Bayesian Structured Additive Distributional Regression with an Application to Regional Income Inequality in Germany

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

Stefan Lang
Department of Statistics
University of Innsbruck

Alexander Sohn
Chair of Statistics
Georg-August-University Göttingen

Abstract

In this paper, we propose a generic Bayesian framework for inference in distributional regression models to study determinants of labour incomes in Germany with a particular focus on regional differences after the German reunification. In a distributional regression model, each parameter of a potentially complex response distribution and not only the mean is related to a structured additive predictor which enables the comparison of several competing distributions for labour incomes and the consideration of effects not only on expected incomes but also on higher order moments of the income distribution. This allows us to study the effects of explanatory variables not only on average incomes but also additional aspects associated with income distributions, such as quantiles or inequality measures, all being derived from one joint model. In our application, a special focus will be on differences of the Gini coefficient to quantify inequality. To incorporate different effect types such as nonlinear effects for age, education and work experience, as well as spatial effects to cover small-scale unobserved spatial heterogeneity, we consider structured additive predictors for all distributional parameters. The different functional effect types are expanded in (possibly non-standard) basis function representations and supplemented with multivariate Gaussian priors to enforce desirable properties such as smoothness. Inference is 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. The case study on labour income allows us to discuss several important practical aspects that will be relevant in any application of distributional regression and in particular model choice including selecting an appropriate response distribution and predictor specification. The analysis of labour incomes therefore also serves as a blueprint for other applications of distributional regression models.

Key words: Generalised additive models for location, scale and shape; income distributions; iteratively weighted least squares proposal; Markov chain Monte Carlo simulation; semiparametric regression; wage inequality
1 Introduction

The analysis of determinants of labour incomes has a long tradition in economics, dating back at least to Mincer (1974). His classical wage equation includes potential labour market experience as well as years of education as the most important determinants of human capital which then translates into expected income (Lemieux, 2006). Additional determinants that are frequently considered in the analysis of incomes are age, actually realised labour market experience, gender, regional information concerning the residence of employees, or area of employment. One considerable restriction of most analyses conducted so far is their sole focus on the impact of these potential explanatory variables on the expected income given covariates, i.e. on the conditional mean. If distributional assumptions are required, like for inequality decomposition (see Morduch and Sicular, 2002) or truncation (see Greene, 2008, Ch. 19), the (log-)normal distribution is often implicitly considered (again with regression effects only on the mean). More flexible types of distributions have so far mostly been used to describe income distributions on a highly aggregated level, normally at the national level (e.g. see Kleiber, 1996).

In this paper, we utilise detailed, longitudinal information on incomes available from the German socio-economic panel (SOEP) to derive a flexible, structured additive distributional regression model for labour incomes of full-time male workers. We consider several candidate distributions for describing the non-negative conditional income distributions including the log-normal distribution, the gamma distribution, the inverse Gaussian distribution and the Dagum distribution. To obtain flexible conditional income distributions, we allow for regression effects on potentially all parameters of the income distribution thereby overcoming the previous concentration on expected incomes. Some earlier attempts to define such distributional regression models comprise Biewen and Jenkins (2005) who suggest a coarse conditioning into subgroups estimating parametric income distributions and thus allowing for variations in the nature of the distribution or Donald et al. (2000) who propose varying location and scale parameters with respect to covariates, with the shape of the distribution fixed over the covariate set. Building on their work, we propose to combine these approaches in the sense that conditional income distributions are modelled parametrically as suggested by Biewen and Jenkins (2005), while allowing for a variation of the whole distribution (not just location and scale) with respect to covariates as specified by Donald et al. (2000).
As a specific research question, we address the differences in conditional income distributions between former East and West Germany. Differences between East and West Germany have received considerable attention in the economic literature (e.g. Biewen [2000], Fuchs-Schündeln et al. [2010], Kohn and Antonczyk [2011]) and also consistently played a major role in the domestic political debate. Instead of solely taking a macroeconomic perspective to look at income inequality in the East and West at a highly aggregated level, we build a microeconomic foundation to the analysis of income inequality. Thereby, we consider the effect of various covariates on the conditional, individual income distribution underlying the aggregate income distribution. It is our hypothesis that not only are there significant differences between East and West in the conditional mean income but also in the conditional income inequality aggrieving the economic divide more than two decades after the reunification.

As a conceptual framework for our analyses, we extend the Bayesian structured additive distributional regression models recently proposed in Klein et al. (2014b) for zero-inflated and overdispersed count data regression to general types of univariate distributions. In this class of regression models, all parameters of a potentially complex response distribution are related to additive regression predictors in the spirit of generalised additive models (GAMs). While the latter assume responses to follow a distribution from the exponential family and focus exclusively on relating the mean of a response variable to covariates (see for example Ruppert et al. [2003], Wood [2004, 2008], Fahrmeir et al. [2004, 2013]), distributional regression enables the consideration of basically any response distribution and allows to specify regression predictors for all parameters of this distribution. The main advantages of Bayesian structured additive distributional regression can be summarised as follows:

• It provides a broad and generic framework for regression models encompassing continuous, discrete and mixed discrete-continuous response distributions and therefore considerably expands the common exponential family framework.

• 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 (Dunn and Smyth [1996]), the deviance information criterion (Spiegelhalter et al. [2002]) and proper scoring rules (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 re-develop the complete inferential machinery.

Distributional regression is in fact closely related to generalised additive models for location, scale and shape (GAMLSS) as suggested by Rigby and Stasinopoulos (2005). We prefer the notion of distributional regression for our approach since in most cases, the parameters of the response distribution are in fact not directly related to location, scale and shape but are general parameters of the response distribution and only indirectly determine location, scale and shape. For example, in case of the Dagum distribution, there are three distributional parameters but none of them is directly related to a measure of location which is jointly determined by all three parameters.

In GAMLSS, inference is commonly based on penalised maximum likelihood estimation achieved in a backfitting loop over the additive components of all predictors in the model. In the implementation provided in the R add-on package gamlss, the score function and observed Fisher information required for the backfitting algorithm are determined by numerical differentiation (at least for distributions with complicated log-likelihood structures). In this paper, we consider 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 use 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. Moreover, the Bayesian approach provides valid credible intervals even without relying on asymptotic arguments, which is typically problematic in the frequentist approach at least with non-identity link functions.

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 shrinkage. A further advantage of the Bayesian formulation is that it automatically yields estimates for the smoothing parameters determining the impact of the prior and allows for a very modular inferential approach. The iteratively weighted least squares proposals yield numerically stable and adaptive proposal densities that do not require manual tuning.

Alternatives to distributional regression are provided by quantile and expectile regression where, in contrast to usual mean regression, 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 class of predictor specifications 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.

Many of the aspects discussed in the remainder of this paper (such as choice of a suitable response distribution and adequate predictor specifications, Bayesian inference, interpretation of estimation results) are of course relevant well beyond the scope of our application on the regional differences of income distributions. We therefore provide an analysis of the proportion of farm outputs achieved by cereals in the online supplement Section C to this paper as a second example on distributional regression.

The remainder of this paper is structured as follows: Section 2 provides a more detailed introduction to distributional regression and Bayesian inference along our case study on labour incomes. Model choice concerning both the type of the response distribution and the specification of the regression predictors is treated in Section 3.
Given the selected models, Section 4 provides empirical results on the regional disparities of conditional incomes in East and West Germany. Section 5 provides a summary and comments on directions of future research. As a guideline for scientists interested in the general application of distributional regression models, we summarise some general aspects of distributional regression in Appendix A. Additional material on the application, derivations of required quantities for the iteratively weighted least squares proposals and simulation studies are provided in the online supplement.

2 Bayesian Structured Additive Distributional Regression for Conditional Income Distributions

2.1 German Labour Income Data

For our empirical analyses, we use data from the German Socio-Economic Panel (Wagner et al., 2007) to study conditional income distributions in Germany. More specifically, we consider real gross annual personal labour income in Germany as defined in Bach et al. (2009) for the years 2001 to 2010. We deflate the incomes by the consumer price index (Statistisches Bundesamt, 2012) setting 2010 as our base year. Thus, all incomes are expressed in real valued 2010 Euros from here on.

Following the standard literature, we only look at the income of males in full time employment (see among others Dustmann et al., 2009; Card et al., 2013) in the age range 20–60. This yielded 7,216 individuals for whom we considered the income trajectories from the ten year-period. For every individual, we used every observation for which all required dependent and independent variables were available, yielding a total of \( n = 40,965 \) observations. Naturally, this implies that for some individuals we do not have full longitudinal coverage over the whole ten-year period.

As covariates, we consider educational level measured as a binary indicator for completed higher education (according to the UNESCO International Standard Classification of Education 1997 classification provided in the SOEP) in effect coding (educ), age in years (age), previous labour market experience in years (lmexp), the calendar time (\( t \)), information on the geographical district (Raumordnungsregionen) representing the area of residence (\( s \)) and a binary indicator in effect coding for districts belonging to the eastern part of Germany (east). A description of the data set is given in Table B1 of the online supplement.

In a mean regression model, a predictor for the log-income of a single observation \( i \)
collected at time point $t_i$ could now be specified as

$$
\eta_i = \beta_i + \text{educ}_{i} \beta_1 + f_1(\text{age}_i) + \text{educ}_{i} f_2(\text{age}_i) + f_3(\text{lcmexp}_i) + f_{\text{spat}}(s_i) + f_{\text{time}}(t_i)
$$

(1)

where $\beta_i$ represents the overall intercept, $\beta_1$ captures the effect of higher education, $f_1(\text{age})$ and $f_2(\text{age})$ are nonlinear effects of age capturing the age-dependency for individuals with low educational level ($f_1(\text{age}) - f_2(\text{age})$) and for individuals with high educational level ($f_1(\text{age}) + f_2(\text{age})$), $f_3(\text{lcmexp})$ is the nonlinear effect of previous labour market experience, $f_{\text{spat}}(s)$ is a spatial effect capturing heterogeneity at the level of the districts $s$, and $f_{\text{time}}(t)$ is an effect specific for the calendar year $t$. In a hierarchical model specification, the spatial effect can further be decomposed into

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

(2)

where $\gamma_1$ captures the difference between the eastern and western part of Germany and $g_{\text{str}}(s)$ and $g_{\text{unstr}}(s)$ represent spatially structured (smooth) or spatially unstructured (unsmooth) district-specific effects. Note that, in addition to the east-west indicator $\text{east}$, more district-specific information could be included if desired.

### 2.2 Distributional Regression

Conditional on all available covariate information summarised in the vector $\mathbf{v}_i$, we assume that the response variables $y_1, \ldots, y_n$ representing individual incomes are independent and that the density $p(y_i|\vartheta_{i1}, \ldots, \vartheta_{iK}) \equiv p_i$ of the conditional distribution of observation $y_i$ given $\mathbf{v}_i$ is indexed by the (in general covariate-dependent) parameters $\vartheta_{i1}, \ldots, \vartheta_{iK}$. Each parameter $\vartheta_{ik}$ is then related to a semiparametric predictor specified in analogy to (1). Similar as in generalised linear models, a suitable (one-to-one) response function is utilised to map the predictor to the parameter of interest, i.e. $\vartheta_{ik} = h_k(\eta_{ik})$ 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.

### 2.3 Potential Response Distributions

One of the great advantages of structured additive distributional regression is the wide range of distribution types which can be modelled. Since labour earnings, which we are considering here, are by definition positive, we will restrict ourselves to four selected non-negative distributions. For a more comprehensive list of distributions supported by the distributional regression framework, see Section A.1 in the appendix.
As already noted in the previous section, the standard conditional distribution type in econometric income analyses is the log-normal distribution. Next to its theoretical appeal from an economic perspective (see Arnold, 2008, p. 122), it also has the advantage that it makes the vast statistical inference machinery built around Gaussian regression available to researchers. However, Atkinson (1975) and others have noted that, at least for the aggregate income distribution, the log-normal distribution fit is problematic at times, especially for the upper tail of the distribution.

Partly as a consequence, various other distribution types have thus been suggested for the modelling of income distributions. Salem and Mount (1974) proposed the gamma distribution as a suitable alternative to the log-normal distribution. One of its advantages is that its estimation is possible within the framework of generalised linear models as the distribution belongs to the exponential family (as long as covariate effects are restricted to the mean).

The third distribution we consider also belongs to the exponential family (if the second parameter is assumed to be independent of covariates). The inverse Gaussian distribution has to our knowledge not been used in the context of modelling income distributions yet. But for other non-negative distributions with a similar economic rationale, like the distribution of claim sizes arising in car insurance (Heller et al., 2006; Klein et al., 2014a), it has shown to perform well due to its flexibility in modelling extreme right skewness. As it is conceivable that some conditional income distributions also portray such extreme skewness, we decided to also consider this distribution type.

The last distribution we consider is the Dagum distribution (Dagum, 1977) which belongs to the beta-type size distributions that have seen considerable attention in the literature on modelling (aggregate) income distributions (see Kleiber and Kotz, 2003). The Dagum distribution is derived from differential equations and proves to be much more adapt for modelling the upper end of the distribution than the log-normal. One of its appealing properties is that towards the upper end of the distribution its shape mirrors the one of the Pareto distribution which is generally assumed to provide a good approximation for the income distribution for the top percentiles of the (aggregate) income distribution (Piketty and Saez, 2007).

In summary, our choice of candidate distributions for modelling conditional incomes is at least partially motivated from the literature on aggregate income distributions. While it is certainly possible that the nature of conditional income distributions is structurally different from aggregate income distributions, we think that these four
distributions provide a good starting point for the analysis of income differences.

2.4 Structured Additive Predictors and Associated Priors

In the following, we will discuss suitable specifications and prior assumptions for the hierarchical predictor defined in (1) and (2). Note that we drop both the dependence on the parameter $\vartheta_k$ and observation index $i$ to simplify notation. A detailed discussion of structured additive predictors (in the context of mean regression) can be found in Fahrmeir et al. (2004) and Fahrmeir et al. (2013, Ch. 9). Hierarchical extensions are treated in detail in Lang et al. (2014).

**Linear Effects.** For all parametric, linear effects, we assume a flat, noninformative prior. This may be considered the limiting case of a multivariate Gaussian prior with high dispersion which can also be used to achieve regularisation in case of high-dimensional parameter vectors. In our analyses, we assume linear effects for the intercept, the educational indicator as well as for the east-west indicator.

**Continuous Covariates.** For the effects of age and previous work experience, assuming a linear effect is probably too restrictive. We therefore consider P(enalised)-splines (Eilers and Marx, 1996) as a flexible device for including potentially nonlinear effects $f(x)$ of a continuous covariate $x$. In a first step, $f(x)$ is approximated by a linear combination of $D$ B-spline basis functions $B_d(x)$ that are constructed from piecewise polynomials of a certain degree $l$ upon an equidistant grid of knots:

$$f(x) = \sum_{d=1}^{D} \beta_d B_d(x).$$

To avoid the requirement of choosing an optimal number of knots together with optimal knot positions, Eilers and Marx (1996) regularise the function estimate by augmenting a difference penalty to the fit criterion. In our Bayesian framework, the stochastic analogue is to assume a first or second order random walk

$$\beta_d = \beta_{d-1} + \varepsilon_d, \quad d = 2, \ldots, D$$

$$\beta_d = 2\beta_{d-1} - \beta_{d-2} + \varepsilon_d, \quad d = 3, \ldots, D$$

with Gaussian errors $\varepsilon_d \sim N(0, \tau^2)$ and noninformative priors for $\beta_1$ or $\beta_1$ and $\beta_2$ as a prior (Lang and Brezger, 2004). The joint prior of all basis coefficients $\beta = (\beta_1, \ldots, \beta_D)'$ can then be shown to be a (partially improper) multivariate Gaussian
distribution with zero mean and precision matrix $K = D'D$, where $D$ is a difference matrix of appropriate order. In our analyses, we will use twenty inner knots, a cubic spline basis and a second order random walk prior as the default specification for penalised splines.

In case of the age effect, we allow for separate functions for individuals with high and low level of education. This is achieved by the inclusion of the varying coefficient term \cite{Hastie1993} $f_2(\text{age})$ such that the age effect is given by $f_1(\text{age}) - f_2(\text{age})$ for individuals with low educational level and $f_1(\text{age}) + f_2(\text{age})$ for individuals with high educational level. In this case, a penalised spline can be assumed for function $f_2(\text{age})$ as well.

**Random Effects.** Penalised splines can in principle also be considered to represent the temporal effect $f_{\text{time}}(t)$ in (1). However, since in economic research temporal effects such as ours are generally considered by year-specific effects, we do not impose the smoothness assumption implied by penalised splines. We therefore consider a random effects specification where separate regression effects $\beta_t = f_{\text{time}}(t)$ are assumed for the distinct time points. An i.i.d. Gaussian prior with random effects variance $\tau^2$ is then placed on the coefficients $\beta = (\beta_1, \ldots, \beta_T)'$. Similarly, random effects priors can be used for any other grouping variable with levels $\{1, \ldots, G\}$ present in the data. Penalised splines can in principle also be considered to represent the temporal effect $f_{\text{time}}(t)$ in (1). However, since in economic research temporal effects such as ours are generally considered by year-specific effects, we do not impose the smoothness assumption implied by penalised splines. We therefore consider a random effects specification where separate regression effects $\beta_t = f_{\text{time}}(t)$ are assumed for the distinct time points. An i.i.d. Gaussian prior with random effects variance $\tau^2$ is then placed on the coefficients $\beta = (\beta_1, \ldots, \beta_T)'$. Similarly, random effects priors can be used for any other grouping variable with levels $\{1, \ldots, G\}$ present in the data.

Note that we have not included individual-specific random effects. The reason for this is that we are specifically interested in the unobserved heterogeneity among individuals with similar covariate sets which finds expression in income inequality among them. In some sense our analysis is thus systematically different from standard regression techniques which pursue to eradicate the stochastic component or at least reduce it to a minimum. The inclusion of individual specific-effects goes a long way towards seemingly achieving this aim, as the share of the variance left to the error term is drastically reduced. However, the inferential gain obtained thereby could
be simply speaking be expressed as follows: Including individual-specific effects, we have found that incomes are largely different because individuals are different. While there are some analyses where such eradication of variance is useful, it sheds little insights on the nature of inequality at the disaggregated level since we are unable to disentangle the differences between individuals in a meaningful way.

Spatial Effects. For the spatial effect \( f_{\text{spat}}(s) \) defined upon the discrete, regional spatial variable \( s \in \{1, \ldots, S\} \) where \( 1, \ldots, S \) denote the different regions in the data set, we assume a hierarchical predictor specification following [Lang et al., 2014]. In fact, equation (2) merely defines a second structured additive predictor where now the distinct spatial regions define the unit of observation. As a consequence, any type of regression effect that is specific for the region can be included on this level. In our case, the east-west indicator is one such example that is assigned a parametric effect with a flat prior.

In addition, we consider the spatially structured and spatially unstructured effects \( g_{\text{str}}(s) \) and \( g_{\text{unstr}}(s) \), respectively. In both cases, separate regression effects \( \beta_{\text{str},s} = g_{\text{str}}(s) \) and \( \beta_{\text{unstr},s} = g_{\text{unstr}}(s) \) are assumed for each of the regions but the effects differ in terms of their prior assumptions. For the structured spatial effect, we assume spatial correlations defined implicitly by assuming a Gaussian Markov random field prior [Rue and Held, 2005] for a suitable neighbourhood structure derived from the spatial orientation of the data. The most common case would be to treat to regions as neighbours of they share a common boundary. If \( \partial_s \) denotes the set of all neighbours of region \( s \), the Markov random field prior then assumes

\[
\beta_{\text{str},s} | \beta_{\text{str},r}, r \neq s, \tau^2 \sim N \left( \sum_{r \in \partial_s} \frac{1}{N_s} \beta_{\text{str},r}, \frac{\tau^2}{N_s} \right),
\]

where \( N_s \) is the number of neighbours of region \( s \). In consequence, the conditional mean of \( \beta_{\text{str},s} \) given all other coefficients is the average of the neighbouring regions. It can be shown that the conditional normal distributions specified in (3) correspond to a multivariate, partially improper Gaussian distribution with zero mean and precision matrix given by the adjacency matrix induced by the assumed neighbourhood structure.

For the unstructured spatial effect, we consider an i.i.d. Gaussian prior, i.e. we assume a random effects prior specification. The rationale for considering both a structured and unstructured part of the spatial effect is that they are surrogates for unobserved spatial heterogeneity which may either be spatially structured (i.e. spatially smooth)...
or unstructured.

**Generic Representation.** All different effects comprised in the predictor \( \eta \) (and a considerable number of extensions) can be cast into a generic framework utilising (possibly non-standard) basis functions as detailed in Appendix A.2. In particular, each effect can be represented by a set of basis coefficients and desirable properties of the effects can be enforced by assuming informative multivariate Gaussian priors. This allows us to rewrite any of the predictors related to a parameter \( \vartheta_k \) as

\[
\eta = \beta_0 \mathbb{1} + Z_1 \beta_1 + \ldots + Z_J \beta_J
\]  

(4)

where again we suppressed the notational dependence on the parameter \( \vartheta_k \) and \( J_k \) denotes the number of different effects contained in the predictor for \( \vartheta_k \). For each of the parameter vectors \( \beta_j \) we can then either assume a hierarchical specification, where \( \beta_j \) is related to another structured additive predictor (as in case of the spatial effect in our example) or we directly assume the multivariate normal prior

\[
p(\beta_j | \tau_j^2) \propto \left( \frac{1}{\tau_j^2} \right)^{\frac{rk(K_j)}{2}} \exp \left( -\frac{1}{2\tau_j^2} \beta_j' K_j \beta_j \right)
\]  

(5)

where \( K_j \) is the (potentially rank-deficient) precision matrix and \( \tau_j^2 \) is the prior smoothing variance. The latter is assigned an inverse gamma hyperprior \( \tau_j^2 \sim IG(a_j, b_j) \) (with \( a_j = b_j = 0.001 \) as a default option) in order to obtain a data-driven amount of smoothness.

### 2.5 Bayesian Inference

To perform Bayesian inference, we consider Markov chain Monte Carlo (MCMC) simulation techniques. Since the complex likelihood structures of non-standard distributions utilised in distributional regression usually results in full conditionals for the unknown regression coefficients that are not analytically accessible, we develop suitable proposal densities based on iteratively weighted least squares (IWLS) approximations to the full conditionals. Note that updating the smoothing variances \( \tau_j^2 \) is always realised by a Gibbs update since the full conditionals follow inverse gamma distributions

\[
\tau_j^2 \sim IG(a'_j, b'_j), \quad a'_j = \frac{rk(K_j)}{2} + a_j, \quad b'_j = \frac{1}{2} \beta'_j K_j \beta_j + b_j
\]  

(6)

with updated parameters \( a'_j, b'_j \).
Approximations to the Full Conditionals. For a typical parameter block $\beta_j$ in the generic predictor (4), the log-full conditional log($p(\beta_j|\cdot)$) is (up to additive constants) given by $l(\eta) - \frac{1}{2}\eta_j\beta_j^T K_j \beta_j$, where $l(\eta)$ denotes the log-likelihood part depending on $\eta$. 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

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

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

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

where $v^{[t]} = \partial l^{[t]}/\partial \eta$ is the score vector and $W^{[t]}$ 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 $\beta_j$ is then constructed from the resulting working model for $z$ as $\beta_j \sim N(\mu_j, P_j^{-1})$ with expectation and precision matrix

$$\mu_j = P_j^{-1} Z_j W (z - \eta_{-j}) \quad P_j = Z_j W Z_j + \frac{1}{\tau_j^2} K_j$$

where $\eta_{-j} = \eta - Z_j \beta_j$ is the predictor without the $j$-th component.

The algorithm resulting from the IWLS-approximation to the full conditionals is summarised in Appendix A.3 together with further implementation-related and theoretical details.

### 2.6 Software

Our Bayesian approach to distributional regression is implemented in the free, open source software BayesX (Belitz et al., 2012). As described in Lang et al. (2014), the implementation makes use of efficient storing mechanisms for large data sets and sparse matrix algorithms for sampling from multivariate Gaussian distributions. An R interface to BayesX is provided in the R add-on package BayesR.
2.7 Empirical Evaluation

We compared the empirical performance of the proposed Bayesian approach to the frequentist GAMLSS framework in two simulation scenarios and also investigated the performance of the deviance information criterion (DIC, Spiegelhalter et al., 2002) for choosing between competing models. The studies and their outcomes are documented in more detail in Section E of the electronic supplement. A summary on the ability of the DIC for model choice is given in Section 3 and for the comparison with the frequentist approach (denoted as ML) in the following:

(1) Comparison with ML in additive models. In pure additive models, i.e. nonlinear effects are simulated on the predictor levels of all parameters of response distributions considered in our application on labour incomes, the point estimates and corresponding posterior means, as well as their mean squared errors (MSEs) are very similar for both approaches. However, coverage rates based on asymptotic maximum likelihood theory for ML are much too narrow in several distribution parameters. In particular, for the Dagum distribution, rates for all three parameters are far from the claimed coverage level while the credible intervals of the Bayesian approach are still reliable (albeit being usually slightly too conservative), compare e.g. Figure E21 of the supplement.

(2) Comparison with ML in geoadditive models. 10% of the estimation runs of ML failed before convergence. A reason might be that spatial effects require the additional add-on package gamlss.add. MSEs of the spatial effect (based on a Markov random field) are slightly smaller for the Bayesian approach compared to ML. While the MSEs of the other effects do not deteriorate for our proposed method, we furthermore observe partly increasing MSEs for ML. Estimates of single replications in Figure E24 indicate furthermore that the proposed approach can be seen as an appropriate competitor of ML.

3 Model Choice for Conditional Income Distributions

In any application of distributional regression, one faces important model choice decisions: choosing the most appropriate out of a set of potential response distributions and selecting adequate predictor specifications for each parameter of these distributions. For our application on conditional income distributions, we consider the inverse
Gaussian (IG), log-normal (LN), gamma (GA) and Dagum (DA) distribution as candidate distributions. A general predictor that could now be utilised for any of the parameters of these distributions was already introduced in equations (1) and (2). Instead of performing a complete stepwise model selection for each distribution, we study the following model specifications:

(M1) All distributional parameters are related to a predictor of type (1). For the spatial effect, we only include the unstructured effect since it turned out in exploratory analyses that the smooth component has only negligible impact on incomes.

(M2) Instead of modelling all parameters in terms of covariates, the model structure of M1 is only applied to the parameters \( \mu \) in case of LN, IG and GA, \( b \) in case of DA. The parameters \( a, c, \sigma, \sigma^2 \) are considered to be equal across all individuals and therefore do not depend on covariates. This corresponds to the usual GAM specification with a focus on conditional means.

(M3) All parameters are modelled in analogy to M1 except that the random effect for calendar time and the complete spatial effect (including the east-west indicator) are not included in the parameters \( a, c, \sigma, \sigma^2 \).

In total, we therefore end up with 12 models to compare. In the following, we will discuss different options for actually conducting this comparison and will also comment on their wider applicability in the context of model choice for distributional regression.

3.1 Deviance Information Criterion

The deviance information criterion (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 \( \theta^{[1]}, \ldots, \theta^{[T]} \) is a MCMC sample from the posterior for the complete parameter vector \( \theta \), the DIC is given by

\[
\text{DIC} = \bar{D}(\theta) + pd = 2\bar{D}(\theta) - D(\bar{\theta})
\]

where \( D(\theta) = -2\log(f(y|\theta)) \) is the model deviance and \( pd = \bar{D}(\theta) - D(\bar{\theta}) \) is an effective parameter count.

The DIC can be used to discriminate between types of response distributions as well as different predictor specifications for a fixed distribution. The latter can also be
implemented in a stepwise model choice strategy. However, since the DIC is sample-based, small differences of DIC values for competing models may induce a region of indecisiveness. If in such a situation sparser models are desired, the DIC based selection of covariate effects can be assisted by only including significant effects.

For count data distributional regression models, the performance of the DIC was positively evaluated in [Klein et al. (2014b)] where several misspecified models have been compared to the true model in terms of the DIC. For distributions and models considered in our applications, we conducted additional simulations which are documented in more detail in the supplement Section E.3. The basic outcome is that the DIC is appropriate to discriminate between competing response distributions although differences can be rather small depending on what distributions are compared.

Concerning the identification of relevant covariates, we focused on spatial effects and found that the DIC usually is in clear favour of the true model if a relevant effect is omitted. In the reverse situation, i.e. irrelevant information is included, the DICs of the true models are only slightly smaller but then the irrelevant covariate mainly yields an insignificant effect and would thus be excluded under the aim of a sparser model.

The DIC values for the 12 income regression models under consideration are documented in Table 1 and indicate a clear preference for the model DA_M1. In general, it is noticeable that the DIC favours our flexible model specifications (M1) compared to the simplified versions (M2, M3).

### 3.2 Quantile Residuals

For continuous random variables, it is a well known result that the cumulative distribution function evaluated at the random variable yields a random variable uniformly distributed on [0,1]. As a consequence, quantile residuals defined as $\hat{r}_i = \Phi^{-1}(u_i)$ with the inverse cumulative distribution function of a standard normal distribution $\Phi^{-1}$ and $F(\cdot|\hat{\theta}_i)$ denoting the cumulative distribution function with estimated parameters plugged in, should at least approximately be standard normally distributed if the correct model has been specified [Dunn and Smyth (1996)]. In practice, the residuals can be assessed graphically in terms of quantile-quantile-plots: the closer the residuals are 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.

Quantile residuals are closely related to the probability integral transform which considers $u_i$ without applying the inverse standard normal cumulative distribution function. If the estimated model is a good approximation to the true data generating process, the $u_i$ will then approximately follow a uniform distribution on $[0, 1]$. As a graphical device, histograms of the $u_i$ are then typically considered. We prefer quantile residuals in the quantile-quantile plot representation since they avoid the requirement to define breakpoints for the construction of the histogram.

Quantile residual plots for the models of type M1 are shown in Figure 1 (for model types M2 / M3, similar figures can be found in Figure B1 of the supplement). 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 $-2$ and $2$ 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. Residuals of the gamma model are also reasonable in the range between $-2$ and $2$ (similar to the Dagum distribution) but deviate more strongly from the diagonal for extreme residuals.
3.3 Proper Scoring Rules

Gneiting and Raftery (2007) propose proper scoring rules as summary measures for the evaluation of probabilistic forecasts, i.e. to evaluate the predictive ability of a statistical model. We consider three common scores, namely the Brier or quadratic score (QS), the logarithmic score (LS) and the spherical score (SPS). For continuous response distributions with density \( p_r(y) = p(y|\vartheta_1, \ldots, \vartheta_K) \) and a given new realisation \( y_{\text{new}} \), these are defined as

\[
\begin{align*}
\text{LS}(p_r, y_{\text{new}}) &= \log(p_r(y_{\text{new}})) \\
\text{SPS}(p_r, y_{\text{new}}) &= \frac{p_r(y_{\text{new}})}{\|p_r\|_2^2} = \frac{p_r(y_{\text{new}})}{(\int |p_r(y)|^2dy)^{1/2}} \\
\text{QS}(p_r, y_{\text{new}}) &= 2p_r(y_{\text{new}}) - \|p_r\|_2^2 = p_r(y_{\text{new}}) - \int |p_r(y)|^2dy.
\end{align*}
\]

Appropriate definitions for discrete as well as mixed discrete continuous responses are provided in the appendix in Section A.4. While all three scoring rules are strictly proper as shown by Gneiting and Raftery (2007), 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. As a fourth alternative, we consider the continuous ranked probability score (CRPS)

\[
\text{CRPS}(p_r, y_{\text{new}}) = -\int_{-\infty}^{\infty} (F_r(y) - 1_{\{y \geq y_{\text{new}}\}})^2 dy
\]

where \( F_r \) is the cumulative distribution function corresponding to the density \( p_r \) (Gneiting and Ranjan, 2011). Laio and Tamea (2007) showed that the CRPS score can also be written as

\[
\text{CRPS}(p_r, y_{\text{new}}) = -2 \int_0^1 \left( 1_{\{y_{\text{new}} \leq F_r^{-1}(\alpha)\}} - \alpha \right) \left( F_r^{-1}(\alpha) - y_{\text{new}} \right) d\alpha
\]

where \( F_r^{-1}(\alpha) \) is the quantile function of \( p_r \) evaluated at the quantile level \( \alpha \in (0, 1) \).

This formulation allows not only to look at the sum of all score contributions (i.e. the whole integral) but also to perform a quantile decomposition and to plot the mean quantile scores versus \( \alpha \) in order to compare fits of specific quantiles (Gneiting and Ranjan, 2011). This decomposition is especially helpful in situations where the quantile score can be interpreted as an economically relevant loss function (Gneiting, 2011).

In practice, we obtain the probabilistic forecasts in terms of predictive distributions \( p_r \) for observations \( y_r \) by cross validation, i.e. the data set is divided into subsets of
Table 1: Comparison of DIC values (calculated based on the complete data set) and average scores obtained from ten-fold cross validation.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>DIC</th>
<th>Quadratic Score</th>
<th>Logarithmic Score</th>
<th>Spherical Score</th>
<th>CRPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>LN_M1</td>
<td>179,090</td>
<td>0.1304</td>
<td>-2.4363</td>
<td>0.3621</td>
<td>-2.1581</td>
</tr>
<tr>
<td>LN_M2</td>
<td>180,533</td>
<td>0.1257</td>
<td>-2.4596</td>
<td>0.3569</td>
<td>-2.1408</td>
</tr>
<tr>
<td>LN_M3</td>
<td>179,451</td>
<td>0.1299</td>
<td>-2.4353</td>
<td>0.3622</td>
<td>-2.1630</td>
</tr>
<tr>
<td>IG_M1</td>
<td>184,614</td>
<td>0.1464</td>
<td>-2.2741</td>
<td>0.3777</td>
<td>-1.6195</td>
</tr>
<tr>
<td>IG_M2</td>
<td>189,702</td>
<td>0.1382</td>
<td>-2.3141</td>
<td>0.3663</td>
<td>-1.6771</td>
</tr>
<tr>
<td>IG_M3</td>
<td>186,494</td>
<td>0.1442</td>
<td>-2.2824</td>
<td>0.3743</td>
<td>-1.6419</td>
</tr>
<tr>
<td>GA_M1</td>
<td>177,453</td>
<td>0.1609</td>
<td>-2.1715</td>
<td>0.3963</td>
<td>-1.2735</td>
</tr>
<tr>
<td>GA_M2</td>
<td>178,736</td>
<td>0.1563</td>
<td>-2.1812</td>
<td>0.3918</td>
<td>-1.2791</td>
</tr>
<tr>
<td>GA_M3</td>
<td>177,971</td>
<td>0.1597</td>
<td>-2.1742</td>
<td>0.3949</td>
<td>-1.2767</td>
</tr>
<tr>
<td>DA_M1</td>
<td>172,421</td>
<td><strong>0.1684</strong></td>
<td><strong>-2.1034</strong></td>
<td><strong>0.4053</strong></td>
<td><strong>-1.2662</strong></td>
</tr>
<tr>
<td>DA_M2</td>
<td>173,791</td>
<td>0.1644</td>
<td>-2.1199</td>
<td>0.4015</td>
<td>-1.2737</td>
</tr>
<tr>
<td>DA_M3</td>
<td>172,790</td>
<td>0.1674</td>
<td>-2.1079</td>
<td>0.4041</td>
<td>-1.2702</td>
</tr>
</tbody>
</table>

approximately equal size and predictions for one of the subsets are obtained from estimates based on all the remaining subsets. Let \(y_1, \ldots, y_R\) be data in a hold-out sample and \(p_r\) the predictive distributions with predicted parameter vectors \(\tilde{\vartheta}_r = (\tilde{\vartheta}_{r1}, \ldots, \tilde{\vartheta}_{rK})'\). Competing forecasts are then ranked by averaged scores

\[
S = \frac{1}{R} \sum_{r=1}^{R} S(p_r, y_r)
\]

such that higher scores deliver better probabilistic forecasts when comparing different models.

In our application, we conducted ten-fold cross validation, i.e. one tenth of the data set is left out and distribution parameters for this part are predicted with estimates obtained from the remaining nine parts. Observations are assigned randomly to the different folds. The scores discussed above are documented in Table I where the values are averages of the ten folds (and scores within the folds are themselves averages over the individual score contributions). In line with the DIC and the residual plots, the scores of the DA_M1 model are the highest and thus deliver the best forecast among the 12 models under consideration. Also similar to the DIC, models of type M2 (the simplest versions) are lower compared to the ones of type M3 and they themselves are inferior compared to the most flexible models of type M1.

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 decompo-
Figure 2: Quantile decomposition of CRPS in the full models DA_M1, LN_M1, IG_M1, GA_M1.

sition of the CRPS over quantile levels (Figure 2) and a decomposition of the scores over the cross validation folds (Figure B2 in the supplement). The quantile level decomposition of the CRPS again indicates a comparable performance of the Dagum distribution and the gamma distribution as compared to the inverse Gaussian distribution which performs somewhat worse and the log-normal distribution which shows a considerably deteriorated behaviour. This ordering holds true over the complete range of quantiles. The fact that the log-normal distribution fails to provide a competing predictive ability is most probably related to the strong impact of the extreme observations. These are hard to capture by the log-normal distribution in general. However, since extreme observations are typically also influential observations, they seem to impact estimates in the log-normal model to such an extent that even predictions for the central part of the distribution are affected negatively. The plot with separate scores per fold (Figure B2) shows a homogeneous average behaviour of the scores across the folds. This provides a basic check for structural differences in the folds which may arise in disadvantageous random assignments.
4 Regional Disparities of the Distribution of Earnings in Germany

As discussed in the introduction, we address the differences in conditional income distributions between former East and West Germany in the first decade of the new millennium. More specifically we focus on the hypothesis that next to significant differences in the conditional mean income between East and West Germany, we also have significant differences in the inequality of the conditional income distribution as measured by the Gini coefficient (Silber, 1999). Thus, based on our model choice of the previous Section 3 we illustrate the estimation results of our data set with model (M1) for the Dagum distribution by contrasting four conditional income distributions with different covariate sets.

In their famous paper, DiNardo, Fortin and Lemieux stress the need to look at differences between the whole conditional income distributions rather than just the conditional mean income, or certain indices (see DiNardo et al., 1996, p. 1026). Using our proposed estimation procedure this is made feasible. Figure 3 displays an exemplary contrast of four conditional income distributions in a ceteris paribus type of analysis within a distributional regression framework. The four distributions have
all but two covariates fixed at their average value. For age (42 years) and labour market experience (19 years) we use the arithmetic mean from the observations in our sample, while we fixed the random effects at their prior expectation, i.e. at zero. Keeping these covariates fixed, we can observe the nature of the change if the regional variable is changed from East to West.

For both educational levels, we can observe from the data that there is a noticeable difference not only in the mean value of the distributions but also for other aspects, like spread, skewness, etc. A simple analysis of sole means thus falls short to portray a comprehensive picture of the differences in income between East and West.

There are various additional aspects of the distribution which can and ought to be considered. In principle it is possible to obtain any distributional measure from the conditional distribution as long as it is defined for the given distribution type and the corresponding parameter set. Here, we will only point to three alternative measures which may be considered next to the mean and the Gini coefficient which we focus on below:

• Using distributional regression it is easily possible to obtain estimates for certain quantiles, like the median, which is an alternative to the mean as a location measure. From that one could easily calculate interquantile ranges as a measure of inequality. Naturally these quantiles can be estimated in a more direct manner using quantile regression. However, in some circumstances distributional regression may be advantageous. See Section B.3 of the online supplement where we contrast distributional regression against quantile regression.

• Next to measures of inequality like the Gini coefficient or the Theil index, which are easily computable, it is also straightforward to calculate measures of polarisation, which have recently received considerable attention in the literature (see among others [Wolfson, 1994; Duclos et al., 2004]). Following [Gradín, 2000] it would be possible to calculate the polarisation between two groups as defined by sets of covariates.

• It is also possible to assess density differences at different income levels or probability mass differences for different income ranges. For example one could consider the probability mass above a certain income, e.g. 48,000€, which according to John Maynard Keynes would suffice to turn one’s mind away from pecuniary worries ([Skidelsky, 2010]). Consequently it could be highlighted that not only the conditional mean income for the average man without higher education is lower in the...
East but also that the probability mass of gross earnings below that threshold is much lower. Such an analysis may be of particular interest for research questions on poverty and vulnerability (see Pudney 1999).

An in-depth discussion on the requirements for interpreting conditional income distributions in an economic context is not the focus of this paper. As mentioned above, we thus restrain ourselves to two exemplary measures which are frequently used in the economic literature. Namely we consider the mean and the Gini coefficient of the estimated conditional income distributions. While the mean provides important information on the location of the income distribution, the Gini coefficient is the most frequently used scalar measure on income inequality (Silber 1999). We thus look at two important aspects of the conditional income distribution and observe how they change over the covariate space.

4.1 The Spatial Effect on Conditional Mean Incomes

Obtaining the conditional means from the conditional income distributions of $y_i$ is straightforward in case of the Dagum distribution. For $a_i > 1$ the expectation is given by

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

Figure 4 displays the posterior mean estimates for the expected incomes by region and education. Following Fox (2003) and as mentioned before, we employ effect displays such that we consider the differences between each region, taking into account the regional random effect and the east-west effect while keeping the other covariates fixed at their mean. For results on additional covariate sets see Section B.1 and B.2 of the online supplement.

Little surprisingly, there is a clearly visible divide between East and West Germany, as expected incomes are much higher in the former Federal Republic of Germany for both education levels and at the expectation of the other covariates. Abstracting from the variations at the district level, we get an expected income of 33,600€ if the average man is living in the East and has no higher education. With higher education the income increases to 55,200€. The corresponding values if a person with the same attributes is living in the West are 48,100€ and 78,300€. The difference between East and West is thus 14,500€ and 23,100€ with and without higher education respectively. The corresponding 95% credible intervals are [19,000€;27,400€]
Figure 4: SOEP data. Estimated posterior means for expected incomes for 42-year-old males with 19 years of working experience. Left: males without higher education. Right: males with higher education.

and \([17,100\,\text{€}; 27,000\,\text{€}]\) respectively. To put these magnitudes into perspective - the standard deviations of the regions’ estimated expected income within East and West are 3,100\,\text{€} and 5,200\,\text{€} for those without higher education while they are 4,700\,\text{€} and 8,900\,\text{€} for those with higher education. Thus the results show that evaluated at the mean of the other covariates, the conditional mean income is not only significantly lower in East compared to the West for both education levels but also that the magnitude of the differences surpasses the standard deviations among the regions within East and West.

4.2 The Spatial Effect on the Conditional Income Inequality

The Gini coefficient is an inequality measure based on the Lorenz curve (see Sarabia, 2008), which can vary between the value 0 (everybody has the same) and 1 (one person has everything). Note that the Gini coefficient is scale invariant such that in standard mean regression on logarithmic incomes it would be postulated as constant across the covariate space. Analogue to the conditional mean income, the Gini coefficient of the conditional income distribution \(G_i\) can easily be obtained from the parameter
estimates of the Dagum distribution.

\[ G_i = \frac{\Gamma(c_i)\Gamma(2c_i + 1/a_i)}{\Gamma(2c_i)\Gamma(c_i + 1/a_i)} - 1. \]

Figure 5 portrays the posterior mean estimates for the Gini coefficients for each region. As we can see the differences are not as clear cut as for the conditional mean incomes. Nonetheless, the pattern emerging indicates that income inequality among 42-year-old males with 19 years of experience is higher in the East for both education levels. Indeed if we only consider the impact of the binary east-west variable on the Gini coefficient as we did for the mean, we get the following difference between East and West. From the resultant conditional income distributions, we obtain a difference of the posterior means of 0.039 and 0.036 for those without higher education and those with higher education respectively. The corresponding 95% credible intervals are [0.015,0.067] and [0.013,0.063] respectively. Thus we have a significantly larger income inequality for 42-year-old males with 19 years of experience, as measured by the Gini-coefficient, in the East than in the West. Putting these differences into perspective again, the standard deviations of the Gini coefficients of the regions’ conditional income distributions within East and West are 0.030 and 0.031 for those
without higher education, and 0.032 and 0.031 for those with higher education. Thus the differences between East and West are not only significant, they also surpass the standard deviations within East and West.

4.3 Economic Consequences Drawn From the Regression of Conditional Income Distributions

Our findings show that keeping other variables fixed at their average level, there are significant differences in income inequalities within East and West Germany. Duclos et al. (2004) have noted the importance of within-groups inequality for levels of alienation and identification within society. The higher income inequality in the East would thereby induce a weakened in-group identity. Lack of in-group identity in turn are likely cause feelings of isolation and mistrust (see Misztal, 2013) and consequently lead a deterioration of well-being which is beyond that captured by solely considering average income, or even distribution-adjusted well-being measures (see Klasen, 2008). Indeed, one may argue that it is the change of (income) inequality at the disaggregated level rather than the growing inequality at the aggregated level which is doing greater harm to the social fabric of the German society.

While a profound answer to this hypothesis must be left to further research and is beyond the scope of this paper, our application already shows that structured additive distributional regression offers a statistical methodology to the analysis of income inequality which goes beyond the analysis at a highly aggregated level and thus allows to start the assessment of this important issue at a microeconomic level.

5 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.

Concerning the specific application of distributional regression to conditional income distributions, we have shown that there are significant differences between men with similar age, work experience and education levels between East and West which go beyond the mean income. Taking the Gini coefficient as an indicator for inequality, we have shown that income inequality among these men is systematically greater in the East than it is in the West, further deepening differences in well-being.

While this study highlights the scope of the new methodology to an application of income analysis and beyond, much work remains to be done on the application of distributional regression techniques such as ours.

Despite the practical solutions outlined in this paper, model choice and variable selection remain relatively tedious and more automatic procedures would be highly desirable, especially if structured additive distributional regression is to be used by applied researchers. 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.

For the context of economic applications, it should be noted that analogously to generalised linear models the additive impact of explanatory variables on the economic measure of interest, like the Gini coefficient, is generally not attained. Consequently, the size and possibly also the direction of the estimated spatial effect, may well be very different for different points in the covariate space. While, it is straightforward to calculate these differences with the corresponding credible intervals for any desired combination of the other covariates to give a more comprehensive assessment of the differences in inequality, more work needs to be done to facilitate the interpretation of the results.

In addition, in-depth-testing is required to find adequate parametric forms for conditional income distributions, as the application of structured additive distributional regression crucially rests on the assumption that the parametric distribution befits
the data. While for the case of full-time working men the Dagum distribution indeed seems to provide a decent fit, further work must be done to allow for an analysis with a less restricted covariate space and thus a more comprehensive analysis of income distributions in Germany and beyond. In addition, other modelling assumptions should be qualified empirically in a diverse number of settings.

Yet despite these issues this paper demonstrates that structured additive distributional regression offers a possible solution to the “daunting estimation challenge” of estimating entire conditional distributions [Fortin et al. 2011, p. 56] and thus offers additional scope for applied statistical analyses in the question of income inequality and beyond.

A Bayesian Structured Additive Distributional Regression

This appendix summarises some general features of Bayesian structured additive distributional regression that should aid in the application of this model class beyond our analysis of regional differences in labour income in Germany. We therefore summarise a number of potential response distributions, details concerning the structured additive predictor specification, Bayesian inference and model choice.

A.1 Response Distributions and Parameterisations

In the following, we discuss some distributions that may be relevant in various applications of distributional regression. An overview is provided in Table 2 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 2.5. Note that $p$ is considered a general density, i.e. we use the same notation for continuous responses, discrete responses and also mixed discrete-continuous responses.

Real-Valued Responses. For continuous responses with an approximately symmetric distribution, the normal distribution is the most common choice. In distributional regression, 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 and therefore heteroskedasticity can be
<table>
<thead>
<tr>
<th>1. Continuous distributions on ( \mathbb{R} )</th>
<th>Density</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>( p(y</td>
<td>\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) )</td>
</tr>
<tr>
<td>( t )</td>
<td>( p(y</td>
<td>\mu, \sigma^2, \nu, \tau) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\sigma^2}} \left(1 + \frac{(y-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} )</td>
</tr>
<tr>
<td>2. Continuous distributions on ( \mathbb{R}^+ )</td>
<td>Density</td>
<td>Parameters</td>
</tr>
<tr>
<td>Log-normal</td>
<td>( p(y</td>
<td>\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(\log(y)-\mu\right)^2}{2\sigma^2}\right) )</td>
</tr>
<tr>
<td>Inverse Gaussian</td>
<td>( p(y</td>
<td>\mu, \sigma^2) = \frac{2}{\sqrt{2\pi\sigma^2}y^{3/2}} \exp\left(-\frac{(\log(y)-\mu)^2}{2\sigma^2}\right) )</td>
</tr>
<tr>
<td>Gamma</td>
<td>( p(y</td>
<td>\mu, \sigma^2) = \left(\frac{\mu}{\sigma}\right)^{-\frac{\sigma^2}{2}} y^{-\frac{\sigma^2}{2}-1} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) )</td>
</tr>
<tr>
<td>Weibull</td>
<td>( p(y</td>
<td>\lambda, \alpha) = \alpha\lambda^{-\alpha} \exp\left(-\frac{y}{\lambda}\right) \lambda^{-\alpha} )</td>
</tr>
<tr>
<td>Pareto</td>
<td>( p(y</td>
<td>b, c) = cb(y + c)^{-c-1} )</td>
</tr>
<tr>
<td>Generalized gamma</td>
<td>( p(y</td>
<td>\mu, \sigma, \tau) = \left(\frac{\mu}{y}\right)^{-\frac{\sigma^2}{2}} y^{-\tau-1} \exp\left(-\left(\frac{\mu}{y}\right)^2\right) )</td>
</tr>
<tr>
<td>Dagum</td>
<td>( p(y</td>
<td>a, b, c) = \frac{b^{\alpha^{-1}} \exp\left(-\frac{y}{ab}\right)}{\Gamma(1+\alpha) \Gamma(1+\beta) \Gamma(1+\gamma) \Gamma(1+\delta)} )</td>
</tr>
<tr>
<td>3. Discrete distributions</td>
<td>Density</td>
<td>Parameters</td>
</tr>
<tr>
<td>Poisson</td>
<td>( g_1(y</td>
<td>\lambda) = \lambda^y \exp(-\lambda) )</td>
</tr>
<tr>
<td>Negative binomial</td>
<td>( g_2(y</td>
<td>\mu, \delta) = \frac{\Gamma(y+\delta)}{\Gamma(\delta) \Gamma(y+\mu)} \left(\frac{\delta}{\pi \mu}\right)^{\delta} \left(\frac{\mu}{y}\right)^y )</td>
</tr>
<tr>
<td>Zero-inflated Poisson</td>
<td>( p(y</td>
<td>\pi, \mu, \delta) = \pi \mathbb{I}_{(y)}(y) + (1-\pi)g_1 )</td>
</tr>
<tr>
<td>Zero-inflated negative binomial</td>
<td>( p(y</td>
<td>\pi, \mu, \delta) = \pi \mathbb{I}_{(y)}(y) + (1-\pi)g_2 )</td>
</tr>
</tbody>
</table>

### 4. Mixed discrete-continuous distributions

<table>
<thead>
<tr>
<th>Zero-adjusted</th>
<th>Density</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p(y</td>
<td>\pi, g(y)) = \begin{cases} 1-\pi &amp; y = 0 \ \pi g(y) &amp; y &gt; 0 \end{cases} )</td>
<td>( g(y) ) a distribution from 2.</td>
</tr>
</tbody>
</table>

### 5. Distributions with compact support

<table>
<thead>
<tr>
<th>Beta</th>
<th>Density</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p(y</td>
<td>\mu, \sigma^2) = \frac{\nu^{\nu/(\nu+\sigma^2)}}{\Gamma(\nu+\mu, \sigma^2)} )</td>
<td>( \mu = \frac{\nu}{\nu + \sigma^2}, \sigma^2 \in (0, 1) )</td>
</tr>
<tr>
<td>Zero-One-inflated Beta</td>
<td>( p(y</td>
<td>\mu, \sigma^2, \nu, \tau) = \begin{cases} \frac{\nu^{(1+y)^\tau}}{(1+y)^\tau \Gamma(\nu, \sigma^2)} &amp; y = 0 \ \frac{(1+y)^\tau}{\nu^{(1+y)^\tau} \Gamma(\nu, \sigma^2)} &amp; y \in (0, 1) \end{cases} )</td>
</tr>
<tr>
<td>Zero-inflated Beta</td>
<td>( p(y</td>
<td>\mu, \sigma^2, \nu) = \begin{cases} \frac{\nu^{(1+y)^\nu}}{(1+y)^\nu \Gamma(\nu, \sigma^2)} &amp; y = 0 \ \frac{(1-y)^\nu}{\nu^{(1+y)^\nu} \Gamma(\nu, \sigma^2)} &amp; y \in (0, 1) \end{cases} )</td>
</tr>
<tr>
<td>One-inflated Beta</td>
<td>( p(y</td>
<td>\mu, \sigma^2, \tau) = \begin{cases} \frac{(1+y)^\tau}{\Gamma(\nu+1+y, \sigma^2)} &amp; y = 0 \ \frac{1-\nu^{(1+y)^\tau}}{(1+y)^\tau \Gamma(\nu, \sigma^2)} &amp; y \in (0, 1) \end{cases} )</td>
</tr>
</tbody>
</table>

Table 2: List of important response distributions in distributional regression.
accounted for. An alternative with heavier tails is the t distribution which is often considered in applications as a robust alternative to the normal which is less affected by extreme values. In addition to effects on the location and scale parameters ($\mu_i$ and $\sigma^2_i$), effects on the degrees of freedom $df_i$ also allow to determine the deviation from normality since the t distribution leads to a normal distribution for $df_i \to \infty$. To account for possible skewness in the response distribution, skewed versions of both the normal and the t distribution can be considered where an additional parameter allows to induce right or left skewness.

**Non-Negative Responses.** In addition to the four distributions that we have considered in the analysis of incomes, general distributional models can be derived from large classes such as the generalised beta family (with up to five parameters allowing for skewness, kurtosis or parameters that affect the shape of the distribution in general), see Kleiber and Kotz (2003), or the generalized gamma family that includes an additional shape parameter $\tau_i$ that allows for extended flexibility compared to the standard 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 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.

If the probability of measuring a particular value varies inversely as a power of that value, i.e. the density of $y_i$ can be written in terms of $k y_i^{-\alpha}$ for some constant $k$, the Pareto distribution is one famous distribution widely used in physics, social sciences, biology, finance and environmental sciences to model different quantities of interest like city populations, sizes of earthquakes, wars, sales of books (see Newman, 2005, for further examples and references). With the parametrisation given in Table 2, the expectation is given by $c_i / (b_i - 1)$, where $c_i > 0$ 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 disease (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. 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

$$p(y_i) = \pi_i I_{\{0\}}(y_i) + (1 - \pi_i)g(y_i),$$

with density $g$ corresponding to any count data distribution. A detailed description of zero inflated and overdispersed count data regression is provided in Klein et al. (2014b).

### Mixed Discrete-Continuous Distributions.

In many applications, e.g. insurances (Klein et al., 2014a; 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

$$p(y_i) = \begin{cases} 
1 - \pi_i & y_i = 0 \\
\pi_i g(y_i) & y_i > 0 
\end{cases}$$

with $g(y_i)$ any parametric density of a positive real random variable and $\pi_i$ 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 $\pi_i$ compared to the expectation of the continuous part.

### Distributions with Compact Support.

Like in the application on output shares of farms with respect to cereal products provided in Section C of the electronic supplement or the analysis of loss given default in finance, 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 2, $E(y_i) = \mu_i \in (0, 1)$ 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 and in one joint model. The additional parameters $\nu_i > 0$ and $\tau_i > 0$ control the probabilities $p_{i1} = P(y_i = 0) = \nu_i / (1 + \nu_i + \tau_i)$ and $p_{i2} = P(y_i = 1) = \tau_i / (1 + \nu_i + \tau_i)$ and are chosen due to the fact that the two probabilities have to be linked such that their sum is always smaller or equal than one. The expectation of $y_i$, compared to a beta distributed variable, is reduced by the factor $1 - p_{i1} - p_{i2}$ and shifted by $p_{i2}$:

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

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 $\frac{\nu_i + \tau_i}{1 + \nu_i + \tau_i}$ and $\frac{\nu_i}{1 + \nu_i}$, respectively.

### A.2 Generic Predictor Specification

While considering a specific instance of a structured additive predictor for the analysis of incomes in most of the main part of this paper, a generic structured additive predictor for parameter $\vartheta_k$ is given by

$$\eta_{\vartheta_k} = \beta_{\vartheta_k} + f_{\vartheta_k}^1(\upsilon_i) + \ldots + f_{\vartheta_k}^{J_k}(\upsilon_i)$$

where $\beta_0$ represents the overall level of the predictor and the functions $f_j(\upsilon_i)$, $j = 1, \ldots, J_k$, relate to different covariate effects defined in terms of the complete covariate vector $\upsilon_i$. Note that of course each parameter vector $\vartheta_k$ may depend on different covariates and especially a different number of effects $J_k$ but we suppress this possibility (as well as 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.

$$f_j(\upsilon_i) = \sum_{d_j=1}^{D_j} \beta_{j,d_j} B_{j,d_j}(\upsilon_i)$$

such that in matrix notation we can write $f_j = (f_j(\upsilon_1), \ldots, f_j(\upsilon_n))' = Z_j \beta_j$ where $Z_j[i,d_j] = B_{j,d_j}(\upsilon_i)$ is a design matrix and $\beta_j$ is the vector of coefficients to be
estimated. The basis function representation then leads to the matrix representation of the generic predictor \( \mathbf{f_j} \) as indicated in equation (4) in Section 2.4.

For regularisation purposes it is common to add a penalty term \( \text{pen}(\mathbf{f_j}) = \text{pen}(\mathbf{\beta_j}) = \mathbf{\beta_j}' \mathbf{K_j} \mathbf{\beta_j} \) that controls specific smoothness properties of the estimates. The Bayesian equivalent to this frequentist formulation is to assign multivariate Gaussian priors as defined in (5) to the regression coefficients \( \mathbf{\beta_j} \) with prior precision matrix \( \mathbf{K_j} \) which corresponds to the penalty matrix in a frequentist formulation and smoothness variance \( \tau_j^2 \).

A brief overview of terms that fit into the generic predictor framework is as follows:

- **P-splines \( f(x) \) for nonlinear effects of a continuous covariate \( x \):** The design matrix is obtained from evaluating the B-spline basis functions at the observations; the penalty matrix is given by \( \mathbf{K} = \mathbf{D_r}' \mathbf{D_r} \), with \( \mathbf{D_r} \) an \( r \)th order difference matrix.

- **Varying coefficient terms \( uf(x) \) for interactions between an interaction variable \( u \) and continuous effect modifier \( x \):** The design matrix is obtained from multiplying the evaluated B-spline basis functions with the interaction variable; the penalty matrix is \( \mathbf{K} = \mathbf{D_r}' \mathbf{D_r} \) as for P-splines.

- **2D P-splines \( f(x_1, x_2) \) for interaction surfaces between two continuous covariates \( x_1 \) and \( x_2 \):** The design matrix arises from evaluating the tensor product B-spline basis \( B_{d_1,d_2}(x_1, x_2) = B_{d_1}(x_1)B_{d_2}(x_2) \); the penalty matrix is \( \mathbf{K} = \mathbf{I} \otimes \mathbf{K_1} + \mathbf{K_2} \otimes \mathbf{I} \) with identity matrix \( \mathbf{I} \) and penalty matrices \( \mathbf{K_1} \) and \( \mathbf{K_2} \) as for univariate P-splines.

- **Kriging for spatial effects \( f_{spat}(s_1, s_2) \) based on coordinate information \( s_1, s_2 \):** The correlation function assumed for the spatial process defines the radial basis; the penalty matrix is \( \mathbf{K} = \mathbf{R} \), with correlation matrix \( \mathbf{R} \).

- **Markov random fields for spatial effects \( f_{spat}(s) \) based on regional data \( s \):** The design matrix is a 0/1-incidence matrix, that links observations and regions; the penalty matrix \( \mathbf{K} \) is the adjacency matrix corresponding to the neighbourhood structure of the regions.

- **Random intercepts for a grouping variable:** The design matrix is a 0/1-incidence matrix, that links observations and clusters; the penalty matrix is given by \( \mathbf{K} = \mathbf{I} \), with identity matrix \( \mathbf{I} \).

A more detailed discussion is provided in Fahrmeir et al. (2013, Chs. 8 and 9)
A.3 Bayesian Inference

Algorithm. Given the IWLS proposals constructed in Section 2.5, 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 $\vartheta_1, \ldots, \vartheta_K$ of length $n$ and for $j = 1, \ldots, J_k$, $k = 1, \ldots, K$, draw a proposal $\beta_j^p$ from the density

$$q \left( \left( \beta_j^{\vartheta_k} \right)^{[t]}, \beta_j^p \right) = N \left( \left( \mu_j^{\vartheta_k} \right)^{[t]}, \left( P_j^{\vartheta_k} \right)^{[t]} \right)^{-1}$$

with expectation $\mu_j^{\vartheta_k}$ and precision matrix $P_j^{\vartheta_k}$ given in (7). Accept $\beta_j^p$ as a new state of $\left( \beta_j^{\vartheta_k} \right)^{[t]}$ with acceptance probability

$$\alpha \left( \left( \beta_j^{\vartheta_k} \right)^{[t]}, \beta_j^p \right) = \min \left\{ \frac{p \left( \beta_j^p, \left( \beta_j^{\vartheta_k} \right)^{[t]} \right) q \left( \beta_j^p, \left( \beta_j^{\vartheta_k} \right)^{[t]} \right)}{p \left( \left( \beta_j^{\vartheta_k} \right)^{[t]}, \beta_j^p \right) q \left( \left( \beta_j^{\vartheta_k} \right)^{[t]}, \beta_j^p \right)}, 1 \right\}.$$ 

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 $A\beta_j = 0$ holds with an appropriate matrix $A$. For updating $\left( \tau_j^{\vartheta_k} \right)^{[t]}$, generate a random number from the inverse Gamma distribution $IG \left( a'_j, \left( b'_j \right)^{[t]} \right)$ with $a'_j$ and $\left( b'_j \right)^{[t]}$ given in (6).

Working Weights. In principle, this algorithm is applicable to all distributions where first and second derivative of the log-likelihood exist. For the diagonal matrix of working weights $W$, we usually work with the 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 challenging in cases of complex likelihood structures. Nevertheless, we observed that it is quite worth computing the expectations because of the positive mixing behaviour 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 $P_j$ of the proposal density is invertible if the design matrices $Z_j$ have full column rank. Explicit derivations of score vectors and working weights for distributions of Table 2 as well as the consideration of their positiveness can be found in Section D of the online supplement.
Propriety of the Posterior  The question whether the posterior distribution in a distributional regression model is proper arises since the model specifications typically include several partially improper priors. Sun et al. (2001) and Fahrmeir and Kneib (2009) treated this question in exponential family regression while Klein et al. (2014b) generalised these results to the distributional framework of structured additive regression in the special case of count data regression. Note that the sufficient conditions derived in Klein et al. (2014b) carry over to all distributions described in Table 2 when assuming that none of the distributional parameters is on the boundary of the parameter space (an assumption that would also have to be made to apply standard maximum likelihood asymptotics).

A.4 Model Choice

In the following, we comment on some required modifications for the model choice tools that we discussed in Section 3 when considering discrete or mixed discrete-continuous responses. For the DIC, no changes are necessary.

Quantile Residuals. When considering discrete responses (or responses with a discrete component), the value \( u_i \) that is plugged into the cumulative distribution function of the standard normal is generated as a random draw from the interval \([F(y_i - 1|\hat{\vartheta}_i), F(y_i|\hat{\vartheta}_i)]\) if the observed value \( y_i \) is from the discrete part of the response. This introduces some randomness in the graphical display of the quantile residuals but in most cases, the impact of redoing the calculations (i.e. drawing new \( u_i \)) on the overall assessment of the fit is rather small.

Scoring Rules. Let \( \Omega \) be a general sample space. Let furthermore \( \mathcal{A} \) be a \( \sigma \)-algebra of subsets of \( \Omega \) and \( \mathcal{F} \) be a class of probability measures on \((\Omega, \mathcal{A})\). A probabilistic forecast is a probability measure \( F \in \mathcal{F} \) and a scoring rule is a real-valued function \( S : \mathcal{F} \to \mathbb{R} = [-\infty, \infty] \) such that \( S(F, \cdot) \) is measurable with respect to \( \mathcal{A} \) for all \( F \in \mathcal{F} \). If \( F \) is the probabilistic forecast the scoring rule is called proper relative to \( \mathcal{F} \) if the expected score \( S(F, F_0) = \mathbb{E}_{F_0}[S(F,Y)] \) under \( F_0 \) fulfils the inequality \( S(F,F_0) \leq S(F_0,F_0) \) for all \( F,F_0 \in \mathcal{F} \). It is strictly proper if equality holds if and only if \( F = F_0 \).

For discrete responses with sample space \( \Omega = \{\omega_1, \omega_2, \ldots\} \), density \( p_r(y) = p(y|\vartheta_{r,1}, \ldots, \vartheta_{r,K}) \) and a given new realisation \( y_{\text{new}} \) the logarithmic, spherical and
quadratic score are defined as

\[
\begin{align*}
\text{LS}(p_r, y_{\text{new}}) &= \log(p_r(y_{\text{new}})) \\
\text{SPS}(p_r, y_{\text{new}}) &= \frac{p_r(y_{\text{new}})}{\left(\sum_{\omega \in \Omega} |p_r(\omega)|^2\right)^{1/2}} \\
\text{QS}(p_r, y_{\text{new}}) &= -\sum_{\omega \in \Omega} (1(y_{\text{new}} = \omega) - p_r(\omega))^2.
\end{align*}
\]

For mixed discrete-continuous variables, let \( \Omega_1 \) and \( \Omega_2 \) (\( \Omega_1 \cap \Omega_2 = \emptyset \)), be the sample space for the continuous and the discrete part, respectively. The sample space for the mixed response is given then by \( \Omega = \Omega_1 \cup \Omega_2 \), the associated \( \sigma \)-algebra is \( \mathcal{A} = \mathcal{B}(\Omega_1) \cup \mathcal{P}(\Omega_2) \) where \( \mathcal{B}(\Omega_1) \) denotes the Borel \( \sigma \)-algebra generated by \( \Omega_1 \) and \( \mathcal{P}(\Omega_2) \) denotes the power set of \( \Omega_2 \). Define the density \( p_r(y) \) of response \( y \in \Omega_1 \cup \Omega_2 \) as

\[
p_r(y) = \begin{cases} 
\pi g_1(y) & y \in \Omega_1 \\
g_2(y) & y \in \Omega_2
\end{cases}
\]

where \( g_1 \) is the square integrable density of the continuous part, \( g_2 \) is the probability mass function of the discrete part and \( 0 \leq \pi = 1 - \sum_{\omega_2 \in \Omega_2} g_2(\omega_2) \leq 1 \) is one minus the sum of all probabilities of the discrete part. The density \( p_r \) is a probability density with respect to the measure

\[
\mu(A) = \begin{cases} 
\sum_{\omega_2 \in A \cap \Omega_2} g_2(\omega_2) + \pi \int_{A \setminus \Omega_2} g_1(\omega_1) \, d\omega_1 & A \in \mathcal{A}, A \cap \Omega_2 \neq \emptyset \\
\pi \int_A g_1(\omega_1) \, d\omega_1 & A \in \mathcal{A}, A \cap \Omega_2 = \emptyset.
\end{cases}
\]

Then, the scores are defined as

\[
\begin{align*}
\text{LS}(p_r, y_{\text{new}}) &= \log(p_r(y_{\text{new}})) \\
\text{SPS}(p_r, y_{\text{new}}) &= \frac{p_r(y_{\text{new}})}{\left(\int |p_r(y)|^2 \mu(dy)\right)^{1/2}} \\
\text{QS}(p_r, y_{\text{new}}) &= 2p_r(y_{\text{new}}) - ||p_r||^2 = p_r(y_{\text{new}}) - \int |p_r(y)|^2 \, dy
\end{align*}
\]

with

\[
\int_\Omega |p_r(\omega)|^2 \mu(d\omega) = \sum_{\omega_2 \in \Omega_2} (g_2(\omega_2))^2 + 2\pi(1 - \pi) \int_{\Omega_1} g(\omega_1) \, d\omega + \pi^2 \int_{\Omega_1} |g(\omega_1)|^2 \, d\omega
\]

\[
= \sum_{\omega_2 \in \Omega_2} (g_2(\omega_2))^2 + 2\pi(1 - \pi) + \pi^2 \int_{\Omega_1} |g(\omega_1)|^2 \, d\omega.
\]

Remark. The scores for continuous variables discussed in Section 3 are obtained as a special case from the scores above when \( \pi = 0 \) and \( \Omega_2 = \emptyset \). Similar, in case of \( \pi = 1 \)
and $\Omega_1 = \emptyset$, we obtain the scores for discrete responses since the quadratic score for discrete responses

$$-\sum_{\omega \in \Omega} (1(y_{\text{new}} = \omega) - p_r(\omega))^2 = 2p_r(y_{\text{new}}) - \sum_{\omega \in \Omega} |p_r(\omega)|^2 - 1.$$ 

is only a linear transformation of the quadratic score for mixed discrete-continuous distributions defined above.

**References**


