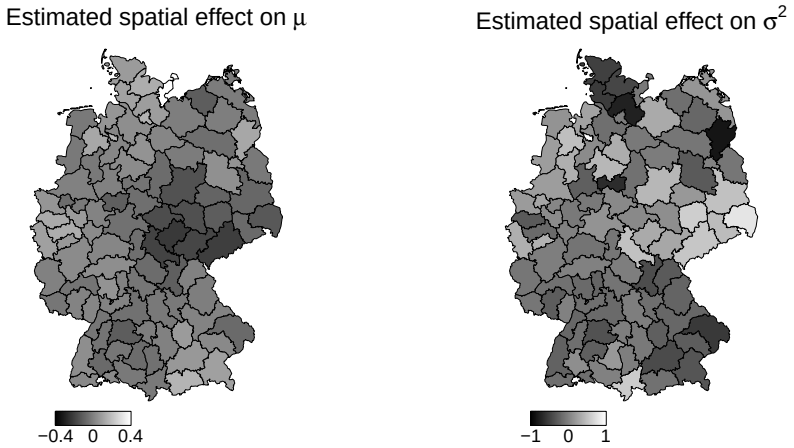

Figure A6: SOEP data. Posterior mean estimates of complete spatial effects in the inverse Gaussian model, centred around zero. Note that axes of sub figures have different ranges to enhance visibility of estimated effects.

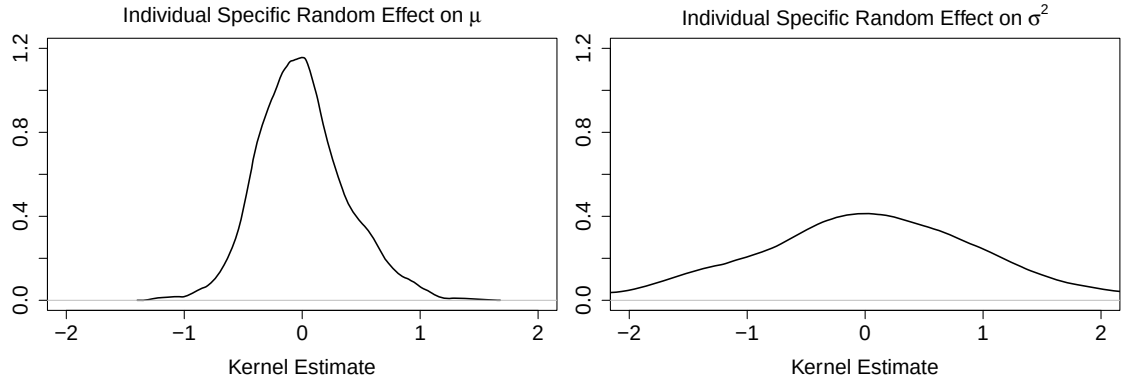

Figure A7: SOEP data. Kernel density estimates of $n = 526$ persons for parameters μ and σ^2 in the inverse Gaussian model.

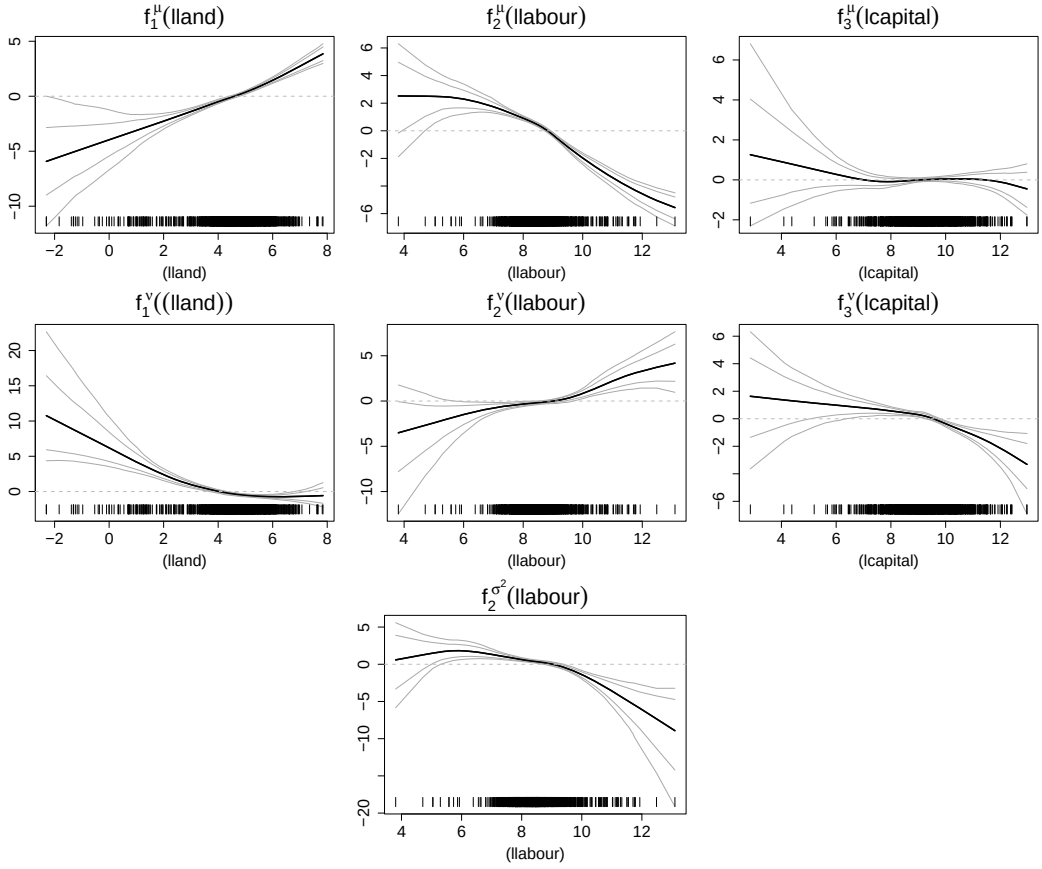

Figure B8: Cereal data. Posterior mean estimates of nonlinear effects on μ (first row), on ν (second row) and on σ^2 (third row) together with pointwise 80% and 95% credible intervals in the zero-one inflated beta model, centred around zero. Note that axes of sub figures have different ranges to enhance visibility of estimated effects.

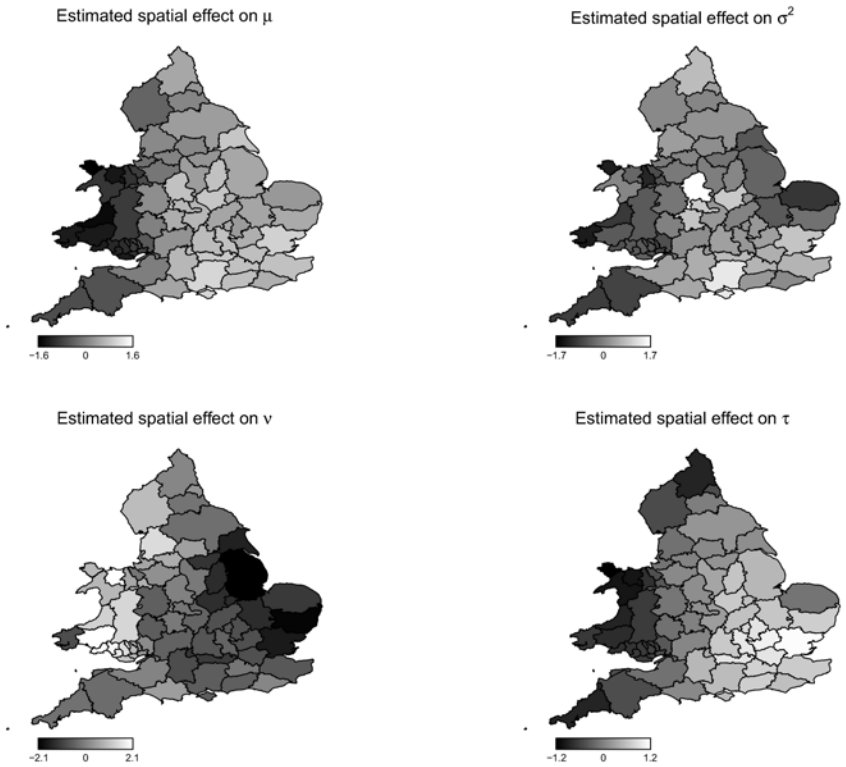

Figure B9: Cereal data. Posterior mean estimates of spatial effects in the zero-one inflated beta model, centred around zero. Note that axes of sub figures have different ranges to enhance visibility of estimated effects.

C Score Vectors and Working Weights

In this section, we provide for each distribution of Table 1 of the main paper the score vectors v_i and the working weights w_i which are needed for the MCMC algorithm of Section 3. We therefore usually compute first and second derivatives of all logarithmic densities $f_i(y_i|\vartheta_{i1},\ldots,\vartheta_{iK})$ with respect to the predictors $\eta_i^{\vartheta_1},\ldots,\eta_i^{\vartheta_K}$, show how to get the expectations of the negative second derivatives, and that the resulting working weights are positive. Further details about positiveness of weights and expectations will be given in the sections of specific distributions. In the following we simply write l for the log-likelihood of y_i as a function of the predictors $\eta_i^{\vartheta_1},\ldots,\eta_i^{\vartheta_K}$.

C.1 Real-Valued Responses

C.1.1 Normal Distribution

Parameters are $\mu_i = \eta_i^{\mu} \in \mathbb{R}$ and $\sigma_i^2 = \exp\left(\eta_i^{\sigma^2}\right) > 0$.

$$l = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\left(\sigma_i^2\right) - \frac{(y_i - \mu_i)^2}{2\sigma_i^2}$$

$$\begin{aligned} E(y_i) &= \mu_i \\ E\left((y_i - \mu_i)^2\right) &= \sigma_i^2 \end{aligned}$$

Score Vectors

$$\begin{aligned} \frac{\partial l}{\partial \eta_i^{\mu}} &= \frac{y_i - \mu_i}{\sigma_i^2} \\ \frac{\partial l}{\partial \eta_i^{\sigma^2}} &= -\frac{1}{2} + \frac{(y_i - \mu_i)^2}{2\sigma_i^2} \end{aligned}$$

Second Derivatives

$$\begin{aligned} \frac{\partial^2 l}{\left(\partial \eta_i^{\mu}\right)^2} &= -\frac{1}{\sigma_i^2} \\ \frac{\partial^2 l}{\left(\partial \eta_i^{\sigma^2}\right)^2} &= -\frac{(y_i - \mu_i)^2}{2\sigma_i^2} \end{aligned}$$

Working Weights

$$\begin{aligned} E\left(-\frac{\partial^2 l}{\left(\partial \eta_i^{\mu}\right)^2}\right) &= \frac{1}{\sigma_i^2} \\ E\left(-\frac{\partial^2 l}{\left(\partial \eta_i^{\sigma^2}\right)^2}\right) &= \frac{1}{2} \end{aligned}$$

C.1.2 t-Distribution

Parameters are $\mu_i = \eta_i^{\mu} \in \mathbb{R}$, $\sigma_i^2 = \exp\left(\eta_i^{\sigma^2}\right) > 0$ and $n_{d,i} = \exp\left(\eta_i^{n_d}\right) > 0$.

$$\begin{aligned} l = {} & \log\left(\Gamma\left(\frac{n_{d,i}+1}{2}\right)\right) - \log\left(\Gamma\left(\frac{n_{d,i}}{2}\right)\right) - \log\left(\Gamma\left(\frac{1}{2}\right)\right) - \frac{1}{2}\log(n_{d,i}) \\ & -\frac{1}{2}\log\left(\sigma_i^2\right) - \frac{n_{d,i}+1}{2}\log\left(1 + \frac{(y_i - \mu_i)^2}{\sigma_i^2 n_{d,i}}\right) \end{aligned}$$

$$E(y_i) = \begin{cases} \mu_i & n_{d,i} > 1 \\ \text{undefined} & n_{d,i} \leq 1. \end{cases}$$

Score Vectors

$$\begin{aligned}
\frac{\partial l}{\partial \eta_i^{\mu}} &= \frac{(n_{d,i}+1)(y_i-\mu_i)}{\sigma_i^2 n_{d,i}+(y_i-\mu_i)^2} \\
\frac{\partial l}{\partial \eta_i^{\sigma^2}} &= -\frac{1}{2}+\frac{(n_{d,i}+1)(y_i-\mu_i)^2}{2(\sigma_i^2 n_{d,i}+(y_i-\mu_i)^2)} \\
\frac{\partial l}{\partial \eta_i^{n_d}} &= \frac{n_{d,i}}{2}\left(\psi\left(\frac{n_{d,i}+1}{2}\right)-\psi\left(\frac{n_{d,i}}{2}\right)-\log\left(1+\frac{(y_i-\mu_i)^2}{\sigma_i^2 n_{d,i}}\right)\right) \\
&\quad -0.5+\frac{(n_{d,i}+1)(y_i-\mu_i)^2}{2(\sigma_i^2 n_{d,i}+(y_i-\mu_i)^2)}
\end{aligned}$$

where $\psi(x)$ is the digamma function of $x>0$.

Second Derivatives

$$\begin{aligned}
\frac{\partial^2 l}{\left(\partial \eta_i^{\mu}\right)^2} &= -\frac{(n_{d,i}+1)\left(\sigma_i^2 n_{d,i}-(y_i-\mu_i)^2\right)}{\left(\sigma_i^2 n_{d,i}+(y_i-\mu_i)^2\right)^2} \\
&= \frac{n_{d,i}+1}{\sigma_i^2 n_{d,i}\left(1+\frac{(y_i-\mu_i)^2}{\sigma_i^2 n_{d,i}}\right)^2}-\frac{2(n_{d,i}+1)}{\sigma_i^2 n_{d,i}\left(1+\frac{(y_i-\mu_i)^2}{\sigma_i^2 n_{d,i}}\right)^2} \\
\frac{\partial^2 l}{\left(\partial \eta_i^{\sigma^2}\right)^2} &= -\frac{\sigma_i^2 n_{d,i}(n_{d,i}+1)(y_i-\mu_i)^2}{2(\sigma_i^2 n_{d,i}+(y_i-\mu_i)^2)^2} \\
&= -\frac{n_{d,i}+1}{2\left(1+\frac{(y_i-\mu_i)^2}{\sigma_i^2 n_{d,i}}\right)}+\frac{n_{d,i}+1}{2\left(1+\frac{(y_i-\mu_i)^2}{\sigma_i^2 n_{d,i}}\right)^2}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 l}{\left(\partial \eta_i^{n_{d,i}}\right)^2} &= \frac{n_{d,i}}{2}\left(\psi\left(\frac{n_{d,i}+1}{2}\right)-\psi\left(\frac{n_{d,i}}{2}\right)-\log\left(1+\frac{(y_i-\mu_i)^2}{\sigma_i^2 n_{d,i}}\right)\right) \\
&\quad +\frac{n_{d,i}(y_i-\mu_i)^2}{\sigma_i^2 n_{d,i}+(y_i-\mu_i)^2}-\frac{\sigma_i^2 n_{d,i}(n_{d,i}+1)(y_i-\mu_i)^2}{2(\sigma_i^2 n_{d,i}+(y_i-\mu_i)^2)^2} \\
&\quad +\frac{n_{d,i}}{2}\left(\frac{n_{d,i}}{2}\psi_1\left(\frac{n_{d,i}+1}{2}\right)-\frac{n_{d,i}}{2}\psi_1\left(\frac{n_{d,i}}{2}\right)\right) \\
&= \frac{n_{d,i}}{2}\left(\psi\left(\frac{n_{d,i}+1}{2}\right)-\psi\left(\frac{n_{d,i}}{2}\right)-\log\left(1+\frac{(y_i-\mu_i)^2}{\sigma_i^2 n_{d,i}}\right)\right) \\
&\quad +\frac{n_{d,i}^2}{4}\left(\psi_1\left(\frac{n_{d,i}+1}{2}\right)-\psi_1\left(\frac{n_{d,i}}{2}\right)\right)+\frac{n_{d,i}(y_i-\mu_i)^2}{(\sigma_i^2 n_{d,i}+(y_i-\mu_i)^2)}+\frac{\partial^2 l}{\left(\partial \eta_i^{\sigma^2}\right)^2}.
\end{aligned}$$

Working Weights In the following we assume $n_{d,i} > 1$.

$$\begin{aligned}
E\left(\frac{1}{1+\frac{(y_i-\mu_i)^2}{\sigma_i^2 n_{d,i}}}\right) &= \frac{\Gamma\left(\frac{n_{d,i}+1}{2}\right)}{\Gamma\left(\frac{n_{d,i}}{2}\right)\Gamma\left(\frac{1}{2}\right)\sqrt{\sigma_i^2 n_{d,i}}}\int_0^\infty \left(1+\frac{(y_i-\mu_i)^2}{\sigma_i^2 n_{d,i}}\right)^{-\frac{(n_{d,i}+2)+1}{2}} \\
&= \frac{\Gamma\left(\frac{n_{d,i}+1}{2}\right)\Gamma\left(\frac{n_{d,i}+2}{2}\right)\Gamma\left(\frac{n_{d,i}+3}{2}\right)}{\Gamma\left(\frac{n_{d,i}}{2}\right)\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n_{d,i}+2}{2}\right)\Gamma\left(\frac{n_{d,i}+3}{2}\right)\sqrt{\frac{n_{d,i}+2}{n_{d,i}}\sigma_i^2 n_{d,i}}} \\
&\quad \int_0^\infty \left(1+\frac{(y_i-\mu_i)^2}{\frac{\sigma_i^2 n_{d,i}}{n_{d,i}+2}(n_{d,i}+2)}\right)^{-\frac{(n_{d,i}+2)+1}{2}} \\
&= \frac{\Gamma\left(\frac{n_{d,i}+1}{2}\right)\Gamma\left(\frac{n_{d,i}+2}{2}\right)\Gamma\left(\frac{1}{2}\right)}{\Gamma\left(\frac{n_{d,i}}{2}\right)\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n_{d,i}+3}{2}\right)} \\
&= \frac{n_{d,i}}{n_{d,i}+1}. \qquad \text{(C.1)}
\end{aligned}$$

In a similar way we obtain

$$E\left(\left(1+\frac{(y_i-\mu_i)^2}{\sigma_i^2 n_{d,i}}\right)^{-2}\right) = \frac{n_{d,i}(n_{d,i}+2)}{(n_{d,i}+1)(n_{d,i}+3)} \qquad \text{(C.2)}$$

and hence

$$E\left(-\frac{\partial^2 l}{(\partial\eta_i^\mu)^2}\right) \underset{\text{(C.2)}}{\overset{\text{(C.1)}}{=}} -\frac{1}{\sigma_i^2}+\frac{2(n_{d,i}+2)}{\sigma_i^2(n_{d,i}+3)}$$

which is obviously greater than zero.

$$\begin{aligned}
E\left(-\frac{\partial^2 l}{\left(\partial\eta_i^{\sigma^2}\right)^2}\right) &\underset{\text{(C.2)}}{\overset{\text{(C.1)}}{=}} \frac{n_{d,i}}{2}-\frac{n_{d,i}(n_{d,i}+2)}{2(n_{d,i}+3)} \\
&= \frac{n_{d,i}}{2(n_{d,i}+3)} > 0. \qquad \text{(C.3)}
\end{aligned}$$

$$\begin{aligned}
E\left(-\frac{\partial^2 l}{\left(\partial\eta_i^{n_{d,i}}\right)^2}\right) \underset{\text{(C.2),(C.3)}}{\overset{\text{(C.1)}}{=}} & -\frac{n_{d,i}^2}{4}\left(\psi_1\left(\frac{n_{d,i}+1}{2}\right)-\psi_1\left(\frac{n_{d,i}}{2}\right)\right) \\
& +\frac{n_{d,i}}{2(n_{d,i}+3)}-\frac{n_{d,i}}{n_{d,i}+1}.
\end{aligned}$$

We note that for $0 < n_{d,i} \leq 4$

$$\begin{aligned}
-\frac{n_{d,i}^2}{4}\left(\psi_1\left(\frac{n_{d,i}+1}{2}\right)-\psi_1\left(\frac{n_{d,i}}{2}\right)\right) &> 0.6 \\
\frac{n_{d,i}}{2(n_{d,i}+3)}-\frac{n_{d,i}}{n_{d,i}+1} &> -0.52
\end{aligned}$$

such that $w_i^{n_d} > 0$ holds on the interval $(0, 4]$. For $n \geq 4$ it follows from Chen and Qi [2003, Lemma 1]

$$\begin{aligned}
-\frac{n_{d,i}^2}{4}\left(\psi_1\left(\frac{n_{d,i}+1}{2}\right)-\psi_1\left(\frac{n_{d,i}}{2}\right)\right) &> \frac{n_{d,i}^2}{2(n_{d,i}-2)}-\frac{n_{d,i}^2}{2(n_{d,i}-2)^2}+\frac{n_{d,i}^2}{3(n_{d,i}-2)^3} \\
&\quad -\frac{4n_{d,i}^2}{15(n_{d,i}-2)^5}-\frac{n_{d,i}^2}{2(n_{d,i}-1)}+\frac{n_{d,i}^2}{2(n_{d,i}-1)^2} \\
&\quad -\frac{n_{d,i}^2}{3(n_{d,i}-1)^3} \\
&= \frac{5n_{d,i}^4-50n_{d,i}^3+196n_{d,i}^2-360n_{d,i}+240}{15(n_{d,i}-2)^5} \\
&\quad +\frac{(n_{d,i}-3)n_{d,i}}{6(n_{d,i}-1)^3}.
\end{aligned}$$

The function $f(x) = \frac{5x^4-50x^3+196x^2-360x+240}{15(x-2)^5} + \frac{(x-3)x}{6(x-1)^3} + \frac{x}{2(x+3)} - \frac{x}{x+1}$ is monoton decreasing for $x > 4$ with $\lim_{x\to\infty} f(x) = 0$. Furthermore $f(x) > 0$ holds for $x \in [4, \infty)$ since $f(4) > 0$ and $x_0 < 4$ for all $x \in \mathbb{R}^+$ with $f(x_0) = 0$. Since f is a lower bound of $E\left(-\frac{\partial^2 l}{\left(\partial \eta_i^{n_{d,i}}\right)^2}\right)$ for $n_{d,i} > 4$ we obtain altogether $w_i^{n_d} > 0$ for $n_{d,i} > 0$ as desired. In conclusion all weights for the t-distribution are positive.

C.2 Non-Negative Responses

Let in this section be $y_i > 0$.

C.2.1 Log-Normal Distribution

Parameters are $\mu_i = \eta_i^{\mu} \in \mathbb{R}$ and $\sigma_i^2 = \exp\left(\eta_i^{\sigma^2}\right) > 0$.

$$l = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\left(\sigma_i^2\right) - \log(y_i) - \frac{(\log(y_i)-\mu_i)^2}{2\sigma_i^2}$$

$$E(y_i) = \exp\left(\mu_i + \frac{\sigma_i^2}{2}\right)$$

$$E\left((\log(y_i)-\mu_i)^2\right) = \sigma_i^2$$

Score Vectors

$$\begin{aligned}
\frac{\partial l}{\partial \eta_i^{\mu}} &= \frac{(\log(y_i)-\mu_i)}{\sigma_i^2} \\
\frac{\partial l}{\partial \eta_i^{\sigma^2}} &= -0.5 + \frac{(\log(y_i)-\mu_i)^2}{2\sigma_i^2}
\end{aligned}$$

Second Derivatives

$$\begin{aligned}\frac{\partial^2 l}{(\partial \eta_i^{\mu})^2} &= -\frac{1}{\sigma_i^2} \\ \frac{\partial^2 l}{\left(\partial \eta_i^{\sigma^2}\right)^2} &= -\frac{(\log(y_i) - \mu_i)^2}{2\sigma_i^2}\end{aligned}$$

Working Weights

$$\begin{aligned}E\left(-\frac{\partial^2 l}{(\partial \eta_i^{\mu})^2}\right) &= \frac{1}{\sigma_i^2} \\ E\left(-\frac{\partial^2 l}{\left(\partial \eta_i^{\sigma^2}\right)^2}\right) &= \frac{1}{2}\end{aligned}$$

Because of $\sigma_i^2 > 0$ we have that all weights of the log-normal distribution are greater than zero.

C.2.2 Inverse Gaussian Distribution

Parameters are $\mu_i = \exp\left(\eta_i^{\mu}\right) > 0$ and $\sigma_i^2 = \exp\left(\eta_i^{\sigma^2}\right) > 0$.

$$l = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\left(\sigma_i^2\right) - \frac{3}{2}\log(y_i) - \frac{(y_i - \mu_i)^2}{2y_i\mu_i^2\sigma_i^2}$$

$$\begin{aligned}E(y_i) &= \mu_i \\ E\left(y_i^2\right) &= \mu_i^2 + \mu_i^3\sigma_i^2 \\ E\left(y_i^{-1}\right) &= \frac{E(y_i^2)}{\mu_i^3} = \frac{1}{\mu_i} + \sigma_i^2\end{aligned}$$

Score Vectors

$$\begin{aligned}\frac{\partial l}{\partial \eta_i^{\mu}} &= \frac{y_i - \mu_i}{\mu_i^2\sigma_i^2} \\ \frac{\partial l}{\partial \eta_i^{\sigma^2}} &= -\frac{1}{2} + \frac{(y_i - \mu_i)^2}{2y_i\mu_i^2\sigma_i^2}\end{aligned}$$

Second Derivatives

$$\begin{aligned}\frac{\partial^2 l}{(\partial \eta_i^{\mu})^2} &= -\frac{2y_i}{\mu_i^2\sigma_i^2} + \frac{1}{\mu_i\sigma_i^2} \\ \frac{\partial^2 l}{\left(\partial \eta_i^{\sigma^2}\right)^2} &= -\frac{(y_i - \mu_i)^2}{2y_i\mu_i^2\sigma_i^2} \\ &= -\frac{y_i}{2\mu_i^2\sigma_i^2} + \frac{1}{\mu_i\sigma_i^2} - \frac{1}{2y_i\sigma_i^2}\end{aligned}$$

Working Weights

$$\begin{aligned} E\left(-\frac{\partial^2 l}{(\partial \eta_i^\mu)^2}\right) &= \frac{1}{\mu_i \sigma_i^2} \\ E\left(-\frac{\partial^2 l}{\left(\partial \eta_i^{\sigma^2}\right)^2}\right) &= \frac{1}{2\mu_i\sigma_i^2} - \frac{1}{\mu_i\sigma_i^2} + \frac{1}{2\mu_i\sigma_i^2} + \frac{1}{2} \\ &= \frac{1}{2} \end{aligned}$$

As for the log-normal distribution it directly follows that all weights for the inverse Gaussian distribution are greater than zero.

C.2.3 Gamma Distribution

Parameters are $\mu_i = \exp(\eta_i^\mu) > 0$ and $\sigma_i = \exp(\eta_i^\sigma) > 0$.

$$l = \sigma_i \log(\sigma_i) - \sigma_i \log(\mu_i) - \log(\Gamma(\sigma_i)) + (\sigma_i - 1)\log(y_i) - \frac{\sigma_i}{\mu_i} y_i$$

where $\Gamma(x) = \int_0^\infty u^{x-1} \exp(-u) \mathrm{d}u$ for $x > 0$ is the gamma function.

$$\begin{aligned} E(y_i) &= \mu_i \\ E(\log(y_i)) &= \psi(\sigma_i) + \log(\mu_i) - \log(\sigma_i) \end{aligned} \tag{C.4}$$

Score Vectors

$$\begin{aligned} \frac{\partial l}{\partial \eta_i^\mu} &= -\sigma_i + \frac{\sigma_i}{\mu_i} y_i \\ \frac{\partial l}{\partial \eta_i^\sigma} &= \sigma_i \left(\log(\sigma_i) + 1 - \log(\mu_i) - \psi(\sigma_i) + \log(y_i) - \frac{y_i}{\mu_i}\right) \end{aligned}$$

Second Derivatives

$$\begin{aligned} \frac{\partial^2 l}{(\partial \eta_i^\mu)^2} &= -\frac{\sigma_i}{\mu_i} y_i \\ \frac{\partial^2 l}{(\partial \eta_i^\sigma)^2} &= \frac{\partial l}{\partial \eta_i^\sigma} + \sigma_i - \sigma_i^2 \psi_1(\sigma_i) \end{aligned}$$

Working Weights

$$\begin{aligned} E\left(-\frac{\partial^2 l}{(\partial \eta_i^\mu)^2}\right) &= \sigma_i \\ E\left(-\frac{\partial^2 l}{(\partial \eta_i^\sigma)^2}\right) &\overset{(C.4)}{=} \sigma_i \left(\sigma_i \psi_1(\sigma_i) - 1\right) \end{aligned}$$

Obviously the weight for μ_i is greater than zero. From $x\psi_1(x) > 1$ for $x > 0$ [Ronning, 1986] it also follows that $w_i^\sigma > 0$ holds such that all weights of the generalized gamma distribution are positive.

C.2.4 Weibull Distribution

Parameters are $\alpha_i = \exp(\eta_i^\alpha) > 0$ and $\lambda_i = \exp\left(\eta_i^\lambda\right) > 0$.

$$l = \log(\alpha_i) + (a_i - 1)\log(y_i) + \alpha_i \log(\lambda_i) - (\lambda_i y_i)^{\alpha_i}$$

$$\begin{aligned} E(y_i) &= \int_0^\infty \alpha_i(\lambda_i y_i)^{\alpha_i} \exp\left(-(\lambda_i y_i)^{\alpha_i}\right) \mathrm{d}y_i \\ &= \int_0^\infty \frac{1}{\lambda_i} u^{\frac{1}{\alpha_i}} \exp(-u)\mathrm{d}u \\ &= \frac{1}{\lambda_i}\Gamma\left(1 + \frac{1}{\alpha_i}\right) \\ \frac{\partial}{\partial \eta_i^\alpha}(\lambda_i y_i)^{\alpha_i} &= \alpha_i \log(\lambda_i y_i)(\lambda_i y_i)^{\alpha_i} \end{aligned}$$

Score Vectors

$$\begin{aligned} \frac{\partial l}{\partial \eta_i^\lambda} &= \alpha_i\left(1 - (\lambda_i y_i)^{\alpha_i}\right) \\ \frac{\partial l}{\partial \eta_i^\alpha} &= 1 + \alpha_i \log(\lambda_i y_i) - \alpha_i \log(\lambda_i y_i)(\lambda_i y_i)^{\alpha_i} \end{aligned} \tag{C.5}$$

Second Derivatives

$$\begin{aligned} \frac{\partial^2 l}{\left(\partial \eta_i^\lambda\right)^2} &= -\alpha_i^2(\lambda_i y_i)^{\alpha_i} \\ \frac{\partial^2 l}{\left(\partial \eta_i^\alpha\right)^2} &= \alpha_i \log(\lambda_i y_i) - \alpha_i \log(\lambda_i y_i)(\lambda_i y_i)^{\alpha_i} - \alpha_i^2 \left(\log(\lambda_i y_i)\right)^2 (\lambda_i y_i)^{\alpha_i} \end{aligned}$$

Working Weights

$$\begin{aligned} E((\lambda_i y_i)^{\alpha_i}) &= \int_0^\infty \alpha_i \lambda_i (\lambda_i y_i)^{2\alpha_i - 1} \exp\left((\lambda_i y_i)^{\alpha_i}\right) \mathrm{d}y_i \\ &= \int_0^\infty u \exp(-u) \mathrm{d}u \\ &= 1 \end{aligned} \tag{C.6}$$

$$\alpha_i E\left(\log(\lambda_i y_i)\left(1 - (\lambda_i y_i)^{\alpha_i}\right)\right) \overset{(C.5)}{=} -1 \tag{C.7}$$

$$\frac{\partial^2 \Gamma(x)}{(\partial x)^2} = \int_0^\infty u^{x-1} \left(\log(u)\right)^2 \exp(-u) \mathrm{d}u \tag{C.8}$$

$$\begin{aligned} \frac{\partial^2 \Gamma(x)}{(\partial x)^2} &= \frac{\partial}{\partial x}\left(\psi(x)\Gamma(x)\right) \\ &= \psi_1(x)\Gamma(x) + \left(\psi(x)\right)^2 \Gamma(x) \end{aligned} \tag{C.9}$$

From (C.8) and (C.9) it follows

$$\begin{aligned} E(\alpha_i^2 \left(\log(\lambda_i y_i)\right)^2 (\lambda_i y_i)^{\alpha_i}) &= \int_0^\infty \alpha_i^3 \lambda_i (\lambda_i y_i)^{2\alpha_i - 1} \left(\log\left(\lambda_i y_i\right)\right)^2) \exp\left((\lambda_i y_i)^{\alpha_i}\right) \mathrm{d}y_i \\ &= \int_0^\infty u \left(\log(u)\right)^2 \exp(-u) \mathrm{d}u \\ &= \Gamma(2)\left(\psi_1(2) + \left(\psi(2)\right)^2\right) \\ &\overset{(C.8)}{\underset{(C.9)}{=}} \left(\psi_1(2) + \left(\psi(2)\right)^2\right. \end{aligned} \tag{C.10}$$

Therefore we get

$$\begin{aligned} E\left(-\frac{\partial^2 l}{\left(\partial \eta_i^\lambda\right)^2}\right) &\overset{(C.6)}{=} \alpha_i^2 > 0 \\ E\left(-\frac{\partial^2 l}{\left(\partial \eta_i^\alpha\right)^2}\right) &\overset{(C.7)}{\underset{(C.10)}{=}} 1 + \left(\psi_1(2) + \left(\psi(2)\right)^2 > 0.\right. \end{aligned}$$

C.2.5 Pareto Distribution

Parameters are $b_i = \exp\left(\eta_i^b\right) > 0$ and $p_i = \exp\left(\eta_i^p\right) > 0$.

$$l = \log(p_i) - p_i \log(b_i) - (p_i + 1)\log(y_i + b_i)$$

$$\begin{aligned} E(y_i) &= p_i b_i^{p_i} \int_0^\infty \frac{y_i}{(y_i + b_i)^{p_i + 1}} \mathrm{d}y_i \\ &= \left[-b_i^{p_i} \frac{y_i}{(y_i + b_i)^{p_i}}\right]_0^\infty + b_i^{p_i} \int_0^\infty \frac{1}{(y_i + b_i)^{p_i}} \mathrm{d}y_i \\ &= \frac{b_i}{p_i - 1} \\ E\left(\frac{y_i}{(y_i + b_i)^2}\right) &= p_i b_i^{p_i} \int_0^\infty \frac{y_i}{(y_i + b_i)^{p_i + 3}} \mathrm{d}y_i \\ &= \frac{p_i}{b_i (p_i + 1)(p_i + 2)} \end{aligned} \tag{C.11}$$

Score Vectors

$$\begin{aligned}\frac{\partial l}{\partial \eta_i^b} &= p_i - \frac{(p_i+1)b_i}{y_i+b_i} \\ &= -1 + \frac{y_i(p_i+1)}{y_i+b_i} \\ \frac{\partial l}{\partial \eta_i^p} &= 1 + p_i\log(b_i) - p_i\log(y_i+b_i) \end{aligned} \tag{C.12}$$

Second Derivatives

$$\begin{aligned}\frac{\partial^2 l}{\left(\partial \eta_i^b\right)^2} &= -\frac{y_i b_i (p_i+1)}{(y_i+b_i)^2} \\ \frac{\partial^2 l}{\left(\partial \eta_i^p\right)^2} &= p_i\log(b_i) - p_i\log(y_i+b_i)\end{aligned}$$

Working Weights From (C.12) it follows $E\left(p_i\left(\log(b_i) - \log(y_i+b_i)\right)\right) = -1$ and hence

$$\begin{aligned}E\left(-\frac{\partial^2 l}{\left(\partial \eta_i^b\right)^2}\right) &\overset{(C.11)}{=} \frac{p_i}{p_i+2} \\ E\left(-\frac{\partial^2 l}{\left(\partial \eta_i^p\right)^2}\right) &\overset{(C.12)}{=} 1\end{aligned}$$

Since $p_i > 0$ holds, the weights for the Pareto distribution are greater than zero as desired.

C.2.6 Generalized Gamma Distribution

Parameters are $\mu_i = \exp\left(\eta_i^\mu\right) > 0$, $\sigma_i = \exp\left(\eta_i^\sigma\right) > 0$ and $\tau_i = \exp\left(\eta_i^\tau\right) > 0$. Since the parametrization is similar to the one of the gamma distribution (compare Section C.2.3) details for parameters μ_i and σ_i will be omitted in the following.

$$l = \log(\tau_i) + \sigma_i\tau_i\log(\sigma_i) - \sigma_i\tau_i\log(\mu_i) - \log\left(\Gamma\left(\sigma_i\right)\right) + \left(\sigma_i\tau_i - 1\right)\log(y_i) - \left(\frac{\sigma_i}{\mu_i}y_i\right)^{\tau_i}$$

$$\begin{aligned}E(y_i) &= \int_0^\infty \frac{\tau_i}{\Gamma(\sigma_i)}\left(\frac{\sigma_i}{\mu_i}y_i\right)^{\sigma_i\tau_i}\exp\left(-\left(\frac{\sigma_i}{\mu_i}y_i\right)^{\tau_i}\right)\mathrm{d}y_i \\ &= \int_0^\infty \frac{\mu_i}{\sigma_i\Gamma(\sigma_i)}u^{\sigma_i - 1 + \frac{1}{\tau_i}}\exp\left(-u\right)\mathrm{d}u \\ &= \frac{\mu_i}{\sigma_i\Gamma(\sigma_i)}\Gamma\left(\sigma_i + \frac{1}{\tau_i}\right)\end{aligned}$$

$$E\left(\left(\frac{\sigma_i}{\mu_i}y_i\right)^{\tau_i}\right) = \sigma_i \tag{C.13}$$

$$E\left(\sigma_i\tau_i\log(y_i)\right) = \sigma_i\tau_i\log\left(\frac{\sigma_i}{\mu_i}\right) + \sigma_i\psi(\sigma_i) \tag{C.14}$$

Score Vectors

$$\begin{aligned}
\frac{\partial l}{\partial \eta_i^\mu} &= -\sigma_i\tau_i + \tau_i\left(\frac{\sigma_i}{\mu_i}y_i\right)^{\tau_i} \\
\frac{\partial l}{\partial \eta_i^\sigma} &= \sigma_i\left(\tau_i\log(\sigma_i) + \tau_i - \tau_i\log(\mu_i) - \psi\left(\sigma_i\right) + \tau_i\log(y_i)\right) - \tau_i\left(\frac{\sigma_i}{\mu_i}y_i\right)^{\tau_i} \\
\frac{\partial l}{\partial \eta_i^\tau} &= 1 + \sigma_i\tau_i\log(\sigma_i) - \sigma_i\tau_i\log(\mu_i) + \sigma_i\tau_i\log(y_i) - \tau_i\left(\frac{\sigma_i}{\mu_i}y_i\right)^{\tau_i}\log\left(\frac{\sigma_i}{\mu_i}y_i\right)
\end{aligned}$$

Second Derivatives

$$\begin{aligned}
\frac{\partial^2 l}{\left(\partial \eta_i^\mu\right)^2} &= -\tau_i^2\left(\frac{\sigma_i}{\mu_i}y_i\right)^{\tau_i} \\
\frac{\partial^2 l}{\left(\partial \eta_i^{\sigma^2}\right)^2} &= \sigma_i\left(\tau_i\log(\sigma_i) + \tau_i - \tau_i\log(\mu_i) - \psi\left(\sigma_i\right) + \tau_i\log(y_i)\right) \\
&\quad + \sigma_i\tau_i - \sigma_i^2\psi_1(\sigma_i) - \tau_i^2\left(\frac{\sigma_i}{\mu_i}y_i\right)^{\tau_i} \\
\frac{\partial^2 l}{\left(\partial \eta_i^\tau\right)^2} &= \frac{\partial l}{\partial \eta_i^\tau} - \tau_i^2\left(\frac{\sigma_i}{\mu_i}y_i\right)^{\tau_i}\left(\log\left(\frac{\sigma_i}{\mu_i}y_i\right)\right)^2 - 1
\end{aligned}$$

Working Weights A similar procedure as for the Weibull distribution in Section C.2.4 yields

$$E\left(\left(\frac{\sigma_i}{\mu_i}y_i\right)^{\tau_i}\left(\log\left(\frac{\sigma_i}{\mu_i}y_i\right)\right)^2\right) = \frac{\sigma_i}{\tau_i^2}\left(\psi_1(\sigma_i+1) + \left(\psi(\sigma_i+1)\right)^2\right) \quad \text{(C.15)}$$

and hence

$$\begin{aligned}
E\left(-\frac{\partial^2 l}{\left(\partial \eta_i^\mu\right)^2}\right) &\overset{\text{(C.13)}}{=} \sigma_i\tau_i^2 \\
E\left(-\frac{\partial^2 l}{\left(\partial \eta_i^\sigma\right)^2}\right) &\overset{\text{(C.13)}}{=} \sigma_i^2\psi_1(\sigma_i) + \sigma_i\tau_i^2 - 2\sigma_i\tau_i \\
E\left(-\frac{\partial^2 l}{\left(\partial \eta_i^\tau\right)^2}\right) &\overset{\text{(C.15)}}{=} \sigma_i\left(\psi_1(\sigma_i+1) + \left(\psi(\sigma_i+1)\right)^2\right) + 1
\end{aligned}$$

Obviously the weight for μ_i and τ_i are greater than zero. For σ_i we have on the one hand that $\sigma_i\psi_1(\sigma_i) > 1$ holds, compare Section C.2.3. On the other hand $\tau_i^2 - 2\tau_i$ is a quadratic function with minimum at -1 such that $w_i^\tau > 0$ is true. Consequently all weights for the generalized gamma distribution are greater than zero.

C.2.7 Dagum Distribution

Parameters are $a_i = \exp\left(\eta_i^a\right) > 0$, $b_i = \exp\left(\eta_i^b\right) > 0$ and $p_i = \exp\left(\eta_i^p\right) > 0$.

$$l = \log(a_i) + \log(p_i) - a_ip_i\log(b_i) + (a_ip_i - 1)\log(y_i) - (p_i+1)\log\left(1 + \left(\frac{y_i}{b_i}\right)^{a_i}\right)$$

$$\begin{aligned}
\frac{\partial}{\partial \eta_i^a}\left(\frac{y_i}{b_i}\right)^{a_i} &= a_i \log\left(\frac{y_i}{b_i}\right)\left(\frac{y_i}{b_i}\right)^{a_i} \\
\frac{\partial}{\partial \eta_i^b}\left(\frac{y_i}{b_i}\right)^{a_i} &= -a_i\left(\frac{y_i}{b_i}\right)^{a_i}
\end{aligned}$$

If $a_i > 1$ we obtain

$$\begin{aligned}
E\left(\frac{\left(\frac{y_i}{b_i}\right)^{a_i}}{\left(1+\left(\frac{y_i}{b_i}\right)^{a_i}\right)^2}\right) &= \int_0^\infty \frac{a_i p_i \left(\frac{y_i}{b_i}\right)^{a_i p_i + a_i - 1}}{b_i\left(1+\left(\frac{y_i}{b_i}\right)^{a_i}\right)^{p_i+3}} \mathrm{d}y_i \\
&= \int_0^\infty \frac{p_i u^{p_i}}{(1+u)^{p_i+3}} \mathrm{d}u \\
&= p_i \,\mathrm{B}\left(p_i+1, 2\right) \\
&= \frac{p_i}{(p_i+1)(p_i+2)} \qquad \text{(C.16)}
\end{aligned}$$

where $\mathrm{B}(x, y)$ is the beta function of $x, y > 0$.

Score Vectors

$$\begin{aligned}
\frac{\partial l}{\partial \eta_i^a} &= 1 + a_i \log(y_i) - a_i p_i \log(b_i) - \frac{a_i(p_i+1)}{1+\left(\frac{y_i}{b_i}\right)^{a_i}} \log\left(\frac{y_i}{b_i}\right)\left(\frac{y_i}{b_i}\right)^{a_i} \\
\frac{\partial l}{\partial \eta_i^b} &= -a_i p_i + \frac{a_i(p_i+1)}{1+\left(\frac{y_i}{b_i}\right)^{a_i}}\left(\frac{y_i}{b_i}\right)^{a_i} \\
&= a_i - \frac{a_i(p_i+1)}{1+\left(\frac{y_i}{b_i}\right)^{a_i}} \\
\frac{\partial l}{\partial \eta_i^p} &= 1 - a_i p_i \log(b_i) + a_i p_i \log(y_i) - p_i \log\left(1+\left(\frac{y_i}{b_i}\right)^{a_i}\right)
\end{aligned}$$

Second Derivatives

$$\begin{aligned}
\frac{\partial^2 l}{\left(\partial \eta_i^a\right)^2} &= a_i \log(y_i) - a_i p_i \log(b_i) - \frac{a_i(p_i+1)}{1+\left(\frac{y_i}{b_i}\right)^{a_i}} \log\left(\frac{y_i}{b_i}\right)\left(\frac{y_i}{b_i}\right)^{a_i} \qquad \text{(C.17)} \\
&\quad - \frac{a_i^2(p_i+1)}{\left(1+\left(\frac{y_i}{b_i}\right)^{a_i}\right)^2}\left(\log\left(\frac{y_i}{b_i}\right)\right)^2\left(\frac{y_i}{b_i}\right)^{a_i} \\
\frac{\partial^2 l}{\left(\partial \eta_i^b\right)^2} &= \frac{a_i^2(p_i+1)}{\left(1+\left(\frac{y_i}{b_i}\right)^{a_i}\right)^2}\left(\frac{y_i}{b_i}\right)^{a_i} \\
\frac{\partial^2 l}{\left(\partial \eta_i^p\right)^2} &= a_i p_i \log(b_i) + a_i p_i \log(y_i) - p_i \log\left(1+\left(\frac{y_i}{b_i}\right)^{a_i}\right) \qquad \text{(C.18)}
\end{aligned}$$

Working Weights

$$
\begin{aligned}
E\left(-\frac{\partial^2 l}{\left(\partial \eta_i^a\right)^2}\right) &\overset{(C.17)}{=} E\left(\frac{a_i^2(p_i+1)}{\left(1+\left(\frac{y_i}{b_i}\right)^{a_i}\right)^2}\left(\log\left(\frac{y_i}{b_i}\right)\right)^2\left(\frac{y_i}{b_i}\right)^{a_i}\right) \\
E\left(-\frac{\partial^2 l}{\left(\partial \eta_i^b\right)^2}\right) &\overset{(C.16)}{=} \frac{a_i^2 p_i}{p_i+2} \\
E\left(-\frac{\partial^2 l}{\left(\partial \eta_i^p\right)^2}\right) &\overset{(C.18)}{=} 1
\end{aligned}
$$

and w_i^b, w_i^p is greater than zero. For a_i there is no obvious way to compute the expectation of $-\frac{\partial^2 l}{\left(\partial \eta_i^a\right)^2}$ for all $a_i > 0$ but taking $\frac{a_i^2(p_i+1)}{\left(1+\left(\frac{y_i}{b_i}\right)^{a_i}\right)^2}\left(\log\left(\frac{y_i}{b_i}\right)\right)^2\left(\frac{y_i}{b_i}\right)^{a_i} > 0$ as working weight has been studied in simulations to work quite well.

C.3 Discrete Responses

For discrete responses see the paper and supplement of Klein et al. [2013].

C.4 Mixed Discrete-Continuous Distributions

C.4.1 Zero-Adjusted Distributions

For mixed discrete-continuous distributions we propose all distributions from Section C.2 as possible candidates for the continuous part. For the point mass we provide the logit and the complementary log-log response functions. For both choices the following quantities hold:

Logit Model In addition to the parameters of the continuous distribution g we have $\pi_i = \frac{\exp\left(\eta_i^\pi\right)}{1+\exp\left(\eta_i^\pi\right)} \in (0,1)$.

$$
l = \log(1-\pi_i))\mathbb{1}_{\{0\}}(y_i) + \left(\log(\pi_i) + \log(g_i)\right)\left(1-\mathbb{1}_{\{0\}}(y_i)\right)
$$

where $\log(g_i) = \log(g(y_i))$ is the log-likelihood of the continuous distribution.

$$
\begin{aligned}
\frac{\partial \pi_i}{\partial \eta_i^\pi} &= \pi_i(1-\pi_i) \\
\frac{\partial}{\partial \eta_i^\pi}(1-\pi_i) &= -\pi_i(1-\pi_i)
\end{aligned}
$$

Score Vector

$$\begin{aligned}\frac{\partial l}{\partial \eta_i^\pi} &= -\pi_i \mathbb{1}_{\{0\}}(y_i) + (1-\pi_i)(1-\mathbb{1}_{\{0\}}(y_i)) \\ &= 1 - \pi_i - \mathbb{1}_{\{0\}}(y_i)\end{aligned}$$

Second Derivative

$$\frac{\partial^2 l}{(\partial \eta_i^{\pi_i})^2} = -\pi_i(1-\pi_i)$$

Working Weight

$$E\left(-\frac{\partial^2 l}{(\partial \eta_i^{\pi_i})^2}\right) = \pi_i(1-\pi_i)$$

which is always greater than zero.

Complementary Log-Log Model In addition to the parameters of the continuous distribution g we have $\pi_i = 1 - \exp(-\exp(\eta_i^\pi)) \in (0,1)$.

$$l = \log(1-\pi_i)\mathbb{1}_{\{0\}}(y_i) + (\log(\pi_i) + \log(g_i))\left(1 - \mathbb{1}_{\{0\}}(y_i)\right)$$

where $\log(g_i) = \log(g(y_i))$ is the log-likelihood of the continuous distribution.

$$E(\mathbb{1}_{\{0\}}(y_i)) = 1 - \pi_i \tag{C.19}$$

$$\begin{aligned}\frac{\partial \pi_i}{\partial \eta_i^\pi} &= (1-\pi_i)\exp(\eta_i^\pi) \\ \frac{\partial}{\partial \eta_i^\pi}(1-\pi_i) &= -(1-\pi_i)\exp(\eta_i^\pi)\end{aligned}$$

Score Vector

$$\frac{\partial l}{\partial \eta_i^\pi} = -\frac{\exp(\eta_i^\pi)}{\pi_i}\mathbb{1}_{\{0\}}(y_i) + \frac{\exp(\eta_i^\pi)(1-\pi_i)}{\pi_i}$$

Second Derivative

$$\begin{aligned}\frac{\partial^2 l}{(\partial \eta_i^{\pi_i})^2} = &-\frac{\exp(\eta_i^\pi)}{\pi_i}\mathbb{1}_{\{0\}}(y_i) + \frac{(\exp(\eta_i^\pi))^2(1-\pi_i)}{\pi_i^2}\mathbb{1}_{\{0\}}(y_i) + \frac{\exp(\eta_i^\pi)(1-\pi_i)}{\pi_i} \\ &-\frac{(\exp(\eta_i^\pi))^2(1-\pi_i)}{\pi_i} - \frac{(\exp(\eta_i^\pi))^2(1-\pi_i)^2}{\pi_i^2}\end{aligned}$$

Working Weight

$$E\left(-\frac{\partial^2 l}{\left(\partial \eta_i^{\pi_i}\right)^2}\right) \overset{(C.19)}{=} \frac{\left(\exp\left(\eta_i^{\pi}\right)\right)^2 (1-\pi_i)}{\pi_i}$$

which is always greater than zero.

C.5 Distributions with Compact Support

C.5.1 Beta Distribution

Parameters are $\mu_i = \frac{\exp\left(\eta_i^{\mu}\right)}{1+\exp\left(\eta_i^{\mu}\right)}$ and $\sigma_i^2 = \frac{\exp\left(\eta_i^{\sigma^2}\right)}{1+\exp\left(\eta_i^{\sigma^2}\right)}$, both in the interval $(0,1)$.

$$\begin{aligned} l &= -\log\left(\Gamma\left(\frac{\mu_i(1-\sigma_i^2)}{\sigma_i^2}\right)\right) - \log\left(\Gamma\left(\frac{(1-\mu_i)(1-\sigma_i^2)}{\sigma_i^2}\right)\right) + \log\left(\Gamma\left(\frac{(1-\sigma_i^2)}{\sigma_i^2}\right)\right) \\ &\quad + \frac{\mu_i(1-\sigma_i^2)}{\sigma_i^2}\log(y_i) + \frac{(1-\mu_i)(1-\sigma_i^2)}{\sigma_i^2}\log(1-y_i) - \log(y_i) - \log(1-y_i) \end{aligned}$$

$$\begin{aligned} E(y_i) &= \mu_i \\ \frac{\partial \mu_i}{\partial \eta_i^{\mu}} &= \mu_i(1-\mu_i) \\ \frac{\partial}{\partial \eta_i^{\mu}}(1-\mu_i) &= -\mu_i(1-\mu_i) \\ \frac{\partial}{\partial \eta_i^{\sigma^2}}\left(\frac{1-\sigma_i^2}{\sigma_i^2}\right) &= -\frac{1-\sigma_i^2}{\sigma_i^2} \end{aligned}$$

Score Vectors

$$\begin{aligned} \frac{\partial l}{\partial \eta_i^{\mu}} &= \mu_i(1-\mu_i)\frac{1-\sigma_i^2}{\sigma_i^2}\left(\psi\left(\frac{(1-\mu_i)(1-\sigma_i^2)}{\sigma_i^2}\right) - \psi\left(\frac{\mu_i(1-\sigma_i^2)}{\sigma_i^2}\right) + \log\left(\frac{y_i}{1-y_i}\right)\right) \\ \frac{\partial l}{\partial \eta_i^{\sigma^2}} &= \frac{1-\sigma_i^2}{\sigma_i^2}\left((1-\mu_i)\psi\left(\frac{(1-\mu_i)(1-\sigma_i^2)}{\sigma_i^2}\right) + \mu_i\psi\left(\frac{\mu_i(1-\sigma_i^2)}{\sigma_i^2}\right) - \psi\left(\frac{1-\sigma_i^2}{\sigma_i^2}\right)\right) \\ &\quad - \frac{1-\sigma_i^2}{\sigma_i^2}\left(\mu_i\log(y_i) - (1-\mu_i)\log(1-y_i)\right) \end{aligned}$$

Second Derivatives

$$\begin{aligned} \frac{\partial^2 l}{\left(\partial \eta_i^{\mu}\right)^2} &= (1-2\mu_i)\frac{\partial l}{\partial \eta_i^{\mu}} \\ &\quad - \left(\frac{1-\sigma_i^2}{\sigma_i^2}\right)^2 \mu_i^2(1-\mu_i)^2\left(\psi_1\left(\frac{(1-\mu_i)(1-\sigma_i^2)}{\sigma_i^2}\right) + \psi_1\left(\frac{\mu_i(1-\sigma_i^2)}{\sigma_i^2}\right)\right) \\ \frac{\partial^2 l}{\left(\partial \eta_i^{\sigma^2}\right)^2} &= -\frac{\partial l}{\partial \eta_i^{\sigma^2}} - \left(\frac{1-\sigma_i^2}{\sigma_i^2}\right)^2\left(-\psi_1\left(\frac{1-\sigma_i^2}{\sigma_i^2}\right) + \mu_i^2\psi_1\left(\frac{\mu_i(1-\sigma_i^2)}{\sigma_i^2}\right)\right) \\ &\quad - \left(\frac{1-\sigma_i^2}{\sigma_i^2}\right)^2 (1-\mu_i)^2\psi_1\left(\frac{(1-\mu_i)(1-\sigma_i^2)}{\sigma_i^2}\right) \end{aligned}$$

Working Weights

$$
\begin{aligned}
E\left(-\frac{\partial^2 l}{\left(\partial\eta_i^{\mu}\right)^2}\right) &= \left(\frac{1-\sigma_i^2}{\sigma_i^2}\right)^2 \mu_i^2(1-\mu_i)^2\left(\psi_1\left(\frac{(1-\mu_i)(1-\sigma_i^2)}{\sigma_i^2}\right)+\psi_1\left(\frac{\mu_i(1-\sigma_i^2)}{\sigma_i^2}\right)\right)\\
E\left(-\frac{\partial^2 l}{\left(\partial\eta_i^{\sigma^2}\right)^2}\right) &= \left(\frac{1-\sigma_i^2}{\sigma_i^2}\right)^2\left(-\psi_1\left(\frac{1-\sigma_i^2}{\sigma_i^2}\right)+\mu_i^2\psi_1\left(\frac{\mu_i(1-\sigma_i^2)}{\sigma_i^2}\right)\right)\\
&\quad+\left(\frac{1-\sigma_i^2}{\sigma_i^2}\right)^2(1-\mu_i)^2\psi_1\left(\frac{(1-\mu_i)(1-\sigma_i^2)}{\sigma_i^2}\right)
\end{aligned}
$$

where w_i^{μ} is greater than zero since the trigamma function only takes positive values. Furthermore $w_i^{\sigma^2} > 0$ holds. Consider Guo and Qi [2013, Lemma 1], saying that $\frac{1}{x}+\frac{1}{2x^2} < \psi_1(x) < \frac{1}{x}+\frac{1}{x^2}$ for $x \in (0,\infty)$. We therefore get with $x = \frac{1-\sigma_i^2}{\sigma_i^2}$:

$$
\begin{aligned}
(1-\mu_i^2)\psi_1\left((1-\mu_i)x\right)+\mu_i^2\psi_1\left(\mu_i x\right)-\psi_1\left(x\right) &> \frac{(1-\mu_i)^2}{(1-\mu_i)x}+\frac{(1-\mu_i)^2}{2(1-\mu_i)^2x^2}\\
&\quad+\frac{\mu_i^2}{\mu_i x}+\frac{\mu_i^2}{2\mu_i^2x^2}-\frac{1}{x}-\frac{1}{x^2}\\
&= \frac{\mu_i}{x}+\frac{1-\mu_i}{x}+\frac{1}{x^2}-\frac{1}{x}-\frac{1}{x^2}\\
&= 0.
\end{aligned}
$$

C.5.2 Zero-One-Inflated Beta Distribution

Let now g denote the density of a beta distribution. Then, additional to the parameters of the beta distribution we obtain $\nu_i = \exp\left(\eta_i^{\nu}\right), \tau_i = \exp\left(\eta_i^{\tau}\right) > 0$.

$$
\begin{aligned}
l &= \left(\log(\nu_i)-\log(1+\nu_i+\tau_i)\right)\mathbb{1}_{\{0\}}(y_i)+\left(\log(\tau_i)-\log(1+\nu_i+\tau_i)\right)\mathbb{1}_{\{1\}}(y_i)\\
&\quad+\left(\log(g_i)-\log(1+\nu_i+\tau_i)\right)\mathbb{1}_{(0,1)}(y_i)
\end{aligned}
$$

where $\log(g_i) = \log(g(y_i))$ is the log-likelihood of the beta distribution from the previous section.

Score Vectors

$$
\begin{aligned}
\frac{\partial l}{\partial\eta_i^{\nu}} &= \left(1-\frac{\nu_i}{1+\nu_i+\tau_i}\right)\mathbb{1}_{\{0\}}(y_i)-\frac{\nu_i}{1+\nu_i+\tau_i}\mathbb{1}_{\{1\}}(y_i)-\frac{\nu_i}{1+\nu_i+\tau_i}\mathbb{1}_{(0,1)}(y_i)\\
&= \mathbb{1}_{\{0\}}(y_i)-\frac{\nu_i}{1+\nu_i+\tau_i}\\
\frac{\partial l}{\partial\eta_i^{\tau}} &= \mathbb{1}_{\{1\}}(y_i)-\frac{\tau_i}{1+\nu_i+\tau_i}
\end{aligned}
$$

Second Derivatives

$$\begin{aligned}\frac{\partial^2 l}{\left(\partial \eta_i^{\nu_i}\right)^2} &= -\frac{\nu_i(1+\tau_i)}{\left(1+\nu_i+\tau_i\right)^2} \\ \frac{\partial^2 l}{\left(\partial \eta_i^{\tau_i}\right)^2} &= -\frac{\tau_i(1+\nu_i)}{\left(1+\nu_i+\tau_i\right)^2}\end{aligned}$$

Working Weights

$$\begin{aligned}E\left(-\frac{\partial^2 l}{\left(\partial \eta_i^{\nu_i}\right)^2}\right) &= \frac{\nu_i(1+\tau_i)}{\left(1+\nu_i+\tau_i\right)^2} \\ E\left(-\frac{\partial^2 l}{\left(\partial \eta_i^{\nu_i}\right)^2}\right) &= \frac{\tau_i(1+\nu_i)}{\left(1+\nu_i+\tau_i\right)^2}\end{aligned}$$

and the weights are obviously greater than zero as desired.

References

C. P. Chen and F. Qi. The best bounds of harmonic sequence, 2003. arXiv:math/0306233.

B. N. Guo and F. Qi. Refinements of lower bounds for polygamma functions. Proceedings of the American Mathematical Society, 141(3):1007–1015, 2013.

N. Klein, T. Kneib, and S. Lang. Bayesian generalized additive models for location, scale and shape for zero-inflated and overdispersed count data. Technical report, 2013.

G. Ronning. On the curvature of the trigamma function. Journal of Computational and Applied Mathematics, 15:397–399, 1986.

University of Innsbruck - Working Papers in Economics and Statistics
Recent Papers can be accessed on the following webpage:

http://eeecon.uibk.ac.at/wopec/

2013-23 **Nadja Klein, Thomas Kneib, Stefan Lang:** Bayesian structured additive distributional regression

2013-22 **David Plavcan, Georg J. Mayr, Achim Zeileis:** Automatic and probabilistic foehn diagnosis with a statistical mixture model

2013-21 **Jakob W. Messner, Georg J. Mayr, Achim Zeileis, Daniel S. Wilks:** Extending extended logistic regression to effectively utilize the ensemble spread

2013-20 **Michael Greinecker, Konrad Podczeck:** Liapounoff's vector measure theorem in Banach spaces *forthcoming in* Economic Theory Bulletin

2013-19 **Florian Lindner:** Decision time and steps of reasoning in a competitive market entry game

2013-18 **Michael Greinecker, Konrad Podczeck:** Purification and independence

2013-17 **Loukas Balafoutas, Rudolf Kerschbamer, Martin Kocher, Matthias Sutter:** Revealed distributional preferences: Individuals vs. teams

2013-16 **Simone Gobien, Björn Vollan:** Playing with the social network: Social cohesion in resettled and non-resettled communities in Cambodia

2013-15 **Björn Vollan, Sebastian Prediger, Markus Frölich:** Co-managing common pool resources: Do formal rules have to be adapted to traditional ecological norms?

2013-14 **Björn Vollan, Yexin Zhou, Andreas Landmann, Biliang Hu, Carsten Herrmann-Pillath:** Cooperation under democracy and authoritarian norms

2013-13 **Florian Lindner, Matthias Sutter:** Level-k reasoning and time pressure in the 11-20 money request game *forthcoming in* Economics Letters

2013-12 **Nadja Klein, Thomas Kneib, Stefan Lang:** Bayesian generalized additive models for location, scale and shape for zero-inflated and overdispersed count data

2013-11 **Thomas Stöckl:** Price efficiency and trading behavior in limit order markets with competing insiders *forthcoming in* Experimental Economics

2013-10 **Sebastian Prediger, Björn Vollan, Benedikt Herrmann:** Resource scarcity, spite and cooperation

2013-09 **Andreas Exenberger, Simon Hartmann:** How does institutional change coincide with changes in the quality of life? An exemplary case study

2013-08 **E. Glenn Dutcher, Loukas Balafoutas, Florian Lindner, Dmitry Ryvkin, Matthias Sutter:** Strive to be first or avoid being last: An experiment on relative performance incentives.

2013-07 **Daniela Glätzle-Rützler, Matthias Sutter, Achim Zeileis:** No myopic loss aversion in adolescents? An experimental note

2013-06 **Conrad Kobel, Engelbert Theurl:** Hospital specialisation within a DRG-Framework: The Austrian case

2013-05 **Martin Halla, Mario Lackner, Johann Scharler:** Does the welfare state destroy the family? Evidence from OECD member countries

2013-04 **Thomas Stöckl, Jürgen Huber, Michael Kirchler, Florian Lindner:** Hot hand belief and gambler's fallacy in teams: Evidence from investment experiments

2013-03 **Wolfgang Luhan, Johann Scharler:** Monetary policy, inflation illusion and the Taylor principle: An experimental study

2013-02 **Esther Blanco, Maria Claudia Lopez, James M. Walker:** Tensions between the resource damage and the private benefits of appropriation in the commons

2013-01 **Jakob W. Messner, Achim Zeileis, Jochen Broecker, Georg J. Mayr:** Improved probabilistic wind power forecasts with an inverse power curve transformation and censored regression

2012-27 **Achim Zeileis, Nikolaus Umlauf, Friedrich Leisch:** Flexible generation of e-learning exams in R: Moodle quizzes, OLAT assessments, and beyond

2012-26 **Francisco Campos-Ortiz, Louis Putterman, T.K. Ahn, Loukas Balafoutas, Mongoljin Batsaikhan, Matthias Sutter:** Security of property as a public good: Institutions, socio-political environment and experimental behavior in five countries

2012-25 **Esther Blanco, Maria Claudia Lopez, James M. Walker:** Appropriation in the commons: variations in the opportunity costs of conservation

2012-24 **Edgar C. Merkle, Jinyan Fan, Achim Zeileis:** Testing for measurement invariance with respect to an ordinal variable *forthcoming in Psychometrika*

2012-23 **Lukas Schrott, Martin Gächter, Engelbert Theurl:** Regional development in advanced countries: A within-country application of the Human Development Index for Austria

2012-22 **Glenn Dutcher, Krista Jabs Saral:** Does team telecommuting affect productivity? An experiment

2012-21 **Thomas Windberger, Jesus Crespo Cuaresma, Janette Walde:** Dirty floating and monetary independence in Central and Eastern Europe - The role of structural breaks

2012-20 **Martin Wagner, Achim Zeileis:** Heterogeneity of regional growth in the European Union

2012-19 **Natalia Montinari, Antonio Nicolo, Regine Oexl:** Mediocrity and induced reciprocity

2012-18 **Esther Blanco, Javier Lozano:** Evolutionary success and failure of wildlife conservancy programs

2012-17 **Ronald Peeters, Marc Vorsatz, Markus Walzl:** Beliefs and truth-telling: A laboratory experiment

2012-16 **Alexander Sebald, Markus Walzl:** Optimal contracts based on subjective evaluations and reciprocity

2012-15 **Alexander Sebald, Markus Walzl:** Subjective performance evaluations and reciprocity in principal-agent relations

2012-14 **Elisabeth Christen:** Time zones matter: The impact of distance and time zones on services trade

2012-13 **Elisabeth Christen, Joseph Francois, Bernard Hoekman:** CGE modeling of market access in services

2012-12 **Loukas Balafoutas, Nikos Nikiforakis:** Norm enforcement in the city: A natural field experiment *forthcoming in European Economic Review*

2012-11 **Dominik Erharter:** Credence goods markets, distributional preferences and the role of institutions

2012-10 **Nikolaus Umlauf, Daniel Adler, Thomas Kneib, Stefan Lang, Achim Zeileis:** Structured additive regression models: An R interface to BayesX

2012-09 **Achim Zeileis, Christoph Leitner, Kurt Hornik:** History repeating: Spain beats Germany in the EURO 2012 Final

2012-08 **Loukas Balafoutas, Glenn Dutcher, Florian Lindner, Dmitry Ryvkin:** The optimal allocation of prizes in tournaments of heterogeneous agents

2012-07 **Stefan Lang, Nikolaus Umlauf, Peter Wechselberger, Kenneth Harttgen, Thomas Kneib:** Multilevel structured additive regression

2012-06 **Elisabeth Waldmann, Thomas Kneib, Yu Ryan Yu, Stefan Lang:** Bayesian semiparametric additive quantile regression

2012-05 **Eric Mayer, Sebastian Rueth, Johann Scharler:** Government debt, inflation dynamics and the transmission of fiscal policy shocks *forthcoming in Economic Modelling*

2012-04 **Markus Leibrecht, Johann Scharler:** Government size and business cycle volatility; How important are credit constraints? *forthcoming in Economica*

2012-03 **Uwe Dulleck, David Johnston, Rudolf Kerschbamer, Matthias Sutter:** The good, the bad and the naive: Do fair prices signal good types or do they induce good behaviour?

2012-02 **Martin G. Kocher, Wolfgang J. Luhan, Matthias Sutter:** Testing a forgotten aspect of Akerlof's gift exchange hypothesis: Relational contracts with individual and uniform wages

2012-01 **Loukas Balafoutas, Florian Lindner, Matthias Sutter:** Sabotage in tournaments: Evidence from a natural experiment *published in Kyklos*

University of Innsbruck

Working Papers in Economics and Statistics

2013-23

Nadja Klein, Thomas Kneib, Stefan Lang

Bayesian structured additive distributional regression

Abstract
In this paper, we propose a generic Bayesian framework for inference in distributional regression models in which each parameter of a potentially complex response distribution and not only the mean is related to a structured additive predictor. The latter is composed additively of a variety of different functional effect types such as nonlinear effects, spatial effects, random coefficients, interaction surfaces or other (possibly non-standard) basis function representations. To enforce specific properties of the functional effects such as smoothness, informative multivariate Gaussian priors are assigned to the basis function coefficients. Inference is then based on efficient Markov chain Monte Carlo simulation techniques where a generic procedure makes use of distribution-specific iteratively weighted least squares approximations to the full conditionals. We study properties of the resulting model class and provide detailed guidance on practical aspects of model choice including selecting an appropriate response distribution and predictor specification. The importance and flexibility of Bayesian structured additive distributional regression to estimate all parameters as functions of explanatory variables and therefore to obtain more realistic models, is exemplified in two applications with complex response distributions.

ISSN 1993-4378 (Print)
ISSN 1993-6885 (Online)