Distributional (Single) Index Models

Alexander Henzi, Gian-Reto Kleger, Johanna F. Ziegel*

May 31, 2021

Abstract

A Distributional (Single) Index Model (DIM) is a semi-parametric model for distributional regression, that is, estimation of conditional distributions given covariates. The method is a combination of classical single index models for the estimation of the conditional mean of a response given covariates, and isotonic distributional regression. The model for the index is parametric, whereas the conditional distributions are estimated non-parametrically under a stochastic ordering constraint. We show consistency of our estimators and apply them to a highly challenging data set on the length of stay (LoS) of patients in intensive care units. We use the model to provide skillful and calibrated probabilistic predictions for the LoS of individual patients, that outperform the available methods in the literature.

*Alexander Henzi is PhD student in Statistics, Johanna F. Ziegel is Professor of Applied Stochastics, Institute of Mathematical Statistics and Actuarial Science, University of Bern, Alpeneggstrasse 22, 3012 Bern, Switzerland (e-mail: alexander.henzi@stat.unibe.ch, johanna.ziegel@stat.unibe.ch); Gian-Reto Kleger, MD, is Head the Division of Intensive Care Medicine, Cantonal Hospital, St. Gallen, Rorschacherstrasse 95, 9007 St.Gallen, Switzerland (e-mail: gian-reto.kleger@kssg.ch). A. Henzi and J. F. Ziegel gratefully acknowledge financial support from the Swiss National Science Foundation. The authors thank the Swiss Society of Intensive Care Medicine for providing the data. The work has greatly benefitted from discussions with Lutz Dümbgen, Tilmann Gneiting, Alexander Jordan, and Alexandre Mösching.

source: https://doi.org/10.48350/164368 | downloaded: 24.10.2022

Keywords: Distributional regression, intensive care unit length of stay, probabilistic forecast, single index model, stochastic ordering constraint

1 Introduction

Regression approaches for the full conditional distribution of an outcome given covariates are gaining momentum in the literature (Hothorn et al., 2014, and the references therein). They have already become an indispensable tool in probabilistic weather forecasting (Gneiting and Katzfuss, 2014; Vannitsem et al., 2018) but also find numerous applications in other fields such as economics, social sciences and medicine; see e.g. Machado and Mata (2000), Chernozhukov et al. (2013), Klein et al. (2015), Duarte et al. (2017) and Silbersdorff et al. (2018).

If the outcome is real-valued, then conditional distributions can be characterized in terms of their cumulative distribution function (CDF) or quantile function, and various techniques for the estimation of these objects have been proposed. Foresi and Peracchi (1995) and Peracchi (2002) build on the extant methods for the estimation of single quantiles or probabilities (Koenker, 2005), and suggest to approximate the conditional distribution by a cascade of regressions for quantiles or for the CDF evaluated at certain thresholds. A drawback of this approach is that the resulting estimates are not necessarily isotonic (the so-called 'quantile crossing problem') and thus require correction, for which remedies have already been developed, see e.g. Dette and Volgushev (2008); Chernozhukov et al. (2010).

A broad class of methods that directly yield well-defined probability distributions are generalized additive models for location, shape and scale (Rigby and Stasinopoulos, 2005, GAMLSS). They build on generalized linear models (McCullagh and Nelder, 1989, GLM) and generalized additive models for the mean (Hastie and Tibshirani, 1990, GAM) but also allow to model shape and scale parameters as functions of covariates. The GAMLSS framework has has been extended to Bayesian statistics (Umlauf et al., 2018) and combined with popular machine learning techniques such as boosting (Thomas et al., 2018), neural

networks (Rasp and Lerch, 2018) and regression forests (Schlosser et al., 2019).

Finally, there are also powerful semi-parametric and nonparametric techniques for the estimation of conditional distributions. Fully nonparametric methods estimate the conditional distribution functions locally, for example by kernel functions (Hall et al., 1999; Dunson et al., 2007; Li and Racine, 2008), or by partitioning of the covariate space, as in quantile random forests (Meinshausen, 2006; Athey et al., 2019). A frequently used semi-parametric distributional regression method is Cox regression (Cox, 1972), which models the hazard rate of the outcome but also allows to derive its survival function. Conditional transformation models (Hothorn et al., 2014) assume a parametric distribution for an unknown monotone transformation of the response, which is estimated along with the model parameters. Hall and Yao (2005); Zhang et al. (2017) propose semi-parametric methods that reduce the dimension of the covariate space by a suitable projection, and then estimate the conditional distributions non-parametrically given the projections by kernel methods.

We introduce a new approach to distributional regression that can be seen as a combination of a single index model with isotonic distributional regression (IDR, Henzi et al., 2019). The dimension reduction of the covariate space achieved by the single index assumption is in the spirit of Hall and Yao (2005); Zhang et al. (2017) but the combination with IDR is new, and has the advantage to be free of any implementation choices or tuning parameters.

Let Y be a real-valued response and X a covariate in some covariate space $\mathcal{X}$. We want to estimate the conditional distribution of Y given X, that is, $\mathcal{L}(Y \mid X)$. To expose the main idea, suppose that $\mathcal{X} = \mathbb{R}^d$. Then, a Distributional (Single) Index Model (DIM) could be

$$\mathbb{P}(Y \leqslant y \mid X) = F_{\alpha_0^\top X}(y), \quad \text{for all } y \in \mathbb{R}, \tag{1}$$

where $\alpha_0 \in \mathbb{R}^d$, $\alpha_0^\top X$ denotes the scalar product between α_0 and X, and $(F_u)_{u\in\mathbb{R}}$ is a family of CDFs such that

$$F_u \preceq_{\text{st}} F_v \quad \text{if } u \leqslant v, \tag{2}$$

where $\preceq_{\text{st}}$ denotes the usual stochastic order, that is $F_u \preceq_{\text{st}} F_v$ if $F_u(y) \geqslant F_v(y)$ for all $y \in \mathbb{R}$. We call $\theta(x) = \alpha_0^T x$ in representation (1) the index (function).

If the parameter α_0 in the previous example (1) is known, then a natural method to estimate the unknown family $(F_u)_u$ of stochastically ordered CDFs is IDR as introduced by Henzi et al. (2019), see also Mösching and Dümbgen (2020). IDR is a nonparametric technique to estimate conditional distributions under stochastic ordering constraints. In brief, IDR works as follows. Given training data $(\vartheta_1, y_1), \ldots, (\vartheta_n, y_n)$, where $\vartheta_i \in \Theta$ for some partially ordered set Θ, IDR yields the unique optimal vector $\hat{\mathbf{F}} = (\hat{F}_1, \ldots, \hat{F}_n)$ of CDFs that minimizes

$$\frac{1}{n} \sum_{i=1}^{n} \mathrm{CRPS}(F_i, y_i),$$

over all vectors $(F_1, \ldots, F_n)$ of CDFs that respect the stochastic ordering constraints $F_i \preceq_{\mathrm{st}} F_j$ if $\vartheta_i \preceq \vartheta_j$, $i, j = 1, \ldots, n$. Here, for any CDF F and $y \in \mathbb{R}$,

$$\mathrm{CRPS}(F, y) = \int_{\mathbb{R}} \left(F(z) - \mathbb{1}\{y \leqslant z\}\right)^2 \, \mathrm{d}z \tag{3}$$

is the widely applied proper scoring rule called the continuous ranked probability score (CRPS, Matheson and Winkler, 1976; Gneiting et al., 2007). If we have a sample (x_1, y_1), $\ldots$, (x_n, y_n) from $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$, we can apply IDR to the training data $(\alpha_0^\top x_1, y_1)$, $\ldots$, $(\alpha_0^\top x_n, y_n)$, that is, we set $\vartheta_i = \alpha_0^\top x_i$, $i = 1, \ldots, n$ and $\Theta = \mathbb{R}$. This yields a distributional regression model for (X, Y) that may be used to provide probabilistic predictions for Y given X, see Henzi et al. (2019, Section 2.5) and Section 4.

DIMs are closely related to generalized linear models, which assume that the conditional distributions $(F_u)_u$ belong to a known exponential family of distributions with mean $\mathbb{E}(Y \mid X = x) = g(\alpha_0^T x)$, where g is a fixed, strictly monotone link function. In fact, the Gaussian, Poisson, Gamma and Binomial GLM can be subsumed under the DIM, since they also satisfy the stochastic ordering constraint on the conditional distributions. Our approach, to leave the conditional distributions $(F_u)_u$ unspecified, is already widely applied in classical regression for the mean, where models of the type $\mathbb{E}(Y \mid X = x) = g(\alpha_0^T x)$ with unknown link function g are called single index models. Typically, g is assumed to be a smooth function and estimated by kernel regression or local polynomial approximation (Härdle et al., 1993) or local polynomial approximation (Carroll et al., 1997; Zou and Zhu, 2014). More recently,

shape constrained single index models have been considered with monotone (Balabdaoui et al., 2019a) and convex (Kuchibhotla et al., 2017) link functions. DIMs directly extend monotone single index models for the mean, since the stochastic ordering assumption on the conditional distributions implies an isotonic conditional mean function.

There is a vast literature on the estimation of the index in single index models, and we refer to Lanteri et al. (2020) for a comprehensive overview. In Section 3, we discuss estimators for the index and the distribution functions in DIMs. Briefly, when IDR is used to estimate the conditional distribution functions, then it is sufficient to know the index function up to isotonic transformations, i.e. to find a pseudo index function that approximates the *ordering* implied by the true index. This approach is supported by the asymptotic analysis in Section 5, which shows that when a monotone transformation of the estimated index function is consistent at the parametric rate, then a DIM with that index estimator is consistent.

A major application of distributional regression techniques is forecasting. It has been recognized in many problems, such as weather prediction or economic forecasting, that point forecasts are unable to account for the full forecast uncertainty and should be replaced by probabilistic forecasts (Gneiting and Katzfuss, 2014). Distributional regression methods are statistical tools to provide such probabilistic forecasts. One fundamental contribution of DIMs is that they allow to associate a natural distributional prediction to point forecasts: If a point forecast from a statistical model is taken as the index in a DIM, for example the estimated conditional expected value, then the DIM naturally extends this deterministic forecast to a probabilistic one. Moreover, the only prerequisite is an isotonic relationship between the point forecast and the outcome in a stochastic ordering sense, which is often a natural and intuitive assumption for reasonable point forecasts.

In Section 6, we use a DIM for predictions in a highly challenging dataset on the length of stay (LoS) of intensive care unit (ICU) patients. Accurate LoS predictions could serve as a tool for ICU physicians, for example to plan the number of available beds, or to identify potential long stay patients at an early stage. Moreover, the same models that are used for prediction may also be used for risk-adjustment and benchmarking across different ICUs. In

the last twenty years, there have been many approaches to find appropriate regression models for LoS, see Zimmerman et al. (2006); Moran and Solomon (2012); Verburg et al. (2014) for some examples and Verburg et al. (2014); Kramer (2017) for literature reviews. The extant methods typically model the conditional mean and are unsatisfactory when applied for single patient predictions, since the distribution of LoS is strongly right-skewed with a large variance even after conditioning on covariates. We therefore argue that LoS predictions should be probabilistic. In Section 6, we derive calibrated and informative probabilistic forecasts for LoS, and show that the DIM outperforms existing distributional regression methods in terms of predictive accuracy.

2 Distributional index models

In this section, we define the DIM in its most general form. Let Y be a real-valued response, and let X be covariates in some general space $\mathcal{X}$. The link between X and Y is the index function $\theta\colon \mathcal{X} \to \mathbb{R}^d$, where $\mathbb{R}^d$ is equipped with some partial order $\preceq$. Let further $(F_u)_{u\in\mathbb{R}^d}$ be a family of CDFs such that $F_u \preceq_{\text{st}} F_v$ if $u \preceq v$. The DIM then assumes that

$$\mathbb{P}(Y \leqslant y \mid X) = F_{\theta(X)}(y). \tag{4}$$

Due to the stochastic ordering assumption, it directly follows that the conditional distributions are ordered in the index, that is, $\theta(x) \preceq \theta(x')$ implies $F_{\theta(x)} \preceq_{\text{st}} F_{\theta(x')}$.

We assume further that the function θ belongs to a finite dimensional vector space $\mathcal{F}$, i.e. a parametric model for θ. If $\theta_1, \dots, \theta_p$ are a basis of $\mathcal{F}$ and if $d = 1$, then we recover the form $\mathbb{P}(Y \leqslant y \mid \tilde{X} = \tilde{x}) = F_{\alpha_0^T\tilde{x}}(y)$, where $\tilde{x} = (\theta_1(x)\dots,\theta_p(x))$, and hence, the analogy to single index models. However, the estimation procedure suggested in the next section can be applied with any dimension d and any partial order $\preceq$ on $\mathbb{R}^d$.

3 Estimation

Having motivated and formalized the DIM, we propose a method for estimation. Assume that a training dataset (x_i, y_i), $i = 1, \dots, n$, of independent realizations of (X, Y) satisfying the model assumption (4) is available.

In principle, it would be desirable to have a simultaneous estimator for both the index and the distribution functions. In Section 5, we show that simultaneous estimation is possible theoretically, but computationally infeasible. The method we propose here, and for which we provide asymptotic results, is a two-stage estimation in which first the index θ is estimated, say by $\hat{\theta}$, and then the conditional CDFs based on pairs $(\hat{\theta}(x_i), y_i)$. This is inspired by the 'plug-in estimators' for monotone single index models suggested in Balabdaoui et al. (2019a). The estimation procedure is straightforward and reads as follows:

1. Estimate θ with some estimator $\hat{\theta}$ on the data $(x_i, y_i)_{i=1}^n$,
2. compute the in-sample predictions $\vartheta_i = \hat{\theta}(x_i)$, $i = 1, \dots, n$,
3. estimate the distribution functions $\hat{F}_u, u \in \mathbb{R}^d$, using $(\vartheta_i, y_i)_{i=1}^n$.

In the next two subsections, we reverse the order of the estimation procedure and first suggest our method for Step 3, because this has important implications for the choice of the index estimators in Step 1.

3.1 Isotonic distributional regression

Because of model assumption (4), we seek an estimator $\hat{F}_u, u \in \mathbb{R}^d$, such that $\hat{F}_u \preceq_{\text{st}} \hat{F}_v$ if $u \preceq v$, i.e. $\hat{F}_u(y) \geqslant \hat{F}_v(y)$ for all $y \in \mathbb{R}$ and given u, v. For fixed y, this suggests to define $\hat{\boldsymbol{F}} = (\hat{F}_{\vartheta_1}, \dots, \hat{F}_{\vartheta_n})$ as

$$\hat{\boldsymbol{F}}(y) \;=\; \underset{\eta_k \geqslant \eta_l \text{ if } \vartheta_k \leq \vartheta_l}{\operatorname{argmin}} \sum_{i=1}^{n} (\eta_i - \mathbb{1}\{y_i \leqslant y\})^2. \tag{5}$$

It turns out that (5) indeed yields a collection of well-defined conditional CDFs, and this estimator is called the IDR in Henzi et al. (2019). By Henzi et al. (2019, Theorem 2.2), IDR

can equivalently be defined in terms of conditional quantile functions, $\hat{\boldsymbol{q}} = (\hat{q}_{\vartheta_1}, \dots, \hat{q}_{\vartheta_n})$, where

$$\hat{\boldsymbol{q}}(\alpha) \;=\; \underset{\beta_k \leqslant \beta_l \text{ if } \vartheta_k \preceq \vartheta_l}{\operatorname{argmin}} \sum_{i=1}^{n} (\mathbb{1}\{y_i \leqslant \beta_i\} - \alpha)(\beta_i - y_i) \tag{6}$$

for any $\alpha \in (0,1)$, and the argmin is defined as the componentwise smallest minimizer if it is not unique. IDR estimates the conditional distributions non-parametrically under the stochastic order constraints. For IDR, the index u can take values in any partially ordered set Θ. The particular choice of the loss functions, i.e. the squared error for the estimation of probabilities in (5) and the classical quantile loss function in (6), is in fact irrelevant here: Any other consistent loss function for the expectation or quantiles would yield the same result (Henzi et al., 2019; Jordan et al., 2019).

The above estimators are defined when the index u (in $\hat{F}_u$ or $\hat{q}_u$) is in $\{\vartheta_1, \dots, \vartheta_n\} \subseteq \Theta$. The CDFs or quantile functions for an arbitrary u can be derived by interpolation of $\hat{F}_{\vartheta_1}, \dots, \hat{F}_{\vartheta_n}$ or $\hat{q}_{\vartheta_1}, \dots, \hat{q}_{\vartheta_n}$ for $\Theta = \mathbb{R}$, and a suitable generalization thereof for general partially ordered Θ (Henzi et al., 2019, Section 2.5).

The following proposition is a direct consequence of the above formulas. It shows invariance properties of IDR, which make it a suitable method for estimating the conditional distributions in DIMs. We use the notation $\hat{F}_u(y; \boldsymbol{\vartheta}, \boldsymbol{y})$ and $\hat{q}_u(\alpha; \boldsymbol{\vartheta}, \boldsymbol{y})$ for the IDR CDFs and quantile functions estimated with training data $\boldsymbol{\vartheta} = (\vartheta_k)_{k=1}^m$ and $\boldsymbol{y} = (y_k)_{k=1}^m$.

Proposition 3.1 (Invariance of IDR)**.** *Let $\boldsymbol{y} = (y_k)_{k=1}^m \in \mathbb{R}^m$ and $\boldsymbol{\vartheta} = (\vartheta_k)_{k=1}^m \in \Theta^m$, and let Θ' be a partially ordered set with order $\preceq'$. Let further $g : \Theta \to \Theta'$ be such that $\vartheta_k \preceq \vartheta_l$ if and only if $g(\vartheta_k) \preceq' g(\vartheta_l)$ and $h : \mathbb{R} \to \mathbb{R}$ be strictly increasing. Define $g(\boldsymbol{\vartheta}) = (g(\vartheta_k))_{k=1}^m$. Then, for $j = 1, \dots, m$, $y \in \mathbb{R}$, $\alpha \in (0,1)$,*

$$\hat{q}_{g(\vartheta_j)}(\alpha;\, g(\boldsymbol{\vartheta}), h(\boldsymbol{y})) = h(\hat{q}_{\vartheta_j}(\alpha; \boldsymbol{\vartheta}, \boldsymbol{y})), \quad \hat{F}_{g(\vartheta_j)}(h(y);\, g(\boldsymbol{\vartheta}), h(\boldsymbol{y})) = \hat{F}_{\vartheta_j}(y; \boldsymbol{\vartheta}, \boldsymbol{y}).$$

Proposition 3.1 shows that when IDR is used to estimate the conditional distributions in Step 3, then it is sufficient to know the index θ up to increasing transformations. Moreover, any isotonic transformation can be applied to the response Y to simplify the estimation of θ

in Step 1, and then reverted by its inverse, without affecting the estimation of the conditional distributions. Hence, the task of estimating the index function θ is simplified to finding an estimator for a pseudo index that induces the same ordering on $\theta(x_i)$, $i = 1, \ldots, n$.

3.2 Index estimators

A simple but effective way to estimate the index in DIMs are classical generalized linear models. This might be surprising, because it seems that a parametric assumption has to be imposed on the distribution functions $(F_u)_u$ for this approach. However, due to the invariance of DIMs under monotone transformations (Proposition 3.1), it is sufficient that such a parametric assumption holds only approximately, in the sense that a monotone transformation of the index estimator converges to the index function; see Assumption (A4) in Section 5. The only requirement is that the linear predictor of the GLM exhibits an isotonic relationship with the outcome. This can be verified by the rank correlation between the index and the outcome, or by plots of the empirical distribution of the outcome stratified according to the index. A further advantage of this approach is that GLMs are well-understood, implemented efficiently in nearly every statistical software, and one can directly build on extant literature from non-distributional regression to find a suitable index estimator. The effectiveness of GLMs in the context of DIMs is demonstrated in the data application in Section 6.

Another powerful tool for index estimation in DIMs is quantile regression (Koenker, 2005). The stochastic ordering of the conditional distributions in DIMs is equivalent to the assumption that the conditional quantile functions $q_{\theta(x)}(\alpha)$ are increasing in the index $\theta(x)$ for every $\alpha \in (0, 1)$. One can thus estimate one or several quantiles by quantile regression, e.g. the median and/or the 90% quantile, and obtain estimates of the complete distribution by taking this (these) quantile(s) as the index (vector) in a DIM. Compared to the direct application of quantile regression for the estimation of conditional distributions, one does not need to specify a grid of quantiles over the whole unit interval and correct quantile crossings, but can focus on the estimation of a small number of quantiles that reveal the ordering of

the conditional distributions.

In the case of a distributional single index model $F_{\theta(X)}(y) = F_{\alpha_0^T X}(y)$, that is a DIM with $d = 1$, one might estimate the index α_0 via methods for single index models. For the monotone single index model, efficient estimators have been developed recently (Balabdaoui et al., 2019b; Balabdaoui and Groeneboom, 2020). Index estimators for the single index model, such the one proposed in Lanteri et al. (2020), also allow for non-monotone relationships between the index function $\alpha_0^T x$ and the response, and hence monotonicity should be checked carefully. Compared to GLMs as a pseudo index, single index models gain flexibility by not assuming any fixed functional form of the relationship between $\alpha_0^T X$ and the outcome Y. The drawbacks are that it is more difficult to accommodate high dimensional categorical variables and to let numeric covariates enter the index-function in a non-linear fashion, e.g. via polynomial or spline expansions, which is essential in our data application on ICU LoS. Since the DIM is already invariant under monotone transformations of the index function, it is questionable whether the benefits of using single index methods surpass these drawbacks. The same concerns are also valid for estimation methods for distributional single index models in the spirit of Hall and Yao (2005), which requires a notion of distance on the covariate space and is hence not directly applicable when categorical covariates are present.

3.3 Extension: Sample splitting and bagging

The estimation procedure suggested so far uses in-sample predictions with the estimated index function, $\hat{\theta}(x_i)$, as covariates for distributional regression with IDR. Depending on the index estimator, this strategy may be prone to overfitting. As a remedy, we propose a procedure in the spirit of (sub)sample aggregation (bagging).

Instead of estimating both the index function and the conditional distributions on the whole dataset, one may split the data (randomly) into two separate parts for these tasks, say $D_1 = \{1, \ldots, \lfloor n\xi \rfloor\}$ and $D_2 = \{\lfloor n\xi \rfloor + 1, \ldots, n\}$ for some $\xi \in (0, 1)$. The index function

is estimated with (x_i, y_i), $i \in D_1$, and the second part of the data with the *out-of-sample* predictions $\hat{\theta}(x_j)$, $j \in D_2$, serves as training data for IDR. To avoid that the estimated distribution functions depend on the random split of the training data, this procedure should be repeated several times, every time with a different split of the training data, and the conditional distribution functions are averaged in the end. The application of (sub-)sample aggregating ((sub-)bagging) has already been suggested in Henzi et al. (2019) in conjunction with IDR, where it yields smoother distribution functions and (in the case of subagging) reduces the computation time for larger datasets with multivariate covariates ($d \geqslant 2$). These advantages can also be expected for the DIM. In addition, the consistency result (Theorem 5.1) still holds under sample splitting when the data is split into D_1 and D_2 at a constant fraction $\xi \in (0, 1)$.

4 Prediction

This section reviews basic tools for the evaluation of probabilistic forecasts, and related properties of DIMs when used for forecasting. We denote by F a generic, random probabilistic forecast for a random variable Y, and all probability statements are understood with respect to the joint distribution of F and Y, which we denote by $\mathbb{P}$. For the distributional index model, the randomness of $F = F_{\theta(X)}$ is fully captured in the index $\theta(X)$.

As argued in Gneiting et al. (2007), *calibration* is a minimal requirement for probabilistic forecasts, meaning that the forecast should be statistically compatible with the distribution of the response. Of particular interest for DIMs is threshold calibration, requiring

$$\mathbb{P}(Y \leqslant y \mid F(y)) = F(y), \quad y \in \mathbb{R}. \tag{7}$$

It is shown in Henzi et al. (2019) that IDR, and hence also the DIM, is always in-sample threshold calibrated, that is, (7) holds when $\mathbb{P}$ is the empirical distribution of the training data used to estimate the distribution functions. Threshold calibration can be assessed by reliability diagrams (Wilks, 2011), in which estimated forecast probabilities $\hat{F}(y)$ are binned

and compared to the observed event frequencies in each bin. Another prominent tool for calibration checks is the probability integral transform (PIT)

$$Z = F(Y-) + V\left(F(Y) - F(Y-)\right), \tag{8}$$

where V is uniformly distributed on $[0,1]$ and independent of F and Y, and $F(y-) = \lim_{z \uparrow y} F(z)$. If Z is uniformly distributed, then the forecast F is said to be probabilistically calibrated. The PIT can be used to identify forecast biases as well as underdispersion and overdispersion (Diebold et al., 1998; Gneiting et al., 2007).

Among different calibrated probabilistic forecasts, the most informative forecast is arguably the one with the narrowest prediction intervals. This property, which only concerns the forecast distribution F, is referred to as *sharpness* (Gneiting et al., 2007). Sharpness and calibration are often assessed jointly by means of proper scoring rules (Gneiting and Raftery, 2007), which map probabilistic forecasts and observations to a numerical score. An important example is the CRPS defined at (3). IDR enjoys in-sample optimality among all stochastically ordered forecasts with respect to a broad class of proper scoring rules, including the CRPS and weighted versions of it, that is,

$$\mathrm{CRPS}_\mu(F, y) = \int_{\mathbb{R}} \left(F(z) - \mathbb{1}\{y \leqslant z\}\right)^2 \, \mathrm{d}\mu(z),$$

where μ is a locally-finite measure. This emphasizes that IDR is a natural way to estimate the probability distributions in DIMs, since it is not tailored to a specific loss function.

5 Consistency

5.1 Two stage estimation

We work with a triangular array of random elements $(X_{ni}, Y_{ni}) \in \mathcal{X} \times \mathbb{R}$, $i = 1, \ldots, n$, and assume that for all n, the following hold:

(A1) The random elements X_{ni}, $i = 1, \dots, n$, are independent and identically distributed, and Y_{ni}, $i = 1, \dots, n$, are independent conditional on $(X_{ni})_{i=1}^n$ with

$$\mathbb{P}(Y_{ni} \leqslant y \mid X_{ni}) = F_{\theta(X_{ni})}(y),$$

where $\theta : \mathcal{X} \to \mathbb{R}$ is a function and $(F_u)_{u \in \mathbb{R}}$ is a family of distributions such that $F_u \preceq_{\mathrm{st}} F_v$ if $u \leqslant v$.

(A2) There exists a constant $L > 0$ such that for all $u, v, y \in \mathbb{R}$,

$$|F_u(y) - F_v(y)| \leqslant L|u - v|.$$

(A3) On an interval I, the random variables $\theta(X_{ni})$ admit a density with respect to the Lebesgue measure which is bounded from below by $C_1 > 0$ and from above by C_2.

(A4) There exist a strictly increasing function $g : \mathbb{R} \to \mathbb{R}$ and a constant $C_0 > 0$ such that

$$\lim_{n \to \infty} \mathbb{P}\left(\sup_{x \in \mathcal{X}} |g(\hat{\theta}_n(x)) - \theta(x)| \geqslant C_0 (\log(n)/n)^{1/2} \right) = 0.$$

We denote by $\hat{F}_{n;u}$ the IDR estimator computed with training data $(\hat{\theta}_n(X_{nj}), Y_{nj})_{j=1}^n$, i.e.

$$\hat{F}_{n;u}(y) = \hat{F}_u(y; (\hat{\theta}_n(X_{nj}))_{j=1}^n, (Y_{nj})_{j=1}^n),$$

with the notation of Section 3.1.

Theorem 5.1 (Consistency of DIM)**.** *Under assumptions (A1)-(A4), there exists a constant* $C > 0$ *such that*

$$\lim_{n \to \infty} \mathbb{P}\left(\sup_{y \in \mathbb{R}, x \in \mathcal{X}_n} |\hat{F}_{n;\hat{\theta}_n(x)}(y) - F_{\theta(x)}(y)| \geqslant C \Big(\frac{\log n}{n} \Big)^{1/6} \right) = 0,$$

where $\mathcal{X}_n = \{x \in \mathcal{X} : [\theta(x) \pm (\log n/n)^{1/6}] \subseteq I\}$.

An analogous result to Theorem 5.1 can be shown for the variant of the DIM with sample splitting described in Section 3.3. The requirements under sample splitting are slightly

weaker, namely, the density of $\theta(X_{ni})$ does not have to be bounded from above in (A2), and in (A4), it is sufficient that the index estimator $\hat{\theta}_n$ converges at a rate of $o((\log(n)/n)^{1/3})$ instead of $n^{-1/2}$. The resulting convergence rate of the DIM with sample splitting is of order at least $(\log(n)/n)^{1/3}$. The proofs of Theorem 5.1, both, with and without sample splitting, rely on the consistency results about the monotonic least squares estimator in Mösching and Dümbgen (2020), and are given in Appendix A.

Assumption (A1) is the basic model assumption of DIMs. The Lipschitz-continuity in (A2) also appears in the monotone single index model for the mean (Balabdaoui et al., 2019a). Since the distributional single index model and the monotone single index model are equivalent when Y is binary, the Lipschitz assumption (A2) is natural in this context; also (A3) can be derived from the assumptions in Balabdaoui et al. (2019a). Assumptions (A2) and (A3) are required for the consistency of the monotone least squares estimator, with (A3) ensuring that the 'design points' $\theta(X_{nj})$ are dense enough in a region of interest, c.f. Mösching and Dümbgen (2020). A parametric model $\theta = \alpha_1\theta_1 + \cdots + \alpha_p\theta_p$ satisfies this assumption when at at least one of the summands $\alpha_i\theta_i$ admits a continuous distribution on I with density bounded away from zero. In (A4), we require uniform consistency of a monotone transformation of the index estimator at a rate of $n^{-1/2}$, i.e. not necessarily consistency of the index estimator itself. In a parametric model $\theta = \alpha_1\theta_1 + \cdots + \alpha_p\theta_p$, uniform consistency is satisfied for any $\sqrt{n}$-consistent estimator of the coefficients $\alpha_1, \ldots, \alpha_p$, when the functions $\theta_1, \ldots, \theta_p$ are bounded. All estimators suggested in Section 3.2 are consistent at the rate $n^{-1/2}$ under suitable conditions.

5.2 Simultaneous estimation

In this subsection, we treat the question to what extent simultaneous estimation of the index and the distribution functions is possible and sensible in the DIM. Currently, the results are of theoretical interest only.

It has been shown in Balabdaoui et al. (2019a) that for the monotone single index model,

there exists a simultaneous minimizer (ψ_0, α_0) of the squared error

$$\sum_{i=1}^{n}(\psi_0(\alpha_0^T x_i) - y_i)^2$$

where $\psi_0 : \mathbb{R} \to \mathbb{R}$ is an increasing function, $\alpha_0 \in \{x \in \mathbb{R}^p : \|x\| = 1\}$ is the index, and $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n$. The minimizer is in general not unique.

A similar result also holds in the distributional index model, when the loss function is defined as

$$l(\hat{\theta}, \hat{\boldsymbol{F}}) = \sum_{i=1}^{n} \mathrm{CRPS}(\hat{F}_{\hat{\theta}(x_i)}, y_i). \tag{9}$$

For basis functions $\theta_1, \ldots, \theta_p$ of the vector space $\mathcal{F}$ containing the true index function θ, every index estimator $\hat{\theta} : \mathcal{X} \to \mathbb{R}^d$ can be written as $\hat{\theta} = \hat{\alpha}_1\theta_1 + \cdots + \hat{\alpha}_p\theta_p$. The loss (9) has a unique minimizer $\hat{\boldsymbol{F}} = (\hat{F}_{\hat{\theta}(x_i)}, \ldots, \hat{F}_{\hat{\theta}(x_n)})$ for fixed $\hat{\theta}$, namely the IDR. This minimizer only depends on $\hat{\theta}$ via the partial order on the points $\hat{\theta}(x_i)$, $i = 1, \ldots, n$. But the number of partial orders on n points is finite, and so there exists a minimizer of (9).

In general, the number of partial orders induced by index functions $\hat{\theta}$ is too large for a direct minimization of (9) to be possible: When $\mathcal{X} = \mathbb{R}^p$ and $\theta_1, \ldots, \theta_p$ are the coordinate projections, then the number of total orders grows at a rate of $n^{2(p-1)}$ (Balabdaoui et al., 2019a). Moreover, when the index space is partially but not totally ordered, trivial solutions (a perfect fit to the training data) may appear, namely if the points $\hat{\theta}(x_i)$, $i = 1, \ldots, n$, are all incomparable in the partial order. Hence, the simultaneous estimation of the index and the distribution functions in DIMs is generally not feasible. A related interesting question for further research is to find a way to directly parametrize and estimate partial orders for isotonic or isotonic distributional regression, instead of indirectly via an index function.

6 Data application

We apply a DIM to derive probabilistic forecasts for intensive care unit (ICU) length of stay (LoS) based on patient information available 24 hours after admission. The main difficulty

of such predictions is that, even conditional on many demographic and physiologic patient specific covariates, there is often great uncertainty in the LoS. In addition to unknown factors (e.g. frailty status, patient or family wishes), the LoS also depends on non-patient-related information such as ICU organization and resources. We therefore model the LoS using data of single ICUs rather than a merged dataset, thus keeping the ICU-related variables fixed. This allows forecasts within each single ICU as well as the comparison of the forecasted LoS of patients across ICUs. The same methodology can also be used on a joint dataset of several ICUs, giving a reference LoS forecast on the combined case-mix. Using these predictions for risk-adjustment and benchmarking is promising but goes beyond the scope of this paper.

All computations in this application were performed in R 4.0 (R Core Team, 2020) using the packages `mgcv` (Wood, 2017) for the estimation of index models and Cox proportional hazards regression, `quantreg` (Koenker, 2020) for quantile regression, and `isodistrreg` (Henzi et al., 2019, https://github.com/AlexanderHenzi/isodistrreg) for IDR.

6.1 Data and variables

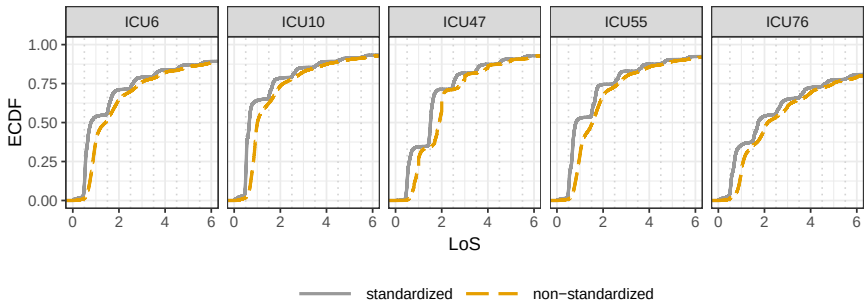
Since 2005, the Swiss Society of Intensive Care Medicine collects ICU key figures and information on patient admissions in the Minimal Dataset of the Swiss Society of Intensive Care Medicine (MSDi). Our analysis is based on a part of this dataset suitable for LoS predictions, namely, we include 18 out of 86 ICUs which, after the application of selection criteria described below, include more than 10'000 patient admissions. The codes used as identifiers for the ICUs were generated randomly. The sample sizes range from 10'041 to 36'865 with an average of 17'181 observations per ICU. The cutoff of 10'000 is based on our experience with IDR and probabilistic forecasts in general, which require sufficiently large datasets for a meaningful and stable evaluation, especially when the models involve large numbers of covariates and a skewed response variable, as it is the case here. However, the prediction methods can also be applied to smaller datasets.

Based on literature review, we identified the variables described in Table 1 as relevant

Table 1: Covariates used for ICU length of stay predictions. Availability of the variables is given by 'admission' (at patient admission) or by the number of hours after admission.

Variable	**Availability**	**Description**
Age	admission	patient age at admission
Sex	admission	male, female
Planned	admission	admission is announced at least 12h in advance (true/false)
Readmission	admission	patient was discharged from the same ICU at most 48 hours ago (true/false)
Admission source	admission	admission source (emergency room; intermediate care unit, high dependency unit, recovery room; hospital ward; surgery; others)
Location before hospital admission	admission	location before *hospital* admission (home; other hospital; others)
Diagnosis	24h	main diagnosis on first day (structured into: cardiovascular, respiratory, gastrointestinal, neurological, metabolic, trauma, others; in total 36 different specific ICU relevant diagnoses)
NEMS	8h	NEMS (Miranda et al., 1997) over first shift after patient admission (8-12h)
SAPS II	24h	Le Gall et al. (1993)
Interventions	24h	interventions 24 hours before until 24 hours after admission (13 categories of interventions, e.g. surgeries, interventions in respiratory system, cardiovascular interventions)

Figure 1: Empirical distribution functions of the standardized and non-standardized LoS for selected ICUs. The standardized LoS is defined as $Y - 1 + h/24$, where h is the admission hour of a patient and Y is the non-standardized LoS, i.e. the time between patient admission and discharge. Only patients with positive standardized LoS are included.



for LoS forecasts (Zimmerman et al. (2006); Verburg et al. (2014, Table S1); Niskanen et al. (2009)). We exclude patients that were transferred from or to another ICU, because their LoS is incomplete. As in Zimmerman et al. (2006), we also remove patients younger than 16 years and patients admitted after transplant operations or because of burns. Patients with missing values in the variables in Table 1 are excluded, too.

Table 1 documents at what time after admission the relevant covariates for LoS predictions are available. While all variables are available 24 hours after patient admission, the information is completed also for patients staying at the ICU less than one day. For example, ICU interventions within the first 24 hours are then only interventions performed until patient discharge, and the SAPS II is computed based on the worst physiological values until discharge instead of the worst values in the first 24 hours at the ICU.

In preliminary tests, we found that for probabilistic LoS forecasts, the usual definition of LoS as the time between patient admission and discharge is problematic, because most ICUs discharge patients during specific time windows, but the admission times are spread throughout the day. As a consequence, it may happen that the predicted LoS for certain patients does not conform with the discharge practice of a ICU, e.g. there might be a high

predicted probability for a patient being discharged around midnight but the ICU actually discharges patients in the early afternoon. To circumvent this problem, we decided to measure the LoS as the *time between the next midnight after patient admission until discharge*, thereby standardizing all admission to the same (day)time and revealing the true pattern in the patient discharge times; see Figure 1. All results in this section use this definition of the LoS. Patients who do not stay over at least two calendar days are excluded, which is unproblematic since in practice, the data required for predictions is only available 24 hours after admission and the forecast should be conditioned on the event that the patient already stayed at the ICU for 24 hours. Forecasts for the non-standardized LoS, i.e. the time between admission and discharge, can be derived via the relation

$$\mathbb{P}(Y > 1 + t | Y > 1) = \frac{\mathbb{P}(\tilde{Y} > t + h/24 | \tilde{Y} > 0)}{\mathbb{P}(\tilde{Y} > h/24 | \tilde{Y} > 0)},$$

where Y and $\tilde{Y} = Y - 1 + h/24$ denote the LoS and the standardized LoS measured in days, respectively, and h the admission hour of a given patient. Since only patients staying at least until midnight of the admission day are used as training data, our LoS forecasts are conditioned on the event $\{\tilde{Y} > 0\}$ in the above equation.

We select the most recent 20% of the observations in each ICU for model validation, thereby mimicking a realistic situation in which past data are used to predict the LoS of present and future patients. This implies that forecasts might be inaccurate if the relationship between the covariates and LoS changes over time, and it is part of our analysis to check to what extent past data can be reasonably used to predict the LoS of future patients. Of the remaining data, randomly selected 75% are used for model fitting and 25% for model selection via out-of-sample predictions. All comparisons of different variants of a distributional regression model were performed by such out-of-sample predictions.

6.2 Derivation of DIM

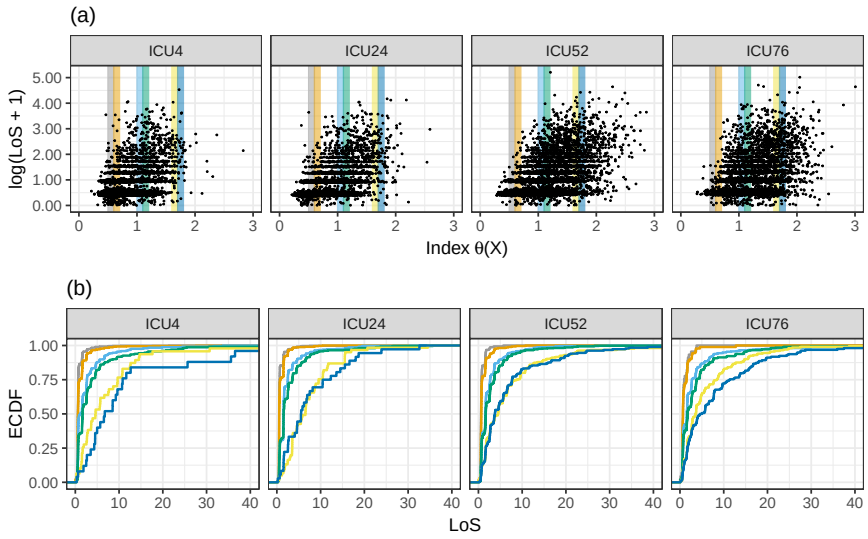
To derive an index estimator for the DIM, we can benefit from the comparisons of regression models for point forecasts for LoS in the extant literature. Moran and Solomon (2012) and

Verburg et al. (2014) found that a Gaussian linear regression for the expected log-LoS is suitable for point forecasts, and we use this as our candidate for the index estimator and will refer to it as the 'lognormal index model'. We use the transformation $y \mapsto \log(y+1)$, which results in more symmetric distributions than the logarithm. All variables from Table 1 were included in the model, and the effects of the continuous variables age, SAPS and NEMS were modeled by cubic regression splines. Interactions of variables were explored but not included in the final model. We also tested whether merging factor levels with few observations improved the model, but the untransformed covariates yielded the best forecasts in out-of-sample predictions on the part of the data used for model selection.

We tested two other index estimators for the expected LoS to investigate the robustness of the DIM with respect to the index. The first one estimates the expected log-LoS under the assumption of a scaled t-distribution. The mean is modeled as a function of the covariates, with the same specification as for the lognormal index model, and the degrees of freedom are estimated, with a minimal threshold of 5 to ensure stability. This model is structurally similar to the lognormal index model, but more robust with respect to outliers, which occur even after the log-transformation. The second alternative is a gamma regression for the untransformed LoS with logarithm as the link function. While the three index models yield different predictions on the scale of the LoS, they largely agree when only the *ordering* of the predictions is considered: Over the 18 ICUs, the rank correlation between predictions by two of the models is 0.98 on average with a minimum of 0.86. As a consequence, there is no significant difference between the corresponding DIM forecasts: Evaluated on the dataset for model selections, the average CRPS over all ICUs of DIM forecasts based on different models only differs by up to 0.01, while the averages are around 1.40. The predictions based on the lognormal index model achieved the best results in most ICUs and were therefore selected for the predictions on the validation data.

Due to the large training datasets, splitting of the training data as described in Section 3.3 only has a marginal effect on the predictions. Estimating the index function on the full training data and the conditional distributions on in-sample predictions only increased the

Figure 2: (a) Index function and log(LoS + 1) for selected ICUs. (b) ECDFs of the LoS stratified into the bins given by the vertical shaded stripes in panel (a).



average CRPS by 0.01 (on 1.40), compared to a bagging approach with 100 random splits of the training data into equally sized parts for the estimation of the index and the CDFs. For the final evaluation, we show the results of the simpler variant without bagging.

Figure 2 illustrates how to perform a check of the stochastic ordering assumption of the DIM: We bin the observed LoS according to the index value, and plot the empirical cumulative distribution functions (ECDFs) of the LoS in each bin. By varying the positions and sizes of the bins, it can be seen that the empirical distributions are indeed sufficiently well ordered. The Spearman correlation between the index and the observed LoS is 0.53 on average over all ICUs (range $0.40-0.65$), which confirms that there is an isotonic relationship between the index and the actual LoS for most ICUs, taking into account the high uncertainty in the LoS of ICU patients even conditional on patient information collected at the first day.

6.3 Alternative regression methods

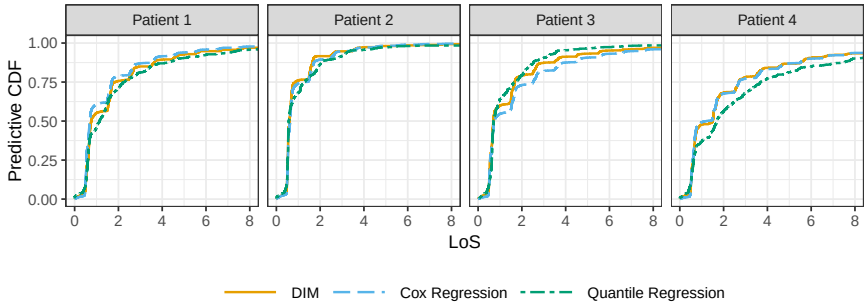
We compare the DIM to two other distributional regression methods: A Cox proportional hazards model (Cox, 1972) and quantile regression with monotone rearrangement (Koenker, 2005; Chernozhukov et al., 2010). For both, we use the same variables and specifications as in the index estimator for the DIM, which was superior compared to other variants tested; detailed results are provided in the supplementary material.

A Cox proportional hazards model is a classical choice for modeling survival times, and it shares some similarities with a DIM. Both models are semi-parametric and based on stochastic order restrictions on the conditional distributions, namely the usual stochastic ordering in the DIM and the hazard rate order in Cox regression, which is stronger than the usual stochastic order (Shaked and Shanthikumar, 2007, Theorem 1.B.1). While the distribution functions are estimated non-parametrically in Cox regression, the relationship between different conditional distributions is modeled parametrically via the hazard ratio, as opposed to the DIM, where only the ordering on the conditional distributions is modeled parametrically by the index function.

Quantile regression, on the other hand, imposes less assumptions on the conditional distributions. The conditional quantiles are modeled separately and satisfy no stochastic order constraints. In particular, if there are strong violations of the stochastic order assumptions of the DIM or Cox regression, we would expect that the more flexible quantile regression achieves better forecasts by fitting crossing quantile curves for different patients. This allows an informal check of the underlying assumptions of Cox regression and the DIM (see Figure S1 in the supplementary material). We use a grid of quantiles from 0.005 to 0.995 with steps of 0.001, which gave better results than a coarser grid with steps of 0.01.

We also tested fully parametric models of GAMLSS type, and kernel methods as implemented in the `np` package in R (Hayfield and Racine, 2008). Unfortunately, we could not find a sufficiently flexible parametric family for a GAMLSS, and the application of kernel methods was not feasible due to computational problems with the large datasets and high

Figure 3: Predictive CDFs for four selected patients based on the training data of the ICU the patients were admitted to.



numbers of covariates. As for the DIM, computation is obviously more demanding than for fully parametric methods, but still fast thanks to the sequential implementation of IDR described in Henzi et al. (2020). On a personal computer with Intel(R) Core i7-8650 CPU, computation with the lognormal index model without bagging takes 3 seconds for the smallest ICU (6'024 observations in training dataset) and 25 seconds for the largest ICU (22'219 observations). Estimation and prediction on the total dataset (all 18 ICUs) require about 2.5 minutes.

6.4 Results

Figure 3 illustrates the probabilistic forecasts for different patients based on the training data of the ICU the patients were admitted to. Patient 1, male, 32 years old, was admitted because of a severe sepsis or septic shock. Patient 2 is a 67 years old female with aortic aneurysm or aortic dissection, Patient 3 is 58 years old, male with a metabolic decompensation, and Patient 4 is a 78 old female admitted from a high dependency unit with subarachnoidal hemorrhage. Patient 2 has the shortest predicted LoS: The DIM and Cox regression predict that she leaves the ICU at the first day after admission with a probability of almost 75%. For the remaining patients, the predictive CDFs are more skewed, and a LoS of more than

Table 2: Summary statistics (mean, median and standard deviation) of numeric variables in the dataset.

ICU	LoS			Age			NEMS			SAPS		
	mean	med.	sd	mean	med.	sd	mean	med.	sd	mean	med.	sd
ICU44	3.9	1.5	7.8	59.0	61	17.6	27.1	27	8.5	34.0	31	18.9
ICU65	1.8	0.6	4.3	67.2	69	13.9	25.5	25	7.9	28.7	28	12.5
ICU76	4.3	1.7	7.2	63.2	66	15.6	30.3	30	8.3	41.2	40	17.2
ICU77	1.8	0.6	3.2	65.0	68	15.9	21.9	18	8.0	31.1	28	16.1

three days is not unlikely. It is immediately visible that the DIM and Cox regression are able to recover the pattern in the ICU discharge times, with flat pieces of the CDFs around midnight. Quantile regression, on the other hand, merely interpolates this pattern.

Here, detailed results are only shown for the best and worst two ICUs with respect to the CRPS of the DIM forecasts; see the supplementary material for tables and figures for all ICUs. Summary statistics of the LoS and other numeric variables for the patients of these ICUs are given in Table 2. All probabilistic regression methods can reliably predict the probability that the LoS exceeds $k = 1, 5, 9, 13$ days; see Figure 4. Figure 5 shows that the forecasts achieve a better probabilistic calibration than the ECDF of the LoS in the training data, which is uninformative as a forecast and does not take into account changes in the ICU-case mix that are reflected in the covariates. Further improvements of calibration may be possible by selecting a tailored training dataset, taking into account organizational changes, and developments in treatments that have an influence on the LoS or on the relationship between covariates and the LoS. Such information is not available in our dataset.

While all three distributional regression methods yield similar results in terms of calibration, there is a clear ranking with respect to forecast accuracy: In all ICUs, the DIM achieves the lowest CRPS, followed by quantile regression in second and Cox regression in third place. For comparison, Table 3 also shows the CRPS of the ECDF forecast, and of

Figure 4: Reliability diagrams of probabilistic forecasts for the predicted probability that the LoS exceeds 1, 5, 9, 13 days. The forecast probability is grouped into the bins $[0, 0.1], (0.1, 0.2], \ldots, (0.9, 1]$ and the observed frequencies are drawn at the midpoints of the bins. Only bins with more than two observations are included.

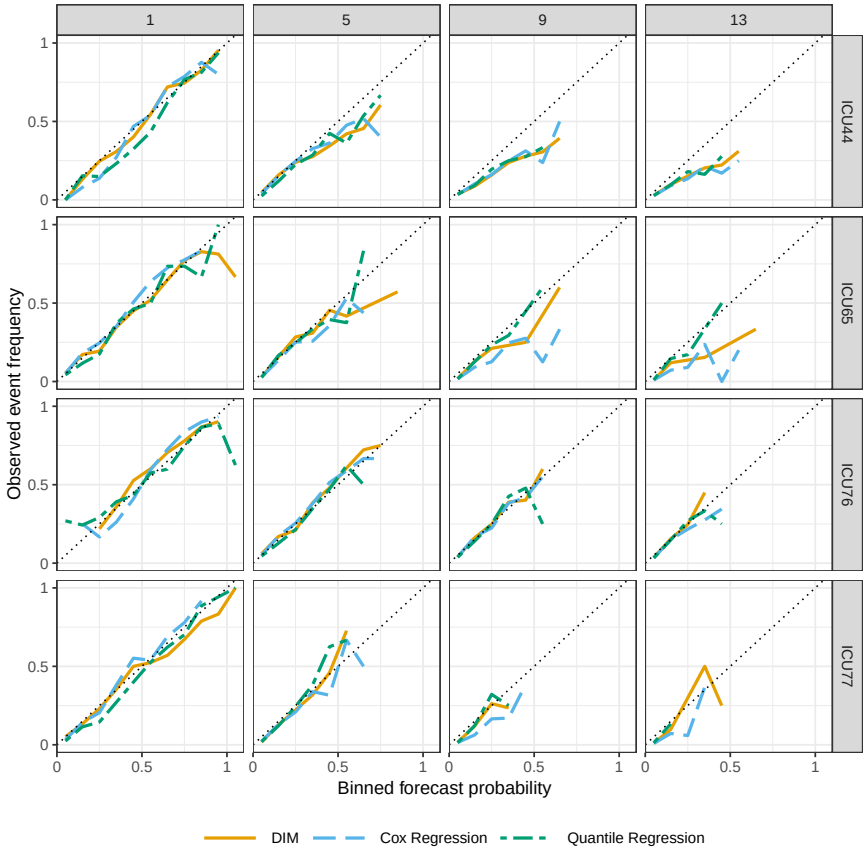
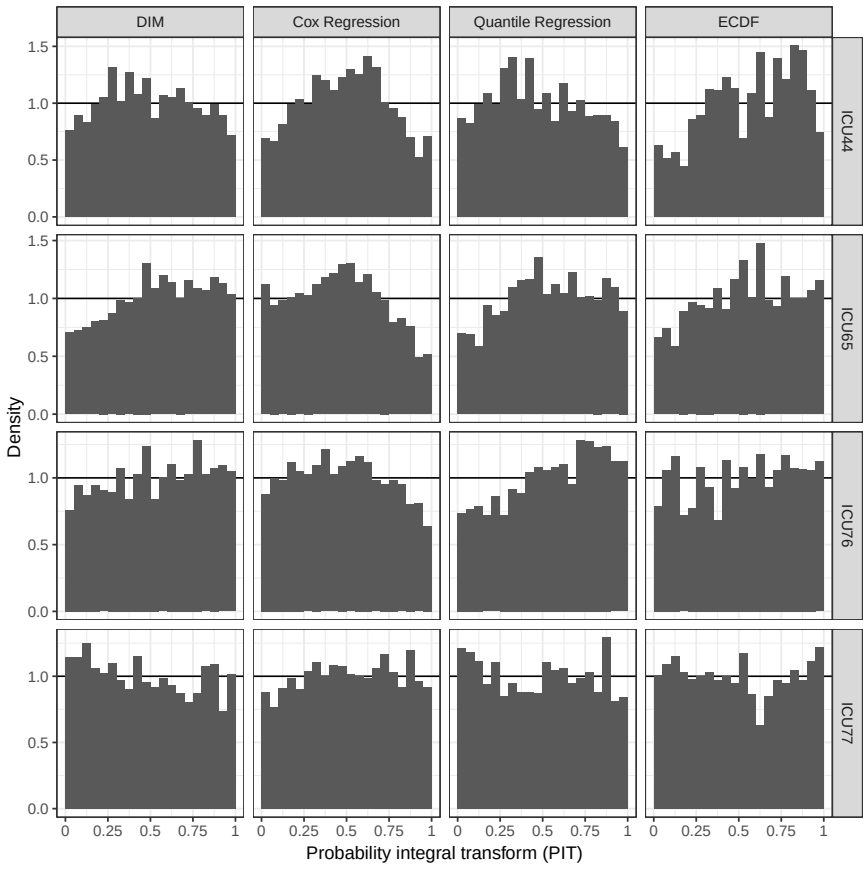


Figure 5: PIT histograms of the probabilistic forecasts with bins of width 1/20.



the deterministic point forecast of the lognormal index model, which is its mean absolute error. Interestingly, the ECDF forecast achieves a lower mean CRPS in all ICUs (average improvement of 13%) than the point forecast, although it does not take any covariate information into account. This highlights the superiority of even simple probabilistic forecast over point forecasts in the context of ICU LoS. A further average improvement of 13% in the mean CRPS is achieved when going from the uninformative ECDF forecast to the worst of the probabilistic regression methods in terms of CRPS, which is Cox regression. The differences in the CRPS of the forecasts using distributional regression methods are smaller, but consistent over the ICUs: In terms of average CRPS, quantile regression outperforms Cox regression in 15 out of 18 ICUs, and the DIM outperforms Cox regression in all and quantile regression in all except 2 ICUs. The difference in CRPS between the DIM and quantile regression is highly significant when tested with Wilcoxon's signed rank test except for the ICUs with identifiers 19 and 33, where the p-values are 0.101 and 0.219 and quantile regression achieves lower average scores. Wilcoxon's signed rank test was applied because the CRPS differences are heavy-tailed, so a t-test is not appropriate (see Figure S6 in the supplementary material).

In conclusion, with distributional regression methods and especially the DIM, it is possible to obtain reliable, reasonably well calibrated, and informative probabilistic forecasts for ICU LoS in a realistic setting. These forecasts are not only more informative than point forecasts, but also reduce the forecast error by more than 25%.

7 Discussion

In this paper, we have introduced DIMs as intuitive and flexible models for distributional regression. Distributional regression approaches provide full conditional distributions of the outcome given covariate information, and are thus more informative than classical regression approaches for the conditional mean, median or specific quantiles. However, specifying a good distributional regression model is usually less intuitive than specifying a regression

Table 3: CRPS of probabilistic forecasts. The column 'Point' shows the mean absolute error of the point forecast obtained from the lognormal index model, and p-values of Wilcoxon's signed rank test for the difference in CRPS between DIM and quantile regression are given in the column labelled p. P-values smaller than 10^{-16} are written as 0.

ICU	p	DIM	Quantile reg.	Cox reg.	ECDF	Point
ICU4	$1.18 \cdot 10^{-11}$	1.074	1.076	1.089	1.191	1.399
ICU6	$3.81 \cdot 10^{-12}$	1.360	1.385	1.386	1.605	1.830
ICU10	0	1.194	1.221	1.209	1.312	1.553
ICU19	$1.01 \cdot 10^{-1}$	1.041	1.032	1.048	1.189	1.350
ICU20	$5.13 \cdot 10^{-6}$	2.216	2.223	2.241	2.505	2.859
ICU24	0	1.099	1.111	1.141	1.265	1.416
ICU33	$2.19 \cdot 10^{-1}$	0.975	0.974	0.983	1.090	1.363
ICU39	$1.38 \cdot 10^{-16}$	1.332	1.352	1.383	1.697	1.872
ICU44	$1.06 \cdot 10^{-3}$	2.256	2.259	2.328	2.480	2.952
ICU47	$3.69 \cdot 10^{-5}$	0.977	0.980	1.036	1.231	1.363
ICU52	$7.40 \cdot 10^{-5}$	1.845	1.866	1.868	2.121	2.580
ICU55	0	1.062	1.085	1.055	1.253	1.445
ICU58	$1.25 \cdot 10^{-15}$	1.393	1.409	1.442	1.763	1.970
ICU65	0	0.908	0.914	0.981	1.062	1.194
ICU76	0	2.420	2.448	2.458	2.783	3.468
ICU77	$1.76 \cdot 10^{-16}$	0.921	0.936	0.938	1.117	1.260
ICU79	$1.86 \cdot 10^{-11}$	1.446	1.457	1.512	2.172	2.228
ICU80	0	0.942	0.971	0.949	1.094	1.253
Mean		1.359	1.372	1.392	1.607	1.853

model for, say, the conditional mean. An appealing feature of DIMs is that for the modeling of the index function classical approaches and intuition for modeling a conditional mean or median can be used. Given the index function, the shape of the full conditional distribution is then learned from training data using IDR, that is, distributional regression under stochastic ordering constraints. The second step does not involve any parameter tuning or implementation choices.

The idea of reducing the complexity of a potentially high-dimensional covariate space by using an index function in distributional regression has also been used in the work of Hall and Yao (2005); Zhang et al. (2017). In these works, the index function has to be univariate and parametrizes a distance on the covariate space that is then used for kernel methods to estimate the conditional distributions. In contrast, the index function in a DIM parametrizes partial orders on the covariate space allowing for stochastic order constrained distributional regression in the second step.

Finding an informative index function is critical and usually requires expertise of the problem at hand. However, in many cases, existing models for the conditional mean or median can be used directly, as demonstrated in the application on ICU LoS. Indeed, it may even happen that a poorly fitting conditional mean model works well for a DIM since it is sufficient that the model is correct up to monotone transformations, or, in other words, that it is a good model for a pseudo index.

The distributional regression approach in Chernozhukov et al. (2020) allows to accomodate continuous, discrete and mixed discrete-continuous outcomes. The same is true for IDR, and thus for DIM models. While the case study in this paper concerns a continuous outcome, IDR has been successfully applied to a mixed discrete-continuous outcome in Henzi et al. (2019). It would be interesting to investigate the different benefits and drawbacks of DIM models versus the methods of Chernozhukov et al. (2020) in particular in the case of discrete outcomes.

Since IDR can be combined well with (sub-)bagging, the same also holds for DIMs. (Sub-)bagging is useful to avoid overfitting, may increase computational efficiency, and lead

to smoother estimated conditional CDFs. We have explored bagging in our data application in Section 6 with relatively at hoc choices for the number of random splits of the training data. A systematic study of optimal choices for subsample sizes and/or iterations is desirable.

A promising future extension of DIMs is to replace the IDR step by distributional regression under a stronger stochastic ordering constraint such as a likelihood ratio ordering constraint, or by a weaker one such as second order stochastic dominance. However, this requires fundamental advances concerning the estimation of distributions under these constraints.

A Proof of Theorem 5.1

The following lemma is Theorem 4.6 in Mösching and Dümbgen (2020), which we state for completeness.

Lemma A.1. *Let $Z_1, Z_2, Z_3, \ldots$ be independent random variables with respective distribution functions $G_1, G_2, G_3, \ldots$. For $k \in \mathbb{N}$, let*

$$\hat{\mathbb{G}}_k(\cdot) = \frac{1}{k}\sum_{i=1}^{k} \mathbb{1}\{Z_i \leqslant \cdot\} \quad \textit{and} \quad \bar{G}_k(\cdot) = \frac{1}{k}\sum_{i=1}^{k} G_i(\cdot).$$

Then there exists a universal constant $M \leqslant 2^{5/2}e$ such that for all $\eta \geqslant 0$,

$$\mathbb{P}\left(\sqrt{k}\|\hat{\mathbb{G}}_k - \bar{G}_k\|_\infty \geqslant \eta\right) \leqslant M\exp(-2\eta^2),$$

where $\|\cdot\|_\infty$ denotes the usual supremum norm of functions.

The results and proofs below use the following definitions. We denote by $\lambda(J)$ the Lebesgue measure of a measurable set $J \subset \mathbb{R}$, and define the events

$$B_n = \left\{\sup_{x\in\mathcal{X}} |g(\hat{\theta}_n(x)) - \theta(x)| < C_0(\log(n)/n)^{1/2}\right\}. \tag{10}$$

For $1 \leqslant r \leqslant s \leqslant n$ and a permutation σ of $\{1, \ldots, n\}$, let

$$w_{rs} = s - r + 1, \qquad \hat{\mathbb{F}}^{\sigma}_{rs} = \frac{1}{w_{rs}} \sum_{i=r}^{s} \mathbb{1}\{Y_{n\sigma(i)} \leqslant \cdot\},$$

$$\bar{F}^{\sigma}_{\theta;rs}(\cdot) = \frac{1}{w_{rs}} \sum_{i=r}^{s} F_{\theta(X_{n\sigma(i)})}(\cdot), \qquad \bar{F}^{\sigma}_{\hat{\theta};rs}(\cdot) = \frac{1}{w_{rs}} \sum_{i=r}^{s} F_{\hat{\theta}_n(X_{n\sigma(i)})}(\cdot).$$

We use π to denote a permutation such that $\hat{\theta}_n(X_{n\pi(1)}) \leqslant \cdots \leqslant \hat{\theta}_n(X_{n\pi(n)})$. The permutation π is a function of $(X_{ni}, Y_{ni})_{i=1}^n$ via $(X_{ni})_{i=1}^n$ and $\hat{\theta}_n$. Let

$$M_n^{\pi} = \max_{1 \leqslant r \leqslant s \leqslant n} w_{rs}^{1/2} \|\hat{\mathbb{F}}^{\pi}_{rs} - \bar{F}^{\pi}_{\theta;rs}\|_{\infty}. \tag{11}$$

Lemma A.2. *Under (A3) and (A4), there exists a constant $s = s(C_0, C_2) > 0$ such that*

$$\lim_{n \to \infty} \mathbb{P}(M_n^{\pi} \geqslant s n^{1/4} \log(n)^{1/4}) = 0.$$

Proof. Define $m = m(n) = \max(1, \lfloor \lambda(I)/(2c_n) \rfloor)$ with $c_n = C_0(\log(n)/n)^{1/2}$, where C_0 is from assumption (A4). Then, for n large enough such that $c_n \leqslant \lambda(I)/4$,

$$2c_n \leqslant \frac{\lambda(I)}{m} \leqslant 4c_n. \tag{12}$$

Slice the interval I from (A3) into m equally sized, disjoint intervals $J_1, \ldots, J_m$ (ordered increasingly). Let $\mathcal{I}_k = \{i \in \{1, \ldots, n\} : \theta(X_{ni}) \in J_k\}$, $n_k = \#\mathcal{I}_k$ for $k = 1, \ldots, m$, and $N_n = \max_{k=1,\ldots,m} n_k$. Define also $\mathcal{I}_j = \varnothing$ for $j \notin \{1, \ldots, m\}$ and $\bigcup_{i=a}^{b} A_i = \varnothing$ for any sets A_i and $a > b$.

Let $r, s \in \{1, \ldots, n\}$, $r \leqslant s$, be indices that attain the maximum in (11), and define the index set $\mathcal{I}^* = \pi(\{r, \ldots, s\})$, so that

$$M_n^{\pi} = \Big\| \frac{1}{(\#\mathcal{I}^*)^{1/2}} \sum_{i \in \mathcal{I}^*} \big(\mathbb{1}\{Y_{ni} \leqslant \cdot\} - F_{\theta(X_{ni})}(\cdot) \big) \Big\|_{\infty}.$$

Note that the indices r and s are (complicated but measurable) functions of (X_i, Y_i), $i = 1, \ldots, n$, and thus random variables. Therefore, the set $\mathcal{I}^*$ is also a random set of indices.

If $i, j \in \mathcal{I}^*$ and $g(\hat{\theta}_n(X_{ni})) < g(\hat{\theta}_n(X_{nj}))$, with g from (A4), then $k \in \mathcal{I}^*$ for all k such that $g(\hat{\theta}_n(X_{ni})) < g(\hat{\theta}_n(X_{nk})) < g(\hat{\theta}_n(X_{nj}))$. This follows from $\hat{\theta}_n(X_{n\pi(1)}) \leqslant \cdots \leqslant \hat{\theta}_n(X_{n\pi(n)})$,

because if $i = \pi(i_0)$, $j = \pi(j_0)$ and $k = \pi(k_0)$, then $g(\hat{\theta}_n(X_{ni})) < g(\hat{\theta}_n(X_{nk})) < g(\hat{\theta}_n(X_{nj}))$ implies that $i_0 < k_0 < j_0$, and $k_0 \in \{i_0, \dots, j_0\} \subseteq \{r, \dots, s\}$ gives $k = \pi(k_0) \in \pi(\{r, \dots, s\}) = \mathcal{I}^*$.

Under the event B_n defined at (10), $i \in \mathcal{I}_k$ and (12) imply that $g(\hat{\theta}(X_{ni})) \in J_t$ for some $t \in \{k-1, k, k+1\}$. Therefore, for $l, k \in \{1, \dots, m\}$ with $l - k > 2$, it follows $g(\hat{\theta}(X_{ni})) < g(\hat{\theta}(X_{nj}))$ for all $i \in \mathcal{I}_k$ and $j \in \mathcal{I}_l$. So if $\mathcal{I}^*$ contains indices $i \in \mathcal{I}_k$ and $j \in \mathcal{I}_l$ with $l - k > 2$, then $\mathcal{I}^*$ must also contain all elements of the sets $\mathcal{I}_t$ for $k + 2 < t < l - 2$. Let $\kappa = \min\{j \in \{1, \dots, m\} : \mathcal{I}_j \cap \mathcal{I}^* \neq \varnothing\}$, $\ell = \max\{j \in \{1, \dots, m\} : \mathcal{I}_j \cap \mathcal{I}^* \neq \varnothing\}$. By the previous considerations, $\mathcal{I}^*$ may contain arbitrary elements of $\mathcal{I}_t$ with $t \in \{\kappa, \kappa+1, \kappa+2, \ell-2, \ell-1, \ell\}$, and it must contain all indices in $\mathcal{I}_j$ for $\kappa + 3 \leqslant j \leqslant \ell - 3$. In conclusion, under B_n, $\mathcal{I}^*$ is almost surely contained in the collection of index sets defined by

$$S_n = \bigcup_{1 \leqslant k \leqslant l \leqslant m} \left\{ \mathcal{J} \cup \left(\bigcup_{t=k+3}^{l-3} \mathcal{I}_t \right) : \mathcal{J} \subseteq \left(\bigcup_{t=k}^{k+2} \mathcal{I}_t \right) \cup \left(\bigcup_{t=l-2}^{l} \mathcal{I}_t \right) \right\}.$$

Indeed, on the event B_n, we know that $\mathcal{I}^*$ must contain all elements of $\mathcal{I}_j$ for $\kappa+3 \leqslant t \leqslant \ell-3$. This explains the part $\bigcup_{t=k+3}^{l-3} \mathcal{I}_t$ in the definition of S_n. As for the $\mathcal{I}_k$ with subscript not in $\{\kappa+3, \dots, \ell-3\}$, $\mathcal{I}^*$ may contain any arbitrary selection from their elements. This arbitrary selection is $\mathcal{J} \subseteq \left(\bigcup_{t=k}^{k+2} \mathcal{I}_t\right) \cup \left(\bigcup_{t=l-2}^{l} \mathcal{I}_t\right)$. For κ and ℓ, all pairs (k, l) with $k \leqslant l$ are possible, which gives the union over $1 \leqslant k \leqslant l \leqslant n$.

Because $\#\mathcal{I}_t \leqslant N_n$ for all t, one can derive from the definition of S_n that

$$\#S_n \leqslant m^2 \cdot 2^{6N_n} = m^2 \exp(6 \log(2) N_n).$$

We now compute an upper bound for N_n, which is a function of $\theta(X_{n1}), \dots, \theta(X_{nn})$ only. Denote by P and G the distribution and the CDF of $\theta(X_{n1})$, and by $\hat{P}$ and $\hat{G}$ the empirical

distribution and the empirical CDF of $\theta(X_{n1}), \ldots, \theta(X_{nn})$. For any $c \geqslant 0$,

$$\begin{aligned}
\mathbb{P}(N_n \geqslant c) &\leqslant \sum_{k=1}^{m} \mathbb{P}(n_k \geqslant c) \\
&\leqslant \sum_{k=1}^{m} \mathbb{P}\Big(\hat{P}(J_k) - P(J_k) \geqslant \frac{c}{n} - P(J_k)\Big) \\
&\leqslant \sum_{k=1}^{m} \mathbb{P}\Big(2\|G - \hat{G}\|_\infty \geqslant \frac{c}{n} - P(J_k)\Big).
\end{aligned}$$

For n sufficiently large, $P(J_k) \leqslant 4C_2C_0c_n = 4C_2C_0(\log(n)/n)^{1/2}$ by (12) and by (A3). Replacing c by $d_n = R\log(n)^{1/2}n^{1/2}$ with $R = \max(2, 8C_2C_0)$ and applying Lemma A.1 and (12) yields

$$\begin{aligned}
\mathbb{P}(N_n \geqslant d_n) &\leqslant \sum_{k=1}^{m} \mathbb{P}\Big(2\|G - \hat{G}\|_\infty \geqslant \frac{d_n}{n} - 4C_2C_0(\log(n)/n)^{1/2}\Big) \\
&\leqslant \sum_{k=1}^{m} \mathbb{P}\Big(2\|G - \hat{G}\|_\infty \geqslant \frac{d_n}{2n}\Big) \\
&\leqslant mM \exp\left(-2n\Big(\frac{d_n}{4n}\Big)^2\right) \\
&\leqslant \frac{\lambda(I)M}{2(\log(n)/n)^{1/2}} \exp\left(-\log(n)/2\right) \\
&\leqslant \frac{\lambda(I)M}{2\log(n)^{1/2}} \exp\left(-\log(n)/2 + \log(n)/2\right) \to 0, \ n \to \infty.
\end{aligned}$$

So with asymptotic probability one,

$$\begin{aligned}
\#S_n \leqslant m^2 \exp\left(6\log(2)R\log(n)^{1/2}n^{1/2}\right) &\leqslant \frac{\lambda(I)^2}{4c_n^2} \exp\left(6R\log(2)\log(n)^{1/2}n^{1/2}\right) \\
&\leqslant r_0 \exp\left(r_1 \log(n)^{1/2}n^{1/2}\right),
\end{aligned}$$

with $r_0 = \lambda(I)^2/(4C_0)$ and $r_1 = 6R\log(2) + 1$. Define $D_n = \{\#S_n \leqslant r_0 \exp(r_1 \log(n)^{1/2}n^{1/2})\}$, let $\mathfrak{S}_n$ be the power set of $\{1, \ldots, n\}$, and, for $\mathcal{J} \in \mathfrak{S}_n$,

$$M_n^{\mathcal{J}} = \Big\| \frac{1}{(\#\mathcal{J})^{1/2}} \sum_{i \in \mathcal{J}} \left(\mathbb{1}\{Y_{ni} \leqslant \cdot\} - F_{\theta(X_{ni})}(\cdot)\right) \Big\|_\infty.$$

Then, for $z_n = s\log(n)^{1/4}n^{1/4}$ with an arbitrary $s > 0$,

$$\begin{aligned}
\mathbb{P}(M_n^\pi \geqslant z_n) &= \mathbb{E}\left(\mathbb{1}\left\{M_n^{\mathcal{I}^*} \geqslant z_n\right\}\right) \\
&= \mathbb{E}\left(\sum_{\mathcal{J}\in\mathfrak{S}_n} \mathbb{1}\{\mathcal{I}^* = \mathcal{J}\}\mathbb{1}\left\{M_n^{\mathcal{J}} \geqslant z_n\right\}\right) \\
&\leqslant \mathbb{P}(B_n^c) + \mathbb{E}\left(\mathbb{1}B_n \sum_{\mathcal{J}\in\mathfrak{S}_n} \mathbb{1}\{\mathcal{I}^* = \mathcal{J}\}\mathbb{1}\left\{M_n^{\mathcal{J}} \geqslant z_n\right\}\right) \\
&= \mathbb{P}(B_n^c) + \mathbb{E}\left(\mathbb{1}B_n \sum_{\mathcal{J}\in S_n} \mathbb{1}\{\mathcal{I}^* = \mathcal{J}\}\mathbb{1}\left\{M_n^{\mathcal{J}} \geqslant z_n\right\}\right) \\
&\leqslant \mathbb{P}(B_n^c) + \mathbb{E}\left(\sum_{\mathcal{J}\in S_n} \mathbb{1}\{\mathcal{I}^* = \mathcal{J}\}\mathbb{1}\left\{M_n^{\mathcal{J}} \geqslant z_n\right\}\right) \\
&\leqslant \mathbb{P}(B_n^c) + \mathbb{P}(D_n^c) + \mathbb{E}\left(\mathbb{1}D_n\mathbb{E}\left[\sum_{\mathcal{J}\in S_n} \mathbb{1}\{\mathcal{I}^* = \mathcal{J}\}\mathbb{1}\left\{M_n^{\mathcal{J}} \geqslant z_n\right\}\middle| X_{n1},\ldots,X_{nn}\right]\right).
\end{aligned}$$

In the last inequality we use the fact that $\mathbb{1}D_n$ is a function of $X_{n1},\ldots,X_{nn}$ and

$$\mathbb{E}\left[\sum_{\mathcal{J}\in S_n} \mathbb{1}\{\mathcal{I}^* = \mathcal{J}\}\mathbb{1}\left\{M_n^{\mathcal{J}} \geqslant z_n\right\}\middle| X_{n1},\ldots,X_{nn}\right] \leqslant 1 \text{ a.s.},$$

since $\mathcal{I}^* = \mathcal{J}$ may only hold for exactly one index set $\mathcal{J}$. Finally,

$$\begin{aligned}
&\mathbb{E}\left(\mathbb{1}D_n\mathbb{E}\left[\sum_{\mathcal{J}\in S_n} \mathbb{1}\{\mathcal{I}^* = \mathcal{J}\}\mathbb{1}\left\{M_n^{\mathcal{J}} \geqslant z_n\right\}\middle| X_{n1},\ldots,X_{nn}\right]\right) \\
&\leqslant \mathbb{E}\left(\mathbb{1}D_n \sum_{\mathcal{J}\in S_n} \mathbb{E}\left[\mathbb{1}\left\{M_n^{\mathcal{J}} \geqslant z_n\right\} \mid X_{n1},\ldots,X_{nn}\right]\right). \\
&= \mathbb{E}\left(\mathbb{1}D_n \sum_{\mathcal{J}\in S_n} \mathbb{P}\left[M_n^{\mathcal{J}} \geqslant z_n \mid X_{n1},\ldots,X_{nn}\right]\right). \\
&\leqslant \mathbb{E}\left(\mathbb{1}D_n(\#S_n)M\exp(-2z_n^2)\right) \\
&\leqslant r_0 M\exp\left(-(2s^2 - r_1)\log(n)^{1/2}n^{1/2}\right) \to 0,\ n\to\infty,
\end{aligned}$$

for $s > \sqrt{r_1/2}$, using Lemma A.1 in the second-last inequality. □

Lemma A.3 shows that for suitable constants D and sequences $(\delta_n)_{n\in\mathbb{N}}$ with limit zero, all subintervals of I with length at least δ_n contain at least $Dn\delta_n$ elements of $\{g(\hat{\theta}_n(X_{nj})) : j = 1,\ldots,n\}$. That is, the pseudo-covariates $g(\hat{\theta}_n(X_{nj}))$ are asymptotically dense in I.

Lemma A.3. *Under (A3) and (A4), with* $\hat{w}(B) = \#\{j \in \{1, \dots, n\} : g(\hat{\theta}_n(X_{nj})) \in B\}$, *for any sequence* $(\delta_n)_{n\in\mathbb{N}}$ *such that* $\delta_n \geqslant 4C_0(\log(n)/n)^{1/2}$, *the event*

$$\left\{\inf\left\{\frac{\hat{w}(I_n)}{n\lambda(I_n)} : \textit{intervals } I_n \subset I \textit{ with } \lambda(I_n) \geqslant \delta_n\right\} \geqslant D\right\} \tag{13}$$

has asymptotic probability one for any $D < C_1/2$.

Proof of Lemma A.3. Similarly to the definition of $\hat{w}$, let $w(B) = \#\{j \in \{1, \dots, n\} : \theta(X_{nj}) \in B\}$ for $B \subseteq I$. Define $c_n = C_0(\log(n)/n)^{1/2}$ with C_0 from (A4). Then on the event B_n defined at (10), for any interval $J \subseteq I$ with $\lambda(J) \geqslant 2c_n$,

$$\begin{aligned}\hat{w}(J) - w(J) &\geqslant -\#\{j \in \{1, \dots, n\} : \hat{\theta}_n(X_{nj}) \notin J,\, \theta(X_{nj}) \in J\} \\ &\geqslant -w(\{z \in J : z + c_n \notin J \text{ or } z - c_n \notin J\}).\end{aligned}$$

This gives $\hat{w}(J) \geqslant w(J\backslash\{z \in J : z + c_n \notin J \text{ or } z - c_n \notin J\})$. The assumption $\delta_n \geqslant 4c_n$ implies that $\delta_n - 2c_n \geqslant \delta_n/2$. For any interval $I_n \subseteq I$ of length at least δ_n, the set $\tilde{I}_n = I_n\backslash\{z \in I_n : z + c_n \notin I_n \text{ or } z - c_n \notin I_n\}$ is in interval of length

$$\lambda(\tilde{I}_n) = \lambda(I_n) - 2c_n \geqslant \lambda(I_n) - \delta_n/2 \geqslant \lambda(I_n) - \lambda(I_n)/2 = \lambda(I_n)/2.$$

This and $\hat{w}(I_n) \geqslant w(\tilde{I}_n)$ yield

$$\begin{aligned}\hat{m}_n &:= \inf\left\{\frac{\hat{w}(I_n)}{n\lambda(I_n)} : \text{intervals } I_n \subset I \text{ with } \lambda(I_n) \geqslant \delta_n\right\} \\ &\geqslant \inf\left\{\frac{w(\tilde{I}_n)}{n\lambda(\tilde{I}_n)} : \text{intervals } \tilde{I}_n \subset I \text{ with } \lambda(\tilde{I}_n) \geqslant \delta_n/2\right\}/2 =: m_n.\end{aligned}$$

Define $A_n = \{\hat{m}_n \geqslant D\}$ and $\tilde{A}_n = \{m_n \geqslant D\}$ for $D < C_1/2$. Then $\tilde{A}_n \subseteq A_n$ and

$$\mathbb{P}(A_n) \geqslant \mathbb{P}(A_n \cap B_n) \geqslant \mathbb{P}(\tilde{A}_n \cap B_n) = \mathbb{P}(\tilde{A}_n) + \mathbb{P}(B_n) - \mathbb{P}(\tilde{A}_n \cup B_n) \to 1,\ n \to \infty,$$

since $\lim_{n\to\infty} \mathbb{P}(B_n) = 1$ by (A4) and $\lim_{n\to\infty} \mathbb{P}(\tilde{A}_n) = 1$ by (A3) and by Equation 4.6 of Mösching and Dümbgen (2020, Section 4.3). □

Proof of Theorem 5.1. Proposition 3.1 implies that for all $u \in \mathbb{R}$,

$$\hat{F}_u(y; (\hat{\theta}_n(X_{nj}))_{j=1}^n, (Y_{nj})_{j=1}^n) = \hat{F}_{g(u)}(y; (g(\hat{\theta}_n(X_{nj})))_{j=1}^n, (Y_{nj})_{j=1}^n).$$

To lighten the notation, we can therefore drop g from (A4) and simply write $\hat{\theta}_n(\cdot)$ instead of $g(\hat{\theta}_n(\cdot))$. Assume that $\hat{\theta}_n(X_{n\pi(1)}) \leqslant \hat{\theta}_n(X_{n\pi(2)}) \leqslant \ldots \leqslant \hat{\theta}_n(X_{n\pi(n)})$ and define $\delta_n = (\log n/n)^{1/6}/2$. Lemma A.3 and (A4) imply that for all $x \in \mathcal{X}_n = \{x \in \mathcal{X} : [\theta(x) \pm 2\delta_n] \subseteq I\}$, the indices

$$r(x) = \min\{j \in \{1, \ldots, n\} : \hat{\theta}_n(X_{n\pi(j)}) \geqslant \hat{\theta}_n(x) - \delta_n\}$$
$$j(x) = \max\{j \in \{1, \ldots, n\} : \hat{\theta}_n(X_{n\pi(j)}) \leqslant \hat{\theta}_n(x)\}$$

are well defined with asymptotic probability one, because $[\hat{\theta}_n(x) - \delta_n, \hat{\theta}_n(x)]$ is of length δ_n and contained in I since $\theta(x) + (\log n/n)^{1/6} \geqslant \hat{\theta}_n(x) \geqslant \hat{\theta}_n(x) - \delta_n \geqslant \theta(x) - \delta_n - C_0 n^{-1/2} > \theta(x) - (\log n/n)^{1/6}$ for n sufficiently large, on the event B_n defined at (10). They satisfy $r(x) \leqslant j(x)$ and $\hat{\theta}_n(x) - \delta_n \leqslant \hat{\theta}_n(X_{nr(x)}) \leqslant \hat{\theta}_n(X_{nj(x)}) \leqslant \hat{\theta}_n(x)$ and, with asymptotic probability one due to Lemma A.3, $w_{r(x)j(x)} = \#\{j \in \{1, \ldots, n\} : \hat{\theta}_n(x) - \delta_n \leqslant \hat{\theta}_n(X_{n\pi(j)}) \leqslant \hat{\theta}_n(x)\} \geqslant Dn\delta_n$ for $0 < D < C_1/2$. Therefore, almost surely with respect to the joint law of (X_{ni}, Y_{ni}),

$i = 1, \dots, n$, for any $y \in \mathbb{R}$,

$$\begin{aligned}
\hat{F}_{n;\hat{\theta}_n(x)}(y) - F_{\theta(x)}(y) &\leqslant \hat{F}_{n;\hat{\theta}_n(X_{nj(x)})}(y) - F_{\theta(x)}(y) \\
&= \min_{r \leqslant j(x)} \max_{s \geqslant j(x)} \hat{\mathbb{F}}^{\pi}_{rs}(y) - F_{\theta(x)}(y) \\
&\leqslant \max_{s \geqslant j(x)} \hat{\mathbb{F}}^{\pi}_{r(x)s}(y) - F_{\theta(x)}(y) \\
&\leqslant w^{-1/2}_{r(x)j(x)} M^{\pi}_n + \max_{s \geqslant j(x)} \bar{F}^{\pi}_{\theta;r(x)s}(y) - F_{\theta(x)}(y) \\
&\leqslant (Dn\delta_n)^{-1/2} M^{\pi}_n \\
&\quad + \max_{s \geqslant j(x)} \big(\bar{F}^{\pi}_{\theta;r(x)s}(y) - \bar{F}^{\pi}_{\hat{\theta};r(x)s}(y) + \bar{F}^{\pi}_{\hat{\theta};r(x)s}(y) \big) - F_{\theta(x)}(y) \\
&\leqslant (Dn\delta_n)^{-1/2} M^{\pi}_n + L \sup_{x \in \mathcal{X}} |\hat{\theta}_n(x) - \theta(x)| + \max_{s \geqslant j(x)} \bar{F}^{\pi}_{\hat{\theta};r(x)s}(y) - F_{\theta(x)}(y) \\
&\leqslant (Dn\delta_n)^{-1/2} M^{\pi}_n + L \sup_{x \in \mathcal{X}} |\hat{\theta}_n(x) - \theta(x)| + F_{\hat{\theta}_n(X_{nr(x)})}(y) - F_{\theta(x)}(y) \\
&\leqslant (Dn\delta_n)^{-1/2} M^{\pi}_n + L \sup_{x \in \mathcal{X}} |\hat{\theta}_n(x) - \theta(x)| + L|\hat{\theta}_n(X_{nr(x)}) - \theta(x)| \\
&\leqslant (Dn\delta_n)^{-1/2} M^{\pi}_n + L \sup_{x \in \mathcal{X}} |\hat{\theta}_n(x) - \theta(x)| + L\delta_n.
\end{aligned}$$

The equality in the second line is the classical min-max formula for monotone regression, see e.g. Equation (2.2) in Mösching and Dümbgen (2020), and the first and the third last inequality use antitonicity of $u \mapsto F_u(y)$. By assumption (A4) and with the constant $s > 0$ from Lemma A.2, the event

$$\{M^{\pi}_n \leqslant s(n \log(n))^{1/4}\} \cap \left\{ \sup_{x \in \mathcal{X}} |\hat{\theta}_n(x) - \theta(x)| < \delta_n \right\}$$

has asymptotic probability one. On this event, the previous considerations imply

$$\sup_{x \in \mathcal{X}_n, y \in \mathbb{R}} (\hat{F}_{n;\hat{\theta}_n(x)}(y) - F_{\theta(x)})(y) \leqslant s(Dn\delta_n)^{-1/2}(n \log(n))^{1/4} + 2L\delta_n \leqslant C \left(\frac{\log(n)}{n} \right)^{1/6},$$

with $C = [s(2D^{-1})^{1/2} + L]$. To finish the proof, we show that $F_{\theta(x)}(y) - \hat{F}_{n;\hat{\theta}_n(x)}(y)$ can be bounded in the same way.

Similar to before, define the indices $r'(x) = \min\{j \in \{1, \dots, n\} : \hat{\theta}_n(X_{nj}) \geqslant \hat{\theta}_n(x)\}$, $j'(x) = \max\{j \in \{1, \dots, n\} : \hat{\theta}_n(X_{nj}) \leqslant \hat{\theta}_n(x) + \delta_n\}$. Then with asymptotic probability one,

also $r'(x) \leqslant j'(x)$ and $\hat{\theta}_n(x) \leqslant \hat{\theta}_n(X_{nr'(x)}) \leqslant \hat{\theta}_n(X_{nj'(x)}) \leqslant \hat{\theta}_n(x) + \delta_n$, $w_{r'(x)j'(x)} \geqslant Dn\delta_n$. Thus,

$$
\begin{aligned}
\hat{F}_{n;\hat{\theta}_n(x)}(y) - F_{\theta(x)}(y) &\geqslant \hat{F}_{n;\hat{\theta}_n(X_{nr'(x)})}(y) - F_{\theta(x)} \\
&= \min_{r \leqslant r'(x)} \max_{s \geqslant r'(x)} \hat{\mathbb{F}}^{\pi}_{rs}(y) - F_{\theta(x)}(y) \\
&\geqslant \min_{r \leqslant r'(x)} \hat{\mathbb{F}}^{\pi}_{rj'(x)}(y) - F_{\theta(x)}(y) \\
&\geqslant -w^{-1/2}_{r'(x)j'(x)} M^{\pi}_n + \min_{r \leqslant r'(x)} \bar{F}^{\pi}_{\theta;rj'(x)}(y) - F_{\theta(x)}(y) \\
&\geqslant -(Dn\delta_n)^{-1/2} M^{\pi}_n \\
&\qquad + \min_{r \leqslant r'(x)} \big(\bar{F}^{\pi}_{\theta;rj'(x)}(y) - \bar{F}^{\pi}_{\hat{\theta};rj'(x)}(y) + \bar{F}^{\pi}_{\hat{\theta};rj'(x)}(y) \big) - F_{\theta(x)}(y) \\
&\geqslant -(Dn\delta_n)^{-1/2} M^{\pi}_n - L \sup_{x \in \mathcal{X}} |\hat{\theta}_n(x) - \theta(x)| + F_{\hat{\theta}_n(X_{nj'(x)})}(y) - F_{\theta(x)}(y) \\
&\geqslant -(Dn\delta_n)^{-1/2} M^{\pi}_n - L \sup_{x \in \mathcal{X}} |\hat{\theta}_n(x) - \theta(x)| - L|\hat{\theta}_n(X_{nj'(x)}) - \theta(x)| \\
&\geqslant -(Dn\delta_n)^{-1/2} M^{\pi}_n - L \sup_{x \in \mathcal{X}} |\hat{\theta}_n(x) - \theta(x)| - L\delta_n. \qquad \square
\end{aligned}
$$

Proof of Theorem 5.1 with sample splitting. Assume that the index estimator $\hat{\theta}_n$ is computed with data $(X_{ni}, Y_{ni})_{i=1}^{\lfloor n\xi \rfloor}$ and the distribution functions with $(\hat{\theta}_n(X_{ni}), Y_{ni})_{i=\lfloor n\xi \rfloor+1}^{n}$. The statement of Lemma A.3 also holds when $C_0(\log(n)/n)^{1/2}$ is replaced by $(\log(n)/n)^{1/3}$. By conditioning on $(X_{ni}, Y_{ni})_{i=1}^{\lfloor n\xi \rfloor}$ and on X_{ni}, $i = \lfloor n\xi \rfloor + 1, \ldots, n$, Corollary 4.7 of Mösching and Dümbgen (2020) implies that M^{π}_n (computed with the data $(\hat{\theta}_n(X_{ni}), Y_{ni})_{i=\lfloor n\xi \rfloor+1}^{n}$) satisfies $\mathbb{P}(M^{\pi}_n \geqslant (R \log(n(1-\xi)))^{1/2}) \to 0$, $n \to \infty$, for any $R > 1$. This requires the fact that the permutation π is constant when conditioned on $(X_{ni}, Y_{ni})_{i=1}^{\lfloor n\xi \rfloor}$. One may now follow exactly the same steps as in the proof for the theorem without sample splitting, but with sample size $\lfloor n(1-\xi) \rfloor$ instead of n, $\delta_n = (n(1-\xi)/\log(n(1-\xi)))^{1/3}/2$ instead of $(n/\log(n))^{1/6}$ and $\{M^{\pi}_n \leqslant (R \log(n(1-\xi)))^{1/2}\}$ instead of $\{M^{\pi}_n \leqslant s(n \log(n))^{1/4}\}$, obtaining an upper bound of $C'(\log(n)/n)^{1/3}$ for the error, where $C' > 0$ also depends on ξ. $\square$

References

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *Annals of Statistics*, 37:1148–1178.

Balabdaoui, F., Durot, C., and Jankowski, H. (2019a). Least squares estimation in the monotone single index model. *Bernoulli*, 25:3276–3310.

Balabdaoui, F. and Groeneboom, P. (2020). Profile least squares estimators in the monotone single index model. *arXiv e-prints*, page arXiv:2001.05454.

Balabdaoui, F., Groeneboom, P., and Hendrickx, K. (2019b). Score estimation in the monotone single-index model. *Scandinavian Journal of Statistics*, 46:517–544.

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92:477–489.

Chernozhukov, V., Fernández-Val, I., and Galichon, A. (2010). Quantile and probability curves without crossing. *Econometrica*, 78:1093–1125.

Chernozhukov, V., Fernández-Val, I., and Melly, B. (2013). Inference on counterfactual distributions. *Econometrica*, 81:2205–2268.

Chernozhukov, V., Fernández-Val, I., Melly, B., and Wüthrich, K. (2020). Generic inference on quantile and quantile effect functions for discrete outcomes. *Journal of the American Statistical Association*, 115:123–137.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, 34:187–202.

Dette, H. and Volgushev, S. (2008). Non-crossing non-parametric estimates of quantile curves. *Journal of the Royal Statistical Society: Series B*, 70:609–627.

Diebold, F. X., Gunther, T. A., and Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, 39:863–883.

Duarte, E., de Sousa, B., Cadarso-Suárez, C., Klein, N., Kneib, T., and Rodrigues, V. (2017). Studying the relationship between a woman's reproductive lifespan and age at menarche using a Bayesian multivariate structured additive distributional regression model. *Biometrical Journal*, 59:1232–1246.

Dunson, D. B., Pillai, N., and Park, J.-H. (2007). Bayesian density regression. *Journal of the Royal Statistical Society: Series B*, 69:163–183.

Foresi, S. and Peracchi, F. (1995). The conditional distribution of excess returns: An empirical analysis. *Journal of the American Statistical Association*, 90:451–466.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B*, 69:243–268.

Gneiting, T. and Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151.

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

Hall, P., Wolff, R. C. L., and Yao, Q. (1999). Methods for estimating a conditional distribution function. *Journal of the American Statistical Association*, 94:154–163.

Hall, P. and Yao, Q. (2005). Approximating conditional distribution functions using dimension reduction. *Annals of Statistics*, 33:1404–1421.

Härdle, W., Hall, P., and Ichimura, H. (1993). Optimal smoothing in single-index models. *The Annals of Statistics*, 21:157–178.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*, volume 43 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, Ltd., London.

Hayfield, T. and Racine, J. S. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27:1–32.

Henzi, A., Mösching, A., and Dümbgen, L. (2020). Accelerating the pool-adjacent-violators algorithm for isotonic distributional regression. Preprint, arxiv.org/abs/2006.05527.

Henzi, A., Ziegel, J. F., and Gneiting, T. (2019). Isotonic distributional regression. *arXiv e-prints*, page arXiv:1909.03725.

Hothorn, T., Kneib, T., and Bühlmann, P. (2014). Conditional transformation models. *Journal of the Royal Statistical Society: Series B*, 76:3–27.

Jordan, A. I., Mühlemann, A., and Ziegel, J. F. (2019). Optimal solutions to the isotonic regression problem. *arXiv e-prints*, page arXiv:1904.04761.

Klein, N., Kneib, T., Lang, S., and Sohn, A. (2015). Bayesian structured additive distributional forecasting with an application to regional income inequality in Germany. *Annals of Applied Statistics*, 9:1024–1052.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press.

Koenker, R. (2020). *quantreg: Quantile Regression*. R package version 5.55.

Kramer, A. A. (2017). Are ICU length of stay predictions worthwhile? *Critical Care Medicine*, 45:379–380.

Kuchibhotla, A. K., Patra, R. K., and Sen, B. (2017). Least squares estimation in a single index model with convex Lipschitz link. *arXiv e-prints*, page arXiv:1708.00145.

Lanteri, A., Maggioni, M., and Vigogna, S. (2020). Conditional regression for single-index models. *arXiv e-prints*, page arXiv:2002.10008.

Le Gall, J.-R., Lemeshow, S., and Saulnier, F. (1993). A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. *JAMA*, 270:2957–2963.

Li, Q. and Racine, J. S. (2008). Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data. *Journal of Business & Economic Statistics*, 26:423–434.

Machado, J. A. F. and Mata, J. (2000). Box–Cox quantile regression and the distribution of firm sizes. *Journal of Applied Econometrics*, 15:253–274.

Matheson, J. E. and Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, 22:1087–1096.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 2nd edition.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, 7:983–999.

Miranda, D. R., Moreno, R., and Iapichino, G. (1997). Nine equivalents of nursing manpower use score (NEMS). *Intensive Care Medicine*, 23:760–765.

Moran, J. L. and Solomon, P. J. (2012). A review of statistical estimators for risk-adjusted length of stay: analysis of the Australian and new Zealand intensive care adult patient data-base, 2008–2009. *BMC Medical Research Methodology*, 12:68.

Mösching, A. and Dümbgen, L. (2020). Monotone least squares and isotonic quantiles. *Electronic Journal of Statistics*, 14:24–49.

Niskanen, M., Reinikainen, M., and Pettilä, V. (2009). Case-mix-adjusted length of stay and mortality in 23 Finnish ICUs. *Intensive Care Medicine*, 35:1060–1067.

Peracchi, F. (2002). On estimating conditional quantiles and distribution functions. *Computational Statistics & Data Analysis*, 38:433–447.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rasp, S. and Lerch, S. (2018). Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, 146:3885–3900.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape (with discussion). *Journal of the Royal Statistical Society: Series C*, 54:507–554.

Schlosser, L., Hothorn, T., Stauffer, R., and Zeileis, A. (2019). Distributional regression forests for probabilistic precipitation forecasting in complex terrain. *Annals of Applied Statistics*, 13:1564–1589.

Shaked, M. and Shanthikumar, J. G. (2007). *Stochastic Orders*. Springer, New York.

Silbersdorff, A., Lynch, J., Klasen, S., and Kneib, T. (2018). Reconsidering the income-health relationship using distributional regression. *Health Economics*, 27:1074–1088.

Thomas, J., Mayr, A., Bischl, B., Schmid, M., Smith, A., and Hofner, B. (2018). Gradient boosting for distributional regression: faster tuning and improved variable selection via noncyclical updates. *Statistics and Computing*, 28:673–687.

Umlauf, N., Klein, N., and Zeileis, A. (2018). Bamlss: Bayesian additive models for location, scale, and shape (and beyond). *Journal of Computational and Graphical Statistics*, 27:612–627.

Vannitsem, S., Wilks, D. S., and Messner, J., editors (2018). *Statistical Postprocessing of Ensemble Forecasts*. Elsevier.

Verburg, I. W., de Keizer, N. F., de Jonge, E., and Peek, N. (2014). Comparison of regression methods for modeling intensive care length of stay. *PloS one*, 9(10):e109684.

Wilks, D. S. (2011). *Statistical Methods in the Atmospheric Sciences*. Elsevier, 3rd edition.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, 2nd edition.

Zhang, J., Chen, Q., Lin, B., and Zhou, Y. (2017). On the single-index model estimate of the conditional density function: Consistency and implementation. *Journal of Statistical Planning and Inference*, 187:56–66.

Zimmerman, J. E., Kramer, A. A., McNair, D. S., Malila, F. M., and Shaffer, V. L. (2006). Intensive care unit length of stay: Benchmarking based on acute physiology and chronic health evaluation (APACHE) IV. *Critical care medicine*, 34:2517–2529.

Zou, Q. and Zhu, Z. (2014). M-estimators for single-index model using B-spline. *Metrika*, 77:225–246.