ANNUAL REVIEWS

Annual Review of Statistics and Its Application

Distributional Regression for Data Analysis

Nadja Klein

Department of Statistics, Technische Universität Dortmund, Dortmund, Germany;
email: nadja.klein@tu-dortmund.de

Annu. Rev. Stat. Appl. 2024. 11:321–46

First published as a Review in Advance on October 13, 2023

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-040722-053607

Keywords

density regression, generalized additive model for location, scale and shape, probabilistic learning, quantile regression, transformation models

Abstract

The flexible modeling of an entire distribution as a function of covariates, known as distributional regression, has seen growing interest over the past decades in both the statistics and machine learning literature. This review outlines selected state-of-the-art statistical approaches to distributional regression, complemented with alternatives from machine learning. Topics covered include the similarities and differences between these approaches, extensions, properties and limitations, estimation procedures, and the availability of software. In view of the increasing complexity and availability of large-scale data, this review also discusses the scalability of traditional estimation methods, current trends, and open challenges. Illustrations are provided using data on childhood malnutrition in Nigeria and Australian electricity prices.

1. INTRODUCTION

GLM: generalized linear model

GAM: generalized additive model

GAMLSS: generalized additive models for location, scale and shape

Hothorn et al. (2014, p. 3) argue that "the ultimate goal of regression analysis is to obtain information about the conditional distribution of a response given a set of explanatory variables." While classical regression models focus on modeling the conditional mean of a response variable as a function of covariates, distributional regression aims at modeling the entire conditional distribution.

For instance, parametric approaches to distributional regression assume a specific parametric distribution. In that vein, generalized additive models for location, scale and shape (GAMLSS; Rigby & Stasinopoulos 2005), also called structured additive distributional regression models (Klein et al. 2015b), allow to relate each distributional parameter of an arbitrary parametric density to be a function of the covariate effects by following the general idea of generalized additive models (GAMs; Hastie & Tibshirani 1990). In fact, GAMLSS can be seen as a generalization of early attempts at distributional regression with specific distributions, such as double exponential family regression (Efron 1986), or the Box–Cox (Box & Cox 1964) and LMS methods (where L stands for the Box–Cox power λ, M for the mean μ, and S for the coefficient of variation σ; Cole 1988). Conditional transformation models (Hothorn et al. 2014) and distribution regression (Foresi & Peracchi 1995, Firpo et al. 2009, Chernozhukov et al. 2013) use flexible transformation functions to map the conditional distribution to a reference distribution, thereby exploiting distributional regression fully without requiring a parametric distribution assumption as with GAMLSS, for instance. Another branch of semiparametric distributional regression models is regression copulas, which rely on an implicit copula construction (Nelsen 2006, Klein & Smith 2019) combined with nonparametric marginals to arrive at a calibrated model. Nonparametric approaches to distributional regression include kernel methods, (finite) mixture models, and dependent Dirichlet process priors, often derived in a Bayesian framework (e.g., Escobar & West 1995, Dunson et al. 2007, Villani et al. 2012).

Quantile and expectile regression (Newey & Powell 1987, Koenker 2005) are alternative functionals to the mean that have been suggested in the literature. However, as pointed out by Henzi et al. (2021), the reduction to a single quantile or expectile can result in a considerable loss of information.

1.1. Distributional Regression: Why and When?

While in some situations it may be sufficient to consider the mean of a response variable—e.g., when the primary interest is in determining the relationship between a covariate and the expected outcome—there are many real examples in which the analyst is more concerned with quantiles, tails of the distribution, or prediction intervals. The following two are used as illustrations throughout this review article (see also the sidebar titled Illustrative Examples for more details).

Consider, for example, childhood malnutrition in developing countries. As described in Fenske et al. (2011), the use of a mean regression model implies that the estimated effects describe the nutritional status of an average child. However, it is of much greater interest to analyze the 5% or 10% quantiles of the response distribution, which relate to the risk of extreme malnutrition.

Another example is that of the relationship between intraday electricity prices and demand. Here, successful bidding on the markets requires accurate prediction of extreme events. As can be seen from **Figure 1**, price distributions are highly skewed (left panel of **Figure 1**) and vary in a complex manner over the day, by time of day, and with respect to demand (right panel of **Figure 1**). As a result, classical Gaussian regression is likely to provide a poor fit and to suffer from low predictive accuracy.

When investigating phenomena such as wage gaps, species diversity, the efficiency of markets, or risks in medicine, accuracy in modeling the entire distribution rather than just the mean

ILLUSTRATIVE EXAMPLES

- *Childhood malnutrition*: According to UNICEF (1998), childhood malnutrition is one of the most urgent public health problems in developing and transition countries, not only affecting child growth directly but also having severe long-term consequences. Illustrations in this review use data from the Nigeria Demographic and Health Survey, conducted in the 36 states and the capital of Nigeria in 2013, with *wasting*, an indicator of acute malnutrition (measured as insufficient weight for height), as response. **Figure 2** depicts the average *wasting* score (top left panel of **Figure 2**) and its empirical standard deviation (top right panel of **Figure 2**). The state of Kano comprises about 1,500 observations. A specific district was selected to keep the illustration simple and to avoid confounding with unobserved spatial effects. The covariates are *gender* and *age* of the child in months, and empirical variation across these covariates in the state of Kano is summarized in the bottom right and left panels of **Figure 2**.
- *Electricity spot prices*: For modeling the distribution of electricity spot prices, hourly data from January 1, 2018, to December 31, 2018, from the Australian National Electricity Market are used (available on **https://www.aemo.com.au**). The response is $y = \log(price + 101)$, where *price* is the market-wide price measured in Australian dollars per MWh; 101 was added because prices can be negative on the wholesale Australian National Electricity Market; and *time of day*, *day*, and *demand* are the covariates. **Figure 1** depicts the skewness of the price distribution (left panel of **Figure 1**) and variation over *time of day* from 1 a.m. (=1) to 12 midnight (=24) (right panel of **Figure 1**).

not only allows for much more realistic modeling assumptions but may also enable a much more comprehensive understanding of the relationship between response and covariates.

1.2. Article Outline

This review provides a selective overview of state-of-the-art methods in distributional regression with a focus on univariate real-valued responses. The field is so broad that the review cannot

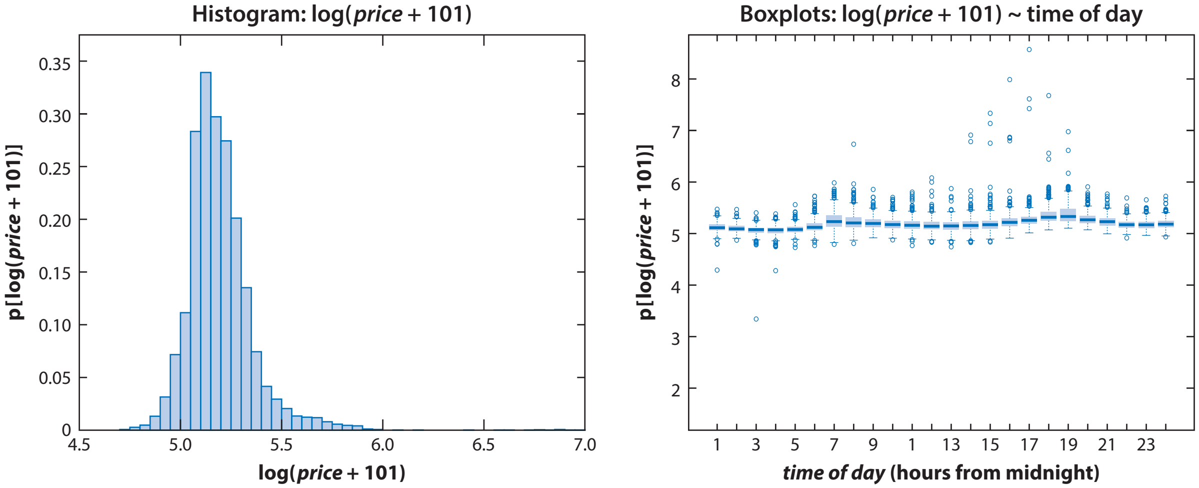

Figure 1

Electricity spot prices: (*Left*) Histogram of log (*price* + 101). (*Right*) Boxplots across *time of day*.

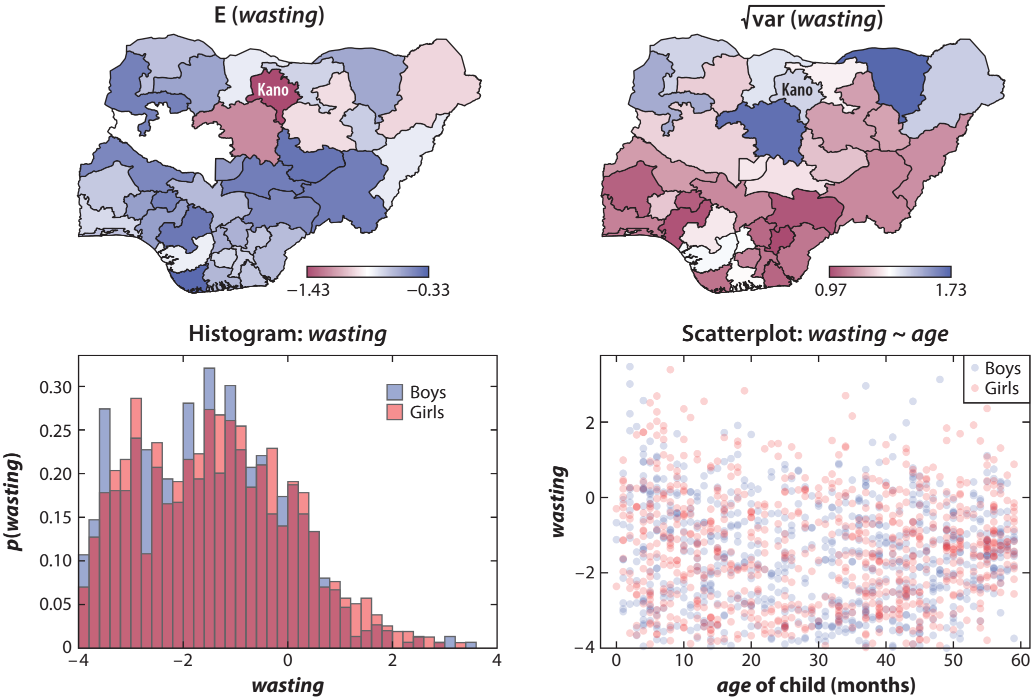

Figure 2

Childhood malnutrition: (*top left*) average and (*top right*) standard deviation of *wasting*. (*Bottom right*) Scatterplots and (*bottom left*) histogram of *wasting* for boys/girls (*blue/red*) in Kano, Nigeria.

appropriately acknowledge all the literature, but the aim is to provide a balanced article that complements the existing material. In Section 2, selected approaches to distributional regression and their developments are reviewed, also touching on inferential schemes, software and extensions, and the benefits and limitations of each approach. Section 3 draws connections to recent developments of density regression at the intersection with machine learning. Selective solutions for scalability of distributional approaches for large-scale data and highly parameterized models are covered in Section 4. The review closes with recent developments and trends for further thought and reading.

2. DISTRIBUTIONAL APPROACHES

Cumulative distribution function (CDF): defined as the function F, where $F(y) = \mathbb{P}(Y \leq y)$

Conditional CDF: the CDF $F(Y \mid \boldsymbol{X} = \boldsymbol{x}) = \mathbb{P}[Y \leq y \mid \boldsymbol{X} = \boldsymbol{x}]$ of a response Y as a function of observed covariates $\boldsymbol{X} = \boldsymbol{x}$

Assume that a real-valued response $Y \in \mathbb{R}$ depends on explanatory variables $\boldsymbol{X} = \boldsymbol{x} \in \mathbb{R}^p$ and that a data set of n realizations from the pairs $(Y, \boldsymbol{X} = \boldsymbol{x})$, denoted $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$, is available. Two modeling decisions have to be made. The first one determines the observation model, which specifies how to estimate the conditional cumulative distribution function (CDF) directly, indirectly, or functionals thereof, from the data:

$$\mathcal{L}(Y \mid \boldsymbol{X} = \boldsymbol{x}). \qquad 1.$$

Specific choices for $\mathcal{L}$ are discussed in the remainder of this section. The second decision determines how $\boldsymbol{x}$ enters the distributional model $\mathcal{L}$. Throughout this section, it is assumed that this is done through predictor(s) $\eta(\boldsymbol{x})$ of the following form:

$$\eta(\boldsymbol{x}) = \beta_0 + \sum_{j=1}^{J} f_j(\boldsymbol{x}_j). \qquad 2.$$

Probability density function (PDF): the density $p_Y = p(y)$

Conditional PDF: $p_{Y \mid X=x} = p(y \mid \boldsymbol{x})$

In the simple Gaussian mean model, the predictor models the conditional mean $\mathbb{E}(Y \mid \boldsymbol{x}) = \mu = \eta(\boldsymbol{x})$, and Equation 1 is defined through a Gaussian distribution with mean μ and variance σ^2, denoted by $\mathrm{N}(\mu(\boldsymbol{x}), \sigma^2)$. In Equation 2, β_0 is the overall level of the predictor and the f_j are smooth functions of subvectors $\boldsymbol{x}_j \in \mathbb{R}^{q_j}$, $q_j \leq p$, of $\boldsymbol{x}$ that are modeled through basis function expansions

$$f_j(\boldsymbol{x}_j) = \sum_{l=1}^{L_j} \beta_{j,l} B_{j,l}(\boldsymbol{x}_j) = \boldsymbol{B}_j(\boldsymbol{x}_j)^\top \boldsymbol{\beta}_j,$$

where $\boldsymbol{B}_j(\boldsymbol{x}_j) = (B_{j,1}(\boldsymbol{x}_j), \ldots, B_{j,L_j}(\boldsymbol{x}_j))^\top$ and $\boldsymbol{\beta}_j = (\beta_{j,1}, \ldots, \beta_{j,L_j})^\top$. The basis functions may correspond to the original covariates for linear effects, or to spline evaluations for nonlinear effects, though random effects, spatial effects, and others are also possible. Each effect may be regularized by a quadratic term of the form $\boldsymbol{\beta}_j^\top \boldsymbol{P}_j \boldsymbol{\beta}_j$, where $\boldsymbol{P}_j$ is a penalty matrix enforcing a data-driven amount of smoothness for the jth effect. Such predictors—also called structured additive predictors—are used in GAMs to model the conditional mean of the response. In Section 3, unstructured extensions of η allowing for even more flexible relationships between $\boldsymbol{x}$ and Y are considered.

2.1. Generalized Additive Models for Location, Scale, and Shape

GAMLSS were introduced by Rigby & Stasinopoulos (2005) and take a parametric approach to Equation 1.

2.1.1. Model specification. These models assume that the conditional CDF stems from a parametric density $p(y \mid \boldsymbol{x})$—the conditional probability density function (PDF)—with distributional parameters $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_K)^\top$. Each ϑ_k is modeled using a predictor η_k; that is, $\vartheta_k \equiv \vartheta_k(\boldsymbol{x}) = h_k[\eta_k(\boldsymbol{x})]$. The functions $h_k : \mathbb{R} \to \Theta_k \subset \mathbb{R}$ are strictly monotonically increasing response functions mapping the predictors into the parameter spaces $\Theta_1, \ldots, \Theta_K$.

The basic idea is similar to that in generalized linear models (GLMs) or GAMs, which can also be interpreted as distributional models, where the distribution is assumed to be an exponential family distribution and only a transformation of the conditional mean is related to covariates. In that vein, consider the malnutrition example with response *wasting* and *gender* and *age* of child as covariates along with a Gaussian varying-coefficient model. Then, the predictor is

$$\mu(gender, age) = h_1[\eta_1(gender, age)] = \beta_{1,0} + \beta_{1,1} gender + f_{1,1}(age) + gender \times f_{1,2}(age),$$

where $h_1(a) = a$ is the identity response function and $\sigma^2 = const$ is a nuisance parameter, implying homoscedastic errors. A more realistic extension that is supported by a lower Akaike information criterion (AIC) value would allow for heteroscedasticity by assuming

$$\sigma(gender, age) = h_2[\eta_2(gender, age)] = \exp[\beta_{2,0} + \beta_{2,1} gender + f_{2,1}(age) + gender \times f_{2,2}(age)],$$

where $h_2(\cdot) = \exp(\cdot)$ ensures positivity of σ, making the model a GAMLS (where L stands for location and S for scale).

The impact of the covariates and the modeling assumption for σ on the estimated probabilities of suffering from wasting and severe wasting, i.e., $-3 < wasting < -2$ and $wasting < -3$, respectively, is depicted in **Figure 3**. The figure shows that the risk of suffering from wasting

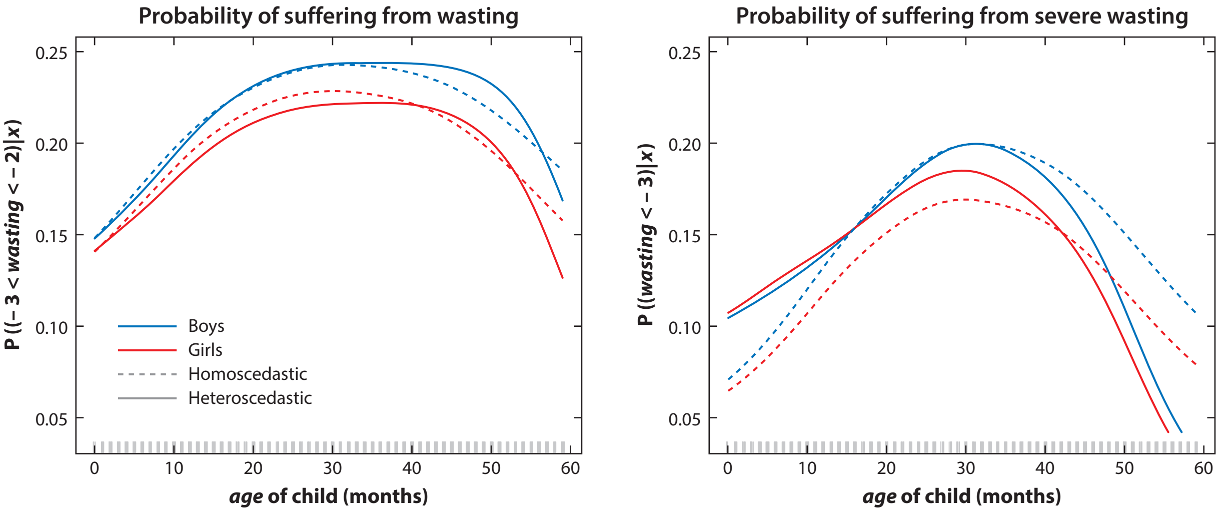

Figure 3

Childhood malnutrition: probabilities of suffering from wasting/severe wasting (*left*/*right*) for boys/girls (*blue*/*red*) using heteroscedastic/homoscedastic Gaussian models (*solid*/*dashed lines*) are shown. The Akaike information criterion supports the heteroscedastic model (5,428 versus 5,462).

is higher for boys than for girls. This difference can be seen as significant for ages between 30 and 40 months, as 95% equal-tailed posterior credible intervals for boys/girls (not shown; see Section 2.1.2 below for details) do not contain the posterior mean estimates of girls/boys. In addition, based on the AIC, the risk of severe wasting is underestimated, particularly for children younger than 10 months, and overestimated in children older than 40 months, when homoscedasticity is assumed.

2.1.2. Estimation and software.

Penalized maximum likelihood estimation (PMLE) (Rigby & Stasinopoulos 2005) for GAMLSS uses first- and second-order derivatives for iterative backfitting (Breiman & Friedman 1985) and provides three algorithms in the R package `gamlss`. Each step of the basic algorithm involves an outer loop over the K distributional parameters and an inner loop over the J_k effects for each ϑ_k. If less generality is needed with respect to Equation 1 or Equation 2, it may be better to use other PMLE-based packages. For instance, Lee & Nelder (2006) proposed a stable and efficient approach to fitting double hierarchical GLMs with random effects. Depending on the needs of analysts, it may be worth looking into the corresponding R packages `dhglm`, which can handle outliers through robust estimation, or `mgcv`. The latter offers numerically stable and convergent computational methods for selected GAMLSS-type distributions along with well-conceived options for smoothing parameter estimation. Numerical and convergence issues can occur with `gamlss` depending on the model and predictor complexities. It may be useful in such cases to review the options in the package manual or to simplify the models.

MLE: maximum likelihood estimation

PMLE: penalized MLE

MCMC: Markov chain Monte Carlo

A Bayesian implementation using Markov chain Monte Carlo (MCMC) simulations with iteratively weighted least squares proposals (Klein et al. 2015b) has been implemented in the standalone software BayesX, as well as in the more user-friendly R package `bamlss`. An advantage of this Bayesian treatment is the direct access to uncertainty quantification through equal-tailed posterior credible intervals using the MCMC samples. Functional gradient boosting (Mayr et al. 2012) for GAMLSS is available in the R package `gamboostLSS`. Boosting can easily handle high-dimensional data settings with $p \gg n$ but does not provide standard errors.

2.1.3. Properties and extensions. Whether asymptotic results for GAMs or heteroscedastic regression (e.g., those of Kauermann et al. 2009, Wang 2013) generalize to GAMLSS has not yet been investigated. Empirical evidence (e.g., Klein et al. 2015b) indicates that the coverage of confidence intervals is often below the nominal level, though this may be connected to the software. However, Klein et al. (2015b) demonstrate with simulations that credible intervals from the MCMC output yield accurate coverage and give sufficient conditions for the propriety of posterior distributions.

GAMLSS are closely related to vector generalized additive models (Yee 2015). Special cases of GAMLSS include the nonlinear heteroscedastic regression (Yau & Kohn 2003). The double exponential family regression with predictors for mean and variance was introduced by Efron (1986) and extended to the smoothing context by Gijbels et al. (2010). The popular LMS method followed shortly after that (see Carroll & Ruppert 1988).

Considerable research in the GAMLSS model class extends the predictors η. Examples include LASSO regularization, functional covariates, and censoring.

2.1.4. Final remarks. On the one hand, GAMLSS make a fully parametric assumption for the conditional CDF, implying a fixed type of response distribution for all observations, which may be too restrictive in some applications. On the other hand, the availability of the parametric likelihood makes it straightforward to implement GAMLSS for noncontinuous data as well.

While GAMLSS can be favorable in applications in which a reasonable parametric choice for the response distribution can be made [e.g., in the context of ensemble methods in meteorology (Gneiting et al. 2005)], the analyst should be aware of the following challenges. First, for many parametric distributions, parameters do not correspond to central moments. Instead, they are often general location, scale, or shape parameters determining the conditional CDF/PDF and functionals thereof, which makes interpretation more difficult. Second, the parameterization of the conditional PDF affects the interpretation of estimated effects [for the Gaussian case, one could model either $\sigma(\boldsymbol{x})$ or $\sigma^2(\boldsymbol{x})$ through a predictor].

Third, model and variable selection are challenging, and it is not feasible to compare all candidate models. Klein et al. (2015b) thus suggest a pragmatic strategy to select a model based on three tools: (*a*) Randomized quantile residuals of fitted models can be used for visual comparison of response distributions with fixed predictor specifications. (*b*) Different predictor specifications can be compared using information criteria. (*c*) Proper scoring rules (Gneiting & Raftery 2007) can be used to evaluate the predictive ability of the models. In `gamlss`, further graphical tools such as worm plots for outlier detection are also available. Klein et al. (2021) extended the use of spike and slab priors to enable automatic Bayesian effect selection for GAMLSS.

2.2. Conditional Transformation Models

Transformation models aim to make the data follow certain modeling assumptions, such as normality or homoscedasticity, through transformation of the response. This idea is exemplified for conditional transformation models (Hothorn et al. 2014) in the following.

2.2.1. Model specification. By specifying a monotonically increasing and covariate-dependent transformation function $h(y\,|\,\boldsymbol{x}) : \mathcal{S} \to \mathbb{R}$ and a reference distribution $F_Z : \mathbb{R} \to [0, 1]$ that is independent of $\boldsymbol{x}$ and does not contain any parameters to be estimated, a conditional transformation model relies on the model formulation

$$F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y) = \mathbb{P}\left[Y \leq y \mid \boldsymbol{X} = \boldsymbol{x}\right] = \mathbb{P}\left[h(Y \mid \boldsymbol{x}) \leq h(y \mid \boldsymbol{x})\right] = F_Z\left[h(y \mid \boldsymbol{x})\right].$$

Following Hothorn et al. (2014), an additive decomposition on the scale of h into J partial transformation functions is assumed—i.e., $h(y\,|\,\boldsymbol{x}) = \beta_0 + \sum_{j=1}^{J} h_j(y\,|\,\boldsymbol{x})$ that takes the role of the predictor (Equation 2). The partial transformations $h_j(y\,|\,\boldsymbol{x})$ can be understood as response-covariate interactions. This is similar to the regression structure of GAMs, but rather than modeling the conditional mean of the response, h acts on the transformed response scale.

To apply conditional transformation models, one must choose the reference distribution F_Z and the transformation function h. Reference distributions with log-concave densities ensure concave likelihoods and thus give unique MLEs under mild regularity conditions. A common choice is the standard Gaussian distribution, but other options can be useful, such as the maximum extreme value distribution to arrive at a Cox-type model.

The transformation function $h(y\,|\,\boldsymbol{x})$ must be a monotonic function of y. This is particularly easy to implement in so-called shift conditional transformation models, in which only the location varies with $\boldsymbol{x}$, and of which the Gaussian linear model $Y|X = x \sim \mathrm{N}(\beta_0 + \beta_1 x, \sigma^2)$ is a special case, with F_Z the standard Gaussian distribution function and the transformation function $h(y\,|\,x) = y/\sigma - (\beta_0 - \beta_1 x)/\sigma$. The covariate-dependent shift $(\beta_0 - \beta_1 x)/\sigma$ can be estimated without monotonicity constraints. It is easy to see how more complex models can be built. For instance, adding the term $\beta_2 yx/\sigma$ to h allows the scale of the response to vary with x, while higher-order interactions enable the modification of other distribution shape features. If h increases strictly monotonically with derivative h', and with p_Z the PDF of Z, the implied conditional PDF is

$$p(y \mid \boldsymbol{x}) = p_Z[h(y \mid \boldsymbol{x})]\left|\frac{\partial h(y \mid \boldsymbol{x})}{\partial y}\right| = p_Z[h(y \mid \boldsymbol{x})]h'(y \mid \boldsymbol{x}). \qquad 3.$$

Conditional transformation models are compared with quantile regression in Section 2.5.

2.2.2. Estimation and software. The original proposal of conditional transformation models by Hothorn et al. (2014) was developed from a sequence of binary indicator regressions for $\mathbb{E}[\mathbb{1}(Y \leq \upsilon)\,|\,\boldsymbol{X} = \boldsymbol{x}]$, similar to the work of Foresi & Peracchi (1995). However, rather than estimating a sequence of models for a grid of υ values, Hothorn et al. (2014) perform joint optimization through scoring rules with estimation using a variant of componentwise gradient boosting. Boosting does not allow for estimation of standard errors without computationally costly resampling-based methods, and the implementation of Hothorn et al. (2014) works for continuous responses only. However, based on the implied conditional PDF in Equation 3, likelihood-based estimation can easily be developed and can allow for discrete or censored responses (Hothorn et al. 2018). These models are implemented in the R package `tram`, which provides formula-based user interfaces to specific likelihood-based transformation models implemented in `mlt`. It is supplemented with a website containing references, a list of available models, and numerous vignettes. Carlan et al. (2023) proposed a Bayesian approach for likelihood-based conditional transformation models using MCMC, for which code is available on GitHub (**https://github.com/manucarl/BCTM**).

2.2.3. Properties and extensions. Hothorn et al. (2014) prove the consistency of boosted conditional transformation models, whereas they practically ignore the monotonicity constraints. Hothorn et al. (2018) prove consistency and asymptotic normality of the MLE estimator. Posterior consistency in the Bayesian approach has not yet been investigated.

Traditionally, transformation models were used for ordered categorical or censored responses. A closely related approach in the context of counterfactual distributions is distribution regression

(Chernozhukov et al. 2013). These models use $\mathbb{P}(Y \leq y \mid \boldsymbol{X} = \boldsymbol{x}) = \mathbb{E}[\mathbb{1}(Y \leq y) \mid \boldsymbol{X} = \boldsymbol{x}]$ to estimate response-varying effects with transformation functions of the form $h(y \mid \boldsymbol{x}) = h_Y(y) - \boldsymbol{x}^\top \boldsymbol{\beta}(y)$ (Foresi & Peracchi 1995) and are used particularly in econometrics (Rothe & Wied 2013, Delgado et al. 2022). Since the popular parametric transformation model, the so-called Box–Cox model (Box & Cox 1964) was proposed, transformation models have been extended and further developed in several directions (for comprehensive literature overviews, see, e.g., Hothorn et al. 2014, Carlan et al. 2023).

Copula C: multivariate CDF with standard uniform margins

Marginal calibration: can be interpreted as equality of forecast and reality; Gneiting et al. (2007) provide a formal definition

2.2.4. Final remarks. Conditional transformation models are a semiparametric approach to distributional regression due to the semiparametric structure of h. Thus, however, interpretation is different to that in Gaussian mean regression models, where signal and noise are decomposed additively directly on the response level.

Each partial transformation function h_j is commonly assumed to increase monotonically, which is sufficient but not necessary for h to be monotone, and Bernstein polynomials or B-splines have been used with appropriate constraints to model h.

Finally, as stated by Hothorn (2018) in the most general form of h, conditional transformation models subsume several simpler models, which can be particularly useful in the context of model choice.

2.3. Regression Copulas

A second very recent semiparametric approach that has potential but is far less well understood is that of regression copulas.

2.3.1. Model specification. Copulas can be used to capture nonlinear dependence structures between multivariate random variables (Nelsen 2006). Sklar's theorem states that every multivariate CDF can be represented by a copula evaluated at the marginal CDFs. Copula regression models use this property and have been implemented widely (Pitt et al. 2006, Song et al. 2009, Craiu & Sabeti 2012, Krämer et al. 2013). Regression copulas proposed by Smith & Klein (2021), however, capture the dependence between multiple observations on a single dependent variable Y, conditional on $\boldsymbol{x}$. Their model defines a copula process (Wilson & Ghahramani 2010) on the covariate space. It is the implicit copula C extracted from the joint distribution of a regression model with latent response $\tilde{Z}$ (for a recent review on implicit copulas, see Smith 2023). The model of Smith & Klein (2021) furthermore assumes invariant marginals for Y to arrive at a distributional regression model that is approximately marginally calibrated.

For data $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ with covariate matrix $\boldsymbol{X} = (\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_n^\top)^\top$, Smith & Klein (2021) model the joint distribution of $\boldsymbol{Y}$ through the density

$$p(\boldsymbol{y} \mid \boldsymbol{X}) = c[F_Y(y_1), \ldots, F_Y(y_n); \boldsymbol{X}, \boldsymbol{\theta}] \prod_{i=1}^{n} p_Y(y_i), \qquad 4.$$

where c is the implicit copula density derived from the model $\tilde{Z}_i = \eta(\boldsymbol{x}_i) + \varepsilon_i$, $\tilde{Z}_i$ is a latent response, and $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$ for $i = 1, \ldots, n$. The parameter vector $\boldsymbol{\theta}$ denotes the copula parameters specified below. The density p_Y is the marginal PDF of Y_i, which is assumed to be invariant with respect to i and $\boldsymbol{X}$ and which can be estimated nonparametrically. Let F_Y be the respective marginal CDF.

The key question is the construction of $c(\cdot; \boldsymbol{X}, \boldsymbol{\theta})$. To derive c, the regression coefficients, say $\boldsymbol{\beta}$, in $\eta(\boldsymbol{X}) = \boldsymbol{B}\boldsymbol{\beta}$ are supplemented with Gaussian prior distributions $\boldsymbol{\beta} \mid \sigma^2, \boldsymbol{\theta} \sim \mathrm{N}[\boldsymbol{0}, \sigma^2 \boldsymbol{P}(\boldsymbol{\theta})^{-1}]$, where $\boldsymbol{B}$ is a design matrix with row-wise basis function evaluations $\boldsymbol{B}_{(i)} = (B_1(\boldsymbol{x}_i), \ldots, B_L(\boldsymbol{x}_i))^\top$ and $\boldsymbol{P}(\boldsymbol{\theta})$ is a prior precision matrix with parameters $\boldsymbol{\theta}$. If $\boldsymbol{P}(\boldsymbol{\theta})$ is of full rank, the distribution of $\tilde{\boldsymbol{Z}} \mid \boldsymbol{X}, \boldsymbol{\theta}$

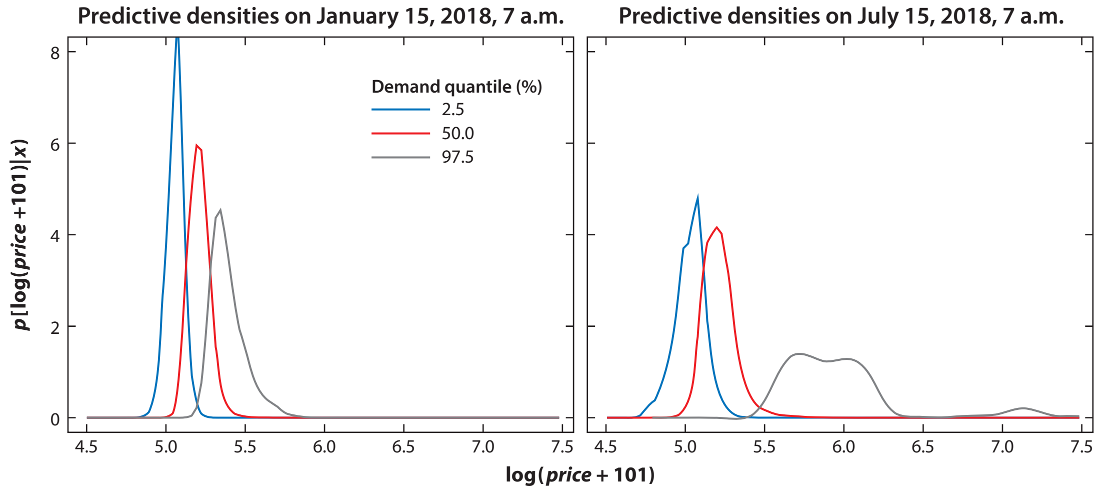

Figure 4

Electricity spot prices: predictive densities on (*left*) January 15 and (*right*) July 15, 2018, at 7 a.m. for three demand quantiles: 2.5% (*blue*), 50% (*red*), and 97.5% (*gray*).

is Gaussian with zero mean vector and covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{B}, \boldsymbol{\theta}, \sigma^2) = \sigma^2(\boldsymbol{I} + \boldsymbol{B}\boldsymbol{P}(\boldsymbol{\theta})\boldsymbol{B}^\top)$. It is easy to show that $\tilde{\boldsymbol{Z}} \,|\, \boldsymbol{X}, \boldsymbol{\theta}$ has a Gaussian copula with correlation matrix $\boldsymbol{R} \equiv \boldsymbol{R}(\boldsymbol{B}, \boldsymbol{\theta}) = \sigma^{-2}\boldsymbol{S}\boldsymbol{\Sigma}\boldsymbol{S}$, which is obtained by standardizing $\tilde{Z}_i$ to have unit variance as follows. Let $\boldsymbol{S} = \text{diag}(s_1, \ldots, s_n)$, be the diagonal matrix with $s_i = (1 + \boldsymbol{B}_{(i)}^\top \boldsymbol{P}(\boldsymbol{\theta})\boldsymbol{B}_{(i)})^{-1/2}$ and $Z_i = s_i\tilde{Z}_i/\sigma$. Then, with $u_i = F_Y(y_i)$ and $z_i = \Phi^{-1}(u_i)$, the copula PDF in Equation 4 is

$$c[F_Y(y_1), \ldots, F_Y(y_n); \boldsymbol{X}, \boldsymbol{\theta}] = \frac{\phi_n(\boldsymbol{z}; \boldsymbol{0}, \boldsymbol{R})}{\prod_{i=1}^n \phi(z_i)},$$

where ϕ and Φ^{-1} are the standard Gaussian PDF and quantile function and $\phi_n(\cdot; \boldsymbol{0}, \boldsymbol{R})$ is the n-dimensional Gaussian PDF with zero mean and correlation matrix $\boldsymbol{R}$, which is a function of $\boldsymbol{x}$ and $\boldsymbol{\theta}$.

To see how Equation 4 defines a distributional model, consider the predictive density for a new response y_{n+1} with covariate vector $\boldsymbol{x}_{n+1}$,

$$p(\boldsymbol{y}_{n+1} \mid \boldsymbol{x}_{n+1}) = p\left[F_Y(Y_{n+1}) \mid F_Y(y_1), \ldots, F_Y(y_n); \boldsymbol{X}, \boldsymbol{x}_{n+1}, \boldsymbol{\theta}\right] p_Y(y_{n+1}),$$

which is a function of $\boldsymbol{X}$ and $\boldsymbol{x}_{n+1}$. Moreover, the entire distribution is a function of $\boldsymbol{x}_{n+1}$. This is illustrated in **Figure 4**, which depicts predictive densities of log(*price* + 101) at 7 a.m. on a day in summer (left panel of **Figure 4**) and winter (right panel of **Figure 4**) in Australia for three demand quantiles. Increases in demand accentuate the upper tail of the distribution. For the chosen days and time, the densities in the summer are sharper than those in the winter. Results were obtained by fitting the heteroscedastic regression copula of Smith & Klein (2021) using a three-dimensional thin plate spline with covariates *day*, *time of day*, and *demand* and a horseshoe prior (Carvalho & Polson 2010), where $\beta_l \,|\, \lambda_l \sim \text{N}(0, \lambda_l^2)$, $\lambda_l \,|\, \tau \sim$ Half-Cauchy(0, τ), $\tau \sim$ Half-Cauchy(0, 1), $l = 1, \ldots, L$, and $\boldsymbol{\theta} = \{\lambda_1, \ldots, \lambda_L, \tau\}$.

2.3.2. Estimation and software.

Although the likelihood derived from Equation 4 is available in closed form, evaluating and inverting the $n \times n$ matrix $\boldsymbol{R}(\boldsymbol{B}, \boldsymbol{\theta})$ is computationally demanding for large n. Instead, it is more efficient to use the likelihood conditional on $\boldsymbol{\beta}$ and integrate out $\boldsymbol{\beta}$ using an MCMC scheme. Better scalable approximate Bayesian methods, namely variational Bayes (Blei et al. 2017; see also Section 4), have been developed by Smith & Klein (2021).

Unfortunately, no R or other package is available to apply this approach yet, though the models can be used through existing Matlab code. The use and visibility of these models in applied research would certainly be helped by the development of a user-friendly package including tutorials and guidelines on when and why to use regression copulas.

2.3.3. Properties and extensions. In Equation 4, the marginals are assumed to be independent of the covariates, while the copula is not. This approach defines a copula process for a univariate response and guarantees the efficacy of regression copulas. However, theoretical results like posterior consistency have not been investigated so far.

The model is semiparametric because the implicit copula c is constructed from the model for the pseudo response $\tilde{Z}$, which has a conditional mean parameterized through the semiparametric predictor $\eta(\boldsymbol{x})$. Due to the invariance assumption, the marginal distribution p_Y can be estimated nonparametrically.

Klein & Smith (2019) suggest forming the regression copula of smoothing models with a large number of basis functions, such as radial or P-spline bases, which they call a copula smoother. Klein & Smith (2021) derive the implicit copula of a linear Bayesian variable selection model, thereby extending the popular Bayesian variable selection approach for linear regression to non-Gaussian responses.

The inherited calibration property has recently been demonstrated to be advantageous for likelihood-free inference (Sisson et al. 2018) and for probabilistic forecasting in time series, when the dependent variable is highly non-Gaussian and skewed (Klein et al. 2023). A comprehensive review on regression copulas is provided by Smith (2023).

2.3.4. Final remarks. This model class is not well understood theoretically, and further investigation is required to determine when regression copulas should be preferred over other distributional approaches. Like conditional transformation models, regression copulas define a semiparametric approach that does not require a parametric distribution. Regression coefficients enter the auxiliary model for $\tilde{Z}$ and are then mapped nonlinearly through the marginal distribution p_Y. This complicates the direct interpretation of covariate effects, even though any quantity of interest can be derived from the global model in Equation 4.

As mentioned above, it is more common to employ copula regression models for multivariate response variables, where the regression margins are specified separately and depend on the covariates (see also Section 2.6). A different approach is copula-based regression (Noh et al. 2013), in which the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{X}$ is derived from assumptions on the joint distribution of Y and a random vector $\boldsymbol{X}$.

2.4. Density Regression

Similar to GAMLSS and regression copulas, density regression involves the formulation of a direct model for the conditional PDF. The focus here is on density regression based on mixture models (dating back to Newcomb 1886), with some alternative approaches presented at the end of this section.

2.4.1. Model specification. Because mixtures of a sufficiently large number of Gaussian densities can approximate any smooth density, consider a mixture of regression models formed of Gaussian densities $p(y \mid \boldsymbol{x}, \boldsymbol{\vartheta}) = \mathrm{N}[y; \mu(\boldsymbol{x}), \sigma(\boldsymbol{x})^2]$. The conditional PDF is then specified as

$$p(y \mid \boldsymbol{x}) = \int_{\Theta} p(y \mid \boldsymbol{x}, \boldsymbol{\vartheta}) \mathrm{d}G(\boldsymbol{x}, \boldsymbol{\vartheta}),$$

where G is a mixture distribution that can vary with $\boldsymbol{x}$ and $\boldsymbol{\vartheta} = (\mu(\boldsymbol{x}), \sigma(\boldsymbol{x})^2)^\top$. Gaussian regression models and finite mixtures of Gaussian regression models (Frühwirth-Schnatter 2006), given by

$$p(y \mid \boldsymbol{x}) = \sum_{h=1}^{H} \pi_h(\boldsymbol{x}) \mathrm{N}(y;\, \mu_h(\boldsymbol{x}), \sigma_h(\boldsymbol{x})^2), \qquad 5.$$

also referred to as latent class regression, with weights $\pi_h(\boldsymbol{x})$ characterized parametrically, are special cases thereof. An elegant way to avoid having to choose the number of components H is to resort to an infinite-dimensional Bayesian model with $H = \infty$ and G a random measure following a Dirichlet process, $G \sim \mathrm{DP}(\alpha G_0)$, with base measure G_0 and precision α. MacEachern & Müller (1998) use a dependent Dirichlet process, which relies on the stick-breaking form $G = \sum_{h=1}^{\infty} \pi_h(\boldsymbol{x}) \delta_{\boldsymbol{\vartheta}_h}$, with degenerate distribution $\delta_{\boldsymbol{\vartheta}_h}$ with all its mass on $\boldsymbol{\vartheta}_h$.

2.4.2. Estimation and software.

Empirical Bayesian density regression has been developed by Dunson (2007), while Dunson et al. (2007) present a comprehensive overview of Bayesian density regression and develop efficient MCMC-based posterior inference using a generalized Pólya urn scheme. This results in a Gibbs sampling algorithm that allows for weights depending on the distance between observed covariates. Generic MCMC algorithms with Metropolis–Hastings updates have been developed by Villani et al. (2012) to allow for covariate-dependent weights, nonlinear regression specifications, and automatic variable selection.

Finite mixture regression models have also been treated in a frequentist framework using the expectation–maximization algorithm (Grün & Leisch 2008). In this approach, the weights are covariate-independent, because parameters need to be optimized under constraints. In addition, it is more difficult to derive measures of uncertainty.

Various software packages are available for density regression with mixture models. For instance, the R packages `BNPmix` and `DPpackage` (no longer available on CRAN, the Comprehensive R Archive Network) implement functions to perform inference via simulation from the posterior distributions for Bayesian nonparametric and semiparametric mixture models, and the R packages `flexmix` and `mixtools` provide point estimates for likelihood-based estimation of finite mixtures of GLMs using the expectation–maximization algorithm.

Figure 5 illustrates predictive densities obtained from a three-component mixture model at two selected ages for boys/girls using `flexmix`. The models have linear specifications in the local means and scales and mixture weights not depending on the covariates; $H = 3$ was selected using the Bayesian information criterion.

2.4.3. Properties and extensions.

Aragam & Yang (2022) prove uniform consistency in nonparametric mixture models and provide a detailed review of further consistency results in the context of density regression through mixture models. References to available consistency results in the Bayesian framework are provided by Frühwirth-Schnatter et al. (2019).

Estimating the number of components H in the finite mixture model has been an ongoing research problem in the literature. In practice, it is common to compare performance under a set of distinct H values or to use a sufficiently large H, arguing that redundant components should empty out (for a recent theoretical investigation and further references, see also Manole & Khalili 2021).

The models can be thought of as infinite mixtures of Gaussian regression models, where both the weights associated to the mixture components and the parameters of each component are covariate dependent. However, such flexibility comes at a computational cost; moreover, there is limited availability of algorithms for posterior inference (Griffin & Steel 2006). As a consequence, the single-weights-dependent Dirichlet process mixture of normals model with covariate-independent weights is very popular in practice. Often, σ_h is assumed to be independent

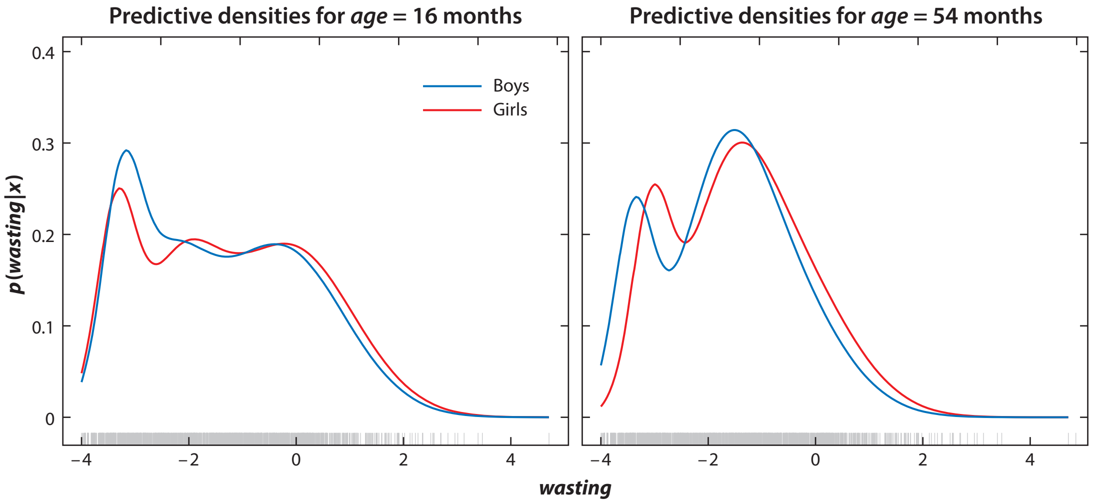

Figure 5

Childhood malnutrition: predictive densities from a finite mixture model with $H = 3$ components for two selected *age* values for boys/girls (*blue*/*red*).

of $\boldsymbol{x}$, and the specification of μ_h is linear (Escobar & West 1995), despite the resulting lack of flexibility (MacEachern & Müller 1998). In addition, Orlandi et al. (2021) state that empirical results suggest that allowing the variance to depend on $\boldsymbol{x}$ can reduce the required number of components.

As a trade-off enabling computationally efficient estimation and flexibility, de Carvalho et al. (2019) propose replacing the linear specification of μ_h by penalized splines and demonstrate that this has considerable advantages over using covariate-dependent weights. This idea could also be extended to σ_h. Further uses of the dependent Dirichlet process include analysis-of-variance type dependence structures (Iorio et al. 2004) or spatial analyses (Gelfand et al. 2005); Quintana et al. (2020) provide a review on dependent Dirichlet processes.

2.4.4. Final remarks. Density regression based on mixture models can be used with many data types, particularly with data that cannot be approximated by simple parametric distributions. However, the regression coefficients of such models are generally harder to interpret. Yet, in some special cases, such as Gaussian mixture regression models with constant weights, the linear regression effects on the means have interpretations as localized means. Although mostly derived for continuous responses and Gaussian densities, extensions to discrete, bounded, or mixed responses are possible. Beyond that, finite mixture models can of course also be used for clustering, e.g., using the R package `mclust`.

Alternatives to mixture models include *n*-nearest-neighbor methods combined with a suitable notion of distance to similarities (e.g., Stone 1977), or kernel smoothing methods (Hyndman et al. 1996, Hall et al. 1999).

2.5. Quantile Regression

Compared with the approaches presented so far that rely on a global model by imposing structure on the conditional CDF or PDF, quantile regression specifies a local model through the conditional quantile function $Q(\tau \mid \boldsymbol{x})$ at quantile level $\tau \in (0, 1)$.

Quantile function: $Q(\tau) = \inf\{y: \tau \leq F(y)\}$

Conditional quantile function: $Q(\tau \mid \boldsymbol{x})$

2.5.1. Model specification. Consider the model $y = \eta_\tau(\boldsymbol{x}) + \varepsilon_\tau$, where both the predictor $\eta_\tau(\boldsymbol{x})$ and the error term ε_τ depend on the quantile τ. Rather than specifying a global CDF for all τ, one assumes that the CDF of the error term at zero is equal to τ; in other words, $Q_{\varepsilon_\tau}(\tau) = 0$. Consequently, the predictor η_τ determines the conditional quantile function of y at τ, $Q(\tau \mid \boldsymbol{x}) = \eta_\tau(\boldsymbol{x})$. The idea is similar to mean regression where $\mathbb{E}(\varepsilon) = 0$ is assumed.

2.5.2. Estimation and software. As proposed by Koenker & Bassett (1978) for linear predictors, estimates for $Q(\tau \mid \boldsymbol{x})$ are obtainable by minimizing the weighted L_1 loss

$$\begin{gathered}\sum_{i=1}^{n} w_{i,\tau} |y_i - Q(\tau \mid \boldsymbol{x}_i)|, \\ w_{i,\tau} = \begin{cases} 1-\tau, & y_i < Q(\tau \mid \boldsymbol{x}_i), \\ \tau, & y_i \geq Q(\tau \mid \boldsymbol{x}_i), \end{cases}\end{gathered} \qquad 6.$$

by linear programming.

A Bayesian approach to quantile regression requires a working likelihood, and the asymmetric Laplace distribution, as suggested by Yu & Moyeed (2001) and further developed by Kozumi & Kobayashi (2011) and Yue & Rue (2011), has become popular. However, plenty of alternatives have been considered as well (see, e.g., Yang et al. 2016). A further alternative to estimate quantile regression is boosting, which requires only the gradients of Equation 6, which are easy to derive (Fenske et al. 2011). The R package quantreg includes an implementation of linear quantile regression and extensions such as quantile smoothing splines. The general framework of Fasiolo et al. (2021) is implemented in the R package qgam. Bayesian and boosting options are available in BayesX, the R package bamlss, and the mboost package.

2.5.3. Properties and extensions. The asymptotic theory for linear quantile regression has been developed under both a fixed and an increasing number of regressors, including the high-dimensional case (e.g., Koenker 2005, Belloni & Chernozhukov 2011, and references therein). Some asymptotic results are also available for longitudinal data (Lamarche & Parker 2023) and special classes of nonparametric quantile regression (Takeuchi et al. 2006). However, these situations are less well understood. Posterior consistency for Bayesian linear quantile regression with asymmetric Laplace density is developed by Sriram et al. (2013), with posterior variance adjustments for small samples and censored data developed by Yang et al. (2016). Consistency results for boosting are currently limited, and existing results in the literature on gradient boosting assume convexity of the objective function (e.g., Biau & Cadre 2021, Vethoen et al. 2023).

It is worth noting that one has to estimate different models for a possibly dense grid of τ values and combine them numerically to arrive at a smoothed estimate of the CDF. Doing so separately does not ensure the natural ordering across quantiles, so quantile crossings can cause inconsistent global models. This issue has generated long-standing interest; for example, He (1997), Bondell et al. (2010), and Rodrigues & Fan (2017) provide solutions in the different estimation frameworks.

The error terms are not assumed to be independent and identically distributed, only conditionally independent. Hence, quantile regression can, in principle, recover any covariate-specific changes in the shape of the conditional CDF on a specific quantile. Of course, misspecifying the regression predictors can still be a concern to be considered.

The Bayesian treatment of Yue & Rue (2011) and Waldmann (2018) allows flexible structured additive predictor specifications and Dirichlet process mixture priors for random effects models

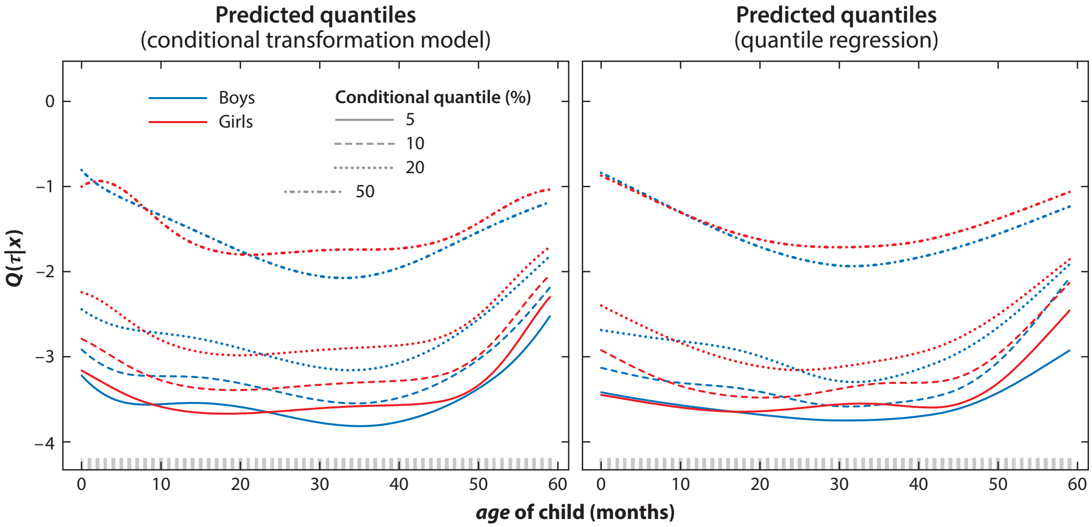

Figure 6

Childhood malnutrition: predicted conditional quantiles $Q(\tau \mid \boldsymbol{x})$ with *gender* and *age* of child (in months; x-axis) as covariates $\boldsymbol{x}$ via (*left*) a conditional transformation model and (*right*) quantile regression. The quantile levels are $\tau = \{0.05, 0.1, 0.2, 0.5\}$.

Flexible non-Bayesian alternatives have also been developed using, for example, refinements of the penalty term or manual tuning of smoothing parameters (for details and further references, see, e.g., Koenker et al. 2017). Fasiolo et al. (2021) developed fast additive quantile regression that offers a flexibility similar to that of GAMs (including estimation of smoothing parameters) using a smoothed version of the quantile loss in Equation 6.

2.5.4. Final remarks. Quantile regression has been used in many applications (particularly in economics; e.g., Rossi & Harvey 2009) and can be appealing when the primary interest is in deriving prediction intervals or in the presence of outliers. The choice between quantile regression and a global model may also be based on other more practical aspects, such as interpretability. When and why to opt for quantile regression are discussed by Waldmann (2018).

To illustrate quantile regression and its close relation to conditional transformation models, **Figure 6** depicts predicted conditional quantiles at levels $\tau = \{0.05, 0.1, 0.2, 0.5\}$ of the *wasting* score for both approaches. Quantiles obtained by fitting separate nonlinear varying coefficient quantile regressions using *gender* and *age* of child as covariates are shown in the right panel of **Figure 6**. Quantile crossing is not an issue here. Quantiles derived from a conditional transformation model with a flexible response-interacting transformation function (left panel of **Figure 6**) yield conceptually similar results, and by construction, quantiles do not cross.

Finally, an alternative to quantile regression is to replace the L_1 distance in Equation 6 by the L_2 loss, which leads to minimizers called expectiles (Newey & Powell 1987). A one-to-one transformation exists between quantiles and expectiles (Waltrup et al. 2015).

2.6. Multivariate Response Models

Recent advances in distributional regression include those for multivariate responses. For GAMLSS with low-dimensional $Y \in \mathbb{R}^D$, $D \leq 3$, using either parametric distributions (Klein et al. 2015a) or one-parameter copulas (e.g., Craiu & Sabeti 2012, Marra & Radice 2013, Vatter & Chavez-Demoulin 2015, Filippou et al. 2019, Hans et al. 2023) has been proposed, with only few

UNCERTAINTIES IN MACHINE LEARNING

In machine learning, a distinction is made between aleatoric and epistemic uncertainties, that is, uncertainties coming from the data-generating process and model or parameter uncertainties, respectively. The distinction between aleatoric and epistemic uncertainty is less common in statistics but relevant in certain machine learning applications (e.g., extrapolating predictions to regions of the covariate space outside the training data; Kendall & Gal 2017).

In statistical regression, distributional modeling approaches provide a way to model the aleatoric uncertainty, whereas statistical inference allows to derive parameter uncertainties. In contrast, deep learning often focuses on modeling the epistemic uncertainty. In addition to distributional approaches toward addressing accurate uncertainty quantification, postprocessing techniques have become popular in the machine learning literature (e.g., Song et al. 2019).

special distributions beyond $D = 3$ (Gioia et al. 2022, Kock & Klein 2023, Muschinski et al. 2022). Multivariate conditional transformation models were proposed by Klein et al. (2022). However, I am unaware of fully distributional versions that extend the approaches in Sections 2.3–2.4 to multivariate responses. To generalize quantile regression to $D > 1$, one has to define a reasonable ordering (Serfling 2002), so multivariate quantile regression cannot easily be developed. Koenker et al. (2017) review possible solutions to this.

3. PROBABILISTIC LEARNING

There is increasing interest in distributional models, not only in statistics but also in machine learning. This section briefly introduces three selected probabilistic approaches that have a strong connection to the methods of Section 2, but which try to make Equation 2 more flexible using algorithms originating from machine learning. The sidebar titled Uncertainties in Machine Learning can be found on top of this page.

3.1. Mixtures of Experts

Hierarchical mixtures-of-experts models (Jordan & Jacobs 1994) are closely related to the mixture model from Section 2.4 but try to increase flexibility through advanced modeling options for μ_h and σ_h^2, p_h, and/or π_h. Bishop (1994) explicitly models all components of a mixture distribution using multi-layer perceptrons, which are vanilla deep neural networks consisting of a fully connected input layer, at least one hidden layer, and the output layer. These models have also been developed with non-Gaussian mixture components for p_h and for noncontinuous data. Estimation uses the expectation–maximization algorithm or other nonlinear optimization routines, such as conjugate gradient or quasi-Newton methods. The R package `CaDENCE` provides a comprehensive framework for fitting such models.

3.2. Distributional Bayesian Additive Regression Trees

BART: Bayesian additive regression tree

The Bayesian additive regression tree (BART; Chipman et al. 2010) consists of the following parts: a sum-of-trees model that can capture complex interactions of a high-dimensional $\boldsymbol{x}$, and a regularization prior on the model parameters that avoids overfitting and enables posterior sampling.

Using the Gaussian model $Y \mid \boldsymbol{x} \sim \mathrm{N}(\eta(\boldsymbol{x}), \sigma^2)$, the original BART models the conditional mean $\mathbb{E}(Y \mid \boldsymbol{x}) = \sum_{j=1}^{J} f_j(\boldsymbol{x}, M_j, T_j)$ as a finite sum of J step functions $f_j(\boldsymbol{x}, M_j, T_j) = \mu_{jl}$, if

$\boldsymbol{x} \in A_{jl}$, which are parameterized through the pairs $\{M_j, T_j\}$ that consist of the terminal node parameters $M_j = \{\mu_{j1}, \ldots, \mu_{jL_j}\}$ and the binary trees T_j, for $j = 1, \ldots, J$, $l = 1, \ldots, L_j$. The tree partitions the covariate space into L_j regions A_{jl}. Each μ_{jl} can capture interaction effects of possibly varying orders, which makes a BART highly flexible (see Chipman et al. 2010 for details).

In the BART model, the parameters of the sum-of-trees model $\{M_j, T_j\}_{j=1}^{J}$ and the variance σ^2 have prior distributions. The general idea of these priors is to regularize and to encourage small trees. Hill et al. (2020) give a deeper exposition of the original BART prior, extensions, and a recent review with current developments.

Summing over sequential weak learners, BARTs have similarities to gradient boosting tree methods (Bartlett et al. 1998). However, the Bayesian approach uses a prior, and yields complete posterior distributions rather than adding small portions of the sequential weak learners and giving point estimates only. Furthermore, the Bayesian framework does not need the tuning of, e.g., the maximum depth tree via cross-validation.

Density regression BART (Orlandi et al. 2021) extends BART to distributional regression via a continuous latent variable representation. The conditional PDF is modeled as

$$p(y \mid \boldsymbol{x}) = \int_0^1 \frac{1}{\sigma(\boldsymbol{x}, u)} \phi\left(\frac{y - f(\boldsymbol{x}, u)}{\sigma(\boldsymbol{x}, u)}\right) \mathrm{d}u. \qquad 7.$$

This defines a flexible location-scale mixture model—again an extension of finite mixtures from Section 2, where $\eta_1 = f$ and $\eta_2 = \sigma$ are modeled as BART.

As stated by Orlandi et al. (2021), Equation 7 has the equivalent formulation $U \sim U(0, 1)$, $Y = f(\boldsymbol{x}, U) + \epsilon, \epsilon \sim \mathrm{N}(0, \sigma^2(\boldsymbol{x}, U))$, such that posterior sampling in density regression BART is straightforward. Conditional on u, Gibbs sampling for the trees can be performed, and u can be generated using efficient slice-sampling.

Fast implementations of BART for regression and classification are given by the R packages `dbarts` and `bartMachine`, which extend the `BayesTree` package. Another option for survival models is the `BART` package. Finally, `drbart` collects code for fitting density regression BART on GitHub (**https://github.com/vittorioorlandi/drbart**).

Figure 7 illustrates posterior mean predictive densities obtained from BART and density regression BART for two selected ages for girls, together with 95% equal-tailed credible intervals for the predictive densities. Density regression BART fits a bimodal density for young girls at *age* = 6 months and is favored by the log-score for both girls (1,330 vs. 1,422) and boys (1,277 vs. 1,317). Log-scores are based on 10-fold cross-validation, and lower values indicate better predictive performance. Owing to the Gaussian assumption, this flexibility cannot be captured by BART or heteroscedastic BART. The estimated posterior means resemble those obtained for girls with a finite mixture model (see **Figure 5**), yet the latter estimates a bimodal predictive density for *age* = 54 months.

While density regression BART seems to provide a very general framework when interest lies in prediction and prediction intervals for data with high-dimensional covariates or complex interactions, the interpretability of some of the approaches in Section 2 is lost. In addition, more empirical investigation may be helpful to better understand the impact of priors. However, it will be interesting to see how density regression BART will be extended further, such as for causal inference (see Hill et al. 2020).

3.3. Quantile Regression Forests

Quantile regression forests (Meinshausen 2006) model conditional quantiles in high-dimensional predictor situations based on random forests. Random forests grow an ensemble of trees, each using a bagged version of the training data. To select the splitting points, only a random subset

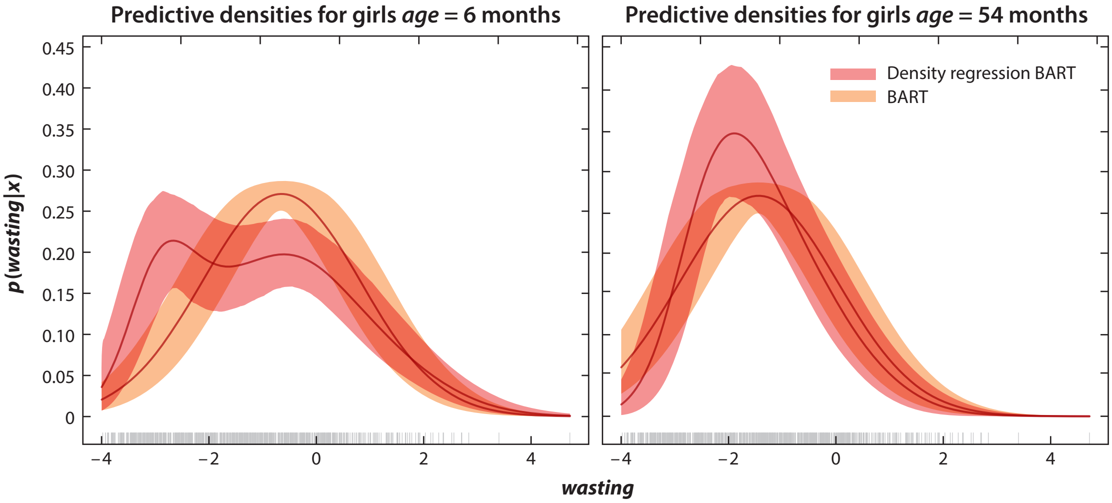

Figure 7

Childhood malnutrition: predictive densities from a Bayesian additive regression tree (BART) and density regression BART for two selected *age* values for girls, showing the posterior mean estimates (*solid lines*) together with 95% equal-tailed credible intervals. 10-Fold cross-validated log-scores favor the density regression BART model for girls.

of the covariates is employed. The size of that random subset is a main tuning parameter, which is often fine-tuned using out-of-bag samples (Breiman 2001). In traditional random forests, the conditional mean is approximated by $\sum_{i=1}^{n} w_i(\boldsymbol{x})Y_i$, where the weights $w_i(\boldsymbol{x}) = J^{-1}\sum_{j=1}^{J} w_i(\boldsymbol{x}, \boldsymbol{\theta}_j)$ are the averages of the weights of the single trees T_j that depend on parameters $\boldsymbol{\theta}_j$. These weight vectors sum to unity and are given by a positive constant if $\boldsymbol{x}_i$ is part of the corresponding leaf and 0 otherwise.

Quantile random forests can easily be created using the relation $F(\upsilon \,|\, \boldsymbol{X} = \boldsymbol{x}) = \mathbb{E}[\mathbb{1}(Y \leq \upsilon) \,|\, \boldsymbol{X} = \boldsymbol{x}]$ to arrive at an estimator $\sum_{i=1}^{n} w_i(\boldsymbol{x})\mathbb{1}(Y_i \leq \upsilon)$ based on the same weights as for traditional random forests (for further details, see Meinshausen 2006). Software is available through the R package `quantregForest`, which builds on the original `randomForest` package.

4. SCALABLE ESTIMATION FOR DISTRIBUTIONAL REGRESSION

Traditional estimation techniques to distributional regression based on (P)MLE or MCMC can be infeasible when using very large data sets and rather complex models with many parameters.

4.1. Efficient Matrix Operations and Batchwise Backfitting

The `mgcv` package has integrated a number of advances toward scalable PMLE estimation with of the order of 10^4 coefficients and over 10^7 data points. These are mostly based on Wood et al. (2017) and Li & Wood (2020). When n and/or p is large, setting up the full covariate matrix $\boldsymbol{X}$ and repeated computation and decompositions of $\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X}$, where $\boldsymbol{W}$ is diagonal or tri-diagonal, is computationally costly. Wood et al. (2017) circumvent this using a pivoted Cholesky decomposition, which can be accumulated blockwise and parallelized. In addition, the effective dimension of $\boldsymbol{X}$ can be drastically reduced by discretization in the covariate space (similar to Lang et al. 2014). Li & Wood (2020) extend the latter to multivariate covariate spaces, thereby further reducing computing time. To the best of my knowledge, these approaches have not yet been used for the generalization to GAMLSS, but the hope is that they will become available in the corresponding software.

Umlauf et al. (2023) propose a batchwise backfitting algorithm for GAMLSS-type models that combines backfitting optimization and stochastic gradient descent. The algorithm provides automatic selection of variables and estimation of smoothing parameters, while maintaining low computational cost and time. Rather than evaluating first- and second-order derivatives $\sum_{k=1}^{K} J_k$ times in each step of the backfitting algorithm, batchwise backfitting does so on a random subset of the data. The updating steps are then stochastic, involving a step length (or learning rate), similar to classical stochastic gradient descent. Umlauf et al. (2023) provide guidelines on choosing batch size and learning rate and demonstrate that, depending on the latter, the convergence behavior is similar to stochastic gradient descent, boosting, or resampling. For a data set with 10^7 observations and a fairly complex location-scale model, estimation with batchwise backfitting takes less than an hour, whereas MCMC would be infeasible for such numbers of observations. The routines can be tested using `bb_optfit` in the `bamlss` package. In addition to efficient handling of large matrices similar to `mgcv`, `bamlss` now supports flat file formats using the `ff` package.

VI: variational inference

4.2. Variational Inference

Variational inference (VI; Blei et al. 2017) has become popular as a scalable technique for approximate Bayesian inference when MCMC is infeasible. The general idea is to turn estimation into an optimization problem, where commonly the variational parameters $\boldsymbol{\lambda}$ are tailored toward a member $q_{\lambda}(\boldsymbol{\theta})$ that is close to $p(\boldsymbol{\theta} \mid \boldsymbol{y})$. Let $\boldsymbol{\theta}$ denote the set of all unknown model parameters. Proximity between $q_{\lambda}(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ is assessed by a divergence measure. The Kullback–Leibler divergence $\mathrm{KL}[q_{\lambda}(\boldsymbol{\theta}) || p(\boldsymbol{\theta} \mid \boldsymbol{y})]$ is commonly employed, and it is straightforward to check that minimizing this is equivalent to maximizing the variational lower bound, also called the evidence lower bound (ELBO; Ormerod & Wand 2010, Blei et al. 2017),

$$\mathcal{L}(\boldsymbol{\lambda}) = \int q_{\lambda}(\boldsymbol{\theta}) \log\left(\frac{p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{q_{\lambda}(\boldsymbol{\theta})}\right) \mathrm{d}\boldsymbol{\theta}.$$

This integral is generally intractable, but it can be optimized using stochastic gradient descent methods. Given an initial value $\boldsymbol{\lambda}^{(0)}$, such methods optimize the ELBO through the updates

$$\boldsymbol{\lambda}^{(t+1)} = \boldsymbol{\lambda}^{(t)} + \boldsymbol{\rho}^{(t)} \circ \nabla_{\lambda} \widehat{\mathcal{L}(\boldsymbol{\lambda}^{(t)})}, \quad t = 1, \ldots,$$

where $\boldsymbol{\rho}^{(t)}$ is a vector of step sizes, $\circ$ denotes the elementwise product of two vectors, and $\nabla_{\lambda} \widehat{\mathcal{L}(\boldsymbol{\lambda}^{(t)})}$ is an unbiased estimate of the gradient of $\mathcal{L}(\boldsymbol{\lambda})$ at $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{(t)}$. For appropriate step sizes, convergence to a local optimum of $\mathcal{L}(\boldsymbol{\lambda})$ is guaranteed.

In addition, variance reduction methods, such as the reparameterization trick, have become popular for gradient estimation to achieve fast convergence and stability (for details and further references, see Kingma & Welling 2014). However, VI algorithms for distributional regression require tedious model-specific calculations and carefully chosen variational approximations in particular. Kucukelbir et al. (2015) propose an automatic VI algorithm called `advi`, which is implemented in Stan and R. The only required input is a well-defined Bayesian model and the data set. In particular, no conjugacy assumptions are made. The general idea of this approach is to transform all model parameters to the real line and approximate the posterior on that transformed model space by a Gaussian variational approximation.

The latter are popular and often provide very accurate approximations to posterior means/modes. However, a big limitation of the Stan implementation of `advi` is that the Gaussian variational approximation assumes a diagonal covariance matrix, which implies that no posterior

CURRENT TRENDS

The literature on distributional regression is currently developing. Examples of current areas of active research include the following:

- Hybrid models are being developed that try to leverage the power of statistics and machine learning simultaneously for distributional regression. The general aim is to combine structured predictors for tabular data with other data modalities to obtain both modeling flexibility through the unstructured deep architecture and interpretability for the structured part. Rügamer et al. (2023) provide a review and a proposal toward so-called semistructured distributional regression within the framework of GAMLSS.
- Progress has been made towards learning causal relations in distributional regression using instrumental variables (Briseño Sanchez et al. 2020), treatment effects (Park et al. 2021), anchor regression (Kook et al. 2022), and distributional random forests (Cévid et al. 2022).
- Automatic visualization tools, such as Shiny apps, have been developed to guide the extraction of results, enhance interpretability, and highlight the merits of distributional regression (e.g., Stadlmann & Kneib 2022).

dependence between parameters exists. This is unlikely to be true in distributional regression models, as argued by Kleinemeier & Klein (2023).

Despite their tractability, Gaussian distributions can be computationally expensive when the dimension of $\boldsymbol{\theta}$ is high. Another challenge is the estimation of the covariance matrix, because when being unrestricted, the number of elements grows quadratically with the parameter dimension. A fruitful way to solve this for Bayesian GAMLSS is to follow Ong et al. (2018) and employ a factor covariance structure, where the covariance is decomposed into a matrix of the form $\boldsymbol{A}^\top\boldsymbol{A}+\boldsymbol{D}^2$, for a diagonal matrix $\boldsymbol{D}$ and a factor matrix $\boldsymbol{A}$ with far fewer columns than rows, considerably reducing the number of variational parameters. Kleinemeier & Klein (2023) illustrate in simulations using GAMLSS that typically only a small number of factors are needed to yield high accuracy of the variational approximation.

To improve the approximation quality of the variational family, An & Jeon (2023) propose a new learning method with a nonparametric distributional assumption for the decoder of a so-called variational auto-encoder (Kingma & Welling 2014). Estimation is made computationally efficient and tractable through a loss function based on the continuous ranked probability score. This could prove a fruitful route to follow in statistical use of VI, not only for distributional regression. For more detail on recent developments in distributional regression, see the sidebar titled Current Trends.

SUMMARY POINTS

1. Transitioning from mean regression to probabilistic learning/distributional regression allows analysts to study the entire distribution of the response in terms of covariates rather than just the mean. Studies may benefit considerably from carefully comparing and selecting an appropriate distributional model.
2. Both in statistics and in machine learning, there is increasing interest in modeling aleatoric uncertainty more realistically.

3. While this review focused on univariate real-valued responses, many distributional approaches have also been extended to univariate discrete or mixed outcomes.

FUTURE ISSUES

1. Tractable yet accurate distributional regression models beyond two dimensions and Gaussianity or linear dependence should be further developed.
2. For some of the approaches, theoretical properties (such as asymptotic theory) need to be investigated to ensure reliability of standard tools, such as the AIC, and to provide uncertainty measures, before the approaches can be recommended to applied researchers.
3. Increased flexibility comes with increased complexity of distributional models and thus requires deeper knowledge of the methodology. Further work should be done to communicate, review, and explain such methods to the applied analyst, along with further tools for model and variable selection.
4. Vignettes and tutorials contrasting the different approaches, along with their disadvantages and advantages and which approach to use for what types of data situation, could help to disseminate the use of distributional regression in general. This would contribute to a better understanding and trust in such methods.
5. As highlighted by Gneiting & Katzfuss (2014, p. 146), "Strong methodological ties between probabilistic forecasting, regression, and the emerging field of uncertainty quantification can be fruitfully explored and utilized." Here, the role of statistics in artificial intelligence (Friedrich et al. 2022) needs more attention, similar to research at the intersection of machine learning approaches with high predictive power and statistics with its inferential machinery (as stated in an interview with Silvia Chiappa of Google DeepMind; Curtis 2017):

   > I would suggest to develop a solid background in machine learning, through learning about the main disciplines underlying it, namely linear algebra, probabilistic reasoning, statistics, and optimization. A solid background makes it easy to understand recent AI advances and make contributions. A big mistake would be, for example, to study deep learning without developing such a background, as this would give a very myopic view about it.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

I would like to thank all my coauthors on distributional regression and related work, who are too numerous to name here. I am especially grateful to Michael Stanley Smith, David Nott, Nikolaus Umlauf, and Andreas Mayr. I am also indebted to Maarten Jung and the entire KleinLab. Critical feedback from the Editorial Committee provided me with very constructive comments for improvements.

LITERATURE CITED

An S, Jeon JJ. 2023. Distributional learning of variational AutoEncoder: application to synthetic data generation. arXiv:2302.11294 [stat.ML]

Aragam B, Yang R. 2022. Uniform consistency in nonparametric mixture models. arXiv:2108.14003 [math.ST]

Bartlett P, Freund Y, Lee WS, Schapire RE. 1998. Boosting the margin: a new explanation for the effectiveness of voting methods. *Ann. Stat.* 26(5):1651–86

Belloni A, Chernozhukov V. 2011. ℓ_1-penalized quantile regression in high-dimensional sparse models. *Ann. Stat.* 39(1):82–130

Biau G, Cadre B. 2021. Optimization by gradient boosting. In *Advances in Contemporary Statistics and Econometrics: Festschrift in Honor of Christine Thomas-Agnan*, ed. A Daouia, A Ruiz-Gazen, pp. 23–44. Switzerland: Springer Intl. Publ.

Bishop CM. 1994. *Mixture density networks*. Tech. Rep., Aston Univ., Birmingham, UK. **https://publications.aston.ac.uk/id/eprint/373/**

Blei DM, Kucukelbir A, McAuliffe JD. 2017. Variational inference: a review for statisticians. *J. Am. Stat. Assoc.* 112(518):859–77

Bondell HD, Reich BJ, Wang H. 2010. Noncrossing quantile regression curve estimation. *Biometrika* 97(4):825–38

Box GEP, Cox DR. 1964. An analysis of transformations. *J. R. Stat. Soc. Ser. B* 26(2):211–52

Breiman L. 2001. Random forests. *Mach. Learn.* 45:5–32

Breiman L, Friedman JH. 1985. Estimating optimal transformations for multiple regression and correlation. *J. Am. Stat. Assoc.* 80(391):580–98

Briseño Sanchez G, Hohberg M, Groll A, Kneib T. 2020. Flexible instrumental variable distributional regression. *J. R. Stat. Soc. Ser. A* 183(4):1553–74

Carlan M, Kneib T, Klein N. 2023. Bayesian conditional transformation models. *J. Am. Stat. Assoc.* **https://doi.org/10.1080/01621459.2023.2191820**

Carroll R, Ruppert D. 1988. *Transformation and Weighting in Regression*. New York: CRC

Carvalho CM, Polson NG. 2010. The horseshoe estimator for sparse signals. *Biometrika* 97:465–80

Cévid D, Michel L, Näf J, Bühlmann P, Meinshausen N. 2022. Distributional random forests: heterogeneity adjustment and multivariate distributional regression. *J. Mach. Learn. Res.* 23(333):1–79

Chernozhukov V, Fernández-Val I, Melly B. 2013. Inference on counterfactual distributions. *Econometrica* 81(6):2205–68

Chipman HA, George EI, McCulloch RE. 2010. BART: Bayesian additive regression trees. *Ann. Appl. Stat.* 4(1):266–98

Cole TJ. 1988. Fitting smoothed centile curves to reference data (with discussion). *J. R. Stat. Soc. Ser. A* 151(3):385–418

Craiu VR, Sabeti A. 2012. In mixed company: Bayesian inference for bivariate conditional copula models with discrete and continuous outcomes. *J. Multivariate Anal.* 110:106–20

Curtis S. 2017. The intersection of probabilistic modeling & deep learning: interview with Silvia Chiappa, Google DeepMind. *LinkedIn Pulse Blog*, May 17. **https://www.linkedin.com/pulse/intersection-probabilistic-modeling-deep-learning-interview-curtis**

Dawid AP. 2007. The geometry of proper scoring rules. *Ann. Inst. Stat. Math.* 59:77–93

de Carvalho V, Rodríguez-Álvarez M, Klein N. 2019. Density regression via penalised splines dependent Dirichlet process mixture of normal models. In *Proceedings of the 34th International Workshop on Statistical Modelling*, Vol. 1, ed. LM Machado, GDCC Soutinho, pp. 184–88. Guimarães, Port.: Stat. Model. Soc.

Delgado MA, García-Suaza A, Sant'Anna PHC. 2022. Distribution regression in duration analysis: an application to unemployment spells. *Econom. J.* 25(3):675–98

Dunson DB. 2007. Empirical Bayes density regression. *Stat. Sin.* 17(2):481–504

Dunson DB, Pillai N, Park JH. 2007. Bayesian density regression. *J. R. Stat. Soc. Ser. B* 69(2):163–83

Efron B. 1986. Double exponential families and their use in generalized linear regression. *J. Am. Stat. Assoc.* 81(395):709–21

Escobar MD, West M. 1995. Bayesian density estimation and inference using mixtures. *J. Am. Stat. Assoc.* 90(430):577–88

Fasiolo M, Wood SN, Zaffran M, Nedellec R, Goude Y. 2021. Fast calibrated additive quantile regression. *J. Am. Stat. Assoc.* 116(535):1402–12

Fenske N, Kneib T, Hothorn T. 2011. Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression. *J. Am. Stat. Assoc.* 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. *J. Stat. Plan. Inference* 199(2):236–48

Firpo S, Fortin NM, Lemieux T. 2009. Unconditional quantile regressions. *Econometrica* 77(3):953–73

Foresi A, Peracchi F. 1995. The conditional distribution of excess returns: an empirical analysis. *J. Am. Stat. Assoc.* 90:451–66

Friedrich S, Antes G, Behr S, Brannath W, Dumpert F, et al. 2022. Is there a role for statistics in artificial intelligence? *Adv. Data Anal. Classif.* 16(4):823–46

Frühwirth-Schnatter S. 2006. *Finite Mixture and Markov Switching Models*. New York: Springer

Frühwirth-Schnatter S, Celeux G, Robert CP. 2019. *Handbook of Mixture Analysis*. Boca Raton, FL: Chapman & Hall/CRC

Gelfand AE, Kottas A, MacEachern SN. 2005. Bayesian nonparametric spatial modeling with Dirichlet process mixing. *J. Am. Stat. Assoc.* 100(471):1021–35

Gijbels I, Prosdocimi I, Claeskens G. 2010. Nonparametric estimation of mean and dispersion functions in extended generalized linear models. *TEST* 19(1):580–608

Gioia V, Fasiolo M, Browell J, Bellio R. 2022. Additive covariance matrix models: modelling regional electricity net-demand in Great Britain. arXiv:2211.07451 [stat.AP]

Gneiting T, Balabdaoui F, Raftery AE. 2007. Probabilistic forecasts, calibration and sharpness. *J. R. Stat. Soc. Ser. B* 69(2):243–68

Gneiting T, Katzfuss M. 2014. Probabilistic forecasting. *Annu. Rev. Stat. Appl.* 1:125–51

Gneiting T, Raftery AE. 2007. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* 102(477):359–78

Gneiting T, Raftery AE, Westveld AH, Goldman T. 2005. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Weather Rev.* 133(5):1098–118

Griffin JE, Steel MFJ. 2006. Order-based dependent Dirichlet processes. *J. Am. Stat. Assoc.* 101(473):179–94

Grün B, Leisch F. 2008. Finite mixtures of generalized linear regression models. In *Recent Advances in Linear Models and Related Areas: Essays in Honour of Helge Toutenburg*, ed. LJ Shalabh, C Heumann, pp. 205–30. Heidelberg: Physica-Verlag

Hall P, Wolff RCL, Yao Q. 1999. Methods for estimating a conditional distribution function. *J. Am. Stat. Assoc.* 94(445):154–63

Hans N, Klein N, Faschingbauer F, Schneider M, Mayr A. 2023. Boosting distributional copula regression. *Biometrics* 79(3):2298–310

Hastie TJ, Tibshirani R. 1990. *Generalized Additive Models*. New York/Boca Raton: Chapman & Hall/CRC

He X. 1997. Quantile curves without crossing. *Am. Stat.* 51(2):186–92

Henzi A, Ziegel JF, Gneiting T. 2021. Isotonic distributional regression. *J. R. Stat. Soc. Ser. B* 83:963–93

Hill J, Linero A, Murray J. 2020. Bayesian additive regression trees: a review and look forward. *Annu. Rev. Stat. Appl.* 7:251–78

Hothorn T. 2018. Top-down transformation choice. *Stat. Model.* 18(3–4):274–98

Hothorn T, Kneib T, Bühlmann P. 2014. Conditional transformation models. *J. R. Stat. Soc. Ser. B* 76(1):3–27

Hothorn T, Möst L, Bühlmann P. 2018. Most likely transformations. *Scand. J. Stat.* 45(1):110–34

Hyndman RJ, Bashtannyk DM, Grunwald GK. 1996. Estimating and visualizing conditional densities. *J. Comput. Graph. Stat.* 5(4):315–36

Iorio MD, Müller P, Rosner GL, MacEachern SN. 2004. An ANOVA model for dependent random measures. *J. Am. Stat. Assoc.* 99(465):205–15

Jordan MI, Jacobs RA. 1994. Hierarchical mixtures of experts and the EM algorithm. *Neural Comput.* 6(2):181–214

Kauermann G, Krivobokova T, Fahrmeir L. 2009. Some asymptotic results on generalized penalized spline smoothing. *J. R. Stat. Soc. Ser. B* 71(2):487–503

Kendall A, Gal Y. 2017. What uncertainties do we need in Bayesian deep learning for computer vision? In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, ed. U von Luxburg, I Guyon, S Bengio, H Wallach, R Fergus, pp. 5580–90. Red Hook, NY: Curran

Kingma DP, Welling M. 2014. Auto-encoding variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14–16, 2014, Conference Track Proceedings*. N.p.: ICLR

Klein N, Carlan M, Kneib T, Lang S, Wagner H. 2021. Bayesian effect selection in structured additive distributional regression models. *Bayesian Anal.* 16(2):545–73

Klein N, Hothorn T, Barbanti L, Kneib T. 2022. Multivariate conditional transformation models. *Scand. J. Stat.* 49(1):116–42

Klein N, Kneib T, Klasen S, Lang S. 2015a. Bayesian structured additive distributional regression for multivariate responses. *J. R. Stat. Soc. Ser. C* 64(4):569–91

Klein N, Kneib T, Lang S. 2015b. Bayesian generalized additive models for location, scale, and shape for zero-inflated and overdispersed count data. *J. Am. Stat. Assoc.* 110(509):405–19

Klein N, Smith MS. 2019. Implicit copulas from Bayesian regularized regression smoothers. *Bayesian Anal.* 14(4):1143–71

Klein N, Smith MS. 2021. Bayesian variable selection for non-Gaussian responses: a marginally-calibrated copula approach. *Biometrics* 77(3):809–23

Klein N, Smith MS, Nott DJ. 2023. Deep distributional time series models and the probabilistic forecasting of intraday electricity prices. *J. Appl. Econom.* 38(4):493–511

Kleinemeier J, Klein N. 2023. *Scalable estimation for structured additive distributional regression through variational inference.* Work. Pap., Dep. Stat., Tech. Univ. Dortumund, Dortmund, Ger.

Kock L, Klein N. 2023. Truly multivariate structured additive distributional regression. arXiv:2306.02711 [stat.ME]

Koenker R. 2005. *Quantile Regression*. Cambridge, UK: Cambridge Univ. Press

Koenker R, Bassett G. 1978. Regression quantiles. *Econometrica* 46(1):33–50

Koenker R, Chernozhukov V, He X, Peng L, eds. 2017. *Handbook of Quantile Regression*. New York: CRC

Kook L, Sick B, Bühlmann P. 2022. Distributional anchor regression. *Stat. Comput.* 32:39

Kozumi H, Kobayashi G. 2011. Gibbs sampling methods for Bayesian quantile regression. *J. Stat. Comput. Simul.* 81(11):1565–78

Krämer N, Brechmann EC, Silvestrini D, Czado C. 2013. Total loss estimation using copula-based regression models. *Insur. Math. Econ.* 53(3):829–39

Kucukelbir A, Ranganath R, Gelman A, Blei D. 2015. Automatic variational inference in Stan. In *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS '15)*, ed. C Cortes, DD Lee, M Sugiyama, R Garnett, pp. 568–76. Cambridge, MA: MIT Press

Lamarche C, Parker T. 2023. Wild bootstrap inference for penalized quantile regression for longitudinal data. *J. Econom.* 235(2):1799–826

Lang S, Umlauf N, Wechselberger P, Harttgen K, Kneib T. 2014. Multilevel structured additive regression. *Stat. Comput.* 24(2):223–38

Lee Y, Nelder JA. 2006. Double hierarchical generalized linear models (with discussion). *J. R. Stat. Soc. Ser. C* 55(2):139–85

Li Z, Wood S. 2020. Faster model matrix crossproducts for large generalized linear models with discretized covariates. *Stat. Comput.* 30:19–25

MacEachern SN, Müller P. 1998. Estimating mixture of Dirichlet process models. *J. Comput. Graph. Stat.* 7(2):223–38

Manole T, Khalili A. 2021. Estimating the number of components in finite mixture models via the Group-Sort-Fuse procedure. *Ann. Stat.* 49(6):3043–69

Marra G, Radice R. 2013. A penalized likelihood estimation approach to semiparametric sample selection binary response modeling. *Electron. J. Stat.* 7:1432–55

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. *J. R. Stat. Soc. Ser. C* 61(3):403–27

Meinshausen N. 2006. Quantile regression forests. *J. Mach. Learn. Res.* 7(35):983–99

Muschinski T, Mayr GJ, Simon T, Umlauf N, Zeileis A. 2022. Cholesky-based multivariate Gaussian regression. *Econom. Stat.* In press

Nelsen R. 2006. *An Introduction to Copulas*. New York: Springer

Newcomb S. 1886. A generalized theory of the combination of observations so as to obtain the best result. *Am. J. Math.* 8(4):343–66

Newey WK, Powell JL. 1987. Asymmetric least squares estimation and testing. *Econometrica* 55(4):819–47

Noh H, Ghouch AE, Bouezmarni T. 2013. Copula-based regression estimation and inference. *J. Am. Stat. Assoc.* 108(502):676–88

Ong VM, Nott D, Smith M. 2018. Gaussian variational approximation with a factor covariance structure. *J. Comput. Graph. Stat.* 27(3):465–78

Orlandi V, Murray J, Linero A, Volfovsky A. 2021. Density regression with Bayesian additive regression trees. arXiv:2112.12259 [stat.ME]

Ormerod JT, Wand MP. 2010. Explaining variational approximations. *Am. Stat.* 64(2):140–53

Park J, Shalit U, Schölkopf B, Muandet K. 2021. Conditional distributional treatment effect with kernel conditional mean embeddings and U-statistic regression. *Proc. Mach. Learn. Res.* 139:8401–12

Pitt M, Chan D, Kohn R. 2006. Efficient Bayesian inference for Gaussian copula regression models. *Biometrika* 93:537–54

Quintana FA, Mueller P, Jara A, MacEachern SN. 2020. The dependent Dirichlet process and related models. arXiv:2007.06129 [stat.ME]

Rigby RA, Stasinopoulos DM. 2005. Generalized additive models for location, scale and shape. *J. R. Stat. Soc. Ser. C* 54(3):507–54

Rodrigues T, Fan Y. 2017. Regression adjustment for noncrossing Bayesian quantile regression. *J. Comput. Graph. Stat.* 26(2):275–84

Rossi GD, Harvey A. 2009. Quantiles, expectiles and splines. *J. Econom.* 152(2):179–85

Rothe C, Wied D. 2013. Misspecification testing in a class of conditional distributional models. *J. Am. Stat. Assoc.* 108(501):314–24

Rügamer D, Kolb C, Klein N. 2023. Semi-structured distributional regression. *Am. Stat.* In press

Serfling R. 2002. Quantile functions for multivariate analysis: approaches and applications. *Stat. Neerl.* 56:214–32

Sisson S, Fan Y, Beaumont M, eds. 2018. *Handbook of Approximate Bayesian Computation*. Boca Raton, FL: Chapman & Hall/CRC

Smith MS. 2023. Implicit copulas: An overview. *Econom. Stat.* 28:81–104

Smith MS, Klein N. 2021. Bayesian inference for regression copulas. *J. Bus. Econ. Stat.* 39(3):712–28

Song H, Diethe T, Kull M, Flach P. 2019. Distribution calibration for regression. *Proc. Mach. Learn. Res.* 97:5897–906

Song PXK, Li M, Yuan Y. 2009. Joint regression analysis of correlated data using Gaussian copulas. *Biometrics* 65(1):60–68

Sriram K, Ramamoorthi R, Ghosh P. 2013. Posterior consistency of Bayesian quantile regression based on the misspecified asymmetric Laplace density. *Bayesian Anal.* 8(2):479–504

Stadlmann S, Kneib T. 2022. Interactively visualizing distributional regression models with distreg.vis. *Stat. Model.* 22(6):527–45

Stone CJ. 1977. Consistent nonparametric regression. *Ann. Stat.* 5(4):595–620

Takeuchi I, Le QV, Sears TD, Smola AJ. 2006. Nonparametric quantile estimation. *J. Mach. Learn. Res.* 7(45):1231–64

Umlauf N, Seiler J, Wetscher M, Simon T, Lang S, Klein N. 2023. Scalable estimation for structured additive distributional regression. arXiv:2301.05593 [stat.CO]

UNICEF. 1998. *The State of the World's Children 1998: Focus on Nutrition*. Oxford: Oxford Univ. Press

Vatter T, Chavez-Demoulin V. 2015. Generalized additive models for conditional dependence structures. *J. Multivariate Anal.* 141:147–67

Velthoen J, Dombry C, Cai JJ, Engelke S. 2023. Gradient boosting for extreme quantile regression. *Extremes* 26:639–67

Villani M, Kohn R, Nott DJ. 2012. Generalized smooth finite mixtures. *J. Econom.* 171(2):121–33

Waldmann E. 2018. Quantile regression: a short story on how and why. *Stat. Model.* 18(3–4):203–18

Waltrup LS, Sobotka F, Kneib T, Kauermann G. 2015. Expectile and quantile regression—David and Goliath? *Stat. Model.* 15(5):433–56

Wang L. 2013. Consistency of posterior distributions for heteroscedastic nonparametric regression models. *Commun. Stat. Theory Methods* 42(15):2731–40

Wilson AG, Ghahramani Z. 2010. Copula processes. In *NIPS'10: Proceedings of the 23rd International Conference on Neural Information Processing Systems*, Vol. 2, ed. JD Lafferty, CKI Williams, J Shawe-Taylor, RS Zemel, A Culotta pp. 2460–68. Red Hook, NY: Curran

Wood SN, Li Z, Shaddick G, Augustin NH. 2017. Generalized additive models for gigadata: modeling the U.K. Black Smoke Network daily data. *J. Am. Stat. Assoc.* 112(519):1199–210

Yang Y, Wang HJ, He X. 2016. Posterior inference in Bayesian quantile regression with asymmetric Laplace likelihood. *Int. Stat. Rev.* 84(3):327–44

Yau P, Kohn R. 2003. Estimation and variable selection in nonparametric heteroscedastic regression. *Stat. Comput.* 13(3):191–208

Yee T. 2015. *Vector Generalized Linear and Additive Models*. New York: Springer

Yu K, Moyeed RA. 2001. Bayesian quantile regression. *Stat. Probab. Lett.* 54:437–47

Yue Y, Rue H. 2011. Bayesian inference for additive mixed quantile regression models. *Comput. Stat. Data Anal.* 55:84–96