Econometrics and Statistics 26 (2023) 99–123

Contents lists available at ScienceDirect

Econometrics and Statistics

journal homepage: www.elsevier.com/locate/ecosta

Rage Against the Mean – A Review of Distributional Regression Approaches

Thomas Kneib*, Alexander Silbersdorff, Benjamin Säfken

Campus Institute Data Science (CIDAS) and Chair of Statistics, Georg-August-Universität Göttingen, Humboldtallee 3, 37073 Göttingen, Germany

ARTICLE INFO

Article history:
Received 23 December 2020
Revised 20 July 2021
Accepted 20 July 2021
Available online 10 August 2021

Keywords:
Conditional transformation models
Density regression
Distribution regression
Expectile regression
Generalized additive models for location
Scale and shape
Quantile regression

ABSTRACT

Distributional regression models that overcome the traditional focus on relating the conditional mean of the response to explanatory variables and instead target either the complete conditional response distribution or more general features thereof have seen increasing interest in the past decade. The current state of distributional regression will be discussed, with a particular focus on the four most prominent model classes: (i) generalized additive models for location, scale and shape, (ii) conditional transformation models and distribution regression, (iii) density regression, and (iv) quantile and expectile regression. Characteristics of the different distributional regression approaches will be provided to establish a structured overview on the similarities and differences with respect to the required assumptions on the conditional response distribution, theoretical properties, and the availability of software implementations. In addition, challenges arising in the interpretability of distributional regression models will be discussed and all four approaches will be illustrated with an application analyzing determinants of income distributions from the German Socio-Economic Panel (GSOEP).

1. Introduction

While the public opinion frequently associates (or even equates) statistics with the calculation of means and averages, anyone seriously conducting an empirical data analysis will agree that statistics offers much more than means. In fact, the structures and associations in a data set can only be fully understood by going beyond the mean, as already emphasized by the originator of regression: “It is difficult to understand why statisticians commonly limit their inquiries to averages, and do not revel in more comprehensive views. Their souls seem as dull to the charm of variety as that of the native of one of our flat English counties, whose retrospect of Switzerland was that, if its mountains could be thrown into its lakes, two nuisances would be got rid of at once.” (Galton, 1889) More specifically, measures reflecting, for example, variability, skewness, kurtosis, or other quantities of interest should be integral part of statistical analyses, let alone the estimation of complete distributions via histograms or kernel density estimates, for instance to identify multimodality. In general, graphical tools play an important role in exploring interesting associations in a given data set.

Despite a presumed general agreement on the insightfulness of aspects beyond the mean, a survey of employed methodologies will display that most empirical modelling is firmly located in the lowlands of the conditional expectation of the distribution of the response, e.g. in generalized linear models. Yet this often equates to the rather presumptive reduction

* Corresponding author.
E-mail address: tkneib@uni-goettingen.de (T. Kneib).

https://doi.org/10.1016/j.ecosta.2021.07.006

of any other potential distributional aspect to a mere nuisance parameter. While this may be acceptable in some set of empirical inquiries, it can also implicate gross negligence of important dynamics regarding the matter at hand. Let us, for example, consider the difference in the remuneration for a person's work depending on their gender and a host of other characteristics stemming from the literature of the so called wage gap. While analyses on these differences are abound covering different time periods, countries and modelling approaches – many of which are statistically intricate and advanced in different ways – many models are found to be focussing on the conditional expectation and its comparison between the respective genders. Yet such narrow minded focus on conditional expectations is rather difficult to justify for an economic assessment of income differentials between selected population groups, when it is known that the individuals from those groups are anything but cold-blooded risk-neutral individuals but rather care deeply about distributional aspects beyond the expectation (e.g. Cowell and Schokkaert, 2001; Engelmann and Strobel, 2004).

More generally speaking, many empirical phenomena such as inequality, efficiency, diversity, risks, etc. are intrinsically linked to distributions rather than means and a fully distributional regression approach offers not only a more realistic but also a much richer framework for an empirical analysis (see, for example Atkinson, 1997; Greene, 1990; Silbersdorff et al., 2018). Even in situations where this is less obvious, investigating regression effects beyond the mean offers the possibility to get a much more comprehensive understanding of the conditional response distribution.

In the following, we will study and review various types of distributional regression models, where "distributional" should emphasize that the conditional distribution of the response variable is modelled in terms of covariates, rather than the mean. Many different types of regression models fit into this broad understanding of distributional regression and have been introduced under very different names in the literature, for example as conditional transformation models (Hothorn et al., 2014), copula regression (Klein and Kneib, 2016), copula additive models (Marra and Radice, 2017), density regression (Dunson et al., 2007), distribution regression (Foresi and Peracchi, 1995; Chernozhukov et al., 2013), distributional regression (Klein et al., 2015c), expectile regression (Newey and Powell, 1987), generalized additive models for location, scale and shape (Rigby and Stasinopoulos, 2005, GAMLSS), or quantile regression (Koenker, 2005), to name only the most prominent examples. The fact that distributional regression comes in so many different flavors and under a variety of names has certainly had a detrimental effect on its widespread usage. In many scientific areas, one or two of the approaches are more or less well established while alternatives have gone unnoticed. While we consider "distributional regression" as an umbrella term in this review, one should in fact, following Manski (1991), understand "regression" as explaining any possible feature of the response variable conditional on the covariates. With this review, we hope to contribute to making this general notion of regression more widespread among statisticians.

We will characterise the different distributional regression approaches to provide a structured overview on their similarities and differences with respect to the required assumptions on the conditional response distribution, theoretical properties, and the availability of software implementations. In addition, we will discuss challenges arising in the interpretability of distributional regression models. With this, we do not only provide a follow up on the discussion paper "Beyond Mean Regression" (Kneib, 2013) which had a more narrow focus on GAMLSS, quantile regression, and expectile regression, but complement the discussion with a structured comparison of the approaches with respect to the aspects mentioned above.

In this review, we will mostly focus on the general model structure and only to a lesser extend on the specific mode of inference or details of computations (although these will also play some role in the reviews of the individual methods). We are also particularly interested in the possibility to accommodate flexible, semiparametric forms of regressions predictors in the spirit of structured additive models (Fahrmeir et al., 2013, Chap. 9). All in all, this implies that the selection of methods as well as the view on the relevance and (dis-)advantages of the models is certainly at least partly subjective.

It is also important to point out some limitations of this review: Our focus is on models for responses that are conditionally independent given the explanatory variables. Of course, dependencies arising from temporal trends, longitudinal structures, or spatial heterogeneity can be accounted for by means of corresponding terms in the regression predictors (e.g. splines for representing a trend, random effects to account for longitudinal data collection or other forms of hierarchical clustering, or spatial effects to account for spatial unobserved heterogeneity), but still no explicit temporally or spatially autoregressive structures are included. We will also mostly focus on structured regression models rather than on prediction-oriented machine learning approaches (although some references in this direction will be provided nonetheless). This decision is driven by the idea that we are interested in understanding the specific impact of the given covariates on the conditional response distribution, although we will see that this is already quite challenging in distributional regression even with relatively simple predictor structures.

The rest of this paper is structured as follows: Section 2 provides the general setup for our considerations. Sections 3, 4, 5, and 6 review the four distributional regression model classes that are central to our discussion and develop aspects for their characterization. Section 7 discusses various challenges arising in the interpretation and application of distributional regression models. The final Section 8 provides a conclusion and discusses promising trends for future research in distributional regression.

2. General Setup

2.1. Observation Model

As the general setup for distributional regression, we assume that observations $(y_i, \boldsymbol{v}_i)$, $i = 1, \ldots, n$ on responses y_i and covariates $\boldsymbol{v}_i$ are given. Furthermore, we assume (conditional) independence of the responses $y_i|\boldsymbol{v}_i$ given the covariates $\boldsymbol{v}_i$.

The covariate vector $\boldsymbol{v}_i$ does not only comprise continuous or categorical covariates, but also non-standard forms of covariate information such as coordinates of observations aligned in space, discrete spatial information in terms of districts or regions, grouping factors associated with random effects, or even more complex types of information such as functional covariates. This allows us to include regression setups featuring random effects, spatial effects, or other regression effects going beyond classical linear regression predictors. Furthermore, this setup allows us to account for certain types of dependencies in the data such that we are in fact not restricted to cross-sectional data but only conditional independence of $y_i|\boldsymbol{v}_i$ is required. For example, the inclusion of spatially correlated (random) effects allows us to compensate at least for specific types of spatial dependence where, conditionally on the spatial effects, the responses are independent of each other while marginally (after integrating out the spatial effects) they are spatially dependent.

2.2. Predictor Structure

More precisely, we use regression predictors of the structured additive form

$$\eta_i = \beta_0 + f_1(\boldsymbol{v}_i) + \ldots + f_J(\boldsymbol{v}_i), \tag{1}$$

where β_0 is an overall intercept while the functions $f_1, \ldots, f_J$ represent different types of effects depending on (subsets of) the covariate $\boldsymbol{v}_i$. In structured additive regression, the different effects are represented in terms of basis function expansions

$$f_j(\boldsymbol{v}_i) = \sum_{l_j=1}^{L_j} \gamma_{jl_j} B_{jl_j}(\boldsymbol{v}_i),$$

where the basis functions $B_{jl_j}(\boldsymbol{v}_i)$ have to be chosen for the specific effect type and $\boldsymbol{\gamma}_j = (\gamma_{j1}, \ldots, \gamma_{jL_j})^\top$ are vectors of basis function coefficients. To ensure desirable properties of the function estimates such as smoothness or shrinkage, suitable penalties (or equivalent priors in a Bayesian setup) have to be supplemented to the effects, (see Fahrmeir et al., 2013, Chs. 8 and 9, for details on various choices and a general introduction to structured additive regression).

2.3. Characteristics of Distributional Regression

To guide our review of different distributional regression models, we structure their discussion along different characteristics representing important properties. In a first part, we introduce the general model structure, including some remarks on the origins of the method and early references. We then illustrate the application and interpretation based on an analysis of the relationship between income and unemployment (see the next section for details). Afterwards, we turn to the required assumptions on the conditional response distribution, focusing on whether the model assumes a parametric, semiparametric or completely non-parametric specification, whether the model targets global or local distributional aspects, which kinds of response distributions (discrete, mixed, continuous, bounded, etc.) the approach supports, and its scalability to multivariate response vectors. Focusing on theoretical properties, we will discuss available modes of statistical inference, possibilities for interpreting estimated models, verifiability of required assumptions, flexibility in specifying the regression predictor(s), and scalability in sample size and model complexity. Furthermore, we review the most important available software implementations. Finally, we provide a short summary for each method, highlighting (dis-)advantages, requirements and our perspective on the relevance of the approaches.

Concerning the nature of the responses, it is important to highlight that we will not restrict ourselves to continuous response variables, but are interested in general distributional regression, encompassing discrete, mixed discrete continuous, and bounded but continuous responses. As we will see later, not all approaches discussed in this paper are equally well designed to accommodate such general types of responses. Another extension of the class of response distributions concerns the dimensionality of the response variable. In particular, distributional regression models with vector-valued, i.e. multivariate responses $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iD})^\top$ instead of scalar responses y_i enable the analyst to not only study the effect of covariates on the marginals of the response vector but also on the dependence between response components.

In our discussion, we will be mostly agnostic about the mode of inference, i.e. no preference is given to specific types of inference, but we only report on available methods and how well the models fit with different types of inference. The same applies to specific implementations and efficient computation. Rather, we will focus on properties of the models themselves and available modelling variants. Note also that our choice of methods to be discussed in this review is guided by a focus on regression models, i.e. models that make the conditional distribution of the responses covariate-dependent. Methods that provide flexible ways of estimating marginal distributions from i.i.d. data (without conditioning on covariates) will therefore be neglected.

2.4. Application: The Income-Scars of Unemployment

As an illustrative example, we consider the relationship between an individuals' gross monthly labour income and the length of their previous spells of unemployment discussed in Sohn (2017). For the illustration of modelling the effect of more than one covariate, we additionally consider the years of schooling. For sake of simplicity and feasible computational requirements, we have reduced and simplified the original data set and only consider a sample of 200 individuals with a

current positive income and an existent previous unemployment experience thus avoiding major point masses both in the dependent variable and the covariate space. For further description of the data used for the illustration see the appendix.

3. Generalized Additive Models for Location, Scale and Shape

3.1. Model Structure

Generalized additive models for location, scale and shape (GAMLSS) assume that the conditional distribution of the responses given covariates can be represented as a parametric distribution with density

$$p(y_i|\boldsymbol{v}_i) = p(y_i|\boldsymbol{\vartheta}(\boldsymbol{v}_i)),$$

where $\boldsymbol{\vartheta}(\boldsymbol{v}_i) = (\vartheta_1(\boldsymbol{v}_i), \ldots, \vartheta_K(\boldsymbol{v}_i))^\top$ is a K-dimensional vector of distributional parameters. This offers huge flexibility with respect to the types and complexity of response distributions that can be considered, see Section 3.3 for details. In particular, the parameters can represent distributional aspects such as location, scale and various types of shape aspects. Each distributional parameter is then potentially a function of the covariates $\boldsymbol{v}_i$, making it specific to the individual observations i. The link between the covariates and the parameters is made explicit via strictly monotonically increasing response functions h_k such that

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

with (parameter-specific) regression predictors $\eta_k(\boldsymbol{v}_i)$ of the form (1). The response functions ensure that the real-valued predictors $\eta_k(\boldsymbol{v}_i)$ are mapped appropriately to the parameter spaces of the distributional parameter, for example via the exponential function for strictly positive parameters or the logit/probit transformation for parameters restricted to the unit interval such as probabilities. For each of the regression predictors, a structured additive combination of regression effects is assumed such that the response function also determines the structural link between the covariate effects and the distributional parameter of interest.

GAMLSS models are quite similar in spirit to generalized linear models (GLMs, Nelder and Wedderburn, 1972) and generalized additive models (GAMs, Hastie and Tibshirani, 1990) and also originate from the same community of researchers. In fact, GLMs themselves form a specific type of distributional regression where a member of the exponential family is assumed for $p(y_i|\boldsymbol{v}_i)$ and only the mean is related to covariates while the scale parameter is fixed and treated as a nuisance parameter. Even more generally, already a likelihood-based treatment of linear models is a special case of a GAMLSS where the normal distribution is assumed for $p(y_i|\boldsymbol{v}_i)$ and, again, only the mean of the normal distribution is related to covariates while the variance is considered a nuisance parameter. Many early references on predecessors of GAMLSS dealt with the problem of modelling heteroscedasticity in linear models, see the discussion contributions to Kneib (2013) for references. Building on these approaches, Efron (1986) introduced double exponential family regression with two predictors for mean and variance and Cole and Green (1992) developed the LMS (for λ, μ, σ) method for models considering location, scale and skewness. An early systematic overview is provided in Carroll and Ruppert (1988).

3.2. Application

Considering the illustrative example of the relationship between an individual's income (y_i) and their unemployment experience (v_i), we assume the conditional distribution to be a log-normal distribution, such that $\log(y_i)|v_i \sim \mathrm{N}(\mu_i(v_i), (\sigma_i(v_i))^2)$ with $\mu_i(v_i) = f^\mu(v_i)$ and $\sigma_i(v_i) = \exp(f^\sigma(v_i))$, where both f^μ and f^σ are smooth functions of the covariate modelled as penalized splines.

Figure 1 displays the obtained parameter estimates for $\mu(v)$ and $\sigma(v)$ for the covered range of unemployment-spell durations using the `gamlss` package in R. For both parameters, we display 95% pointwise confidence bands on the basis of a parametric bootstrap using the implementation provided in the `acid` package (see Sohn, 2016). Given the restricted direct informative value of the parameters, we display the resultant conditional distribution for $v = 0, 2, 4, 6$ and 8, respectively, as well as a set of derived percentiles, namely the 5$^{\text{th}}$, the 25$^{\text{th}}$, the 50$^{\text{th}}$, the 75$^{\text{th}}$ and the 95$^{\text{th}}$ percentile. As can be observed, the estimated distributions feature a general downward trend and a diminishing spread. This finding is coherent with various economic lines of thought arguing for a negative effect of unemployment experiences on income. The arguably most prominent line of argument is that unemployment experiences reduce a person's human capital stock, thus lowering that individual's potential productivity (Pissarides, 1992). Another potentially complementary explanation is the idea of negative signalling effects presuming a stigmatisation of those with unemployment history (Kübler and Weizsäcker, 2003).

3.3. Assumptions on the Conditional Response Distribution

The GAMLSS framework relies on a completely parametric approach for specifying the conditional response distribution, assuming not only universally valid, global effects on the parameters of these distributions but also a fixed type of response distribution for all observations. While this seems rather restrictive at first sight, it still offers considerable flexibility due to the wide range of available response distributions supporting various types of response structures, in particular (i) count distributions such as Poisson, negative binomial, zero-inflated Poisson and negative binomial, or the Sichel distribution,

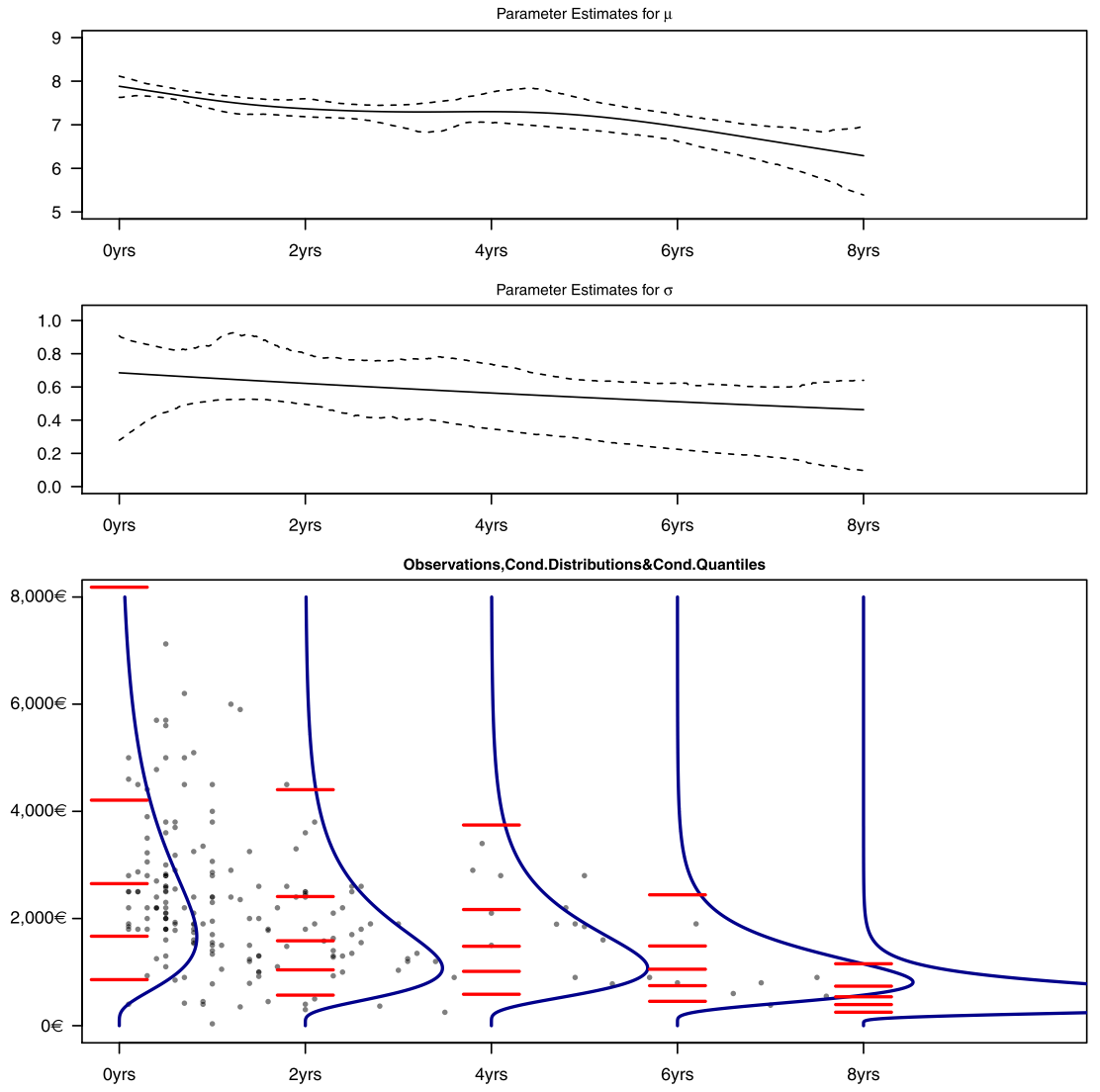

Fig. 1. Parameter estimates with confidence bands (dotted) and derived conditional distributions (blue) and quantiles (red) as estimated by GAMLSS.

(ii) continuous non-negative distributions such as log-normal, gamma, inverse Gauß, Dagum, Box-Cox power exponential, or Sing-Maddala distribution, (iii) skewed distributions on the real line such as skew normal and skew t distribution, (iv) bounded continuous distributions such as the beta distribution, (v) mixed discrete continuous distributions allowing to add discrete point masses to a continuous distribution e.g. in zero adjusted gamma or zero-one inflated beta distributions, (vi) multiple continuous fractions summing to one based on the Dirichlet distribution, or (vii) multivariate response distributions based on copula specifications, see Stasinopoulos et al. (2020) for an in-depth treatment of univariate distributions available with GAMLSS.

Importantly, the GAMLSS framework immediately scales towards multivariate distributions, if an appropriate multivariate response distribution can be defined (which admittedly limits the applicability beyond a moderate dimension of the response vector due to the curse of dimensionality unless being very restrictive in terms of the number of parameters characterising the dependence structure). While some results can be achieved with familiar multivariate distributions such as the multivariate normal, the multivariate t, and the Dirichlet distribution (Klein et al., 2015a) or multivariate count data distributions (van der Wurp et al., 2020), copula specifications turned out to be particularly promising for multivariate GAMLSS since marginal distributions can be flexibly combined with different types of dependence, see for example Klein and Kneib (2016), Vatter and Chavez-Demoulin (2015), or Marra and Radice (2017). A general framework for GAMLSS including extensions to multivariate responses based on vector generalized additive models (VGAMs) is developed in Yee (2015). Most of these developments are in fact restricted to bivariate settings while a trivariate model for binary responses is suggested in Filippou et al. (2019).

Formally, even more flexibility with respect to the response distribution can be achieved by considering mixture specifications, but this makes identification quite challenging. This will, however, be revisited in density regression in Section 5 which is based on mixtures of relatively simple parametric components.

3.4. Theoretical Properties

While the restriction to parametric response distributions may be considered a drawback in terms of possible model miss-specifications, it has the distinct advantage that the model is inherently likelihood-based such that any type of statistical inference building on likelihoods can be applied to conduct inference in GAMLSS. In particular, (penalized) maximum likelihood (Rigby and Stasinopoulos, 2005; Stasinopoulos and Rigby, 2007; Wood et al., 2016), Bayesian approaches utilizing Markov chain Monte Carlo simulations with iteratively weighted least squares proposals (Klein et al., 2015b; 2015c; Umlauf and Kneib, 2018), and functional gradient descent boosting with penalized least squares base-learners (Mayr et al., 2012) have been applied to GAMLSS. Since first and second derivatives of the log-likelihood feature prominently in all these approaches, it is also quite easy to incorporate regularization penalties and other flexible forms of structured additive regression components. Beyond nonlinear effects of continuous covariates, random effects, or spatial effects of different types, LASSO-regularisation in GAMLSS has been suggested in Groll et al. (2019), potentially complex forms of interactions have been constructed in Kneib et al. (2019), and functional covariates have been dealt with in Stöcker et al. (2021). GAMLSS also scale well to large models, both in terms of sample size and complex model specifications involving several thousands of model parameters. It is also to be expected that asymptotic results for GAMs as developed, for example, in Kauermann et al. (2009) should carry over to GAMLSS, although in-depth investigations are still missing.

Although GAMLSS can be seen as an extension to GLMs or GAMs, such that one may be tempted to interpret estimated effects in a similar way, this is more challenging in GAMLSS. The reasons for this include the involvement of response functions, but also the fact that the same covariate may impact different parameters of the response distribution simultaneously. We will discuss these points more extensively in Section 7. Tools for checking models and verifying assumptions are well developed for GAMLSS and often rely on likelihood principles. The most important options are quantile residuals for checking the overall model adequacy and information criteria for comparing rival model specifications, see Stasinopoulos et al. (2017) for details. Predictive evaluations based on cross validation are also facilitated by the availability of various types of proper scoring rules based on the estimated distribution, see for example Klein et al. (2015c).

3.5. Software

Various implementations for GAMLSS exist, with the `gamlss` package in R as the most prominent example based on penalized likelihood inference. Other likelihood-based R packages with a stronger focus on multivariate models are `GJRM` and `VGAM`. Bayesian inference based on Markov chain Monte Carlo simulations is available in the R package `bamlss` and the stand-alone software `BayesX` (www.bayesx.org). Functional gradient descent boosting is implemented in `gamboostLSS`. The well-known R package `mgcv` offers numerically stable and convergent computational methods for likelihood-based inference for some specific GAMLSS-type specifications including various options for smoothing parameter estimation.

3.6. Conclusion

GAMLSS can build on a long history in flexible regression modelling and provides a familiar look and feel for researchers that have experience with GLMs and GAMs. It provides a very rich collection of response models, predictor terms, and inferential approaches, benefiting from its embedding in a likelihood framework. The latter does not only pay off in terms of model fit and evaluation, but also allows for extensions to models featuring censoring or sample selection where the likelihood is modified to incorporate the required adaptations. Finally, there is a variety of software implementations available.

Based on this assessment, GAMLSS is ready for routine use in various areas of applications. What is still missing the most are tools that facilitate the interpretation and visualization for applied researchers. While the modelling concept provides an intuitive appeal appreciated by such applied researchers, they often struggle when working with estimated models, where interpretation is much more involved than in single predictor GLMs (see Section 7 below).

4. Conditional Transformation Models and Distribution Regression

4.1. Model Structure

A common approach to make empirical data adhere to given model assumptions such as normality is to apply suitable transformations, e.g. the logarithmic transformation or the Box-Cox transformation (Box and Cox, 1964) for strictly positive responses. Conditional transformation models as suggested in Hothorn et al. (2018) extend this basic idea by considering covariate-dependent, flexible transformation functions to map the conditional transformation of the response variable to a reference distribution, i.e.

$$h_{\boldsymbol{\nu}_i}(y_i) \overset{\mathcal{D}}{=} z_i, \quad z_i \overset{\text{i.i.d.}}{\sim} p_{\text{ref}}, \tag{2}$$

where $h_{\boldsymbol{\nu}_i}(\cdot)$ denotes the (strictly monotonically increasing) transformation function (indexed by the covariates), $\overset{\mathcal{D}}{=}$ represents equality in distribution, and z_i, $i = 1, \ldots, n$ are i.i.d. realisations from a reference distribution with density p_{ref}. The reference distribution is assumed to be completely specified without any unknown parameters (e.g. the standard normal distribution) such that all unknown parameters of the model are contained in the transformation function.

The likelihood implied by a conditional transformation (with differentiable transformation function) is immediately available from the transformation theorem for densities as

$$p(y_i|\boldsymbol{\nu}_i) = p_{\text{ref}}(h_{\boldsymbol{\nu}_i}(y_i)) \left| \frac{\partial h_{\boldsymbol{\nu}_i}(y_i)}{\partial y_i} \right|. \tag{3}$$

Note that, by assumption, the transformation function is invertible such that the model Eq. (2) can be inverted to generate covariate-specific response distributions from the reference distribution. By construction, any strictly continuous distribution can be represented as conditional transformation model if the class of transformation functions is chosen flexibly enough.

An alternative perspective on conditional transformation models arises when interpreting the model in terms of its cumulative distribution function (CDF)

$$F(y_i|\boldsymbol{\nu}_i) = F_{\text{ref}}(h_{\boldsymbol{\nu}_i}(y_i)),$$

where F_{ref} is the CDF corresponding to the reference density p_{ref}, which implies

$$\mathbb{P}(y_i \leq c|\boldsymbol{\nu}_i) = \mathbb{P}(z_i \leq h_{\boldsymbol{\nu}_i}(c)).$$

Rather then deriving a likelihood-based estimate for the transformation function from (3), we can then assess the shape of the CDF by running a sequence of binary regression models with the indicators $I_i = \mathbb{I}(y_i \leq c)$ as responses and for various values of c. This is an approach proposed in Foresi and Peracchi (1995) as distribution regression that has received considerable interest in the econometrics literature, see for example Chernozhukov et al. (2013), Rothe and Wied (2013), or Van Kerm et al. (2016), in particular for evaluating distributional policy effects. Motivated from a similar idea, Manuguerra and Heller (2010) establish a link between ordinal cumulative logit models and conditional transformation models. In fact, the original proposal of conditional transformation models in Hothorn et al. (2014) was also developed from a sequence of binary exceedance indicator regressions with joint optimisation achieved in a boosting approach.

To practically apply conditional transformation models, we have to choose the reference distribution and parameterise the transformation function. While the choice of the former is basically arbitrary, reference distributions with log-concave densities have the particular advantage to yield unique maximum likelihood estimates (under mild regularity conditions), see Hothorn et al. (2018). A useful default is the standard normal distribution, but other choices can be of interest as well, for example to achieve specific ways of interpreting the model. For example, choosing the maximum extreme value distribution yields a relation to the Cox proportional hazards model, see again Hothorn et al. (2018).

Concerning the transformation function, there are various ways of achieving flexible, covariate-dependent transformations of the response variable. The main difficulty in coming up with a sensible transformation is the restriction that $h_{\boldsymbol{\nu}_i}(y_i)$ has to be strictly monotonically increasing in y_i to make the transformation unique and invertible. This is particularly easy in a simple shift conditional transformation model

$$h_{\boldsymbol{\nu}_i}(y_i) = h_0(y_i) + \eta(\boldsymbol{\nu}_i), \tag{4}$$

where the covariates only induce a marginal shift $\eta(\boldsymbol{\nu}_i)$ while the shape of the conditional response distribution is determined by the transformation function $h_0(y_i)$ that is common to all responses regardless of the covariates. In this additive decomposition between location effects contained in the predictor $\eta(\boldsymbol{\nu}_i)$ and the common transformation function $h_0(y_i)$, monotonicity constraints only have to be applied to the latter while the covariate effects do not have to be restricted in any way. However, the scenario is also somewhat simplistic since the covariates only impact the location and not any other shape parameter of the conditional response distribution. In fact, the linear regression model results, when further restricting $h_0(y_i)$ to be linear in the response y_i.

Including interactions of different complexity allows to deviate from the simplicity of shift models. For example, linear interaction effects such as $y_i g(\nu_i)$ imply dependence of the scale of the response distribution on the covariate ν_i while higher order interactions enable modification of other shape features. Obviously, implementing the monotonicity constraints in such scenarios requires much more care. The original development of conditional transformation models in Hothorn et al. (2018) relies on Bernstein polynomial bases for representing the transformation function since the monotonicity constraints then translate into linear constraints on the polynomial coefficients that can easily be integrated into the estimation process. However, polynomials also imply considerable restrictions on the flexibility of the transformation function and therefore the class of distributions that can reasonably be approximated with a conditional transformation model. More flexibility can be achieved with transformation forests (Hothorn, 2020b) or based on penalized spline specifications (Carlan et al., 2020).

4.2. Application

Considering the illustrative example again, we now assume a conditional transformation function that is allowed to vary along $\nu \in [0, 8]$. Following Jogesh Babu et al. (2002) we use Bernstein polynomial basis of order 4 for modelling the transformation function. The choice of the order is based on Hothorn et al. (2018) with the main impetus directed towards the

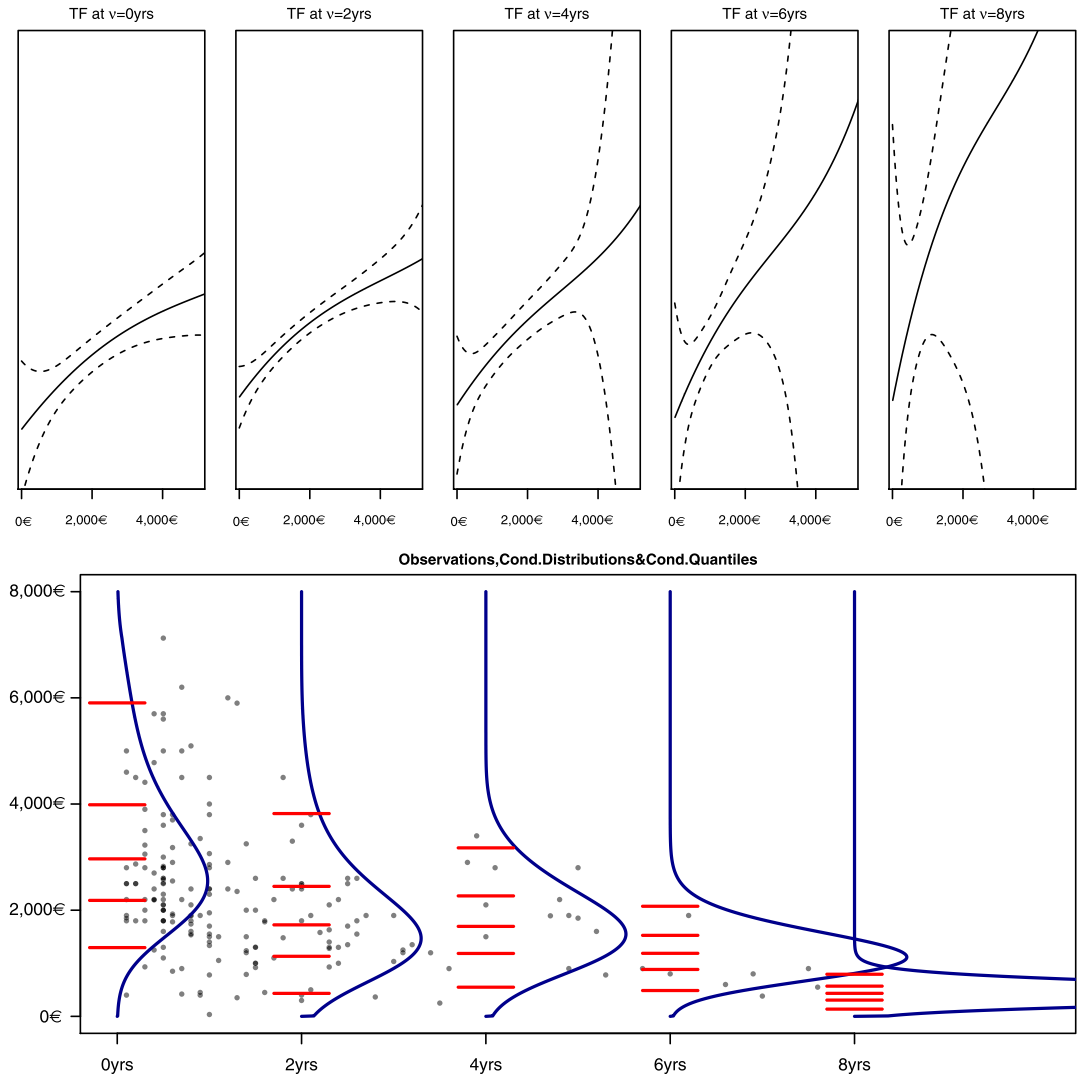

Fig. 2. Transformation function (TF) estimates with confidence bands (dotted) and derived conditional distributions (blue) and quantiles (red) as estimated by CTM.

computational stability while allowing for sufficient flexibility of the polynomial. Figure 2 depicts the resultant transformation functions for $\nu = 0, 2, 4, 6$ and 8, respectively, as well as the resultant conditional distributions and derived quantiles analogously to the illustrations above. We find that, especially in the tails, the estimates provided by CTMs differ from those obtained by GAMLSS, owing to the additional flexibility offered by CTMs for deviating from the shape of a log-normal distribution.

4.3. Assumptions on the Conditional Response Distribution

CTMs provide a semiparametric type of distributional regression where a very flexible model specification in terms of the transformation function and a parametric reference distribution are combined. This still implies a globally valid model specification, but various types of local properties can be derived from the model setup. While the original model formulation is tailored towards univariate, continuous distributions, more general transformation functions (e.g. monotonically increasing step functions) enable the treatment of discrete data (Siegfried and Hothorn, 2020) as well and the approach has also recently been extended to multivariate specifications, scaling well to moderately large dimensions (5 – 10) of the response vector (Klein et al., 2021). The latter provide a flexible extension to copula regression, where the marginal distributions are estimated by CTMs that are then combined in terms of a (covariate-dependent) Gaussian copula (in a representation enabling its treatment as a multivariate transformation function).

4.4. Theoretical Properties

Since CTMs provide direct access to the likelihood, the same modes of inference are available as with GAMLSS. While most current research relies on likelihood-based inference, Bayesian variants (Carlan et al., 2020) and boosting approaches (Hothorn et al., 2014; Hothorn, 2020b) have also been considered.

An inherent challenge in CTMs is the interpretation of estimated effects since these act on the scale of the transformation function rather than on the scale of the response. Hence, unless when the structure of the transformation function is quite simple (as in the shift model (4) discussed above), visualization and effect displays for quantities of interest are the easiest way to assess the effect of covariates in a CTM. Model verification and checking are not yet well explored for CTMs, although all likelihood-based principles such as quantile residuals, information criteria, likelihood-based tests, or scoring rules for predictive evaluation are in principle conceivable. Similarly, the flexibility of predictors is still relatively limited in available implementations. This is particularly true with respect to interactions between the response and covariates. Simple shifts could easily be extended to the full complexity of structured additive predictors, but the main virtue of CTMs only comes into play when also allowing for interactions with the response variable to make distributional features beyond location depend on covariates as well. At the moment, tree-based methods provide most flexibility in this respect. Hothorn (2020b) develops a general boosting approach for CTMs of varying complexity with random forests as the most flexible option including extensions for censored and truncated responses.

CTMs also exhibit an interesting connection to normalizing flows developed in the machine learning community where the transformation function is derived from iterative updating in the spirit of artificial deep neural nets, see Kobyzev et al. (2020) and Papamakarios et al. (2019) for details on normalizing flows and Sick et al. (2020) for relations to CTMs.

4.5. Software

At the moment, implementations for CTMs are mostly in a demonstrator / experimental stage. The R packages `mlt` and `tram` provide implementations for various types of likelihood-based transformation models, see also Hothorn (2020a) for a description of the `mlt` package. We are not aware of dedicated software for distribution regression, but in principle this can be straightforwardly approached by a sequence of binary regression models.

4.6. Conclusion

Currently, the main virtue of CTMs is their encompassing, unifying perspective on a variety of model types from which novel model variants as well as new insights can be generated. They are embedded in likelihood-based inference but provide considerable flexibility in terms of the transformation function. On the downside, this flexibility makes interpretation a real challenge in particular for applied researchers. This difficulty is exacerbated by the (currently) limited support of complex modelling variants and software. For the future, we expect the development of multivariate CTMs to be of particular interest since it enables the specification of covariate-dependent multivariate models well beyond the bivariate case. The connection to normalizing flows and machine learning will be particularly valuable to integrate unstructured types of covariate effects in CTMs.

5. Density Regression

5.1. Model Structure

Implicitly the previous approaches define conditional densities of the response distributions. Density regression explicitly aims at estimating these densities allowing to flexibly change with the covariates employing methods from density estimation, such as mixture models that date back to early years of statistics, see for instance Newcomb (1886). More precisely, in density regression the conditional densities of the responses are represented as

$$p(y_i|\boldsymbol{v}_i) = \sum_{k=1}^{K} \omega_k(\boldsymbol{v}_i) p_k(y_i|\boldsymbol{\vartheta}_k(\boldsymbol{v}_i)),$$

with mixture densities $p_k(y_i|\boldsymbol{\vartheta}_k(\boldsymbol{v}_i))$ and mixture weights $\omega_k(\boldsymbol{v}_i)$. If we assume normal distributions $\mathrm{N}(\mu_k(\boldsymbol{v}_i), \sigma_k^2(\boldsymbol{v}_i))$ for the mixture components, the dependency on the covariates is modelled through the means $\mu_k(\boldsymbol{v}_i)$, the variances $\sigma_k^2(\boldsymbol{v}_i)$ and possibly the weights $\omega_k(\boldsymbol{v}_i)$. This representation is motivated by the useful result that any continuous distribution can be approximated accurately by a sufficiently large numbers of Gaussian mixture components, although the approach is not limited to neither Gaussian nor continuous densities. Identifiability issues are inherent to finite mixtures of distributions and have been analyzed in Titterington et al. (1985). An early reference for finite mixtures of regression models is Wang et al. (1996). An overview of finite mixtures of generalized linear regression models and possible extensions is presented in Grün and Leisch (2008).

Density regression is often derived in a Bayesian non-parametric framework which allows for a countably infinite number of components, i.e. $K = \infty$ by employing a Dirichlet process formulation. Mean-based density regression for instance was

proposed in Escobar and West (1995) and Müller et al. (1996). The Dirichlet process formulation is especially convenient as it circumvents the need of having to choose the number of mixture components K. In case of weights without covariate-dependence, i.e. $\omega_k(\boldsymbol{\nu}_i) = \omega_k$, the mixture model can be represented as

$$p(y_i|\boldsymbol{\nu}_i) = \int p(y_i|\boldsymbol{\vartheta}, \boldsymbol{\nu}_i) G(d\boldsymbol{\vartheta}) = \sum_{k=1}^{\infty} \omega_k p_k(y_i|\boldsymbol{\vartheta}_k(\boldsymbol{\nu}_i)),$$

with G following a Dirichlet process with base measure G_0 and precision α, i.e. $G \sim \text{DP}(\alpha, G_0)$. The mixture becomes apparent by writing G in stick breaking form $G = \sum_{k=1}^{\infty} \omega_k \delta_{\boldsymbol{\vartheta}_k}$ with the distribution $\delta_{\boldsymbol{\vartheta}}$ having a point mass in $\boldsymbol{\vartheta}$. Notice that by integrating out the infinite dimensional G, the model can be linked to a finite mixture predictive density based on the n observations as the conditional prior of $\boldsymbol{\vartheta}_i|\boldsymbol{\vartheta}_{-i}$ chooses new values from the base distribution G_0 with probability $\alpha/(\alpha + n - 1)$ and otherwise chooses from $\boldsymbol{\vartheta}_{-i} = (\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_{i-1}, \boldsymbol{\vartheta}_{i+1}, \ldots, \boldsymbol{\vartheta}_n)$. Empirical Bayesian density regression is presented in Dunson (2007) while Dunson et al. (2007) propose a framework for fully Bayesian inference.

5.2. Application

We now consider a conditional Bayesian non-parametric regression model as implemented in Jara et al. (2011). This model induces covariate-dependent weights following a Dirichlet process stick-breaking construction and linear parameters for the Gaussian mixtures, i.e. $\boldsymbol{\vartheta}_k(\boldsymbol{\nu}) = (\boldsymbol{\beta}_k'\boldsymbol{\nu}, \sigma_k^2)$. Figure (3) shows the conditional densities of the individual income for 0,2,4,6 and 8 years respectively together with the pointwise 95% credible bands. For comparison, the conditional densities and the quantiles are plotted together with the data in the same way as for the other three distributional regression approaches. Notice that the quantiles are no generic output of the estimated model but are derived through the fitted conditional densities.

Compared to the former approaches (GAMLSS and CTMs) the conditional densities are less smooth and fit local characteristics of the data much better. In return, no telling parameter estimates of the model are available (and the approach is computationally comparatively demanding).

5.3. Assumptions on the conditional response distribution

The proposed non-parametric model is highly flexible and encompasses a large amount of sub-models. This flexibility comes at the price of less interpretable parameters. However in certain settings, for instance assuming constant weights in a Gaussian mixture and with linear regression parameters $\boldsymbol{\vartheta}_k(\boldsymbol{\nu}_i) = (\boldsymbol{\beta}_k'\boldsymbol{\nu}_i, \sigma_k^2)$, the conditional density can be interpreted meaningfully. As such the model is most easily understood as a localized density and not by global parameters.

Although density regression is commonly used with mixtures of normal densities, also discrete, bounded and even mixed discrete-continuous response distributions are applicable in general. Even scalability to multivariate response vectors could be possible, though no such approach has been fully investigated to the knowledge of the authors.

5.4. Theoretical properties

Next to the Bayesian framework presented here, finite mixture regression models can also be considered in a frequentist framework (Grün and Leisch, 2007). In this framework, the weights are mostly considered as independent of the covariates. Different mixture components have been proposed with generalized linear models being most prominent. However the lack of identifiability poses a problem for model estimation and parameter interpretation (Grün and Leisch, 2007). Moreover the parameters need to be optimized under constraints such that the expectation maximization (EM) algorithms need to be used. This makes it difficult to derive measures of uncertainty (and other modes of inference) such as standard errors or confidence intervals unless bootstrap methods are employed. The Bayesian approach however relies on Markov chain Monte Carlo and Gibbs sampling methods for posterior computation. Therefore the full posterior distribution is computed making uncertainty measures available naturally. Visualization and effect displays can be obtained best by plotting densities for different outcomes of the quantities of interest. Although also specific measures such as the mean or quantiles can be derived from the estimated densities for different outcomes. Model choice and variable selection have not been examined extensively for density regression, making model diagnostics difficult. Due to the high flexibility of the mixture itself and that the overall number of parameters is scaled by the number of the mixture components, the predictor is mostly restricted to linear effects. Furthermore a high-dimensional predictor could induce identifiability issues complicating computation and interpretation of the model (Jara et al., 2011; Grün and Leisch, 2007). However, the possibly high number of parameters make this approach interesting for scalability in sample size. Up to date there are no sophisticated algorithms for density regression for big data available. Similarly, multivariate extensions are in principle conceivable by considering mixtures of multivariate densities, but this has not yet been thoroughly explored.

5.5. Software

Neal (2000) developed Markov chain methods for sampling from the posterior distribution of a Dirichlet process mixture model. These are used in the R package `DPpackage`. The package offers functions for Bayesian density regression models

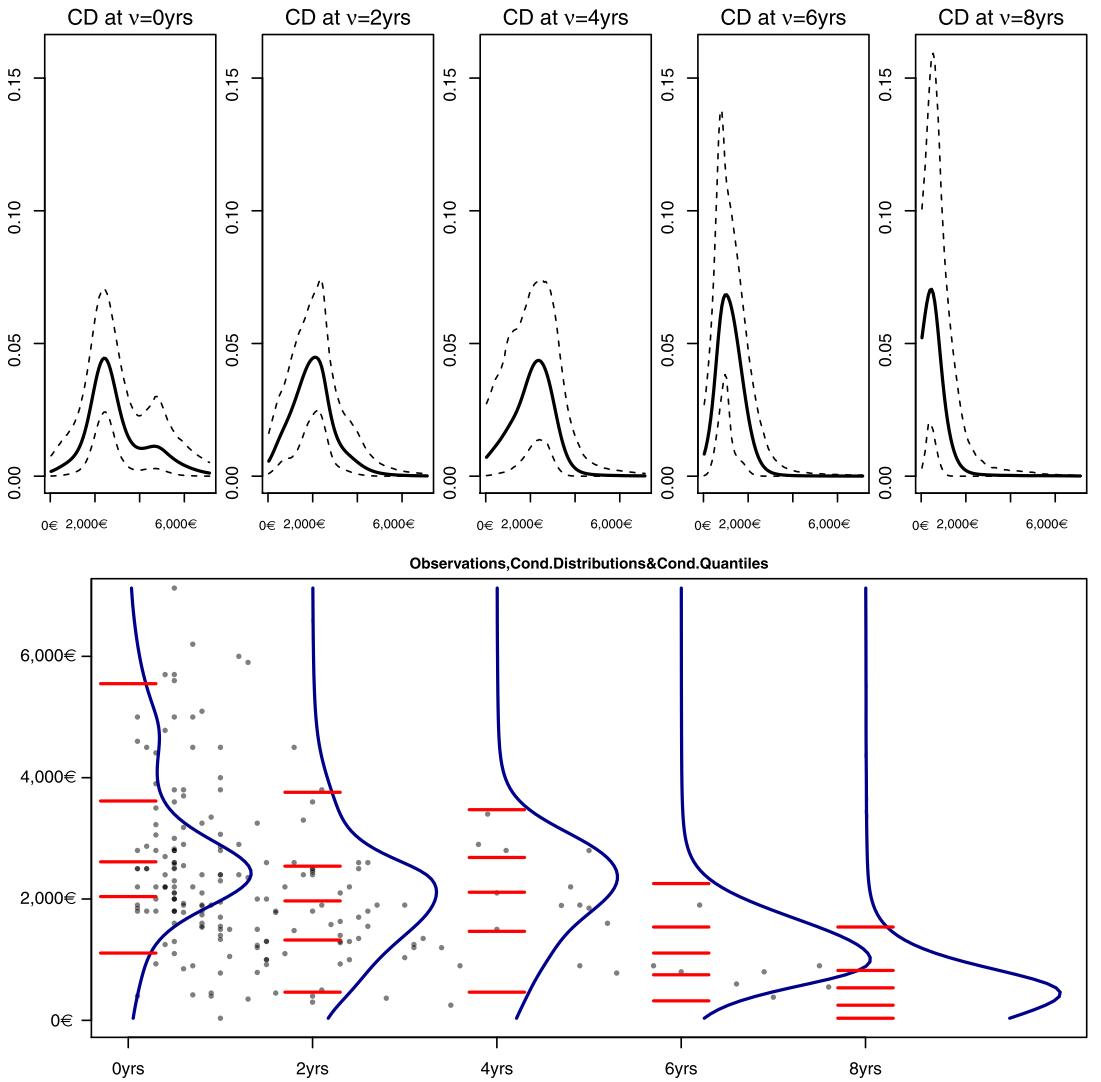

Fig. 3. Conditional density functions with credible bands (dotted) and derived conditional distributions (blue) and quantiles (red).

with continuous predictors using a Dirichlet process mixture of normal models. However this package is no longer available on CRAN. The `flexmix` package provides a general framework for likelihood-based estimation of finite mixtures of (generalized) linear regression models. Leisch (2004) describe the computational aspects of estimating finite mixtures of generalized linear regression models with EM algorithms.

5.6. Conclusion

Density regression, especially in the Bayesian non-parametric setting, has a certain natural mathematical beauty. As such, the framework is broadly applicable due to its non-parametric nature and the entire conditional densities as outcomes. Thus this model is particularly sensible for data that can not be approximated by standard parametric distributions such as multimodal response data. However, the parameters of such models are in general hard to interpret and make the method less appealing for applications focusing on structured parametric effects on the conditional response distribution. For Bayesian density regression, this is for instance the case for the representation of the infinite number of mixture distributions. The difficult overall interpretation of the parameters is also reflected in the identification problems mentioned above.

In recent years, density regression somehow fell into oblivion. This could be due to the lack of the availability of an actively maintained implementation. Also fitting such models to a moderately large data set of about 2,000 observations was difficult with the available implementations which shows that this method is not well equipped for use in times of big data. There have neither been any recent contributions towards multivariate responses that would be an obvious extension

and are an active field of research for the other approaches discussed in this review. In the light of the theoretical elegance of the approach, the authors would be highly interested in future developments that overcome the discussed obstacles and make it fit predictable challenges.

6. Quantile and Expectile Regression

6.1. Model Structure

While all approaches discussed so far assume a global model specified for the complete conditional distribution of the response, there are also approaches that rather focus on localized properties of this distribution. Often these aspects are indexed by some kind of "extremeness" parameter $\tau \in [0, 1]$ that describes how far in the tail the property of interest should be located. To formalize this setup, let us consider the model equation

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

with both the regression predictor $\eta_\tau(\boldsymbol{v}_i)$ and the error terms $\varepsilon_{i\tau}$ being specific to the index τ. We now do not assume a specific distribution for the error terms but rather that $G_{\varepsilon_{i\tau}}(\tau) = 0$ where $G_\bullet(\tau)$ represents a functional corresponding to the local quantity of interest. In standard regression scenarios with estimation based on least squares, we would choose the quantity of interest to be the mean such that $\mathbb{E}(\varepsilon_i) = 0$ and, as a consequence, $\mathbb{E}(y_i) = \eta(\boldsymbol{v}_i)$. More generally, applying the assumption to the quantile function

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

yields regression models for the conditional τ-quantile where

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

i.e. the regression predictor determines the τ-quantile of the conditional response distribution. Quantile regression with linear predictor $\eta_\tau(\boldsymbol{v}_i)$ has originally been proposed by Koenker and Bassett (1978).

Replacing the quantile function with the expectile function $E_{\varepsilon_{i\tau}}(\tau)$ yields expectile regression models that generalize least squares estimation towards the tails of the distribution. More specifically, the τ-expectile of $\varepsilon_{i\tau}$ is defined as the solution to

$$\tau = \frac{\int_{-\infty}^{E_{\varepsilon_{i\tau}}(\tau)} |\varepsilon_{i\tau} - E_{\varepsilon_{i\tau}}(\tau)| p_{\varepsilon_{i\tau}}(\varepsilon_{i\tau}) d\varepsilon_{i\tau}}{\int_{-\infty}^{\infty} |\varepsilon_{i\tau} - E_{\varepsilon_{i\tau}}(\tau)| p_{\varepsilon_{i\tau}}(\varepsilon_{i\tau}) d\varepsilon_{i\tau}},$$

and the mean is obtained as a special case for $\tau = 0.5$. Expectiles have first been considered in Aigner et al. (1976) in the context of stochastic frontier analysis albeit without explicitly using the term expectiles. The latter was introduced in Newey and Powell (1987), but expectiles did not receive as much attention as quantile regression early on. Recent years have seen a reviving interest, in particular due to the immediate possibility to combine corresponding weighted least squares estimation schemes with quadratic penalties for additive (Schnabel and Eilers, 2009) and geoadditive (Sobotka and Kneib, 2012) regression specifications.

The main advantage of quantile and expectile regression is that they avoid specific distributional assumptions on the conditional response distribution, except for the assumption on the local property of interest and conditional independence. In particular, the error terms are not assumed to be i.i.d., but only conditionally independent, such that features such as heteroscedasticity or covariate-dependent skewness do not prevent the application of quantile and expectile regression. More precisely, the presence of such covariate-specific changes in the shape of the conditional response distribution makes it particularly interesting to apply quantile and expectile regression. If the shape of the response distribution stays the same apart from location shifts, the regression effects will also stay approximately constant over the asymmetry levels τ, with the intercept as the sole exception (and differences arising from finite sample uncertainty).

Without assuming a global model distribution, estimation in quantile and expectile regression resorts to the minimization of an appropriate (lack of) fit criterion. For independent data, the asymmetrically weighted L_p, $p \geq 0$, loss

$$\sum_{i=1}^{n} w_\tau(y_i, \eta_{i\tau}(\boldsymbol{v}_i)) |y_i - \eta_{i\tau}(\boldsymbol{v}_i)|^p \tag{5}$$

with weights

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

provides us with a convenient way of determining regression quantiles (for $p = 1$) and regression expectiles (for $p = 2$). Depending on the value chosen for p, different ways of minimizing the criterion have to be employed. In particular, solving the minimization problem gets increasingly intricate for $p \to 0$ when the criterion will face multiple minima. For $p = 1$, linear programming provides easy and numerically stable access to the minimizers at least for linear model specifications and some generalisations (see Koenker, 2005, for details). For $p = 2$, iteratively weighted least squares approaches similar to

the well known Fisher scoring iterations in generalized linear models provide numerically efficient estimates that can also immediately be extended to structured additive predictors (Schnabel and Eilers, 2009; Sobotka and Kneib, 2012). Alternatives to the direct optimisation of the loss criterion (5) have been developed within the context of gradient descent boosting for additive quantile regression models in Fenske et al. (2011) and for geoadditive expectile regression in Sobotka and Kneib (2012).

Unlike approaches based on optimisation of (5), a Bayesian approach to quantile or expectile regression requires the assumption of a working likelihood. Bayesian quantile regression can be based on the asymmetric Laplace distribution as working model as suggested in Yu and Moyeed (2001). Kozumi and Kobayashi (2011) and Yue and Rue (2011) both introduced location-scale mixture representations of the asymmetric Laplace distribution that considerably facilitated Bayesian inference for quantile regression by enabling the construction of Gibbs sampling steps. This is utilized in Waldmann et al. (2013) to develop Bayesian inference for complex, geoadditive quantile regression models including semiparametric random effects specifications with Dirichlet process mixture priors. A Bayesian version of expectile regression based on the asymmetric normal distribution is suggested in Waldmann et al. (2017). There has been considerable debate about whether utilizing a working model for estimating quantile or expectile regression models is a sensible strategy. While formally providing the same point estimates, any inferences relying on the working model should be interpreted with care although simulation results in the papers cited above often yield remarkably good approximate results.

For both quantile and expectile regression, it is common practice to estimate models for a (possibly dense) set of asymmetry values τ. Note, however, that all these models are estimated separately of each other such that the natural ordering of the estimated quantiles/expectiles is not ensured in such a setup. More specifically, it is likely to observe crossing quantiles or expectiles for some values of the asymmetry parameter, in particular when making the set of asymmetry levels very fine. The issue of crossing quantile (or expectile) curves has long-lasting interest in the corresponding literature, see for example Bondell et al. (2010) for suggestions to circumvent crossing in quantile regression and Schulze Waltrup and Kauermann (2017) for expectile regression. General approaches based on rearrangements have been developed in Chernozhukov et al. (2009) and Rodrigues and Fan (2017).

6.2. Application

Given the more challenging nature of the interpretation of expectile-based models, we will solely focus on the illustration of the application by means of quantile regression. In Figure 4, the illustrative example is analysed by additive quantile regression as suggested in Fasiolo et al. (2020) where inference about the conditional quantiles and inherent smoothing parameter estimation are based on similar methodology as in the well-known R package `mgcv`. Specifically, we model the quantiles for the set $\{0.01, 0.02, \ldots, 0.99\}$, while using thin-plate regression splines (Wood, 2003). Next to smooth functions of two more or less haphazard quantiles, including pointwise 95% confidence bands again, the same plot as for the three other approaches is presented showing the conditional densities and quantiles for five different values of the covariate, i.e. unemployment experience. The densities are derived somewhat artificially from a large number of quantiles. As quantile crossings occur, negative values for the densities are possible in general. Crudely amending for this problem, negative values were excluded and a kernel density estimation was performed in order to yield a smooth density. Nevertheless, the resulting densities still feature considerable instability. This highlights that quantile regression is less appropriate for analysing a full conditional distribution but rather suited for investigating a single or low number of quantiles in dependence of covariates.

6.3. Assumptions on the Conditional Response Distribution

Different from the other three distributional regression approaches discussed so far, quantile and expectile regression rely on a non-parametric specification that does not require a specific type of response distribution (except for the Bayesian approaches utilizing working likelihoods). Rather, in addition to the restriction on the quantile/expectile of the error distribution, only independence (and existence of first moments in case of expectile regression) is required. This considerably reduces the risk of model miss-specification concerning the conditional response distribution although, of course, the regression structure as such may still be miss-specified.

Quantile and expectile regression results are indexed by an asymmetry index, where quantile regression is indeed only locally defined while in case of expectiles, the definition of the τ-expectile relies on information from the complete distribution albeit still characterising local variation. The locality of quantiles also gives rise to their well known robustness with respect to outliers. On the other hand, in the context of assessing tail risks, expectiles make use of the additional information provided by the size of losses that are not taken into account by quantile regression. This intuition is further substantiated in the investigations in Ziegel (2016) where expectiles show better properties as a risk measure than quantiles. Schulze Waltrup et al. (2015) supplement these results with a detailed study of the statistical properties of quantile and expectile regression, highlighting, for example, potential efficiency gains for expectiles under certain types of data generating processes.

The interpretation of regression quantiles is usually based on the central property that

$$\mathbb{P}(\varepsilon_{i\tau} \leq Q_{\varepsilon_{i\tau}}(\tau)) = \frac{\int_{-\infty}^{Q_{\varepsilon_{i\tau}}(\tau)} p_{\varepsilon_{i\tau}}(\varepsilon_{i\tau}) d\varepsilon_{i\tau}}{\int_{-\infty}^{\infty} p_{\varepsilon_{i\tau}}(\varepsilon_{i\tau}) d\varepsilon_{i\tau}} = \tau.$$

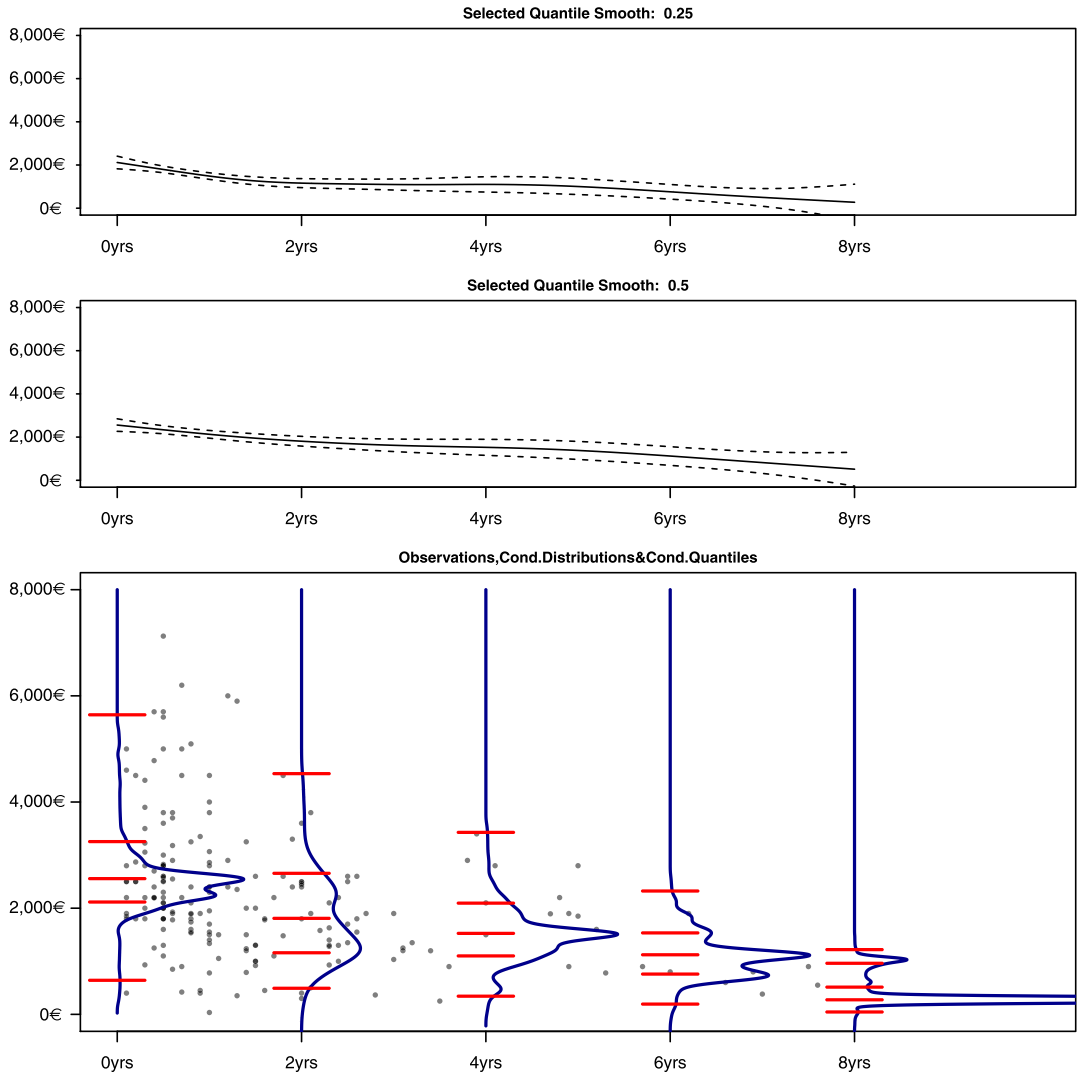

Fig. 4. Selected quantile functions with confidence bands (dotted) and derived conditional distributions (blue) and quantiles (red).

One can also derive a similar property for expectiles, where

$$\frac{\int_{-\infty}^{E_{\varepsilon_{i\tau}}(\tau)} |\varepsilon_{i\tau} - E_{\varepsilon_{i\tau}}(\tau)| p_{\varepsilon_{i\tau}}(\varepsilon_{i\tau}) d\varepsilon_{i\tau}}{\int_{-\infty}^{\infty} |\varepsilon_{i\tau} - E_{\varepsilon_{i\tau}}(\tau)| p_{\varepsilon_{i\tau}}(\varepsilon_{i\tau}) d\varepsilon_{i\tau}} = \tau,$$

i.e. the average distance of $\varepsilon_{i\tau}$ from $E_{\varepsilon_{i\tau}}(\tau)$ that falls below $E_{\varepsilon_{i\tau}}(\tau)$ is given by the expectile level τ. This further substantiates our considerations from above on the incorporation of the size of risks when utilizing expectiles as risk measures. In general, the comparison between quantiles and expectiles can be very well exemplified by the comparison of the median and the mean.

Both quantile and expectile regression are most useful for scalar, continuous responses. While some generalizations to discrete data have been explored (in particular for quantile regression, see Manski (1985) for an early reference), these did not see much response in applications. A notable exception is, for example, Li and Racine (2008) dealing with the derivation of quantile curves from discrete data.

For multivariate extensions, one faces the challenge that no obvious ordering exists in the multivariate response space such that the definition of asymmetric weights is not straightforward, see Serfling (2002) for a review of different possibilities of defining multivariate quantiles. One approach that proved very promising for generalising quantile regression to multivariate responses utilizes the notion of half space depth which allows to reduce the multivariate quantile regression setup to a sequence of univariate directional quantile regressions (Hallin et al., 2010). This approach has received considerable attention in the last years, see for example Paindaveine and Šiman (2012) for a relation of directional quantile regression

with projection regression quantiles, Hallin et al. (2015) for additional flexibility with respect to dependence structures that can be recovered, and Boček and Šiman (2017) for nonlinear effects represented by local polynomial regression. A Bayesian version of multivariate directional quantile regression, including non-crossing restrictions for the multivariate quantile contours, has been suggested in Santos and Kneib (2020). A different multivariate quantile regression approach is proposed in Waldmann and Kneib (2015) based on correlating two asymmetric Laplace distributions in their location-scale mixture representation. A recent overview on multivariate quantile regression is provided in the corresponding chapter in Koenker et al. (2020). Only recently, multivariate expectiles have also seen increasing interest, see Herrmann et al. (2018) for geometric multivariate expectiles and Daouia and Paindaveine (2019) for multiple output expectiles.

6.4. Theoretical Properties

In their classic forms, quantile and expectile regression are based on minimizing the asymmetrically weighted loss criterion (5). However, as discussed above, the loss can also be considered the likelihood from an appropriate working model specifications, more specifically the asymmetric Laplace distribution for quantile regression and an asymmetric normal distribution for expectile regression. Of course, these working likelihoods by no means provide a valid distributional model for the response of interest, but can nonetheless be utilized to derive point estimates, e.g. in Bayesian approaches. Despite the on-going debate about whether this is a sensible idea or not, the representation of the asymmetric Laplace distribution as a location-scale mixture of normals has the advantage of enabling Bayesian inference based on simple Gibbs sampling steps and therefore provides access to quite complex predictors specifications. These are less accessible in the standard form of quantile regression. In particular asymptotic results and finite sample suggestions as summarized in Koenker (2005) are only valid for model specifications with linear predictors.

For expectile regression, inference can be conducted based on iteratively weighted least squares which facilitates the combination with structured additive regression specifications relying on penalization approaches with quadratic penalties. Sobotka et al. (2013) and Guo and Härdle (2012) develop the (asymptotic) theory for confidence intervals and simultaneous confidence bands. Spiegel et al. (2017) propose various approaches for model selection in expectile regression.

6.5. Software

Most statistical software packages nowadays include an implementation of linear quantile regression. The R package `quantreg` furthermore includes several extensions such as quantile smoothing splines while expectile regression is available through the package `expectreg`. A recent fast and stable implementation for fitting additive quantile regression models based on `mgcv` is available in the R package `qgam`, see Fasiolo et al. (2020). Multivariate multiple-output directional quantile regression can be accessed via the R package `modQR`.

6.6. Conclusion

Quantile regression is probably the distributional regression approach that has received most interest in applications (in particular in economics), see for example De Rossi and Harvey (2009). While many quantile regression aficionados tend to dislike the idea of expectile regression (mostly due their lacking robustness properties), in our (humble) opinion they deserve more attention in statistics and econometrics. Besides the advantages discussed above, we would like to re-emphasize that the robustness of quantiles can actually be considered neglecting relevant information that can be important in a statistical analysis.

As a major distinction to the three other distributional regression approaches discussed in this paper, quantile and expectile regression are based on a localized assessment of the conditional response distribution rather than a global model. This brings the advantage of avoiding a global model structure, reducing the risk for model miss-specification. It is, however, important to stress that this does not rule out miss-specifications in the regression predictor.

While the localized fit of quantile and expectile curves has intuitive graphical appeal in simple scatter plot cases with one continuous covariate and continuous covariate, interpretation can be a challenge in more complex data. A similar statement pertains to model verification. One possibility for the latter is checking the subgradient condition as exemplified in Santos and Kneib (2020).

7. Application and Interpretation of Distributional Regression

While offering appealing additional flexibility in terms of statistical modelling, distributional regression yet has to make its way into the mainstream of statistical analysis. One particular difficulty encountered frequently when starting to apply distributional regression to a given data set is the challenge of interpreting the resulting regression effects. In particular, all models discussed in this review require the analyst to move beyond the simple and convenient ceteris paribus type of interpretation, where the effect of differences in one covariate of interest can be interpreted while keeping all other covariates fixed. In conditional transformation models and density regression, it is most obvious that a direct interpretation of the estimated regression effects is hardly possible since they act on the scale of the transformation function on the one hand and on parameters of the components in the mixture densities (and possibly the mixture weights) on the other hand.

In essence, this is the price to pay for the semiparametric model specification that offers considerable flexibility but limits interpretability. At first sight, interpretation in GAMLSS and quantile/expectile regression seems to be more straightforward, but the multiple parameter setup with response functions for GAMLSS and the local modelling strategy in quantile/expectile regression also pose their own challenges as discussed below.

In the following, we will first describe effect displays as one powerful way of assessing the results of distributional regression models, will then differentiate between conditional and marginal effect interpretations, and will afterwards discuss specific challenges arising from the impact of response functions, additive predictor specifications, and the involvement of multiple predictors in one model specification. Finally, we will emphasize specific aspects of interpreting effects on local model properties as in quantile/expectile regression.

7.1. Effect Displays and the Evaluation of Scenarios

Effect displays (Fox, 2003) and the visualization of model properties of interest for given covariate scenarios are an easy way to approach and quantify the impact of covariates in a distributional regression setup. While they have the disadvantage to abandon the usual ceteris paribus type of interpretation, they offer rich and (arguably) more realistic assessments of the way covariates shape features of the response distribution.

For effect displays, the analyst has to choose one covariate of main interest, say ν_j while fixing all other covariates at pre-specified reference values, say $\boldsymbol{\nu}_{-j}^{(0)}$ where the subscript $-j$ indicates that the jth covariate is removed. In addition, one picks some distributional property of interest $g(\mathcal{D})$ where $\mathcal{D}$ denotes the distribution under consideration while $g(\cdot)$ represents the functional calculating the property of interest from that distribution. For the effect display, we now derive that quantity for the estimated distribution $\hat{\mathcal{D}}(y|\nu_j, \boldsymbol{\nu}_{-j}^{(0)})$ obtained from any of the distributional regression methods. The final plot then consists of the points $\left\{\nu_j, g\left(\hat{\mathcal{D}}(y|\nu_j, \boldsymbol{\nu}_{-j}^{(0)})\right)\right\}$ for a grid of values for the covariate of interest ν_j.

This generic framework allows us to consider a wide range of distributional features (e.g. measures of location, variability, skewness, quantiles, the Gini coefficient, etc.), all derived from one common distributional specification such that the conclusions about different distributional aspects are inherently consistent with each other. Of course, informative effect displays should not only consider one set of reference values $\boldsymbol{\nu}_{-j}^{(0)}$ but multiple reference value configurations corresponding to stylized types of "representative" observations from the population of interest.

The main difficulties with effect displays are (i) the multitude of reference value configurations that could be considered (and that may lead to quite distinct forms of effects for the covariate of interest) and (ii) the aggregation of the effect displays to general conclusions. On the positive side, effect displays portray a more realistic picture of the complexity of the data generating process and foster the careful derivation of quantities related to the main research question. For example, in the analysis of income inequality, a multitude of measures reflecting different aspects of inequality can be computed from a given model and can be related to the explanatory variables of interest.

In order to illustrate the use of effect displays for the application and interpretation of distributional regression, let us build on the simple univariate set-up used for illustration in the previous section and add a second covariate: the educational attainment measured in years of schooling. Simplistically, we will assume an additive relation between the duration of the unemployment spell (ν_{i1}) and the educational attainment (ν_{i2}) on the one hand and the predictors on the other, i.e. $\eta(\nu_{i1}, \nu_{i2}) = \beta + f_1(\nu_{i1}) + f_2(\nu_{i2})$. Using effect displays, we thus consider the "effect" of one covariate, namely the duration of the unemployment spell, for given levels of the other covariate. Here, we contemplate the income-unemployment relation for 9 years of schooling (the lowest common educational attainment for Germany), 11 years of schooling (the median schooling level in Germany) and 18 years of schooling (the highest common educational attainment for Germany).

In Figure 5, we portray the estimated conditional distributions and the estimated 10^{th}, 50^{th} and 90^{th} percentile on the left hand side. The former is estimated by GAMLSS, while for the latter we use quantile regression. In terms of observations, we display the sample for $\nu_2 \in 9 \pm 1$, $\nu_2 \in 11 \pm 1$ and $\nu_2 \in 18 \pm 1$ respectively.

In the middle column of Figure 5, we portray commonly used location measures. The blue line depicts the conditional expectation, which is derived from the estimated conditional distributions, while the conditional median can be obtained directly from the quantile regression output. In addition, we display some inequality measures which are frequently used in the economic literature and which should be conceived as within-group inequalities (Silbersdorff, 2017, pp. 17–21). On the one hand, we display the Gini coefficient, which is derived from the estimated conditional distributions and depicts the overall inequality associated with the full conditional distribution, and some commonly used quantile ratios $\left(\frac{Q_{y|\nu}(\tau=0.9)}{Q_{y|\nu}(\tau=0.1)}, \frac{Q_{y|\nu}(\tau=0.9)}{Q_{y|\nu}(\tau=0.5)}, \frac{Q_{y|\nu}(\tau=0.5)}{Q_{y|\nu}(\tau=0.1)}\right)$ depicting the inequality associated with certain domains of the distribution.

Another difficulty arising with the derivation of effect displays is the uncertainty assessment for the quantities derived from the original model parameters to conduct statistical inference. Depending on the chosen mode of inference, this can be approached in different ways, for example

- based on Markov chain Monte Carlo samples in Bayesian inference, where samples from the quantity of interest can easily be obtained by transforming the samples of the original model parameters. This enables routine and exact inferences without having to resort to asymptotic considerations (assuming that the model has been correctly specified).
- utilizing different variants of the bootstrap, e.g. bootstrapping samples of the model parameters from their asymptotic distribution (if asymptotic results are available) and applying the transformation to those samples, a parametric fixed

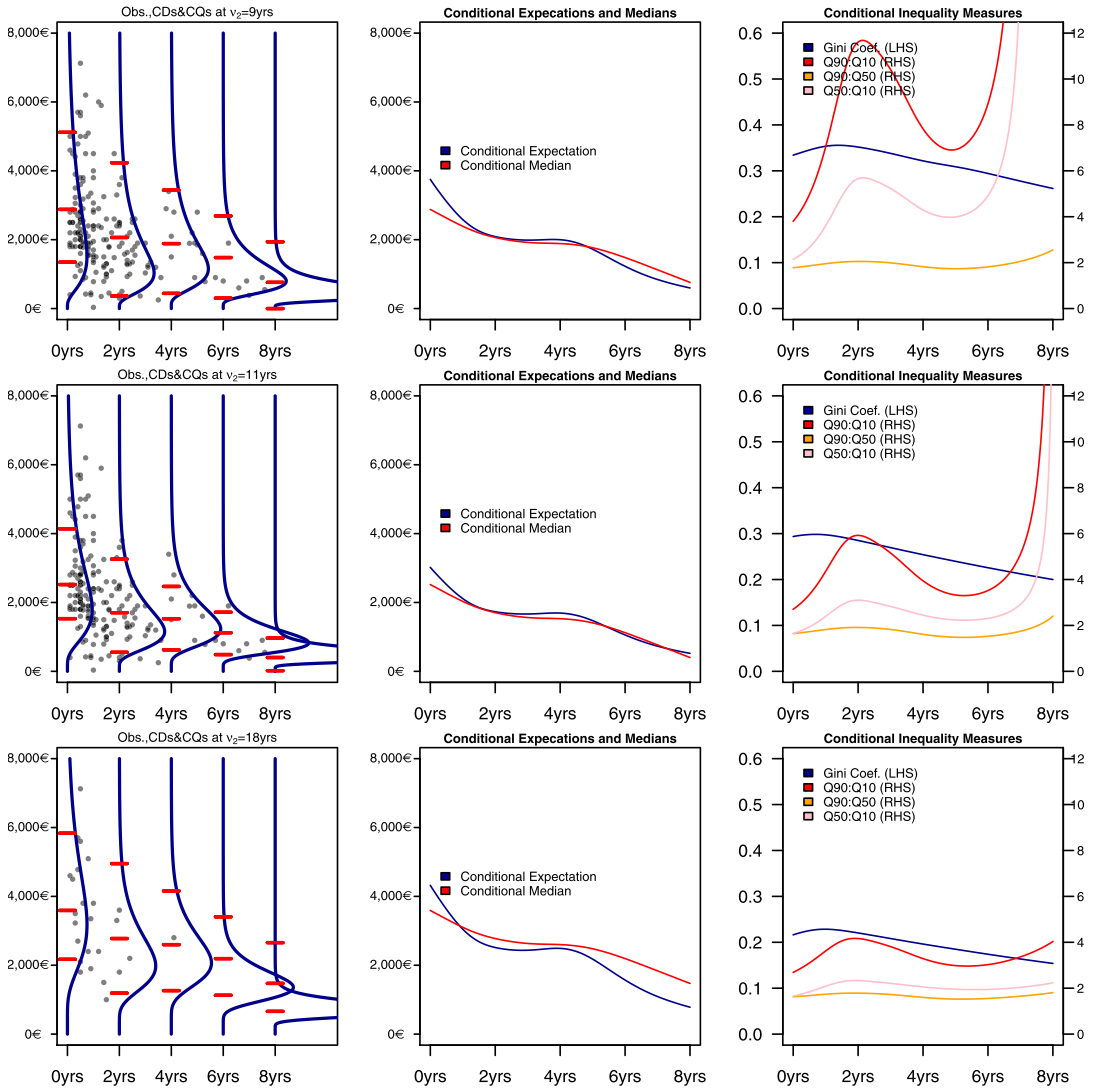

Fig. 5. Effect displays for $\nu_2 = 9(\pm 1)$, $\nu_2 = 11(\pm 1)$ and $\nu_2 = 18(\pm 1)$ displaying the conditional distributions (blue) and quantiles (red) and the derived conditional expectation & Gini coefficient (both blue) and the conditional median and quantile ratios (red/orange/pink) respectively.

regressor bootstrap where new responses are simulated from the estimated conditional response distribution, keeping the regressors fixed at their observed values, or a non-parametric bootstrap resampling both covariates and responses simultaneously from the original sample.

- applying the Δ-rule to transfer asymptotic results for the model parameters to the quantity of interest (if the required transformation is differentiable in the model parameters).

7.2. Conditional vs. Marginal Effects

While the distributional regression models considered in this paper first and foremost fit conditional response distributions $\mathcal{D}(y|\boldsymbol{\nu})$ such that estimated effects impact this conditional distribution, we can also derive effects on marginal distributional aspects for $\mathcal{D}(y)$ without conditioning on the covariates $\boldsymbol{\nu}$. For example, when analyzing inequality, the conditional perspective of deriving implied Gini coefficients from $\mathcal{D}(y|\boldsymbol{\nu})$ refers to inequality within the sub-population defined by covariate characteristics $\boldsymbol{\nu}$ (e.g. a group of individuals with similar education, labour market experience, gender and age). In contrast, the Gini coefficient derived from $\mathcal{D}(y)$ refers to inequality in the complete population. If we are interested in the impact of a policy intervention that, for example, changes the education sector, we would not necessarily be interested in the effect this has on inequality within subgroups with similar education, but on the overall effect on inequality on the population level or some hypothetical population with a pre-specified covariate distribution.

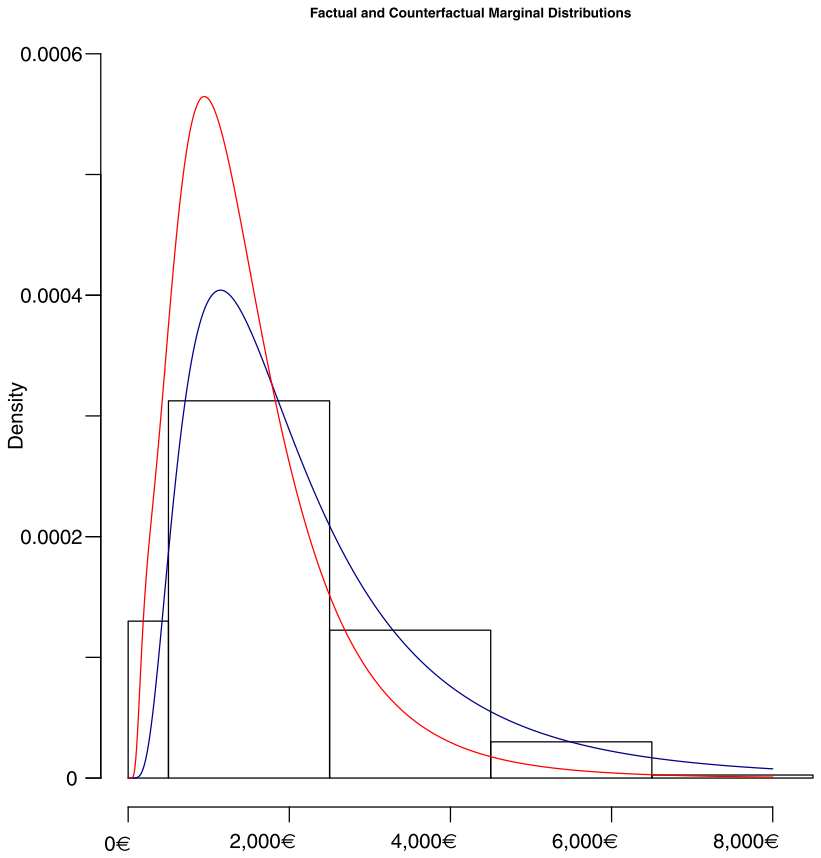

Fig. 6. Empirical marginal distribution (black), estimated factual marginal distribution (blue) and estimated counterfactual distribution (red).

Marginal effects can be obtained by integrating the conditional distribution with respect to the covariate distribution $p(\boldsymbol{\nu})$, i.e. we derive the density of the marginal distribution as

$$p_{p(\boldsymbol{\nu})}(y) = \int p(y|\boldsymbol{\nu})p(\boldsymbol{\nu})d\boldsymbol{\nu},$$

where the index $p(\boldsymbol{\nu})$ emphasizes that the marginal distribution was obtained under a certain covariate distribution. When a policy intervention changes the distribution of the covariates from $p_1(\boldsymbol{\nu})$ to $p_2(\boldsymbol{\nu})$, we can evaluate $p_{p_1(\boldsymbol{\nu})}(y)$ and $p_{p_2(\boldsymbol{\nu})}(y)$ and compare them with respect to different aspects representing inequality. One obvious choice for estimating the marginal response distribution for the given population under the assumption of a representative sample is to replace the covariate distribution $p(\boldsymbol{\nu})$ by the empirical distribution, leading to

$$p_{\text{ECDF}}(y) = \frac{1}{n}\sum_{i=1}^{n} p(y|\boldsymbol{\nu}_i),$$

where each observation point $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_n$ receives a weight of $1/n$.

As illustrated above, marginal effects are of particular interest in the economic literature with a focus on treatment evaluation, see for example Rothe (2012). So far, they have mostly been investigated for quantile treatment effects with the concept of recentred influence functions (Firpo et al., 2009) arguably the most prominent example. Naturally, the marginal perspective can also be applied for other distributional regression setups.

Considering our univariate illustrative example from Section 2 again, we can derive the marginal income distribution by integrating over the distribution of the duration of unemployment spells found in our sample. This estimated marginal distribution is displayed in Figure 6 in blue, with the empirical distribution displayed by the underlying histogram in black coarsely mirroring the estimate derived from the conditional distributions. Let us now assume a hypothetical scenario not (yet) portrayed in our data. For example, let's consider a scenario where due to some economic shock – say an economically devastating pandemic – the distribution of the duration of unemployment spells changes such that everyone undergoes three months of "lockdown-unemployment". By integrating over the estimated conditional distribution on the basis of this counterfactual distribution of unemployment spell durations, we can construct the counterfactual marginal distribution displayed in the figure, featuring little surprisingly an increased share of low incomes.

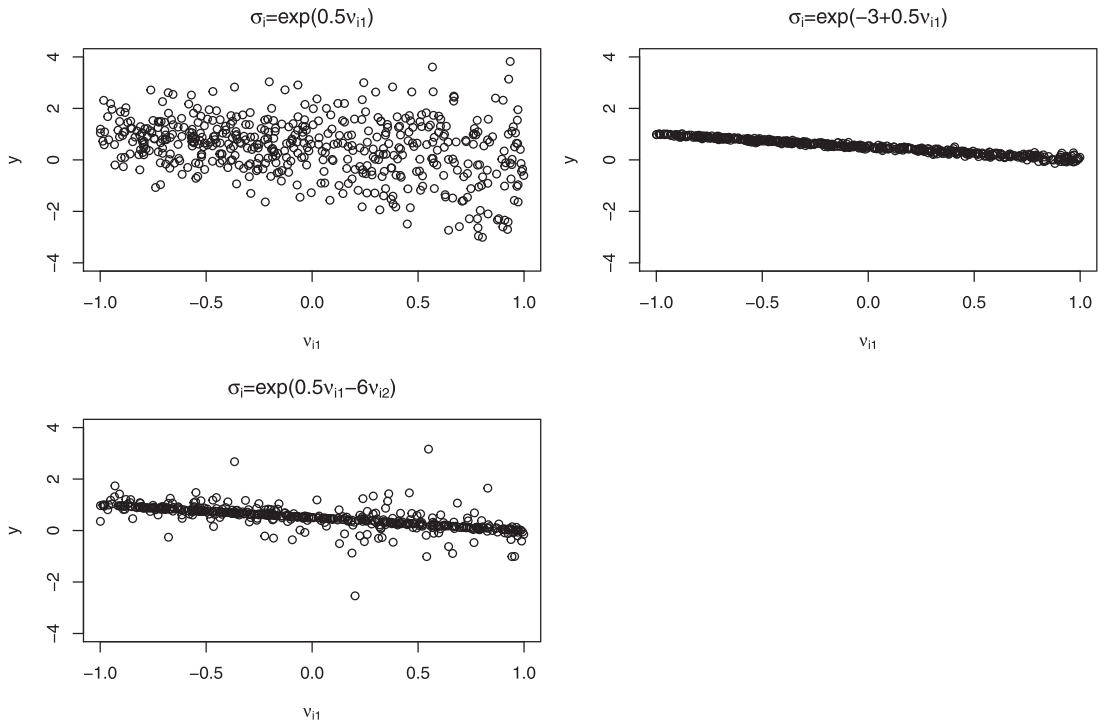

Fig. 7. Impact of the response function: Scatterplots of data from three different data generating processes with the same relative effect of covariate ν_1 on the standard deviation of the normally distributed response.

7.3. The Impact of Response Functions

In particular for GAMLSS, but also for the other distributional regression approaches, interpretation of the estimated model coefficients poses additional challenges due to the fact that the coefficients usually do not directly impact the parameters of the response distribution or properties derived from it. In GAMLSS, this is a consequence already from the specification of response functions $h_k(\cdot)$ mapping the regression predictors η_k to the distributional parameters $\vartheta_k = h_k(\eta_k)$. The nonlinearity of most of these response functions means that differences in the predictor arising from differences in a covariate do not imply ceteris paribus changes in the distributional parameters. For example, even for a linear predictor $\eta_k = \beta_{0k} + \beta_{1k}\nu_1 + \beta_{2k}\nu_2$, the change from ν_2 to $\nu_2 + 1$ implies different changes in the corresponding parameter ϑ_k, depending on the value of ν_1. While for some response functions such as the exponential, a relative ceteris paribus interpretation is still possible due to the implied multiplicative model structure where

$$\exp(\beta_{0k} + \beta_{1k}\nu_1 + \beta_{2k}(\nu_2 + 1)) = \exp(\beta_{0k} + \beta_{1k}\nu_1 + \beta_{2k}\nu_2)\exp(\beta_{2k})$$

it is often relevant to supplement the relative with an absolute interpretation. This can be achieved with the effect displays discussed above when choosing appropriate references values for the other covariates.

Figure 7 illustrates this point with scatter plots for $n = 500$ observations simulated from the data generating process $y_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$ with $\mu_i = 0.5 - 0.5\nu_{i1}$ and three different specifications for the standard deviation: (i) $\sigma_i = \exp(0.5\nu_{i1})$, (ii) $\sigma_i = \exp(-3 + 0.5\nu_{i1})$, and (iii) $\sigma_i = \exp(0.5\nu_{i1} - 6\nu_{i2})$ where $\nu_{i1} \sim \mathrm{U}(-1, 1)$ and $\nu_{i1} \sim \mathrm{U}(0, 1)$. All three models share the same effect size for ν_{i1} (which is also estimated very precisely from the data) such that we obtain the same relative interpretation for the multiplicative effect of the effect on the standard deviation: Ceteris paribus, a difference of one unit in ν_{i1} leads to the same multiplicative effect of $\exp(0.5) \approx 1.64$. However, the absolute effect certainly differs a lot across the scenarios, due to the differences in the level of the predictor arising from a large negative effect in model (ii) and the large negative coefficient for ν_2 (which as a non-zero mean) in model (iii).

This discussion should emphasize that particular care has to be taken in interpreting effects that do not linearly relate to model quantities of interest, but rather require a nonlinear transformation. This does not only apply to GAMLSS, but similarly to CTMs (where effects on the response require the evaluation of the, usually nonlinear, inverse transformation function) and density regression if effects are not only on the means of the mixture components.

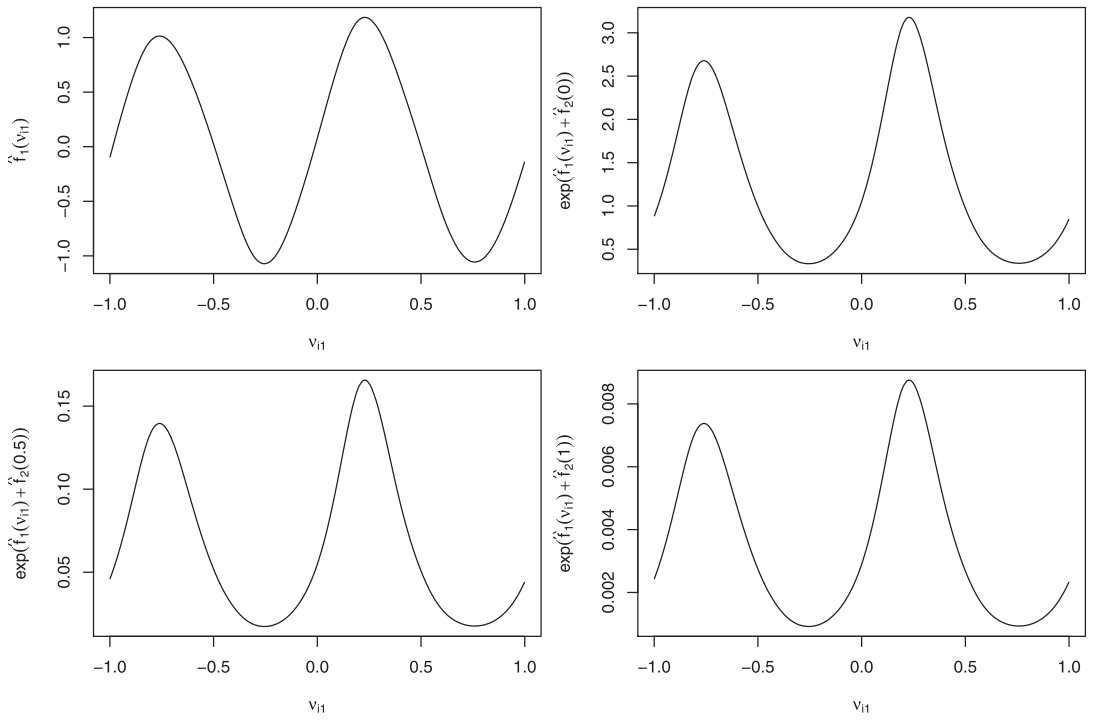

Fig. 8. Additive predictor specification: Estimated nonlinear effect of ν_{i1} (top left) and estimated standard deviations for three different reference values for ν_{i2} (top right: $\nu_{i2} = 0$, bottom left: : $\nu_{i2} = 0.5$, bottom right: : $\nu_{i2} = 1$

7.4. Additive Predictor Specifications

An interpretational challenge that does not stem from the consideration of distributional regression models per se relates to additive predictor specifications of the form

$$\eta(\boldsymbol{\nu}_i) = \beta_0 + f_1(\boldsymbol{\nu}) + \ldots + f_J(\boldsymbol{\nu}).$$

To ensure identifiability, centering constraints such as $\sum_i f_j(\boldsymbol{\nu}_i) = 0$ are often applied to the nonlinear effects resulting in limited interpretation of the level of the estimated effects. Importantly, the centering restriction is by no means a natural or easily interpretable constraint but just one, arbitrary way of fixing the level. Yet, it is tempting to interpret "positive" and "negative" parts of nonlinear effects in displays visualizing their estimates. This complication is exacerbated further when considering the effect of transforming the estimated effects with response functions as discussed in the previous section, where we found that absolute effect interpretation can only be achieved when taking the values of the other covariates into account.

Again, the interpretation of estimated nonlinear effects can either be done in effect displays (which also allows to deal with the fact that one covariate of interest may show up in multiple of the nonlinear effects, e.g. when incorporating interactions) or by comparing the estimated effect for different covariate values. More specifically, differences of the form $\hat{f}_j(\boldsymbol{\nu}_1) - \hat{f}_j(\boldsymbol{\nu}_2)$ for two pre-specified values of the covariate $\boldsymbol{\nu}_1$ and $\boldsymbol{\nu}_2$ remain unaffected by the centering constraint and can therefore be directly interpreted.

Figure 8 shows the difficulties of interpreting a centered effect for estimates obtained from $n = 500$ observations simulated from the data generating process $y_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$ with $\mu_i = 0.5 - 0.5\nu_{i1}$ and $\sigma_i = \exp(\sin(2\pi\,\nu_{i1}) - 6\nu_{i2})$ where $\nu_{i1} \sim \mathrm{U}(-1, 1)$ and $\nu_{i1} \sim \mathrm{U}(0, 1)$. The top left panel shows the centered estimate for the nonlinear effect of ν_{i1} obtained from a penalized spline specification. The estimate resembles the true model specification with cyclic variations around zero, which may tempt the analyst to interpret these as areas where the standard deviation is larger or smaller than one (due to the exponential response function). However, when taking the effect of ν_{i2} into account to produce three effect displays of the estimated standard deviation, we see that the actual range of the estimated standard deviation depends strongly on the level induced by the chosen reference value for ν_{i2}.

7.5. Multiple Predictor Models

Regression models involving multiple regression predictors such as GAMLSS, disable the ceteris paribus interpretation of single regression coefficients if the same covariate enters multiple predictors. This can be illustrated along the example of

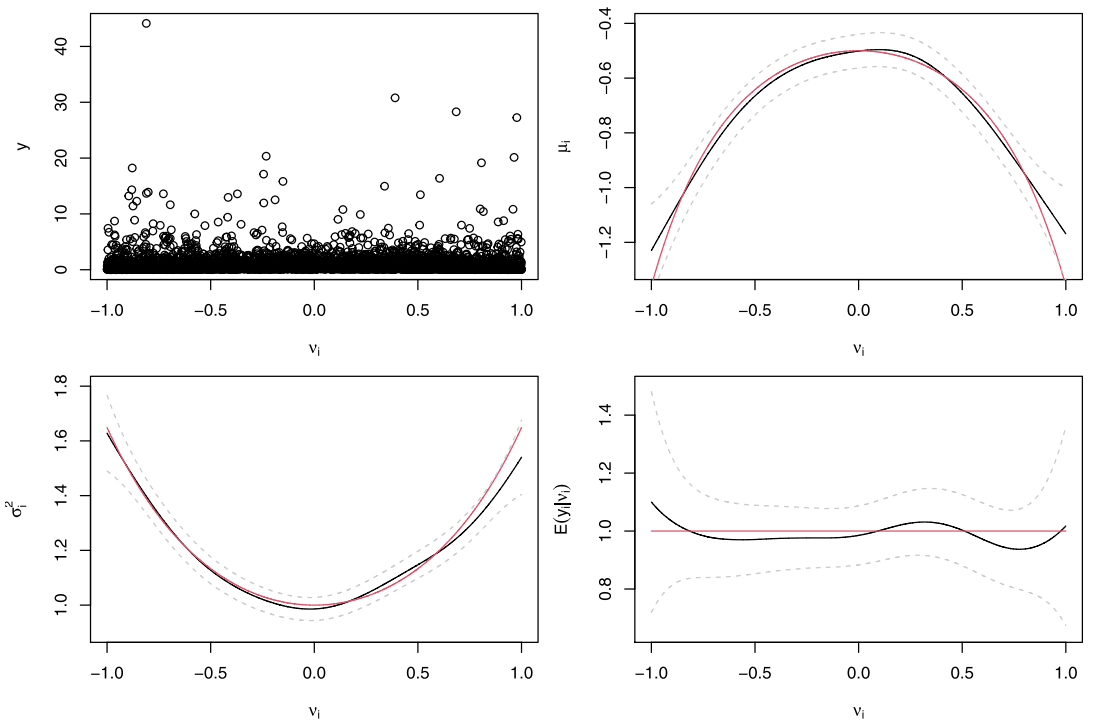

Fig. 9. Multiple predictor models: Simulated data from a log-normal model (top right) together with estimates (solid black line) and 95% confidence intervals (dashed gray lines) for the location parameter μ_i (top right), the scale parameter σ_i^2 (bottom left) and the conditional expectation $\mathbb{E}(y_i|\nu_i)$ (bottom right). True effects are represented as solid red lines.

the log-normal distribution where two predictors are specified for the expectation and the variance (on the log-scale) of the predictor. More precisely, we consider $n = 500$ observations simulated from $y_i \sim \text{LN}(\mu_i, \sigma_i^2)$ where $\mu_i = \mathbb{E}(\log(y_i)|\nu_i)$, $\sigma_i^2 = \text{Var}(\log(y_i)|\nu_i)$ and therefore

$$\mathbb{E}(y_i|\nu_i) = \exp(\mu_i + 0.5\sigma_i^2).$$

Assuming the model specification

$$\mu_i = -0.5\exp(\nu_i^2) \quad \text{and} \quad \sigma_i^2 = \exp(\nu_i^2),$$

the data (visualized in the top left panel of Figure 9) indicate strong nonlinear effects of the continuous covariate ν_i on both parameters of the log-normal distribution when those are modelled as penalized splines (top right and bottom left panels in Figure 9). In contrast, both effects cancel each other out when evaluating the impact of ν_i on the expectation on the original scale of the response (bottom right panel in Figure 9). While, of course, ν_i indeed has a strong impact on the two parameters characterising the log-normal distribution, the impact on derived quantities such as the conditional expectation can only be meaningfully interpreted when combining both effects together, e.g. in an effect display.

7.6. Interpretation of Effects on Local Model Properties

The interpretation of effects on local model properties is particularly relevant in quantile and expectile regression but the comments below can also be useful when considering effect displays of quantile curves (or other local feature curves) derived from any of the other distributional regression approaches. In the following, we will discuss issues for the special case of quantile curves, but similar comments apply to other local model properties.

One very common miss-perception in the interpretation of quantile curves is the intuition that a covariate leading to large conditional quantiles is particularly relevant for observations ranking high in the response distribution. However, there is no such link between the ranking of individual observations in the respective conditional distribution and the size of estimated effects. Rather, the regression effects relate to properties of the conditional distribution, not to individual observations realized from that conditional distribution. For example, in a linear quantile regression specification with only one single covariate ν and associated regressions coefficients $\beta_{\tau_1} > \beta_{\tau_2} > 0$ for $\tau_1 > \tau_2$, we find that the difference between conditional quantiles $Q_\tau(y|\nu + 1)$ and $Q_\tau(y|\nu)$ is positive for both values of the quantile level τ, but the difference is larger for $\tau = \tau_1$ then for $\tau = \tau_2$. As a consequence, the conditional distribution of the response widens as ν increases (at least in the comparison of τ_1 and τ_2).

Another important aspect to take into account when interpreting quantile regression results, is the supposedly likely yet wrong intuition that a fraction of $1-\tau$ observations falls below an estimated quantile regression line while a fraction of τ is located above the quantile line. While this of course holds true for empirical quantiles derived from an i.i.d. sample, the quantile curves in quantile regression are derived from a model specification that imposes restrictions on the shape of these curves and makes the quantiles conditional on covariates. As a consequence, the quantile level applies conditionally, i.e. the fraction of observations below/above the quantile line is itself to be understood conditionally, relating to observations with the given covariate value. Furthermore, model assumptions such as a linear specification of covariate effects limit the flexibility of the quantile curve such that the fraction of τ will not be achieved for all given covariate values. On the other hand, assuming a restrictive structure for the quantile curve enables borrowing information from covariate configurations that are close to each other, such that quantile curves can also be estimated in areas with sparse information.

8. Summary and Future Trends

In this paper, we have summarized the current status of distributional regression, mainly focusing on the four most popular approaches in this area of research well worth consideration both to statisticians and applied researchers. We firmly believe that distributional regression specifications have high relevance for all areas of regression modelling and should belong to the toolbox of any applied statistician. In fact, following Manski (1991), the term regression should be understood in the general sense of explaining any property of the (conditional) distribution of the response variable by available covariate information. It is certainly unfortunate that regression nowadays is understood by many exclusively as mean regression and we hope that this will change in the future. Once starting to think about the chances and possibilities of going beyond the mean in regression modeling, the world is in fact full of interesting challenges, from an applied or a methodological perspective and we very much look forward to seeing many such developments in the future. From our own, subjective perspective, three areas deserve particular attention: (i) Understanding and interpreting the results achieved with distributional regression, (ii) distributional response models for multivariate responses (in particular models going beyond the bivariate response case), and (iii) exploring the link between distributional regression and causal inference:

One major hurdle for the application of distributional regression for applied researchers is the increased difficulty to understand and interpret their results. While many researchers immediately feel attracted by the ability to study different phenomena or to get a more detailed picture concerning the association between variables with distributional regression, this enthusiasm is often curbed when they realize that standard ways of interpreting regression output (ceteris paribus interpretation of changes in covariates, individual significances, etc.) do not easily carry over to distributional regression. The main reasons for these difficulties include (i) the fact that the regression effects are not directly related to the distributional quantity of interest, (ii) the models involve complex transformations of the regression predictors, (iii) many model variants involve multiple predictors, and (iv) many aspects of the response distribution can not be easily separated from each other (e.g. for many distributions there is a natural relation between the mean and the variance such that one can not interpret effects on the mean without taking corresponding changes in the variance into account). To deal with these challenges, graphical tools such as effect displays and scenario analyses can be very helpful to grasp the main outcomes of the analysis, but more tools are clearly needed. One specific point is the automatisation of such visualisations, for example in Shiny apps (see for example Stadlmann and Kneib, 2021, for GAMLSS-type models), to make them accessible also to applied researches not familiar with the technical details of a model. In addition, quantitative measures are needed to supplement the graphical exploration, for example representing overall variables importance measures, simultaneous significances, etc. This will also require changing the way we report results of distributional regression analyses, where a simple table of regression effects is certainly not sufficient. Finally, detailed case studies presenting best practice cases will be needed to provide blueprints for sensible analyses with the various types of distributional regression models.

Concerning methodological developments, we believe that multivariate extensions are still underexplored and underdeveloped. Partly, this can be explained by the difficulties arising from the curse of dimensionality, i.e. the fact that data become increasingly sparse in higher dimensions such that naively estimating high-dimensional distributions including their dependence structure is usually prohibitive. In addition, introducing tractable and interpretable parameterizations of multivariate distributions is certainly challenging. For example, in case of the multivariate normal distribution a useful parameterization of the dependence structure can be obtained in terms of the Cholesky factor of the precision matrix, but interpreting regression effects on elements of the Cholesky factor is a challenge. The multivariate conditional transformation models presented in Klein et al. (2021) and the multiple output quantiles and expectiles of Hallin et al. (2010) and Daouia and Paindaveine (2019), respectively, scale well beyond two dimensions (see also Santos and Kneib, 2020) and some trivariate GAMLSS have also been proposed (e.g. Filippou et al., 2019). However, to circumvent the curse of dimensionality, it is probably necessary to impose additional structural assumptions in truly multivariate cases even of moderate dimension.

While distributional regression, similar as many other regression techniques, can be extremely helpful in exploratory research, the rigorous quantification of causal effects from observational data is often required, for example to evaluate the efficacy of treatments or policy interventions. For mean regression, such tools are well developed (especially in econometrics), for example based on instrumental variables, matching, sample selection models, or regression discontinuity designs. While quantile regression models (e.g. Firpo et al., 2009) and distribution regression (e.g. Rothe and Wied, 2013; Chernozhukov et al., 2013; Van Kerm et al., 2016) have already been linked with methods for causal identification (probably, again, due to their more widespread use in econometrics), this does not hold true for the other distributional regression approaches (but

see Hohberg et al., 2020; Briseño Sanchez et al., 2020, for some recent work for GAMLSS). We firmly believe that the ability to conduct causal inference will significantly contribute to the more routine use of distributional regression, in particular for questions with policy relevance. This also concerns the definition of appropriate measures reflecting treatment effects beyond the mean, see for example Park et al. (2021) and Hohberg et al. (2020).

Acknowledgements

The authors thank two referees and an associate editor for their very valuable and insightful comments that helped us to considerably improve upon the first submission. Financial support from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) grant KN 922/9-1 is gratefully acknowledged. We are indebted to various collaborators and colleagues for fruitful and enjoyable discussions on various aspects of distributional regression, in particular Luisa Barbanti, Elisabeth Bergherr (neé Waldmann), Guillermo Briseño Sanchez, Peter Bühlmann, Manuel Carlan, Nora Fenske, Panagiota Filippou, Claudia Flexeder, Andreas Groll, Julien Hambuckers, Gillian Heller, Benjamin Hofner, Maike Hohberg, Torsten Hothorn, Göran Kauermann, Stephan Klasen, Nadja Klein, Stefan Lang, Julia Lynch, Giampiero Marra, Andreas Mayr, Peter Pütz, Natalya Pya, Rosalba Radice, Bob Rigby, Bruno Santos, Matthias Schmid, Linda Schulze Waltrup, Elmar Spiegel, Stanislaus Stadlmann, Mikis Stasinopoulos, Nikolaus Umlauf, Hendrik van der Wurp, Simon Wood and Yu Ryan Yue. The title for the paper was initially suggested by Paul Wiemann. We dedicate this paper to the memory of Stephan Klasen (1966 – 2020), whose friendliness, curiosity, and encouragement was a constant source of inspiration to all of us.

Appendix A. Data

A1. Data origins

As source of our data, we use the SOEP database (Wagner et al., 2007). We use samples A to J, i.e. all available samples excluding only the last refreshment sample 2012 and the migration sample 2013. Concerning the waves, we only use information only from wave BD, i.e. the wave for 2013. Only taking those values for which we have the full set of variables, as described below, this yields 8,344 observations (3780 males and 4564 females). For further discussion see Sohn (2017).

A2. Variables used

As an income variable, we use the gross-market labour income obtained in the previous months (`bdp7701`). The education level is taken on grounds of the variable `BILZEIT` from the person-related status and generated variables (`PGEN`) available in the SOEP. For the length of the personal unemployment spell, we use the available information from the spell data (`PBIOSPE` and `ARTKALEN`) for each individual to construct the duration of the unemployment spells.

A3. Downsizing the dataset

From the SOEP sample, we select 200 observations randomly from a subset containing only those individuals with a positive income and positive unemployment durations of up to 8 years as well positive years of schooling – i.e. we select a sample that only entails observations for which all three variable are available and for which the covariate space is curtailed in order to reduce the sparcity of observations.

References

Aigner, D.J., Amemiya, T., Poirier, D.J., 1976. On the estimation of production frontiers: Maximum likelihood estimation of the parameters of a discontinuous density function. International Economic Review 17, 377–396.
Atkinson, A.B., 1997. Bringing Income Distribution in from the Cold. Economic Journal 107 (441), 297–321.
Boček, P., Šiman, M., 2017. On weighted and locally polynomial directional quantile regression. Computational Statistics 32, 929–946.
Bondell, H., Reich, B., Wang, H., 2010. Noncrossing quantile regression curve estimation. Biometrika 97, 825–838.
Box, G.E.P., Cox, D.R., 1964. An analysis of transformations. Journal of the Royal Statistical Society. Series B (Methodological) 26 (2), 211–252.
Briseño Sanchez, G., Hohberg, M., Groll, A., Kneib, T., 2020. Flexible instrumental variable distributional regression. Journal of the Royal Statistical Society: Series A (Statistics in Society) 183 (4), 1553–1574.
Carlan, M., Kneib, T., Klein, N., 2020. Bayesian Conditional Transformation Models. Technical Report arXiv:https://arxiv.org/abs/2012.11016.
Carroll, R.J., Ruppert, D., 1988. Transformation and Weighting in Regression. CRC Press.
Chernozhukov, V., Fernández-Val, I., Galichon, A., 2009. Improving point and interval estimators of monotone functions by rearrangement. Biometrika 96 (3), 559–575.
Chernozhukov, V., Fernández-Val, I., Melly, B., 2013. Inference on counterfactual distributions. Econometrica 81 (6), 2205–2268.
Cole, T.J., Green, P.J., 1992. Smoothing reference centile curves: the LMS method and penalized likelihood. Statistics in Medicine 11, 1305–1319.
Cowell, F.A., Schokkaert, E., 2001. Risk Perceptions and Distributional Judgments. European Economic Review 45 (4), 941–952.
Daouia, A., Paindaveine, D., 2019. From halfspace m-depth to multiple-output expectile regression.
De Rossi, G., Harvey, A., 2009. Quantiles, expectiles and splines. Journal of Econometrics 152, 179–185.
Dunson, D., 2007. Empirical bayes density regression. Statistica Sinica 17, 481–504.
Dunson, D.B., Pillai, N., Park, J.-H., 2007. Bayesian density regression. Journal of the Royal Statistical Society. Series B (Statistical Methodology) 69 (2), 163–183.
Efron, B., 1986. Double exponential families and their use in generalized linear regression. Journal of the American Statistical Association 81, 709–721.
Engelmann, S., Strobel, M., 2004. Inequality Aversion, Efficiency and Maximising Preferences in Simple Distribution Experiments. American Economic Review 94 (4), 857–869.

Escobar, M.D., West, M., 1995. Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association 90 (430), 577–588.
Fahrmeir, L., Kneib, T., Lang, S., Marx, B., 2013. Regression. Springer Berlin Heidelberg, Berlin, Heidelberg.
Fasiolo, M., Wood, S.N., Zaffran, M., Nedellec, R., Goude, Y., 2020. Fast calibrated additive quantile regression. Journal of the American Statistical Association to appear.
Fenske, N., Kneib, T., Hothorn, T., 2011. Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression. Journal of the American Statistical Association 106 (494), 494–510.
Filippou, P., Kneib, T., Marra, G., Radice, R., 2019. A trivariate additive regression model with arbitrary link functions and varying correlation matrix. Journal of Statistical Planning and Inference 199, 236–248.
Firpo, S., Fortin, N.M., Lemieux, T., 2009. Unconditional Quantile Regressions. Econometrica 77 (3), 953–973.
Foresi, S., Peracchi, F., 1995. The conditional distribution of excess returns: An empirical analysis. Journal of the American Statistical Association 90 (430), 451–466.
Fox, J., 2003. Effect displays in R for generalised linear models. Journal of Statistical Software, Articles 8 (15), 1–27.
Galton, F., 1889. Natural Inheritance. Macmillan, London.
Greene, W.H., 1990. A Gamma-distributed Stochastic Frontier Model. Journal of Econometrics 46 (1), 141–163.
Groll, A., Hambuckers, J., Kneib, T., Umlauf, N., 2019. Lasso-type penalization in the framework of generalized additive models for location, scale and shape. Computational Statistics & Data Analysis 140, 59–73.
Grün, B., Leisch, F., 2007. Fitting finite mixtures of generalized linear regressions in r. Computational Statistics & Data Analysis 51 (11), 5247–5252. Advances in Mixture Models
Grün, B., Leisch, F., 2008. Finite Mixtures of Generalized Linear Regression Models. Physica-Verlag HD, Heidelberg, pp. 205–230.
Guo, M., Härdle, W., 2012. Simultaneous confidence bands for expectile functions. AStA Advances in Statistical Analysis 96, 517–541.
Hallin, M., Lu, Z., Paindaveine, D., Šiman, M., 2015. Local bilinear multiple-output quantile/depth regression. Bernoulli 21, 1435–1466.
Hallin, M., Paindaveine, D., Šiman, M., 2010. Multivariate quantiles and multiple-output regression quantiles: From L_1 optimization to halfspace depth. The Annals of Statistics 38, 635–669.
Hastie, T.J., Tibshirani, R.J., 1990. Generalized Additive Models. Chapman & Hall / CRC.
Herrmann, K., Hofert, M., Mailhot, M., 2018. Multivariate geometric expectiles. Scandinavian Actuarial Journal 7, 629–659.
Hohberg, M., Pütz, P., Kneib, T., 2020. Treatment effects beyond the mean using distributional regression: Methods and guidance. PLOS One 15 (2), e0226514.
Hothorn, T., 2020. Most likely transformations: The mlt package. Journal of Statistical Software 92 (1), 1–68.
Hothorn, T., 2020. Transformation boosting machines. Statistics and Computing 30, 141–152.
Hothorn, T., Kneib, T., Bühlmann, P., 2014. Conditional transformation models. Journal of the Royal Statistical Society: Series B 76 (1), 3–27.
Hothorn, T., Möst, L., Bühlmann, P., 2018. Most likely transformations. Scandindavian Journal of Statistics 45 (1), 110–134.
Jara, A., Hanson, T., Quintana, F.A., Müller, P., Rosner, G.L., 2011. Package: Bayesian semi- and nonparametric modeling in R. Journal of Statistical Software, Articles 40 (5), 1–30.
Jogesh Babu, G., Canty, A.J., Chaubey, Y.P., 2002. Application of Bernstein Polynomials for smooth estimation of a distribution and density function. Journal of Statistical Planning and Inference 105 (2), 377–392.
Kauermann, G., Krivobokova, T., Fahrmeir, L., 2009. Some asymptotic results on generalized penalized spline smoothing. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71 (2), 487–503.
Klein, N., Hothorn, T., Barbanti, L., Kneib, T., 2021. Multivariate conditional transformation models. Scandinavian Journal of Statistics to appear.
Klein, N., Kneib, T., 2016. Simultaneous inference in structured additive conditional copula regression models: a unifying Bayesian approach. Statistics and Computing 26 (4), 841–860.
Klein, N., Kneib, T., Klasen, S., Lang, S., 2015a. Bayesian structured additive distributional regression for multivariate responses. Journal of the Royal Statistical Society: Series C 64 (4), 569–591.
Klein, N., Kneib, T., Lang, S., 2015b. Bayesian generalized additive models for location, scale, and shape for zero-inflated and overdispersed count data. Journal of the American Statistical Association 110 (509), 405–419.
Klein, N., Kneib, T., Lang, S., Sohn, A., 2015c. Bayesian structured additive distributional regression with an application to regional income inequality in Germany. The Annals of Applied Statistics 9 (2), 1024–1052.
Kneib, T., 2013. Beyond mean regression. Statistical Modelling 13 (4), 275–303.
Kneib, T., Klein, N., Lang, S., Umlauf, N., 2019. Modular regression - a lego system for building structured additive distributional regression models with tensor product interactions. TEST 28 (1), 1–39.
Kobyzev, I., Prince, S., Brubaker, M., 2020. Normalizing flows: An introduction and review of current methods. IEEE Transactions on Pattern Analysis and Machine Intelligence.
Koenker, R., 2005. Quantile Regression. Cambrigde University Press, New York. Economic Society Monographs
Koenker, R., Bassett, G., 1978. Regression quantiles. Econometrica 46, 33–50.
, 2020. In: Koenker, R., Chernozhukov, V., He, X., Peng, L. (Eds.), Handbook of Quantile Regression. CRC Press.
Kozumi, H., Kobayashi, G., 2011. Gibbs sampling methods for bayesian quantile regression. Journal of Statistical Computation and Simulation 81, 1565–1578.
Kübler, D., Weizsäcker, G., 2003. Information Cascades and the Labor Market. Journal of Economics 80 (3), 211–229.
Leisch, F., 2004. Flexmix: A general framework for finite mixture models and latent class regression in R. Journal of Statistical Software, Articles 11 (8), 1–18.
Li, Q., Racine, J.S., 2008. Nonparametric estimation of conditional cdf and quantile functions with mixed categorical and continuous data. Journal of Business and Economic Statistics 26, 423–434.
Manski, C.F., 1985. Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. Journal of Econometrics 27 (3), 313–333.
Manski, C.F., 1991. Regression. Journal of Economic Literature 29, 34–50.
Manuguerra, M., Heller, G.Z., 2010. Ordinal regression models for continuous scales. The International Journal of Biostatistics 6 (1).
Marra, G., Radice, R., 2017. Bivariate copula additive models for location, scale and shape. Computational Statistics & Data Analysis 112, 99–113.
Mayr, A., Fenske, N., Hofner, B., Kneib, T., Schmid, M., 2012. Generalized additive models for location, scale and shape for high dimensional data—a flexible approach based on boosting. Journal of the Royal Statistical Society: Series C 61 (3), 403–427.
Müller, P., Erkanli, A., West, M., 1996. Bayesian curve fitting using multivariate normal mixtures. Biometrika 83 (1), 67–79.
Neal, R.M., 2000. Markov chain sampling methods for dirichlet process mixture models. Journal of Computational and Graphical Statistics 9 (2), 249–265.
Nelder, J.A., Wedderburn, R.W.M., 1972. Generalized linear models. Journal of the Royal Statistical Society, Series A 135, 370–384.
Newcomb, S., 1886. A generalized theory of the combination of observations so as to obtain the best result. American Journal of Mathematics 8 (4), 343–366.
Newey, W.K., Powell, J.L., 1987. Asymmetric least squares estimation and testing. Econometrica: Journal of the Econometric Society 55 (4), 819–847.
Paindaveine, D., Šiman, M., 2012. Computing multiple-output regression quantile regions. Computational Statistics & Data Analysis 56, 840–853.
Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., Lakshminarayanan, B., 2019. Normalizing flows for probabilistic modeling and inference, arXiv:1912.02762.
Park, J., Shalit, U., Schölkopf, B., Muandet, K., 2021. Conditional distributional treatment effect with kernel conditional mean embeddings and u-statistic regression, arXiv:https://arxiv.org/abs/2102.08208
Pissarides, C., 1992. Loss of Skill during Unemployment and the Persistence of Unemployment Shocks. Quarterly Journal of Economics 107, 1371–1391.

Rigby, R.A., Stasinopoulos, D.M., 2005. Generalized additive models for location, scale and shape (with discussion). Journal of the Royal Statistical Society. Series C (Applied Statistics) 54, 507–554.
Rodrigues, T., Fan, Y., 2017. Regression adjustment for noncrossing Bayesian quantile regression. Journal of Computational and Graphical Statistics 26 (2), 275–284.
Rothe, C., 2012. Partial distributional policy effects. Econometrica 80 (5), 2269–2301.
Rothe, C., Wied, D., 2013. Misspecification testing in a class of conditional distributional models. Journal of the American Statistical Association 108 (501), 314–324.
Santos, B., Kneib, T., 2020. Noncrossing structured additive multiple-output Bayesian quantile regression models. Statistics and Computing 30 (4), 855–869.
Schnabel, S.K., Eilers, P., 2009. Optimal expectile smoothing. Computational Statistics & Data Analysis 53, 4168–4177.
Schulze Waltrup, L., Kauermann, G., 2017. Smooth expectiles for panel data using penalized splines. Statistics and Computing 27, 271–282.
Schulze Waltrup, L., Sobotka, F., Kneib, T., Kauermann, G., 2015. Expectile and quantile regression – David and Goliath? Statistical Modelling 15 (5), 433–456.
Serfling, R., 2002. Quantile functions for multivariate analysis: approaches and applications. Statistica Neerlandica 56, 214–232.
Sick, B., Hothorn, T., Dürr, O., 2020. Deep transformation models: Tackling complex regression problems with neural network based transformation models. arXiv:https://arxiv.org/abs/2004.00464.
Siegfried, S., Hothorn, T., 2020. Count transformation models. Methods in Ecology and Evolution 11 (7), 818–827.
Silbersdorff, A., 2017. Analysing Inequalities in Germany. Springer, Berlin.
Silbersdorff, A., Lynch, J., Klasen, S., Kneib, T., 2018. Reconsidering the income-health relationship using distributional regression. Health Economics 27 (7), 1074–1088.
Sobotka, F., Kauermann, G., Schulze Waltrup, L., Kneib, T., 2013. On confidence intervals for semiparametric expectile regression. Statistics and Computing 23 (2), 135–148.
Sobotka, F., Kneib, T., 2012. Geoadditive expectile regression. Computational Statistics & Data Analysis 56, 755–767.
Sohn, A., 2016. acid: R-Package for Analysing Conditional Income Distributions. https://cran.r-project.org/web/packages/acid/.
Sohn, A., 2017. The Regression of an Equitable Market Economy. Universität Göttingen, Göttingen Ph.D. thesis.
Spiegel, E., Sobotka, F., Kneib, T., 2017. Model selection in semiparametric expectile regression. Electronic Journal of Statistics 11 (2), 3008–3038.
Stadlmann, S., Kneib, T., 2021. Interactively visualizing distributional regression models with distreg.vis. Statistical Modelling. To appear
Stasinopoulos, D.M., Rigby, R.A., 2007. Generalized additive models for location, scale and shape (GAMLSS) in R. Journal of Statistical Software 23 (7).
Stasinopoulos, M.D., Rigby, R.A., Heller, G.Z., Bastiani, F.D., 2020. Distributions for Modeling Location, Scale and Shape: Using GAMLSS in R. Chapman & Hall/CRC.
Stasinopoulos, M.D., Rigby, R.A., Heller, G.Z., Voudouris, V., Bastiani, F.D., 2017. Flexible Regression and Smoothing: Using GAMLSS in R. Chapman & Hall/CRC.
Stöcker, A., Brockhaus, S., Schaffer, S.A., von Bronk, B., Opitz, M., Greven, S., 2021. Boosting functional response models for location, scale and shape with an application to bacterial competition. Statistical Modelling to appear.
Titterington, D., Smith, A., Makov, U., 1985. Statistical Analysis of Finite Mixture Distributions. Wiley, New York.
Umlauf, N., Kneib, T., 2018. A primer on Bayesian distributional regression. Statistical Modelling 18, 1–39.
van der Wurp, H., Groll, A., Kneib, T., Marra, G., Radice, R., 2020. Generalised joint regression for count data: a penalty extension for competitive settings. Statistics and Computing 30, 1419–1432.
Van Kerm, P., Yu, S., Choe, C., 2016. Decomposing quantile wage gaps: a conditional likelihood approach. Journal of the Royal Statistical Society: Series C (Applied Statistics) 65 (4), 507–527.
Vatter, T., Chavez-Demoulin, V., 2015. Generalized additive models for conditional dependence structures. Journal of Multivariate Analysis 141, 147–167.
Wagner, G.G., Frick, J.R., Schupp, J., 2007. The German socio-economic panel study (SOEP) - SCOPE evolution and enhancements. Schmollers Jahrbuch 127 (1), 139–169.
Waldmann, E., Kneib, T., 2015. Bayesian bivariate quantile regression. Statistical Modelling 15 (4), 326–344.
Waldmann, E., Kneib, T., Yue, Y.R., Lang, S., Flexeder, C., 2013. Bayesian semiparametric additive quantile regression. Statistical Modelling 13 (3), 223–252.
Waldmann, E., Sobotka, F., Kneib, T., 2017. Bayesian regularisation in geoadditive expectile regression. Statistics and Computing 27 (6), 1539–1553.
Wang, P., Puterman, M.L., Cockburn, I., Le, N., 1996. Mixed poisson regression models with covariate dependent rates. Biometrics 52 (2), 381–400.
Wood, S.N., 2003. Thin plate regression splines. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65 (1), 95–114.
Wood, S.N., Pya, N., Säfken, B., 2016. Smoothing parameter and model selection for general smooth models. Journal of the American Statistical Association 111 (516), 1548–1563.
Yee, T.W., 2015. Vector Generalized Linear and Additive Models: With an Implementation in R. Springer-Verlag, New York, U.S.A..
Yu, K., Moyeed, R.A., 2001. Bayesian quantile regression. Statistics & Probability Letters 54, 437–447.
Yue, Y., Rue, H., 2011. Bayesian inference for additive mixed quantile regression models. Computational Statistics and Data Analysis 55, 84–96.
Ziegel, J.F., 2016. Coherence and elicitability. Mathematical Finance 26 (4), 901–918.