Isotonic Distributional Regression (IDR)

Leveraging Monotonicity, Uniquely So!

Tilmann Gneiting
Heidelberger Institut für Theoretische Studien (HITS)
Karlsruher Institut für Technologie (KIT)

Alexander Henzi   Johanna F. Ziegel
Universität Bern

Wirtschaftsuniversität Wien

15 January 2021

Karlsruhe Institute of Technology

Heidelberg Institute for Theoretical Studies

HITS

WAVES TO WEATHER

Isotonic Distributional Regression (IDR)

Isotonic Distributional Regression (IDR)

Origins of Regression

regression originates from arguably the most notorious priority dispute in the history of mathematics and statistics

between Carl-Friedrich Gauss (1777–1855) and Adrien-Marie Legendre (1752–1833) over the method of least squares

- Stigler (1981): "Gauss probably possessed the method well before Legendre, but [. . .] was unsuccessful in communicating it to his contemporaries"

Current Views: Distributional Regression

Wikipedia notes (actually: noted until recently) that

- "commonly, regression analysis estimates the conditional expectation [...] Less commonly, the focus is on a quantile [...] of the conditional distribution [...] In all cases, a function of the independent variables called the regression function is to be estimated"
- "it is also of interest to characterize the variation of the dependent variable around the prediction of the regression function using a probability distribution"

Hothorn, Kneib and Bühlmann (2014) argue forcefully that the

- "ultimate goal of regression analysis is to obtain information about the conditional distribution of a response given a set of explanatory variables"

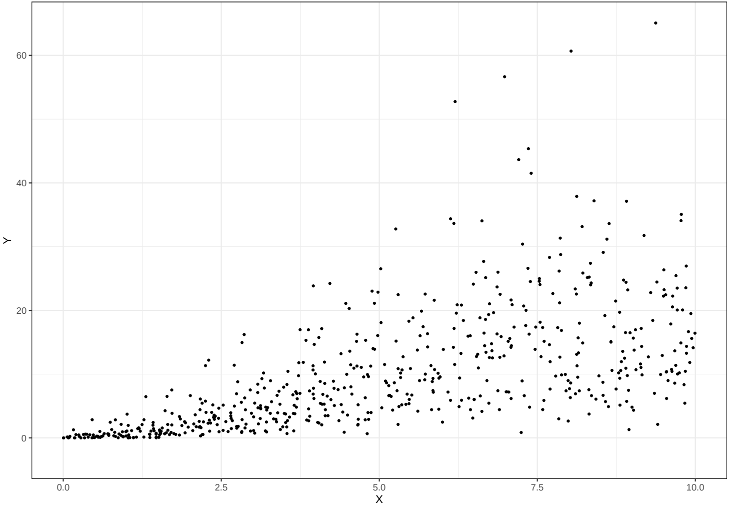
in a nutshell, distributional regression

- uses training data

  $$\{(x_i, y_i) \in \mathcal{X} \times \mathbb{R} : i = 1, \dots n\}$$

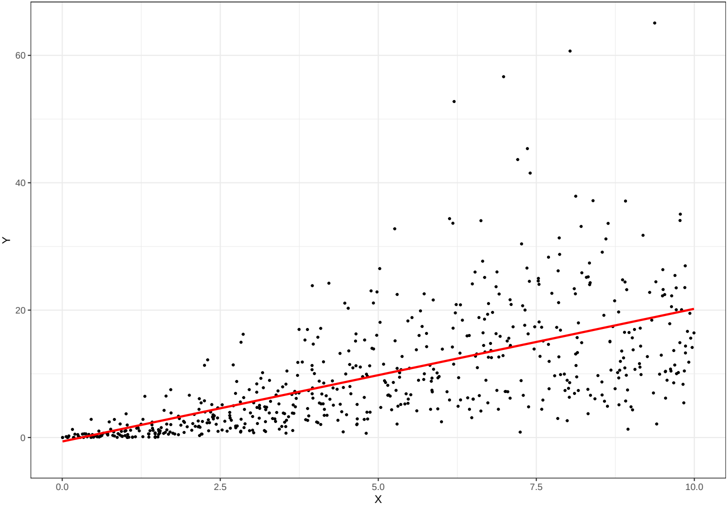
  to estimate the conditional distribution of the response variable, $y \in \mathbb{R}$, given the explanatory variables or covariates, $x \in \mathcal{X}$
- isotonic distributional regression (IDR) uses monotonicity relations to find nonparametric conditional distributions

Isotonic Distributional Regression (IDR) . . . in Pictures



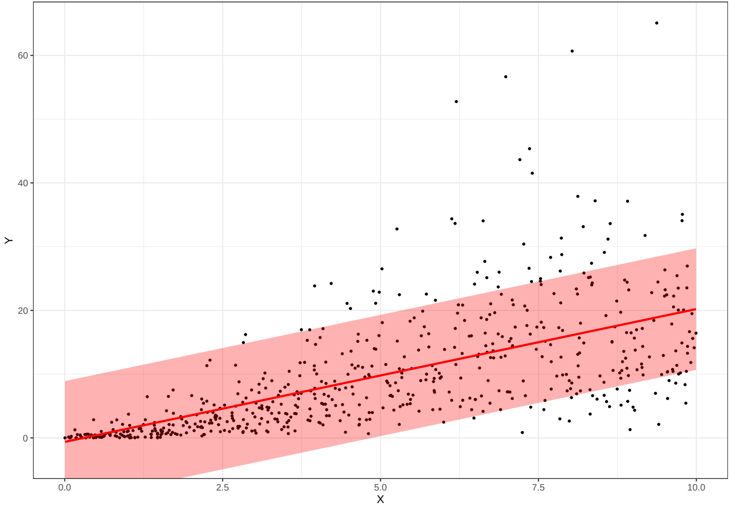
bivariate point cloud — regression of Y on X

Isotonic Distributional Regression (IDR) . . . in Pictures



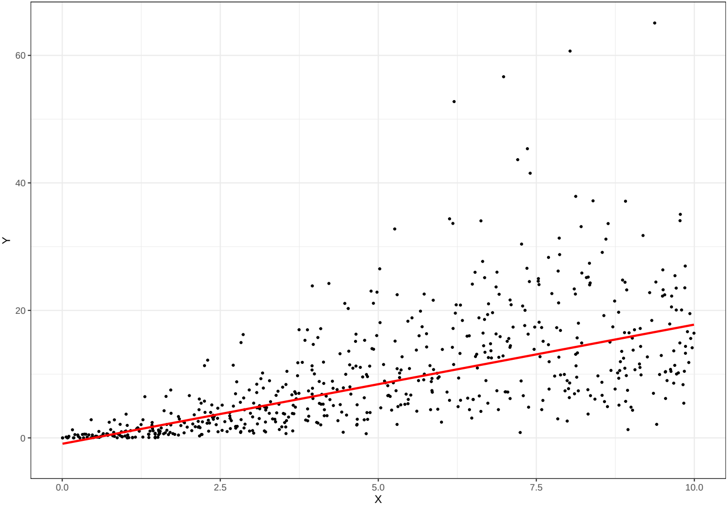
linear ordinary least squares (OLS; L_2) regression line

Isotonic Distributional Regression (IDR) ... in Pictures



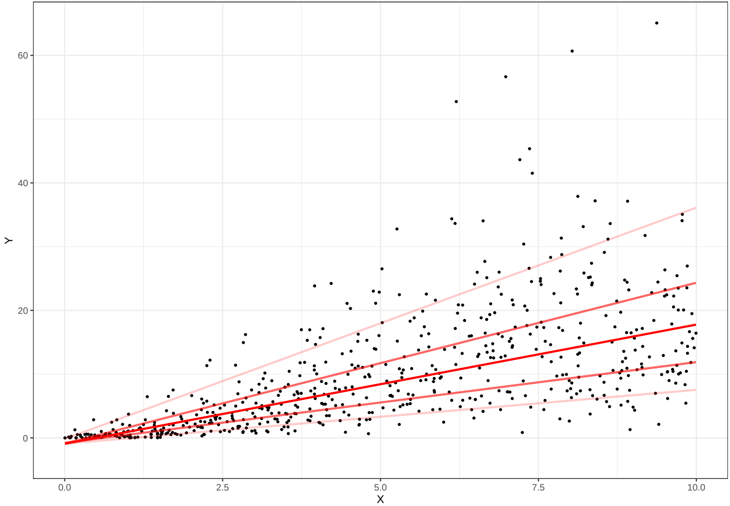
linear L_2 regression line with 80% prediction intervals

Isotonic Distributional Regression (IDR) ... in Pictures



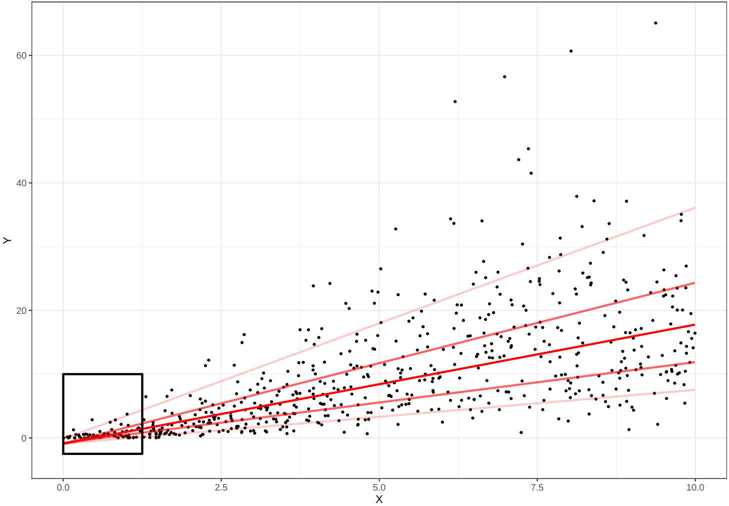
linear L_1 regression line — median regression

Isotonic Distributional Regression (IDR) . . . in Pictures



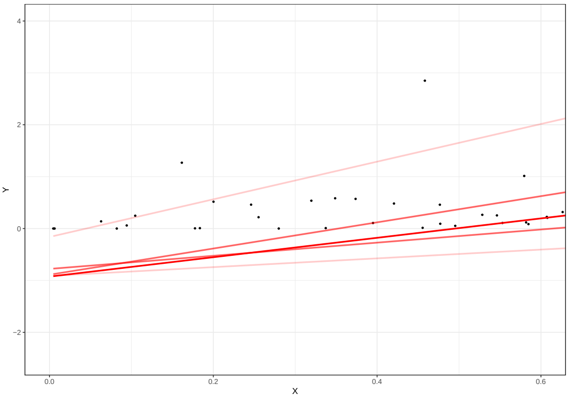
linear quantile regression — levels 0.10, 0.30, 0.50, 0.70, 0.90

Isotonic Distributional Regression (IDR) . . . in Pictures



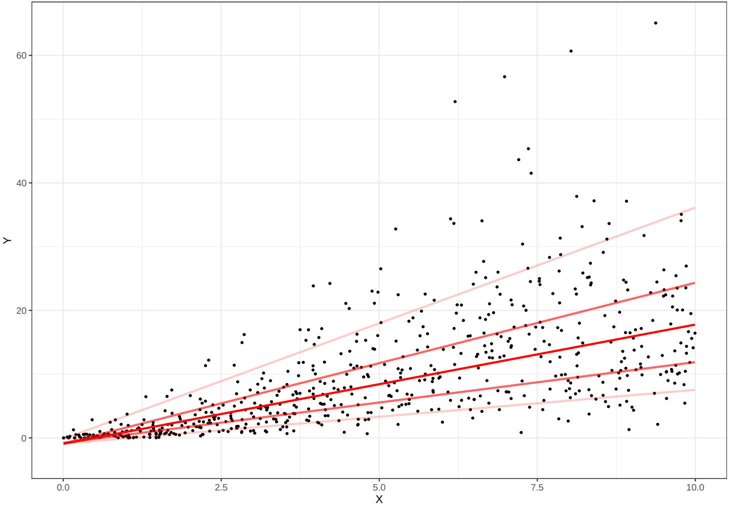
linear quantile regression — zoom in

Isotonic Distributional Regression (IDR) . . . in Pictures



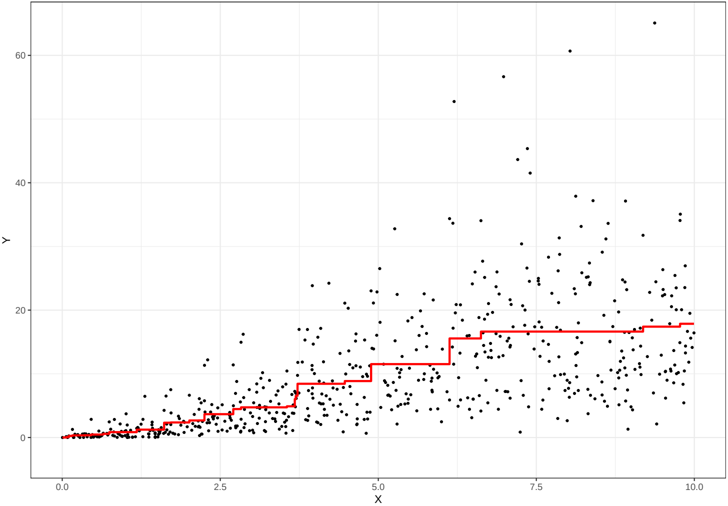
linear quantile regression — beware quantile crossing

Isotonic Distributional Regression (IDR) . . . in Pictures



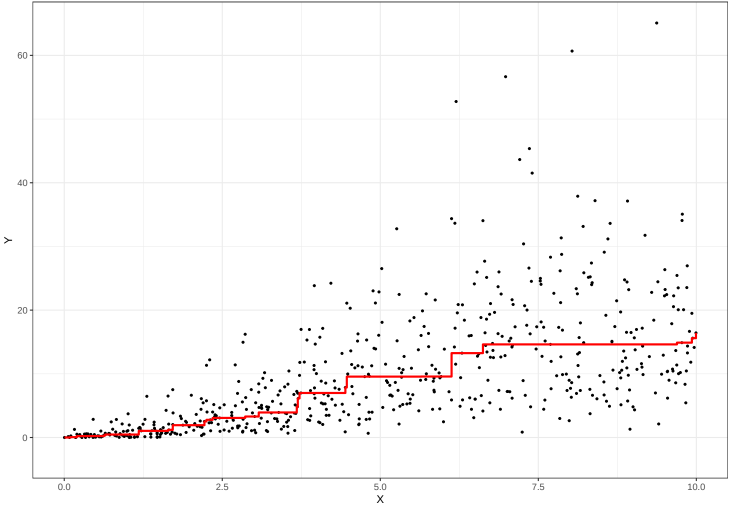
linear quantile regression

Isotonic Distributional Regression (IDR) . . . in Pictures



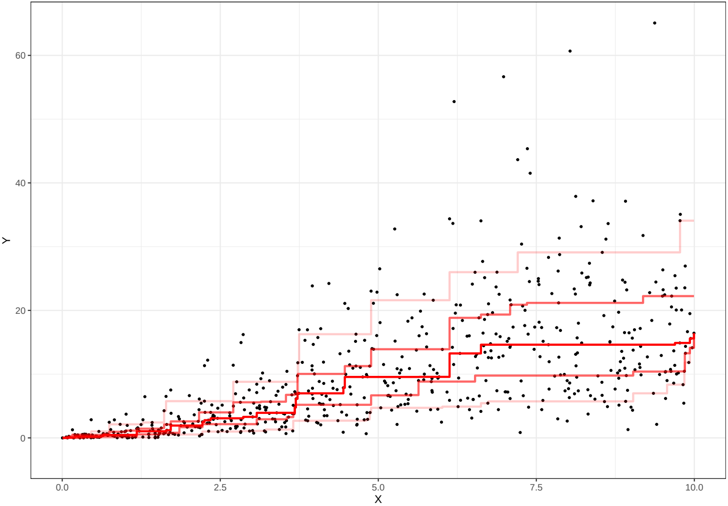
nonparametric isotonic mean (L_2) regression

Isotonic Distributional Regression (IDR) . . . in Pictures



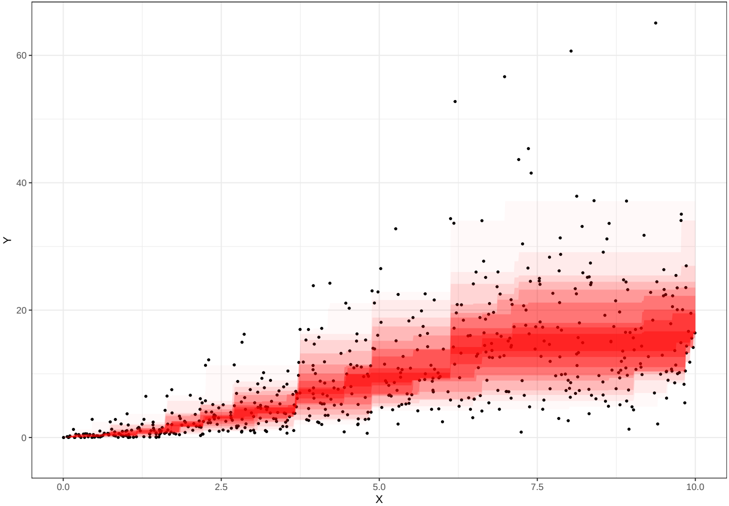
nonparametric isotonic median (L_1) regression

Isotonic Distributional Regression (IDR) . . . in Pictures



nonparametric isotonic quantile regression

Isotonic Distributional Regression (IDR) . . . in Pictures



isotonic distributional regression (IDR)

Isotonic Distributional Regression (IDR) ... the Details

isotonic distributional regression (IDR) uses training data of the form

$$\{(x_i, y_i) \in \mathcal{X} \times \mathbb{R} : i = 1, \ldots n\}$$

to estimate a conditional distribution of the response variable or outcome, $y \in \mathbb{R}$, given the explanatory variables or covariates, $x \in \mathcal{X}$

takes advantage of known or assumed nonparametric monotonicity relations between the covariates, x, and the real-valued outcome, y

has primary uses in prediction and forecasting, where we know the covariates x, but do not know the outcome y

a full understanding relies on a number of (partly, rather recent) mathematical concepts and developments, namely,

- proper scoring rules, and
- partial orders

Isotonic Distributional Regression (IDR)

Scoring Rules

scoring rules seek to quantify predictive performance, assessing calibration and sharpness simultaneously

a scoring rule is a function

$$\mathrm{S}(F, y)$$

that assigns a negatively oriented numerical score to each pair (F, y), where F is a probability distribution, represented by its cumulative distribution function (CDF), and y is the real-valued outcome

a scoring rule S is proper if

$$\mathbb{E}_{Y \sim G}\left[\mathrm{S}(G, Y)\right] \leq \mathbb{E}_{Y \sim G}\left[\mathrm{S}(F, Y)\right] \quad \text{for all} \quad F, G,$$

and strictly proper if, furthermore, equality implies $F = G$

truth serum: under a proper scoring rule truth telling is an optimal strategy in expectation

characterization results relate closely to convex analysis (Gneiting and Raftery 2007)

Continuous Ranked Probability Score (CRPS)

the widely used, proper continuous ranked probability score (CRPS) is defined as

$$\begin{aligned} \mathrm{CRPS}(F, y) &= \int_{-\infty}^{\infty} \left[F(x) - \mathbb{1}(x \geq y)\right]^2 \mathrm{d}x \\ &= \mathbb{E}_F |X - y| - \frac{1}{2} \mathbb{E}_F |X - X'|, \end{aligned}$$

where X and X' are independent with CDF F

for all customary distributions, closed form expressions are available (Jordan et al. 2019); e.g.,

$$\mathrm{CRPS}(\mathcal{N}(\mu, \sigma^2), y) = \sigma \left(\frac{y-\mu}{\sigma} \left(2\,\Phi\left(\frac{y-\mu}{\sigma}\right) - 1 \right) + 2\,\phi\left(\frac{y-\mu}{\sigma}\right) - \frac{1}{\sqrt{\pi}} \right)$$

the CRPS is reported in the same unit as the outcomes, and it generalizes the absolute error, to which it reduces if F is a point measure

reduces to the Brier score when the outcome is binary

Mixture Representations of the CRPS

the CRPS can be represented equivalently as

$$\begin{aligned}\mathrm{CRPS}(F,y) &= 2\int_{(0,1)} \mathrm{QS}_\alpha(F,y)\,\mathrm{d}\lambda(\alpha) \\ &= 2\int_{(0,1)}\int_{\mathbb{R}} \mathrm{S}^Q_{\alpha,\theta}(F,y)\,\mathrm{d}\lambda(\theta,\alpha) \\ &= \int_{\mathbb{R}}\int_{(0,1)} \mathrm{S}^P_{z,c}(F,y)\,\mathrm{d}\lambda(c,z)\end{aligned}$$

in terms of the asymmetric piecewise linear loss QS_α, or the elementary or extremal scoring functions $\mathrm{S}^Q_{\alpha,\theta}$ for the α-quantile functional, or $\mathrm{S}^P_{z,c}$ for probability assessments of the binary outcome $\mathbb{1}(y \le z)$, namely

$$\mathrm{QS}_\alpha(F,y) = \begin{cases} (1-\alpha)\,(F^{-1}(\alpha) - y), & y \le F^{-1}(\alpha), \\ \alpha\,(y - F^{-1}(\alpha)), & y \ge F^{-1}(\alpha), \end{cases}$$

$$\mathrm{S}^Q_{\alpha,\theta}(F,y) = \begin{cases} 1-\alpha, & y \le \theta < F^{-1}(\alpha), \\ \alpha, & F^{-1}(\alpha) \le \theta < y, \\ 0, & \text{otherwise}, \end{cases} \qquad \mathrm{S}^P_{z,c}(F,y) = \begin{cases} 1-c, & F(z) < c,\ y \le z, \\ c, & F(z) \ge c,\ y > z, \\ 0, & \text{otherwise}, \end{cases}$$

respectively (Ehm et al. 2016)

Partial Orders

a partial order relation $\preceq$ on a general set $\mathcal{X}$

- has the same properties as a total order, namely reflexivity, antisymmetry and transitivity
- except that the elements need not be comparable, i.e., there might be elements $x \in \mathcal{X}$ and $x' \in \mathcal{X}$ such that neither $x \preceq x'$ nor $x' \preceq x$
- a key example is the componentwise order on $\mathbb{R}^d$

of particular importance in our context are partial orders on the set $\mathcal{P}$ of the Borel probability measures on $\mathbb{R}$, which we identify with their respective CDFs

- stochastic order ($\leq_{\text{st}}$) $G \leq_{\text{st}} H$ if, and only if, $G(y) \geq H(y)$ for $y \in \mathbb{R}$
- increasing convex order ($\leq_{\text{icx}}$) $G \leq_{\text{icx}} H$ if, and only if,

$$\mathbb{E}[\phi(X_G)] \leq \mathbb{E}[\phi(X_H)]$$

whenever ϕ is increasing and convex and the expectations exist

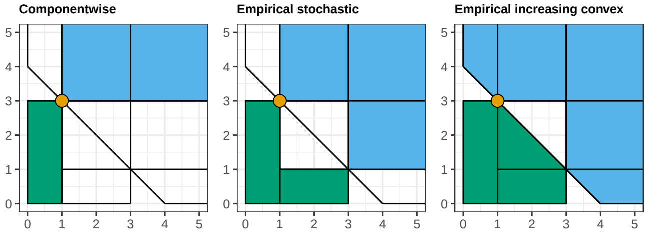
Partial Orders on $\mathbb{R}^d$

in our case study $\mathcal{X} = \mathbb{R}^d$, and we consider the

- componentwise order ($\preceq$)

$$x \preceq x' \iff x_i \leq x_i' \text{ for } i = 1, \ldots, d$$

- empirical stochastic order ($\preceq_{\text{st}}$) induced by the stochastic order on the associated empirical distributions, and equivalent to the componentwise order on the sorted elements
- empirical increasing convex order ($\preceq_{\text{icx}}$) induced by the increasing convex order on the associated empirical distributions



Isotonic Distributional Regression (IDR)

Isotonic Distributional Regression (IDR): Basic Concept

basic concept

- we use training data

$$\{(x_i, y_i) \in \mathcal{X} \times \mathbb{R} : i = 1, \ldots, n\}$$

  to estimate the conditional distribution of the response variable or outcome, $y \in \mathbb{R}$, given the explanatory variables or covariates, $x \in \mathcal{X}$

- formally, distributional regression generates a mapping from a covariate vector $x \in \mathcal{X}$ to a probability measure F_x, which serves to model the conditional distribution of the outcome, y, given x

- given a partial order $\preceq$ on the covariate space $\mathcal{X}$, this mapping is isotonic if

$$x \preceq x' \quad \Rightarrow \quad F_x \leq_{\text{st}} F_{x'},$$

  where $\leq_{\text{st}}$ denotes the usual stochastic order on the space $\mathcal{P}$ of the Borel probability measures in $\mathbb{R}$

IDR: Definition, Existence and Uniqueness

formal setting

- covariate space $\mathcal{X}$ equipped with partial order $\preceq$
- training data $\{(x_i, y_i) \in \mathcal{X} \times \mathbb{R} : i = 1, \ldots n\}$
- the stochastic order $\leq_{\text{st}}$ on the space $\mathcal{P}$ of the Borel probability measures on $\mathbb{R}$
- proper scoring rule S

Definition (isotonic S-regression) An element $\hat{\boldsymbol{F}} = (\hat{F}_1, \ldots, \hat{F}_n) \in \mathcal{P}^n$ is an isotonic S-regression if it is a minimizer of the empirical loss

$$\ell_{\mathrm{S}}(\boldsymbol{F}) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{S}(F_i, y_i)$$

over all $\boldsymbol{F} = (F_1, \ldots, F_n) \in \mathcal{P}^n$, subject to the condition that $F_i \leq_{\text{st}} F_j$ if $x_i \preceq x_j$, for $i, j = 1, \ldots, n$.

Theorem (existence and uniqueness) There exists a unique isotonic CRPS-regression $\hat{\boldsymbol{F}} \in \mathcal{P}^n$.

Terminology We refer to this unique $\hat{\boldsymbol{F}}$ as the isotonic distributional regression (IDR) solution.

Isotonic Distributional Regression (IDR): Universality

Theorem (universality) The IDR solution $\hat{F}$ is an isotonic S-regression under just any scoring rule of the form

$$\mathrm{S}(F, y) = \int_{(0,1)\times\mathbb{R}} \mathrm{S}^{Q}_{\alpha,\theta}(F, y)\,\mathrm{d}H(\alpha, \theta)$$

or

$$\mathrm{S}(F, y) = \int_{\mathbb{R}\times(0,1)} \mathrm{S}^{P}_{z,c}(F, y)\,\mathrm{d}M(z, c),$$

where $\mathrm{S}^{Q}_{\alpha,\theta}$ and $\mathrm{S}^{P}_{z,c}$ are the elementary quantile and probability scoring functions, and H and M are locally finite Borel measures.

Proof relies on results and techniques in Ehm et al. (2016) and Jordan et al. (2019)

Consequence (theoretical) IDR is optimal under just any proper scoring rule that depends on quantile or binary probability assessments only.

Consequence (practical) IDR subsumes extant approaches to non-parametric isotonic regression as special cases, including but not limited to quantile regression and binary regression.

Estimation / Learning

the IDR solution exists and, by definition, is the solution to a constrained optimization problem in $\mathcal{P}^n$ ... but can we actually compute it?

yes — universality and the method of least squares come to the rescue!

- by universality ($M = \delta_z \otimes \lambda_1$), the IDR solution $\hat{\boldsymbol{F}}$ satisfies

$$\hat{\boldsymbol{F}}(z) = \arg\min_{\eta \in [0,1]^n} \sum_{i=1}^{n} \left(\eta_i - \mathbb{1}(y_i \leq z)\right)^2,$$

  at every threshold $z \in \mathbb{R}$, subject to the condition that $\eta_i \geq \eta_j$ if $x_i \preceq x_j$, for $i, j = 1, \ldots, n$
- at any fixed threshold, the IDR CDFs yield a quadratic programming problem, which we tackle with the OSQP solver (Stellato et al. 2020)
- the target function is constant for z inbetween the unique values of $y_1, \ldots, y_n$, and so it suffices to consider these points only
- the overall cost may reduce to $\mathcal{O}(n \log n)$ (Henzi et al. 2020)

Prediction

by construction, the IDR solution $\hat{\boldsymbol{F}} = (\hat{F}_1, \ldots, \hat{F}_n)$ is defined at the training covariate values $x_1, \ldots, x_n \in \mathcal{X}$ only

a key task in practice is to make a prediction at a new covariate value $x \in \mathcal{X}$ where $x \notin \{x_1, \ldots, x_n\}$, for which we proceed as follows

- define the sets $p(x)$ and $s(x)$ of the indices of immediate predecessors and successors of x among $x_1, \ldots, x_n$ as

$$p(x) = \{i \in \{1, \ldots, n\} : x_i \preceq x_j \preceq x \implies x_j = x_i, j = 1, \ldots, n\}$$
$$s(x) = \{i \in \{1, \ldots, n\} : x \preceq x_j \preceq x_i \implies x_j = x_i, j = 1, \ldots, n\},$$

- any predictive CDF F that is consistent with $\hat{\boldsymbol{F}}$ must satisfy

$$\max_{i \in s(x)} \hat{F}_i(z) \le F(z) \le \min_{j \in p(x)} \hat{F}_j(z)$$

  at all threshold values $z \in \mathbb{R}$

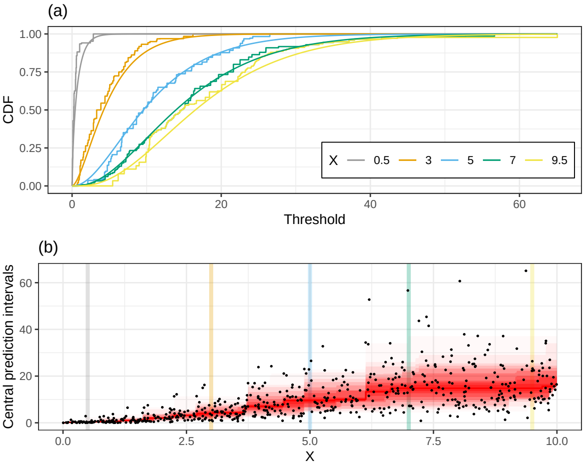
- if both $p(x)$ and $s(x)$ are nonempty, we let F be the pointwise arithmetic average of these bounds, i.e.,

$$F(z) = \frac{1}{2}\left(\max_{i \in s(x)} \hat{F}_i(z) + \min_{j \in p(x)} \hat{F}_j(z)\right)$$

Synthetic Example

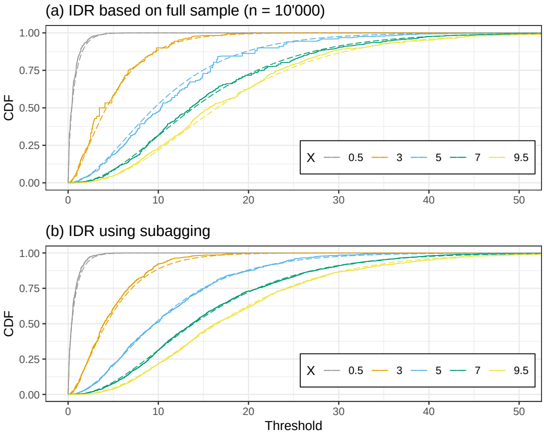
we compute the IDR solution based on a training sample of size $n = 600$ from a population where $X \sim \text{Unif}_{(0,10)}$ and

$$Y \mid X \sim \text{Gamma}(\text{shape} = \sqrt{X}, \text{scale} = \min\{\max\{X, 1\}, 6\})$$



Synthetic Example: Subset Aggregation

same setting as before, but now for a training sample of size $n = 10\,000$

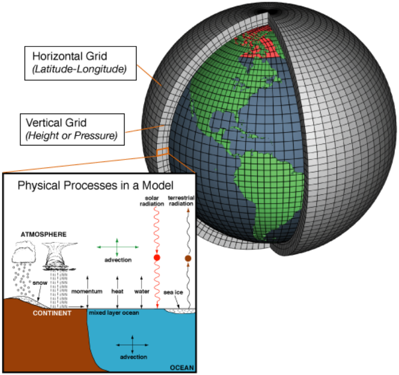


linear aggregation of IDR estimates on 100 subsamples of size 1 000 each (subagging, panel (b)) is superior to using the full training sample (panel (a)) in terms of both computational costs and estimation accuracy

Isotonic Distributional Regression (IDR)

Numerical Weather Prediction (NWP)

modern weather forecasts rely on numerical weather prediction (NWP) models that represent physical processes in the atmosphere



Source: NOAA

run operationally on supercomputers, with huge success

nevertheless, major sources of uncertainty remain (initial conditions, representation of sub-grid scale processes, . . .)

ensemble simulations seek to quantify uncertainty and provide distributional forecasts

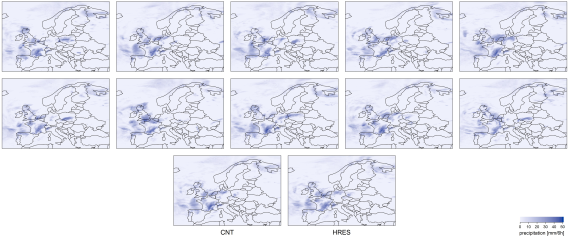
despite continuous improvement, NWP ensemble forecasts remain subject to systematic deficiencies

https://celebrating200years.noaa.gov/breakthroughs/climate_model/AtmosphericModelSchematic.png

ECMWF Ensemble System

the 52-member ensemble system operated by the European Centre for Medium-Range Weather Forecasting (ECMWF) comprises

- a high-resolution member (x_{hres}) at 9 km horizontal grid spacing
- a control member (x_{ctr}) at 18 km horizontal grid spacing
- 50 perturbed members ($x_1, \ldots, x_{50}$) at the same lower resolution but with perturbed initial conditions, to be considered exchangeable



systematic deficiencies call for postprocessing of the raw ensemble output via distributional regression, with covariate vector

$$x = (x_{\text{hres}}, x_{\text{ctr}}, x_1, \ldots, x_{50})$$

Case Study: Precipitation Forecasts

our weather data comprise

- 52-member ECMWF ensemble forecasts and associated observations of 24-hour accumulated precipitation
- at prediction horizons of 1 to 5 days ahead
- from 6 January 2007 to 1 January 2017
- at weather stations on airports in London, Brussels, Zurich and Frankfurt
- precipitation is a particularly challenging variable, due to its nonnegativity and mixed discrete-continuous character with a point mass at zero and a right skewed component on $(0, \infty)$

we perform an out-of-sample evaluation and comparison of distributional regression forecasts

- years 2015 and 2016 as test period
- prior years serve to provide training data
- generally, IDR uses all available training data, whereas parametric competitors benefit from smaller, rolling training periods

Out-of-sample Comparison of Predictive Performance

systematic deficiencies call for postprocessing of the raw ensemble output via distributional regression, with covariate vector

$$x = (x_{\text{hres}}, x_{\text{ctr}}, x_1, \ldots, x_{50})$$

we compare IDR to the raw ensemble and state-of-the-art distributional regression techniques developed specifically for the purpose

- ENS ECMWF raw ensemble forecast, i.e., the empirical distribution of the 52 ensemble members
- BMA Bayesian Model Averaging (Sloughter et al. 2007)
  - semi-parametric, based on mixtures of Bernoulli and power-transformed Gamma components
  - plenty of implementation decisions to be made
- EMOS Ensemble Model Output Statistics (Scheuerer 2014)
  - parametric, predictive CDFs from the three-parameter family of left-censored generalized extreme value (GEV) distributions
  - location and scale parameters linked to covariates, numerous implementation decisions to be made

Choice of Partial Order for IDR

IDR applies readily in this setting

- without any need for adaptations due to the mixed discrete-continuous character of precipitation, nor requiring data transformations

however, the partial order on the elements $x = (x_{\text{hres}}, x_{\text{ctr}}, x_1, \ldots, x_{50})$ of the covariate space $\mathcal{X} = \mathbb{R}^{52}$ needs to be selected thoughtfully

- considering that the elements of $x_{\text{ptb}} = (x_1, \ldots, x_{50})$ are exchangeable

we apply IDR in three variants

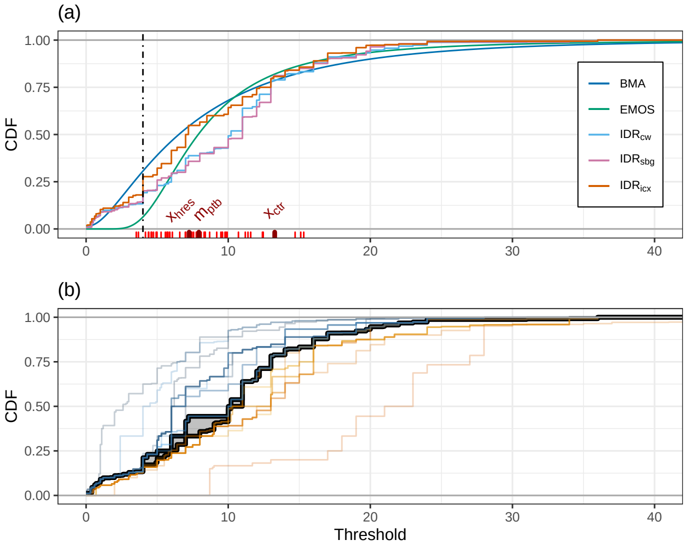
- IDR_{cw} based on x_{hres}, x_{ctr} and $m_{\text{ptb}} = \frac{1}{50} \sum_{i=1}^{50} x_i$ and the componentwise order on $\mathbb{R}^3$, so that

$$x \preceq x' \iff x_{\text{hres}} \le x'_{\text{hres}},\; x_{\text{ctr}} \le x'_{\text{ctr}},\; m_{\text{ptb}} \le m'_{\text{ptb}},$$

- IDR_{sbg} same as IDR_{cw} but combined with subset aggregation
- IDR_{icx} invokes the empirical increasing convex order on x_{ptb}, so that

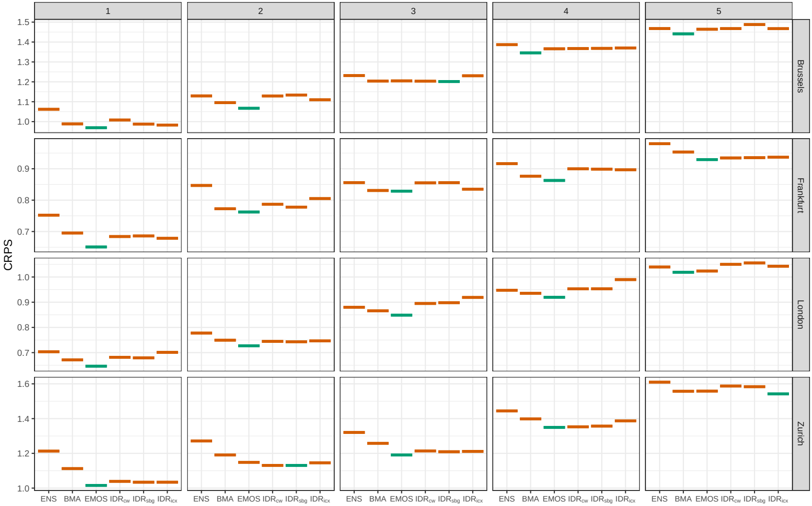
$$x \preceq x' \iff x_{\text{hres}} \le x'_{\text{hres}},\; x_{\text{ptb}} \preceq_{\text{icx}} x'_{\text{ptb}}$$

Example: Predictive CDFs for Brussels, 16 December 2015

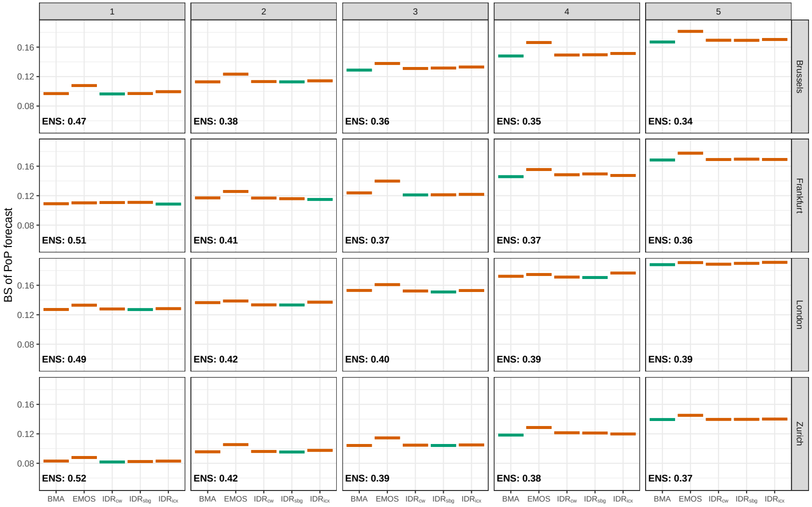


prediction horizon: two days

CRPS


1
2
3
4
5
Brussels
Frankfurt
London
Zurich
CRPS
1.5
1.4
1.3
1.2
1.1
1.0
0.9
0.8
0.7
1.0
0.9
0.8
0.7
1.6
1.4
1.2
1.0
ENS BMA EMOS IDR_{cw} IDR_{sbg} IDR_{icx}

Brier Score



1
2
3
4
5
Brussels
Frankfurt
London
Zurich
BS of PoP forecast
0.16
0.12
0.08
ENS: 0.47
ENS: 0.38
ENS: 0.36
ENS: 0.35
ENS: 0.34
ENS: 0.51
ENS: 0.41
ENS: 0.37
ENS: 0.37
ENS: 0.36
ENS: 0.49
ENS: 0.42
ENS: 0.40
ENS: 0.39
ENS: 0.39
ENS: 0.52
ENS: 0.42
ENS: 0.39
ENS: 0.38
ENS: 0.37
BMA
EMOS
IDR_{cw}
IDR_{sbg}
IDR_{icx}

Isotonic Distributional Regression (IDR)

Summary

in regression analysis

- we are witnessing a transition from conditional mean estimation to conditional distribution estimation
- nonparametric distributional regression techniques such as IDR or techniques based on modern neural networks (SQF-RNN, Gasthaus et al. 2019) are in strong demand

isotonic distributional regression (IDR) learns conditional distributions under order restrictions

- the IDR solution is simultaneously optimal relative to comprehensive classes of proper scoring rules
- IDR provides a unified treatment of all types of real-valued outcomes
- is entirely generic and fully automated, and does not require implementation decisions, except for the choice of a partial order
- shows strongly competitive predictive performance in challenging and important applications
- code for the implementation of IDR in R is available at `https://github.com/AlexanderHenzi/isodistrreg`

Selected References

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

Jordan, A. I., Mühlemann, A., Ziegel, J. F. (2019), **Optimal solutions to the isotonic regression problem**, preprint, `https://arxiv.org/abs/1904.04761`.

Henzi, A., Ziegel, J. F., Gneiting, T. (2019), **Isotonic distributional regression**, preprint, `https://arxiv.org/abs/1909.03725`.