

# Isotonic distributional regression

A. Henzi, J. F. Ziegel, T. Gneiting

10/03/2025

# Introduction

Setting:

- $\mathcal{X}$ , covariate space equipped with partial order " $\preceq$ ";
- $Y$ , real-valued response;
- mostly, probability measures will be identified by their CDFs.



**DISTRIBUTIONAL REGRESSION:** mapping from  $x \in \mathcal{X}$  to a probability measure  $F_x$  which models the conditional distribution  $\mathcal{L}(Y|X = x)$ .

**ISOTONICITY:** the mapping is *isotonic* if  $x \preceq x'$  implies  $F_x \leq_{st} F'_{x'}$ .

# IDR in a nutshell

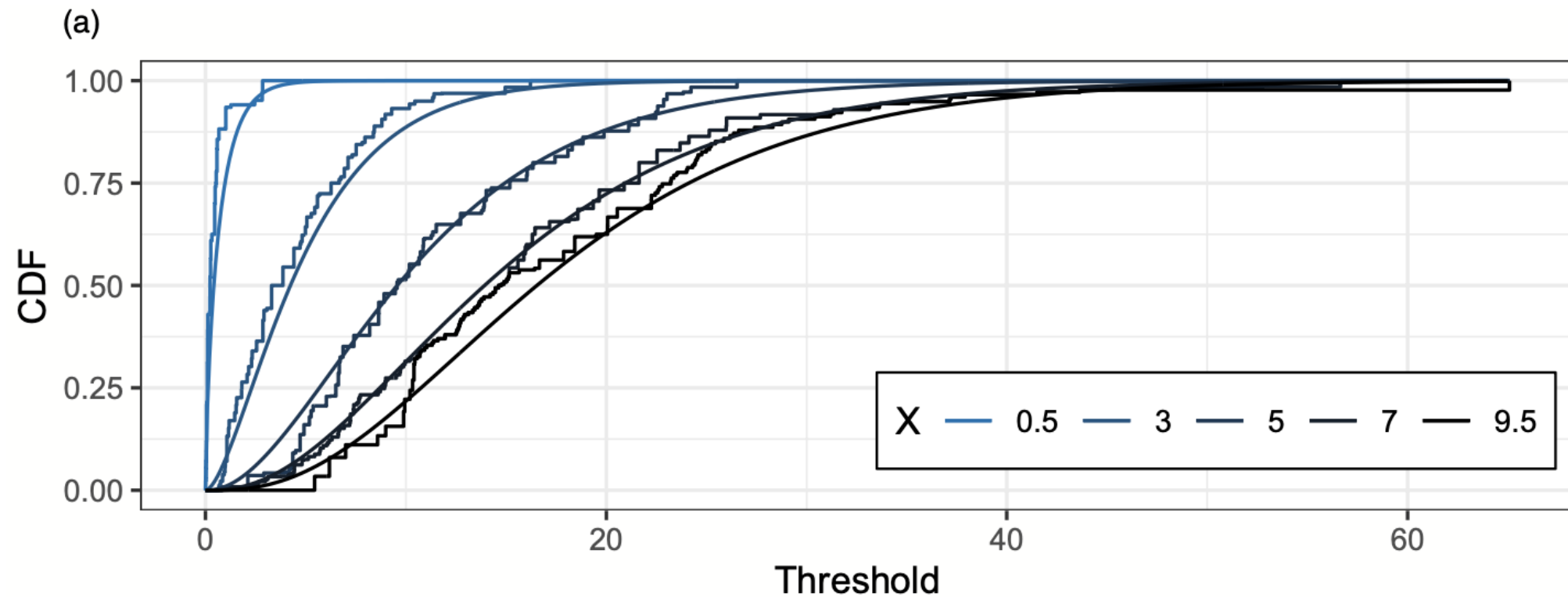
## **IDR:**

- 1) is a non-parametric technique;
- 2) does not need any implementation choices except for a partial order on the covariate space;
- 3) learns conditional distributions that are calibrated and optimal to a comprehensive class of loss functions;
- 4) should be used as a generic benchmark technique in probabilistic forecast problems;
- 5) competitive with other state-of-the-art techniques;
- 6) can be combined with subagging to reduce computational cost and improve predictive performance.

# A first example

$$X \sim U(0, 10)$$

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



1.Preliminaries

2.Existence, uniqueness and universality

3.Consistency

4.Partial orders

5.Implementation

6.Simulation study

7.Case study and discussion

1.Preliminaries

2.Existence, uniqueness and universality

3.Consistency

4.Partial orders

5.Implementation

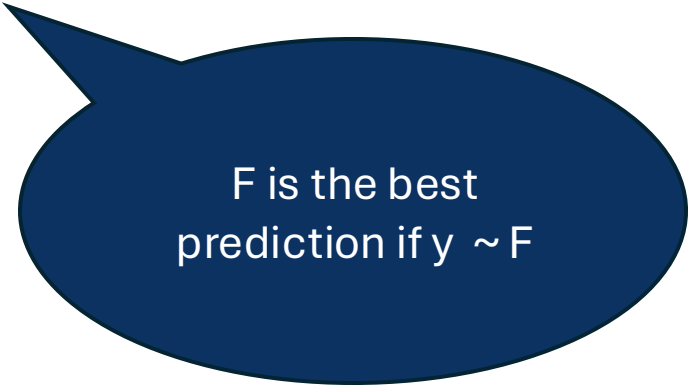
6.Simulation study

7.Case study and discussion

# Proper Scoring Rule

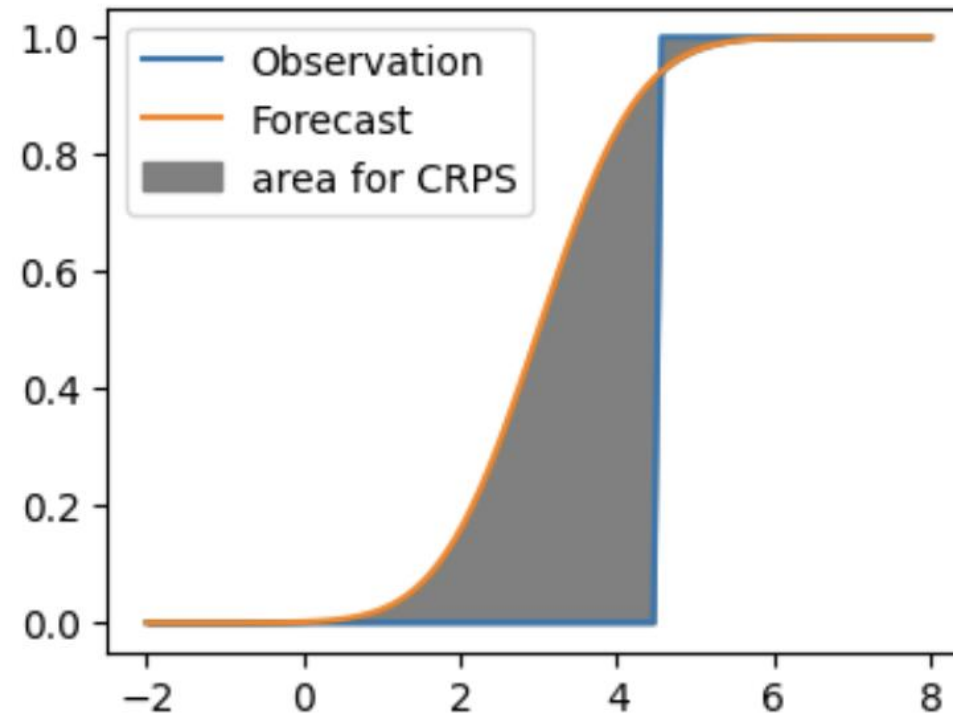
**Proper scoring rule:** a function  $S : P \times \mathbb{R} \rightarrow \overline{\mathbb{R}}$ , where  $P$  is a suitable class of probability measures on  $\mathbb{R}$ , such that  $S(F, \cdot)$  is measurable for any  $F \in P$ , the integral  $\int S(G, y) dF(y)$  exists, and for all  $F, G \in P$

$$\int S(F, y) dF(y) \leq \int S(G, y) dF(y)$$



$F$  is the best  
prediction if  $y \sim F$

$$\text{CRPS}(F, y) = \int_{\mathbb{R}} (F(z) - \mathbf{1}\{y \leq z\})^2 dz.$$

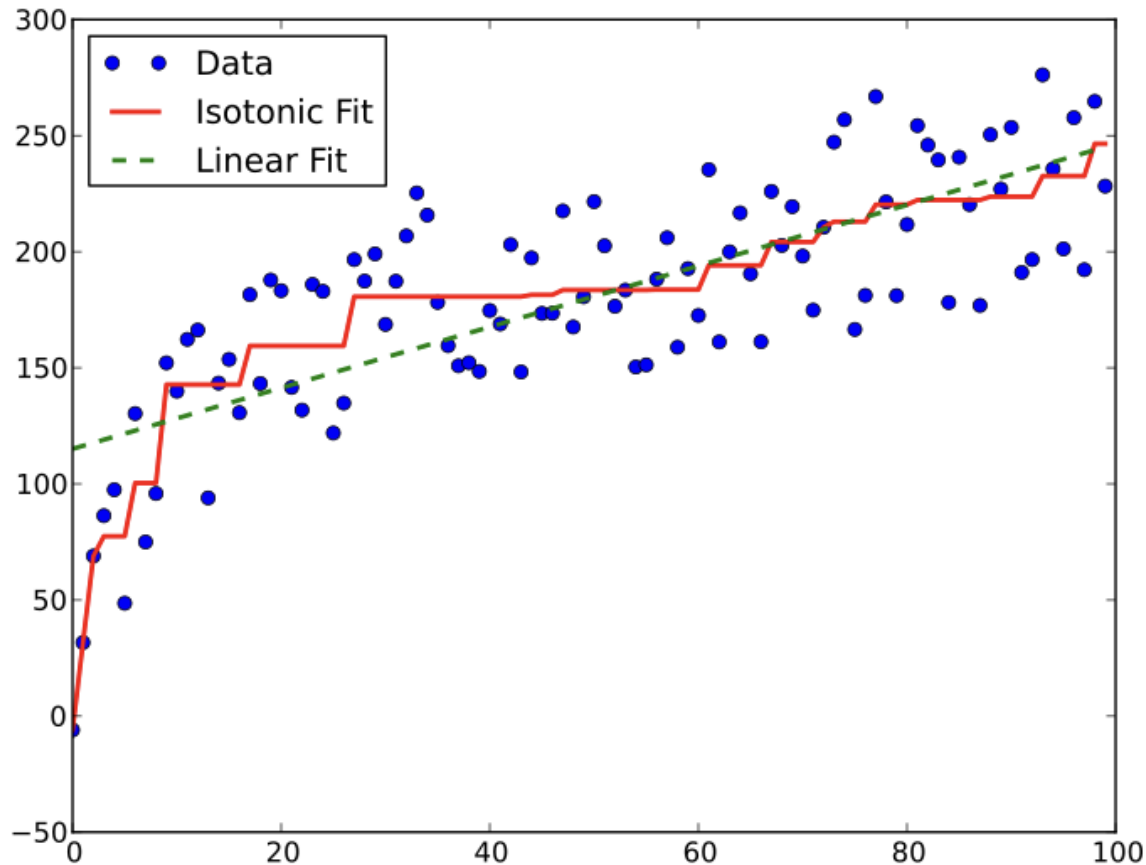




# Calibration

TYPE OF CALIBRATION	“DEFINITION”	COMMENTS
Probabilistic Calibration	$PIT \sim U(0,1)$ $PIT = F_X(Y_-) + V(F_X(Y) - F_X(Y_-))$	Useful to study dispersion
Marginal Calibration	$P(Y \leq y) = \mathbb{E}(F_X(y))$	The expected predicted probability matches the true probability
Threshold calibration	$\mathbb{P}(Y \leq y \mid F_X(y)) = F_X(y) \text{ a.s } \forall y \in \mathbb{R}$	Implies marginal calibration

# Isotonic regression



Source: [https://en.wikipedia.org/wiki/Isotonic\\_regression](https://en.wikipedia.org/wiki/Isotonic_regression)

$$\text{QP: } \min \sum_{i=1}^n w_i (\hat{y}_i - y_i)^2$$

subject to  $\hat{y}_i \leq \hat{y}_j$  for all  $(i, j) \in E = \{(i, j) : x_i \leq x_j\}$

$$f(x) = \begin{cases} \hat{y}_1, & \text{if } x \leq x_1 \\ \hat{y}_i + \frac{x - x_i}{x_{i+1} - x_i} (\hat{y}_{i+1} - \hat{y}_i), & \text{if } x_i \leq x \leq x_{i+1} \\ \hat{y}_n, & \text{if } x \geq x_n \end{cases}$$

1.Preliminaries

2.Existence, uniqueness and universality

3.Consistency

4.Partial orders

5.Implementation

6.Simulation study

7.Case study and discussion

# Ordered Tuples

Let  $n \in \mathbb{N}$ ,  $\mathcal{X}$  and  $\mathcal{Q}$  partially ordered sets,  
 $\mathcal{X}_{\uparrow}^n = \{x = (x_1, \dots, x_n) \in \mathcal{X}^n : x_1 \preceq \dots \preceq x_n\},$   
 $\mathcal{X}_{\downarrow}^n = \{x = (x_1, \dots, x_n) \in \mathcal{X}^n : x_1 \succeq \dots \succeq x_n\}$



$$\mathcal{Q}_{\uparrow, x}^n = \{q = (q_1, \dots, q_n) \in \mathcal{Q}^n : q_i \preceq q_j \text{ if } x_i \preceq x_j\},$$
$$\mathcal{Q}_{\downarrow, x}^n = \{q = (q_1, \dots, q_n) \in \mathcal{Q}^n : q_i \succeq q_j \text{ if } x_i \preceq x_j\}$$

# S-based Regression

**S-based regression:** an element  $\hat{\mathbf{F}} = (\hat{F}_1, \dots, \hat{F}_n) \in \mathcal{P}^n$  is an S-based isotonic regression of  $\mathbf{y} \in I^n$  on  $\mathbf{x} \in \mathcal{X}^n$ , if it is a minimizer of the empirical loss

$$\ell_S(\mathbf{F}) = \frac{1}{n} \sum_{i=1}^n S(F_i, y_i)$$

over all  $\mathbf{F} = (F_1, \dots, F_n)$  in  $\mathcal{P}_{\uparrow, \mathbf{x}}^n$ .



Similar to M-estimation  
But more complex  
space

# Existence and Uniqueness

**Theorem:**  $\exists!$  CRPS-based isotonic regression  $\hat{\mathbf{F}} \in \mathcal{P}^n$  of  $\mathbf{y}$  on  $\mathbf{x}$ .

We refer to this unique  $\hat{\mathbf{F}}$  as the IDR of  $\mathbf{y}$  on  $\mathbf{x}$ .

In the particular case of a total order on the covariate space, and assuming that  $x_1 < \dots < x_n$ , for each  $z \in I$  the solution  $\hat{\mathbf{F}}(z) = (\hat{F}_1(z), \dots, \hat{F}_n(z))$  is given by

$$\hat{F}_i(z) = \min_{k=1, \dots, i} \max_{j=k, \dots, n} \frac{1}{j - k + 1} \sum_{l=k}^j \mathbf{1}\{y_l \leq z\}$$

for  $i = 1, \dots, n$ .

# Alternative representations CRPS

$$\text{CRPS}(F, y) = 2 \int_{(0,1)} \text{QS}_\alpha(F, y) d\alpha$$

$$= 2 \int_{(0,1)} \int_{\mathbb{R}} S_{\alpha,\theta}^Q(F, y) d\theta d\alpha$$

$$= \int_{\mathbb{R}} \int_{(0,1)} S_{z,c}^P(F, y) dc dz$$

$$\text{QS}_\alpha(F, y) = \begin{cases} (1 - \alpha) (F^{-1}(\alpha) - y), & y \leq F^{-1}(\alpha), \\ \alpha (y - F^{-1}(\alpha)), & y \geq F^{-1}(\alpha), \end{cases}$$

$$S_{\alpha,\theta}^Q(F, y) = \begin{cases} 1 - \alpha, & y \leq \theta < F^{-1}(\alpha), \\ \alpha, & F^{-1}(\alpha) \leq \theta < y, \\ 0, & \text{otherwise.} \end{cases}$$

$$S_{z,c}^P(F, y) = \begin{cases} 1 - c, & F(z) < c, y \leq z, \\ c, & F(z) \geq c, y > z, \\ 0, & \text{otherwise.} \end{cases}$$

# Universality

[Universality] The IDR solution  $\hat{F}$  of  $\mathbf{y}$  on  $\mathbf{x}$  is threshold calibrated and has the following properties.

*i) The IDR solution  $\hat{\mathbf{F}}$  is an  $S$ -based isotonic regression of  $\mathbf{y}$  on  $\mathbf{x}$  under any scoring rule of the form*

$$S(F, y) = \int_{(0,1) \times \mathbb{R}} S_{\alpha, \theta}^Q(F, y) dH(\alpha, \theta) \quad (1)$$

*or*

$$S(F, y) = \int_{\mathbb{R} \times (0,1)} S_{z, c}^P(F, y) dM(z, c), \quad (2)$$

*where  $S_{\alpha, \theta}^Q$  is the elementary quantile scoring function,  $S_{z, c}^P$  is the elementary probability scoring rule, and  $H$  and  $M$  are locally finite Borel measures on  $(0, 1) \times \mathbb{R}$  and  $\mathbb{R} \times (0, 1)$ , respectively.*



ii) For every  $\alpha \in (0, 1)$  it holds that  $\hat{\mathbf{F}}^{-1}(\alpha) = (\hat{F}_1^{-1}(\alpha), \dots, \hat{F}_n^{-1}(\alpha))$  is a minimizer of

$$\frac{1}{n} \sum_{i=1}^n s_{\alpha}(\theta_i, y_i) \tag{3}$$

over all  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n) \in I_{\uparrow, x}^n$ , under any function  $s_{\alpha} : I \times I \rightarrow \bar{\mathbb{R}}$  which is left-continuous in both arguments and such that  $S(F, y) = s_{\alpha}(F^{-1}(\alpha), y)$  is a proper scoring rule on  $\mathcal{P}$ .

iii) For every threshold value  $z \in I$ , it is true that

$$\hat{F}(z) = (\hat{F}_1(z), \dots, \hat{F}_n(z))$$

is a minimizer of

$$\frac{1}{n} \sum_{i=1}^n s(\eta_i, 1\{y_i \leq z\}) \tag{4}$$

over all ordered tuples  $\eta = (\eta_1, \dots, \eta_n) \in [0, 1]_{\downarrow, x}^n$ , under any function  $s : [0, 1] \times \{0, 1\} \rightarrow \bar{\mathbb{R}}$  that is a proper scoring rule for binary events, which is left-continuous in its first argument, satisfies  $s(0, y) = \lim_{p \rightarrow 0} s(p, y)$ , and is real-valued, except possibly  $s(0, 1) = -\infty$  or  $s(1, 0) = -\infty$ .

1.Preliminaries

2.Existence, uniqueness and universality

3.Consistency

4.Partial orders

5.Implementation

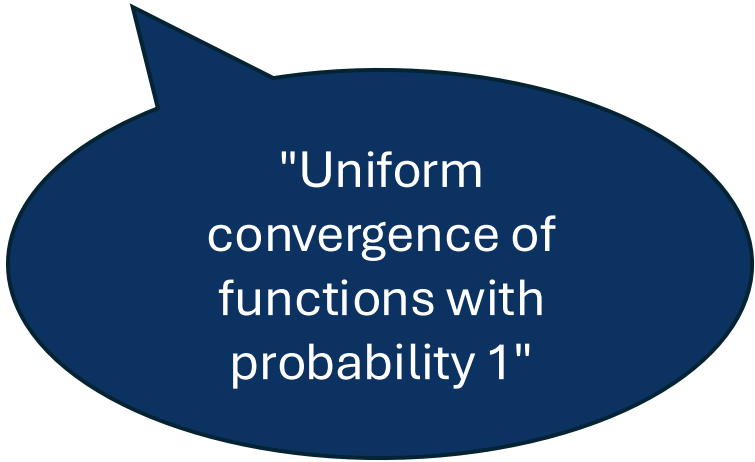
6.Simulation study

7.Case study and discussion

# Consistency

**Consistency:**  $\forall \epsilon > 0$  and  $\delta > 0$

$$\lim_{n \rightarrow \infty} \mathbb{P} \left( \sup_{x \in [\delta, 1-\delta]^d, y \in \mathbb{R}} \left| \hat{F}_x(y) - F_x(y) \right| \geq \epsilon \right) = 0.$$



"Uniform  
convergence of  
functions with  
probability 1"

# Consistency

We assume that  $\widehat{F}_x(y)$  is some value between the bounds given by

$$\max_{i \in \textcolor{red}{s}(\textcolor{red}{x})} \widehat{F}_i(y) \leq \widehat{F}_x(y) \leq \min_{i \in \textcolor{blue}{p}(\textcolor{blue}{x})} \widehat{F}_i(y).$$

**Direct predecessors:**

$$\textcolor{blue}{p}(\textcolor{blue}{x}) = \{i \in \{1, \dots, n\} : X_i \preceq X_j \preceq x \Rightarrow X_j = X_i, j = 1, \dots, n\}$$

**Direct successors:**

$$\textcolor{red}{s}(\textcolor{red}{x}) = \{i \in \{1, \dots, n\} : x \preceq X_j \preceq X_i \Rightarrow X_j = X_i, j = 1, \dots, n\}$$

# Consistency

**Theorem** (uniform consistency). Let  $\mathcal{X} = [0, 1]^d$  be endowed with the componentwise partial order and the norm  $\|u\| = \max_{i=1, \dots, d} |u_i|$ . Let further  $(X_{ni}, Y_{ni}) \in [0, 1]^d \times \mathbb{R}$ ,  $n \in \mathbb{N}$ ,  $i = 1, \dots, n$ , be a triangular array such that  $(X_{n1}, Y_{n1}), \dots, (X_{nn}, Y_{nn})$  are independent and identically distributed random vectors for each  $n \in \mathbb{N}$ , and let  $S_n = \{X_{n1}, \dots, X_{nn}\}$ .

Assume that

(i) for all non-degenerate rectangles  $J \subseteq \mathcal{X}$ , there exists a constant  $c_J > 0$  such that

$$\#(S_n \cap J) > nc_J$$

with asymptotic probability one, i.e., if  $A_n$  denotes the event that  $(S_n \cap J) > nc_J$ , then  $\mathbb{P}(A_n) \rightarrow 1$  as  $n \rightarrow \infty$ ;

(ii) for some  $\gamma \in (0, 1)$ ,

$$\max \{ \#A : A \subset S_n \text{ is antichain} \} \leq n^\gamma$$

with asymptotic probability one.

# Consistency

Assume further that the true conditional CDFs  $F_x(y) = P(Y_{ni} \leq y \mid X_{ni} = x)$  satisfy

- (iii)  $F_x(y)$  is decreasing in  $x$  for all  $y \in \mathbb{R}$ ;
- (iv) for every  $\eta > 0$ , there exists  $r > 0$  such that

$$\sup\{|F_x(y) - F_{x'}(y)| : x, x' \in [0, 1]^d, \|x - x'\| \leq r, y \in \mathbb{R}\} < \eta.$$

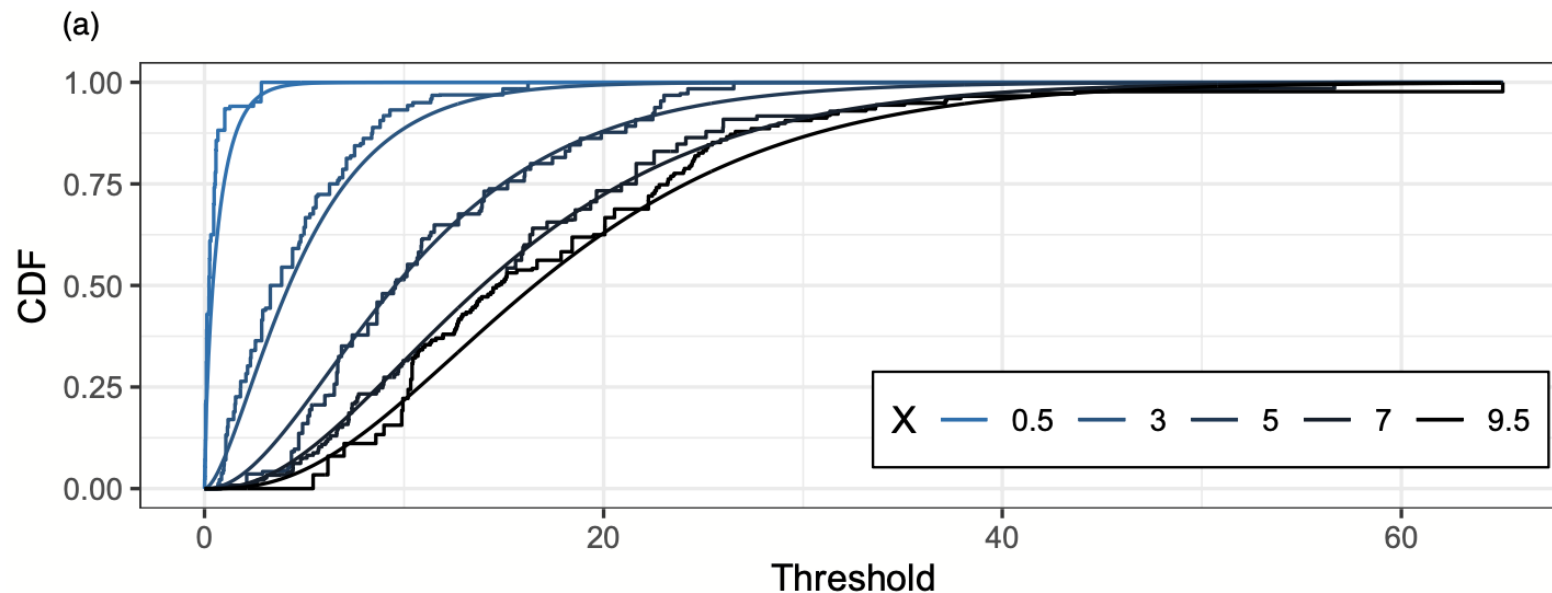
Then for every  $\epsilon > 0$  and  $\delta > 0$ ,

$$\lim_{n \rightarrow \infty} \mathbb{P} \left( \sup_{x \in [\delta, 1-\delta]^d, y \in \mathbb{R}} \left| \hat{F}_x(y) - F_x(y) \right| \geq \epsilon \right) = 0.$$

# Prediction when $X=\mathbb{R}$

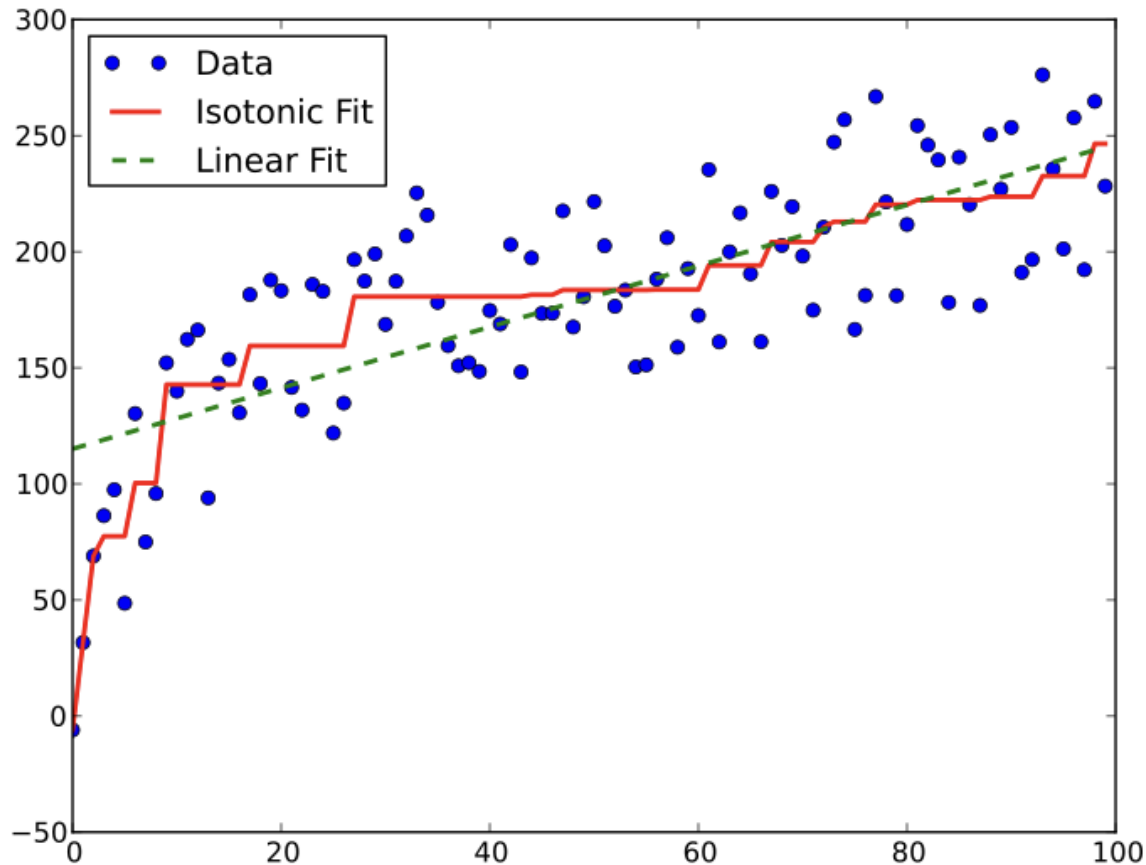
- if  $x < x_1$ , we set  $F = F_1$ ,
- if  $x > x_n$ , we set  $F = F_n$ ,
- if  $x \in (x_i, x_{i+1})$  for some  $i \in \{1, \dots, n-1\}$ , we interpolate linearly:

$$F(z) = \frac{x - x_i}{x_{i+1} - x_i} F_i(z) + \frac{x_{i+1} - x}{x_{i+1} - x_i} F_{i+1}(z).$$





# Isotonic regression



Source: [https://en.wikipedia.org/wiki/Isotonic\\_regression](https://en.wikipedia.org/wiki/Isotonic_regression)

$$\text{QP: } \min \sum_{i=1}^n w_i (\hat{y}_i - y_i)^2$$

subject to  $\hat{y}_i \leq \hat{y}_j$  for all  $(i, j) \in E = \{(i, j) : x_i \leq x_j\}$

$$f(x) = \begin{cases} \hat{y}_1, & \text{if } x \leq x_1 \\ \hat{y}_i + \frac{x - x_i}{x_{i+1} - x_i} (\hat{y}_{i+1} - \hat{y}_i), & \text{if } x_i \leq x \leq x_{i+1} \\ \hat{y}_n, & \text{if } x \geq x_n \end{cases}$$

# Prediction

In the general case of a partial order, considering both  $p(x)$  and  $s(x)$  non-empty:

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

If  $x$  is not comparable to any of  $x_1, \dots, x_n$ , we may set  $F$  to the empirical distribution of the response values  $y_1, \dots, y_n$ .

1.Preliminaries

2.Existence, uniqueness and universality

3.Consistency

4.Partial orders

5.Implementation

6.Simulation study

7.Case study and discussion

# Partial orders

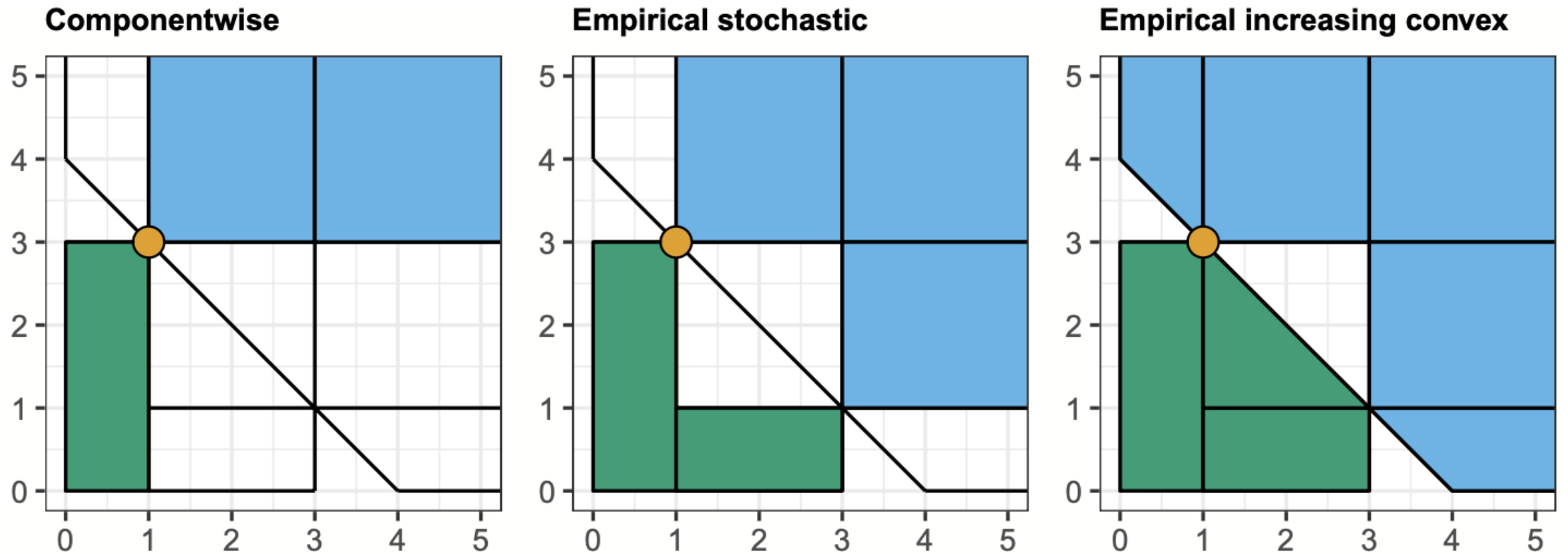
The choice of a sufficiently informative partial order on the covariate space is critical to any successful application of IDR.

In the extreme case of distinct, totally ordered covariate values  $x_1, \dots, x_n \in \mathcal{X}$  and a perfect monotonic relationship to the response values  $y_1, \dots, y_n$ , the IDR distribution associated with  $x_i$  is the point measure in  $y_i$ , for  $i = 1, \dots, n$ .



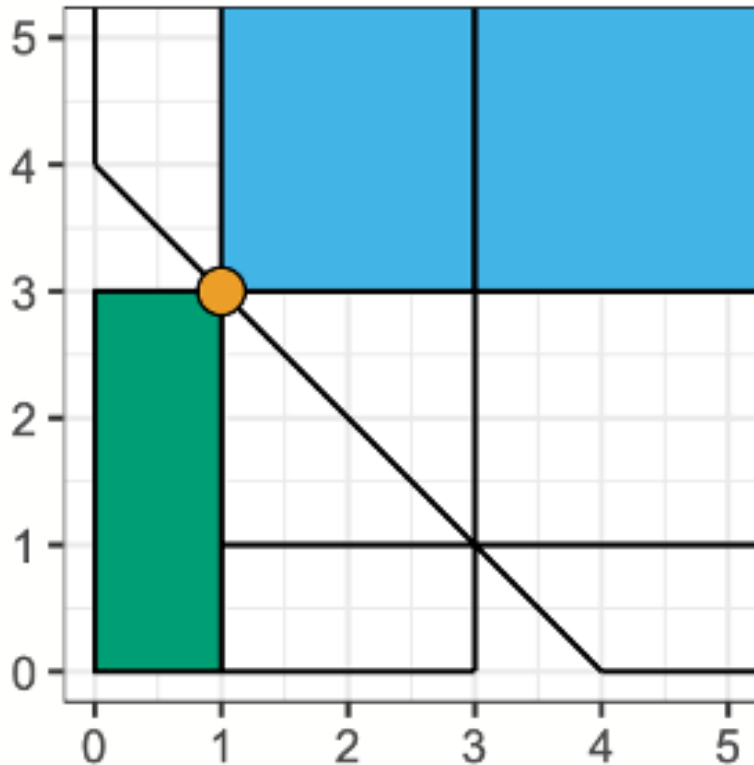
The partial order serves to regularize the IDR solution

# Partial orders



# Componentwise order

Componentwise



Setting  $x = (x_1, \dots, x_d)$  and  $x' = (x'_1, \dots, x'_d) \in \mathbb{R}^d$ .

$$x \preceq x' \iff x_i \leq x'_i \quad \text{for } i = 1, \dots, d.$$

This order becomes weaker as  $d$  increases, but:

**Proposition 1** *Let  $x = (x_1, \dots, x_n)$  and  $x^* = (x_1^*, \dots, x_n^*)$  have components  $x_i = (x_{i1}, \dots, x_{id}) \in \mathbb{R}^d$  and  $x_i^* = (x_{i1}, \dots, x_{id}, x_{i,d+1}) \in \mathbb{R}^{d+1}$  for  $i = 1, \dots, n$ , and let  $S$  be a proper scoring rule. Then if  $\mathbb{R}^d$  and  $\mathbb{R}^{d+1}$  are equipped with the componentwise partial order, and  $\hat{F}$  and  $\hat{F}^*$  denote  $S$ -based isotonic regressions of  $y$  on  $x$  and  $x^*$ , respectively, it is true that*

$$\ell_S(\hat{F}^*) \leq \ell_S(\hat{F}).$$

New covariates can only improve the fit!

# Empirical stochastic order

Setting  $x = (x_1, \dots, x_d)$  and  $x' = (x'_1, \dots, x'_d) \in \mathbb{R}^d$ .

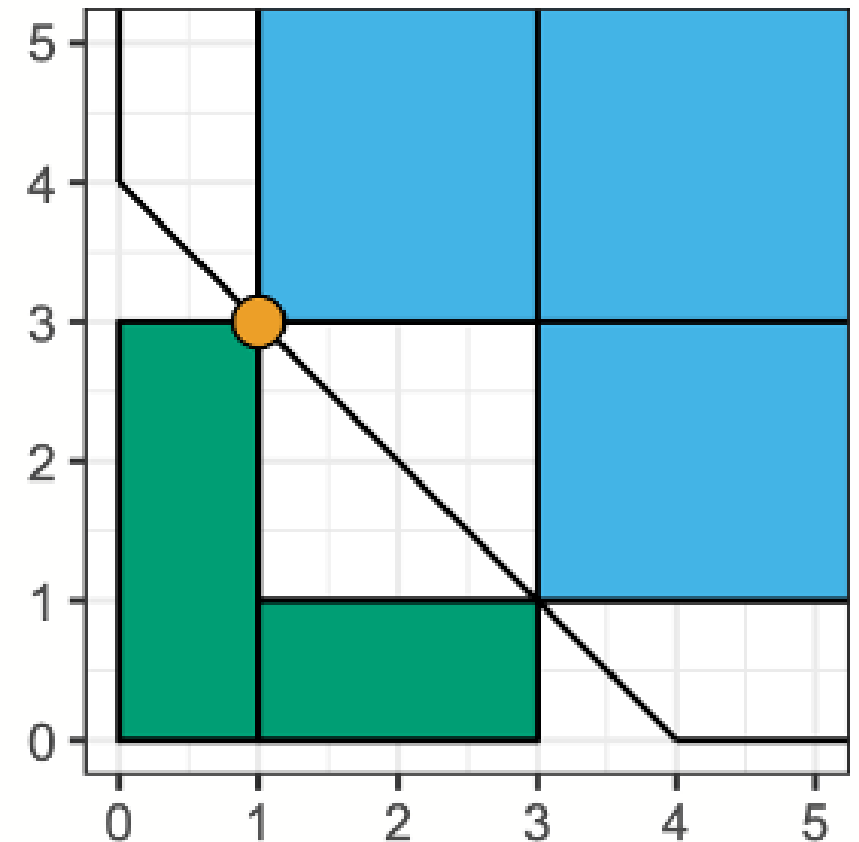
$x \preceq_{st} x' \iff$  the empirical CDF of  $x$  lies *above* the empirical CDF of  $x'$

Componentwise order  
on the sorted elements!

## Proposition 2

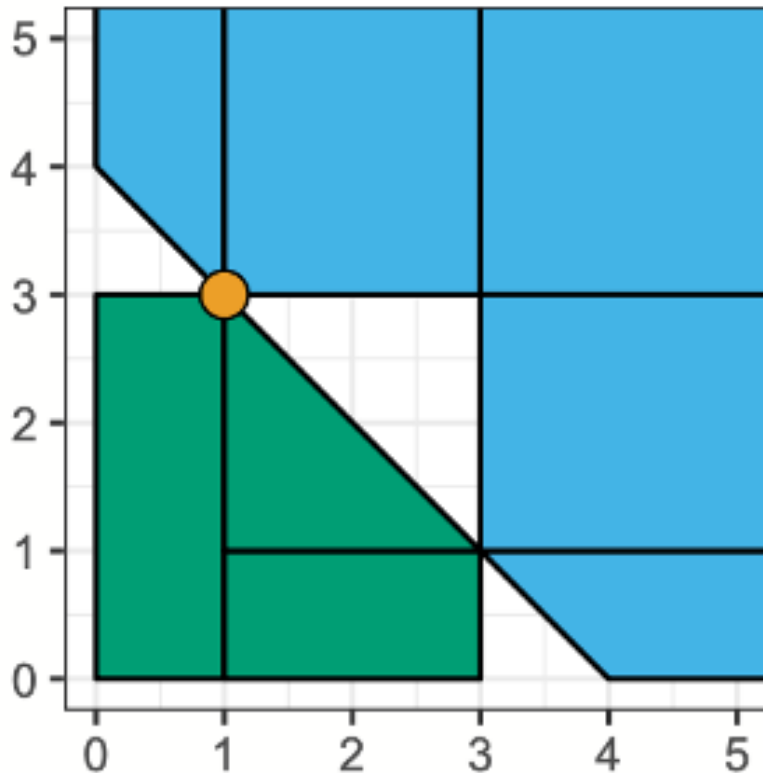
- (a) The relation  $x \preceq_{st} x'$  is equivalent to  $x_{(i)} \leq x'_{(i)}$  for  $i = 1, \dots, d$ .
- (b) If  $x \preceq x'$  then  $x \preceq_{st} x'$ .
- (c) If  $x \preceq_{st} x'$  and  $x$  and  $x'$  are comparable in the componentwise partial order, then  $x \preceq x'$ .

## Empirical stochastic



# Empirical increasing convex order

**Empirical increasing convex**



Let  $X \sim F$  and  $X' \sim F'$ .  $F$  is smaller than  $F'$  if *increasing convex order* if  $\mathbb{E}[\phi(X)] \leq \mathbb{E}[\phi(X')] \quad \forall \phi$  increasing convex function.

$x \preceq_{icx} x' \iff$  the empirical CDF of  $x$  is smaller than the empirical CDF of  $x'$  in icx

## Proposition 3

(a) The relation  $x \preceq_{icx} x'$  is equivalent to

$$\sum_{i=j}^d x_{(i)} \leq \sum_{i=j}^d x'_{(i)} \quad \text{for } j = 1, \dots, d.$$

(b) If  $x \preceq_{st} x'$  then  $x \preceq_{icx} x'$ .



1.Preliminaries

2.Existence, uniqueness and universality

3.Consistency

4.Partial orders

5.Implementation

6.Simulation study

7.Case study and discussion

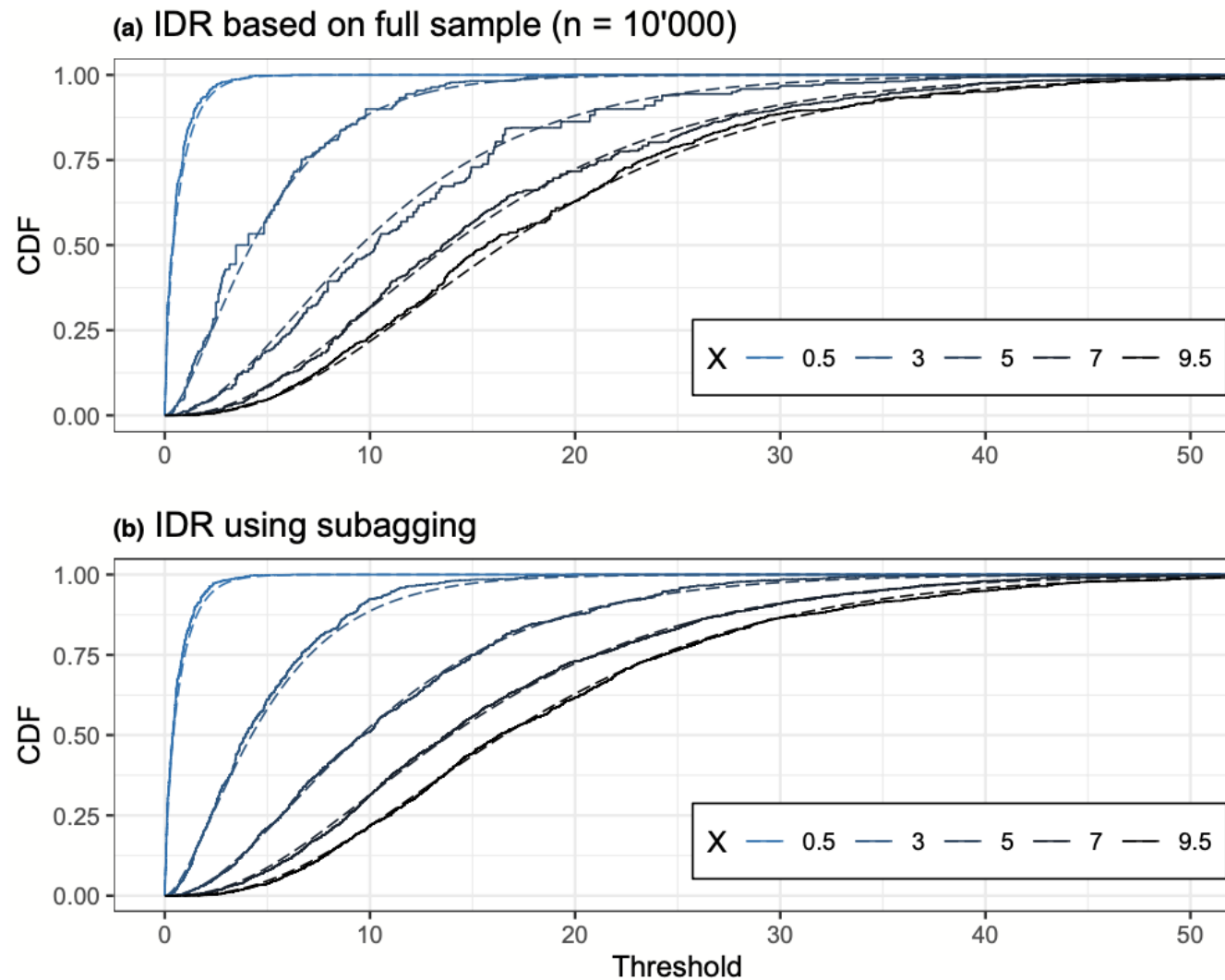
# Implementation via QP

Special case seen before suggests that  $\hat{F} \in \mathcal{P}^n$  of  $\mathbf{y} \in \mathbb{R}^n$  on  $\mathbf{x} \in \mathcal{X}^n$  satisfies:

$$\hat{F}(z) = \arg \min_{\eta \in [0,1]_{\downarrow, \mathbf{x}}^n} \sum_{i=1}^n (\eta_i - \mathbb{I}\{y_i \leq z\})^2$$

The above target function is constant in between the unique values of  $y_1, \dots, y_n$ , say  $\tilde{y}_1 < \dots < \tilde{y}_m$ , and so it suffices to estimate the CDFs at these points only.

# Improvements with subagging



1.Preliminaries

2.Existence, uniqueness and universality

3.Consistency

4.Partial orders

5.Implementation

6.Simulation study

7.Case study and discussion

# Simulation study

Since IDR is primarily seen as a tool for prediction, we compare it to other distributional regression methods in terms of predictive performance (using CRPS) on univariate simulation examples (both continuous and discrete):

*the CRPS links asymptotically to L2 estimation error,  
so it captures both prediction and estimation quality for large test sets.*

Here, our simulation scenarios build on the illustrating example in the introduction. Specifically, we draw a covariate  $X \sim \text{Unif}(0,10)$  and then

$$Y_1 | X \sim \text{Gamma}(\text{shape} = \sqrt{X}, \text{scale} = \min\{\max\{X, 1\}, 6\}), \quad (23) \quad \text{-smooth}$$

$$Y_2 | X = Y_1 | X + 10 \cdot \mathbb{1}\{X \geq 5\}, \quad (24) \quad \text{-discontinuous}$$

$$Y_3 | X = Y_1 | X - 2 \cdot \mathbb{1}\{X \geq 7\}, \quad (25) \quad \text{-non-isotonic}$$

$$Y_4 | X \sim \text{Poisson}(\lambda = \min\{\max\{X, 1\}, 6\}). \quad (26) \quad \text{-discrete}$$

# Simulation study

**For comparison with IDR (and its subagging variant,  $IDR_{sbg}$ ), we use:**

1. A non-parametric kernel (or nearest neighbour) smoothing technique (NP)
2. semiparametric quantile regression + monotone rearrangement (SQR)
3. conditional transformation models (TRAM)
4. distributional or quantile random forests (QRF)

These methods have been chosen since they are not subject to any restrictive assumption on the distribution of the response variable. We also include the ideal forecast which is the conditional distribution of  $Y$  given the covariate.

# Mean CRPS Across Methods & Scenarios

**TABLE 1** Mean CRPS in smooth (23), discontinuous (24), non-isotonic (25), and discrete (26) simulation scenarios with training sets of size  $n$

$n$	500	1000	2000	4000	500	1000	2000	4000
	Smooth (23)				Discontinuous (24)			
NP	3.561	3.542	3.532	3.525	3.614	3.582	3.562	3.549
SQR	3.571	3.543	3.530	3.524	3.647	3.619	3.606	3.600
TRAM	3.560	3.543	3.535	3.531	3.642	3.625	3.616	3.612
QRF	3.589	3.561	3.555	3.553	3.614	3.576	3.561	3.556
IDR	3.604	3.568	3.548	3.535	3.628	3.581	3.555	3.540
IDR <sub>sbg</sub>	3.595	3.561	3.543	3.532	3.620	3.577	3.551	3.537
Ideal	3.516	3.516	3.516	3.516	3.516	3.516	3.516	3.516
	Non-isotonic (25)				Discrete (26)			
NP	3.564	3.544	3.534	3.527	1.136	1.131	1.128	1.126
SQR	3.574	3.546	3.533	3.527	1.129	1.121	1.116	1.114
TRAM	3.566	3.549	3.543	3.539	1.115	1.110	1.107	1.106
QRF	3.587	3.560	3.555	3.553	1.121	1.113	1.112	1.112
IDR	3.605	3.569	3.549	3.536	1.130	1.119	1.113	1.109
IDR <sub>sbg</sub>	3.597	3.564	3.545	3.534	1.128	1.118	1.112	1.109
Ideal	3.516	3.516	3.516	3.516	1.104	1.104	1.104	1.104

# Results

- **Case 23 (smooth):** NP, SQR, and TRAM outperform IDR and QRF.
- **Case 24 (discontinuous):** IDR performs best, highlighting the importance of continuity assumptions.
- **Case 25 (non-Isotonic):** IDR maintains acceptable performance even when assumptions are violated.
- **Case 26 (Poisson/discrete):** IDR is outperformed only by TRAM—strong performance for discrete data.

## Takeaway:

- IDR serves as a universal benchmark in probabilistic forecasting and DR problems.
- With large training sets, IDR remains competitive across various types of outcomes.



1.Preliminaries

2.Existence, uniqueness and universality

3.Consistency

4.Partial orders

5.Implementation

6.Simulation study

7.Case study and discussion

# Case study: Probabilistic Precipitation Forecasts

In the past decades, there have been significant advancements in weather prediction. Along with the culture change from point forecasts to distributional forecasts, the most radical change is the implementation of ensemble systems.

- An ensemble system runs multiple NWP (Numerical Weather Prediction) models under varied initial conditions.
- However, raw ensembles often exhibit biases and dispersion errors, therefore cannot be directly interpreted as a sample from the conditional distribution of future states of the atmosphere.
- Hence, they require some sort of statistical post-processing, typically done by fitting a distributional regression model, such as IDR.

# Case study: Probabilistic Precipitation Forecasts

In this case study, IDR is applied to the statistical post-processing of ensemble forecasts of accumulated precipitation. This variable is notoriously difficult to handle due to its mixed discrete-continuous character, which requires both a point mass at zero and a right skewed continuous component on the positive half-axis.

As competitors to IDR, we consider:

- **BMA** (Bayesian Model Averaging)
- **EMOS** (Ensemble Model Output Statistics)
- **HCLR** (Heteroscedastic Censored Logistic Regression)

While BMA, EMOS and HCLR are parametric methods that have been developed for probabilistic quantitative precipitation forecasting, IDR is a generic technique and fully automatic, once the partial order on the covariate space has been specified.

- 24h accumulated precipitation forecasts and observations from 6/1/2007 to 1/1/2017 at four European airports (London, Brussels, Zurich, Frankfurt).
- Days with missing data are removed for each station.
- Observations are typically in millimeters.
- The covariates come from the ECMWF ensemble of 52 members:
- 1 high-resolution ( $x_{\text{HRES}}$ ), 1 control ( $x_{\text{CTR}}$ ), and 50 perturbed members ( $x_1, \dots, x_{50}$ ).

To summarize, the covariate vector in the distributional regression is:

$$\mathbf{x} = (x_1, \dots, x_{50}, x_{\text{CTR}}, x_{\text{HRES}}) = (x_{\text{PTB}}, x_{\text{CTR}}, x_{\text{HRES}}) \in \mathbb{R}^{52}$$

# EMOS, BMA and HCLR

Before describing the IDR implementation, we revise its leading competitors.

## **EMOS (Ensemble Model Output Statistics):**

Techniques of ensemble model output statistics type can be interpreted as parametric instances of generalized additive models for location, scale and shape. The specific variant used here is based on the three-parameter family of left-censored generalized extreme value (GEV) distributions.

- The left-censoring generates a point mass at zero (no precipitation)
- The shape parameter allows for flexible skewness on the positive half-axis, (rain, hail or snow accumulations)

# EMOS, BMA and HCLR

## BMA (Bayesian Model Averaging):

The general idea of the Bayesian model averaging approach is to employ a mixture distribution, where each mixture component is parametric and associated with an individual ensemble member forecast, with mixture weights that reflect the member's skill.

Here, the BMA predictive CDF is of the form:

$$F_x(y) = w_{HRES}G(y | x_{HRES}) + w_{CTR}G(y | x_{CTR}) + w_{PTB}G(y | x_{PTB})$$

- The component CDFs  $G(y | \cdot)$  are parametric
- The weights sum to one
- Specifically,  $G(y | x_{HRES})$  models the logit of the point mass at zero as a linear function of  $\sqrt[3]{x_{HRES}}$  and  $p_{HRES} = 1\{x_{HRES} = 0\}$  and the distribution for positive accumulations as a gamma density with mean and variance being linear in  $\sqrt[3]{x_{HRES}}$  and  $x_{HRES}$ .
- Analogous approach for  $G(y | x_{CTR})$  and  $G(y | x_{PTB})$

Estimation relies on a two-step procedure:

1. The logit and mean models are fitted
2. Maximum likelihood estimation of the weight parameters and the (joint) variance model via the EM algorithm

## **HCLR (Heteroscedastic Censored Logistic Regression):**

Heteroscedastic censored logistic regression originates from the observation that conditional CDFs can be estimated by dichotomizing the random variable of interest at given thresholds and estimating the probability of threshold exceedance via logistic regression.

The HCLR model used here assumes that the square-root transformed precipitation follows a logistic distribution:

- Censored at zero
- With location parameter linear in  $\sqrt{x_{HRES}}$ ,  $\sqrt{x_{CTR}}$  and the mean of the square-root transformed perturbed forecasts
- With scale parameter linear in the standard deviation of the square-root transformed perturbed forecasts.

# 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 accumulation.
- However, the partial order on the elements of the covariate space  $\mathbf{X}$ , needs to be selected thoughtfully, considering that the perturbed members are exchangeable.

Here, IDR is applied in three variants:

- The first variant ( $\text{IDR}_{\text{cw}}$ ) is based on  $x_{\text{HRES}}$ ,  $x_{\text{CTR}}$  and  $m_{\text{PTB}}$ , along with the usual component-wise order on  $\mathbb{R}^3$

$$x \leq x' \iff m_{\text{PTB}} \leq m'_{\text{PTB}}, x_{\text{CTR}} \leq x'_{\text{CTR}}, x_{\text{HRES}} \leq x'_{\text{HRES}}$$

- The second implementation ( $\text{IDR}_{\text{sbg}}$ ) uses the same variables and partial order but combined with a simple subagging approach: Before applying IDR, the training data is split into the two disjoint subsamples of training observations, and we average the predictions based on these two subsamples.
- The third one ( $\text{IDR}_{\text{icx}}$ ) combines the empirical increasing convex order for  $x_{\text{PTB}}$  with the total order on  $\mathbb{R}$  for  $x_{\text{HRES}}$ .

$$x \leq x' \iff x_{\text{PTB}} \leq_{\text{icx}} x'_{\text{PTB}}, x_{\text{HRES}} \leq x'_{\text{HRES}}$$



# Selection of training periods

- The selection of the training period is a crucial step in the statistical post-processing of NWP output as the assumption of a stationary relationship between the forecasts and the observations may not hold in practice.
- A careful selection of the training period, ensures that post-processing models remain accurate and robust despite these variations

## For BMA and EMOS:

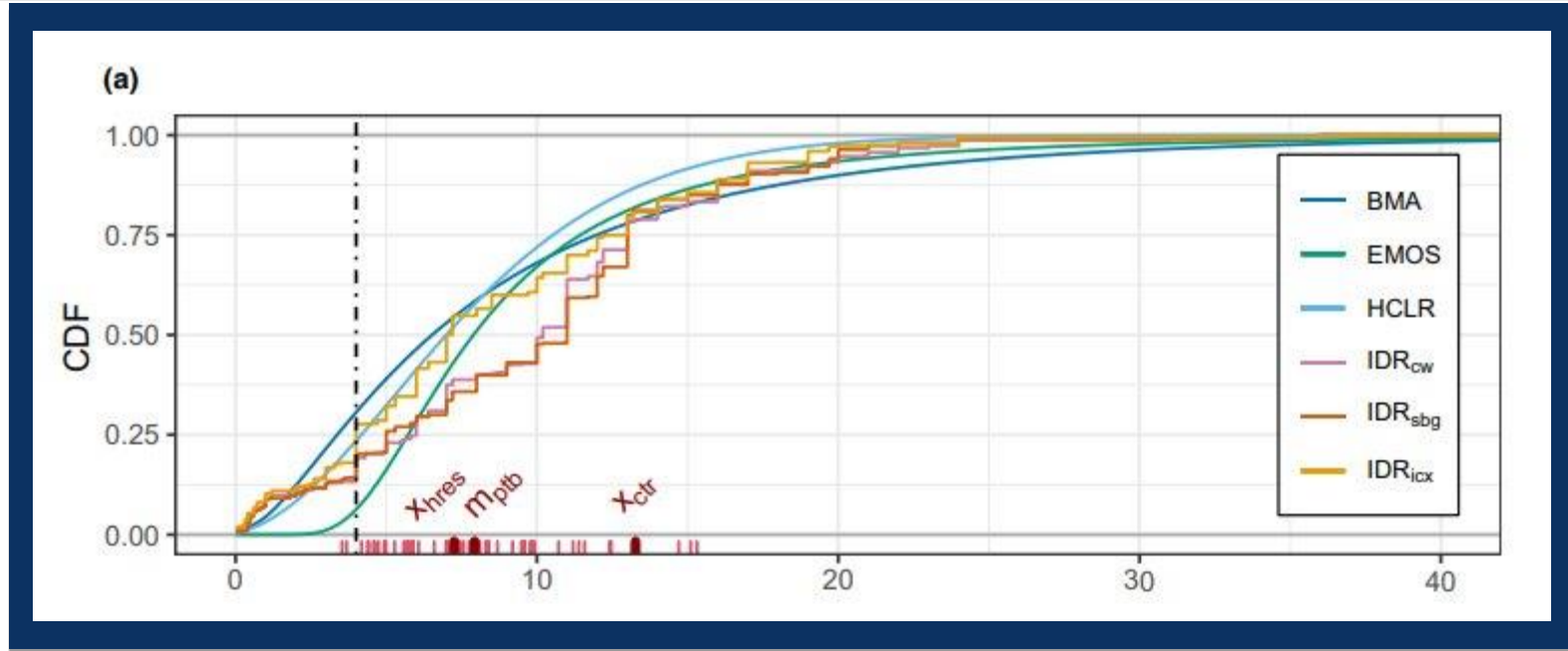
- A training period over a rolling window of the latest available 720 days at the time of forecasting.
- In general, it is preferable to select training data seasonally, however in this case the positive effect of using seasonal training data does not outweigh the negative effect of a smaller sample size.

## For IDR:

- As a non-parametric technique, IDR requires larger sets of training data than BMA or EMOS.
- All data available at the time of forecasting (about 2500 to 3000 days for Frankfurt, Brussels and Zurich, and 1500 days for London) was used. The same training periods are also used for HCLR.

For evaluation, the years 2015 and 2016 (and 01 January 2017) was used for all post-processing methods and the raw ensemble. This test dataset consists of roughly 700 instances for each station and lead time.

# Example of predictive CDFs



The above image shows predictive CDFs for accumulated precipitation in Brussels on 16 December 2015, at a prediction horizon of 2 days.

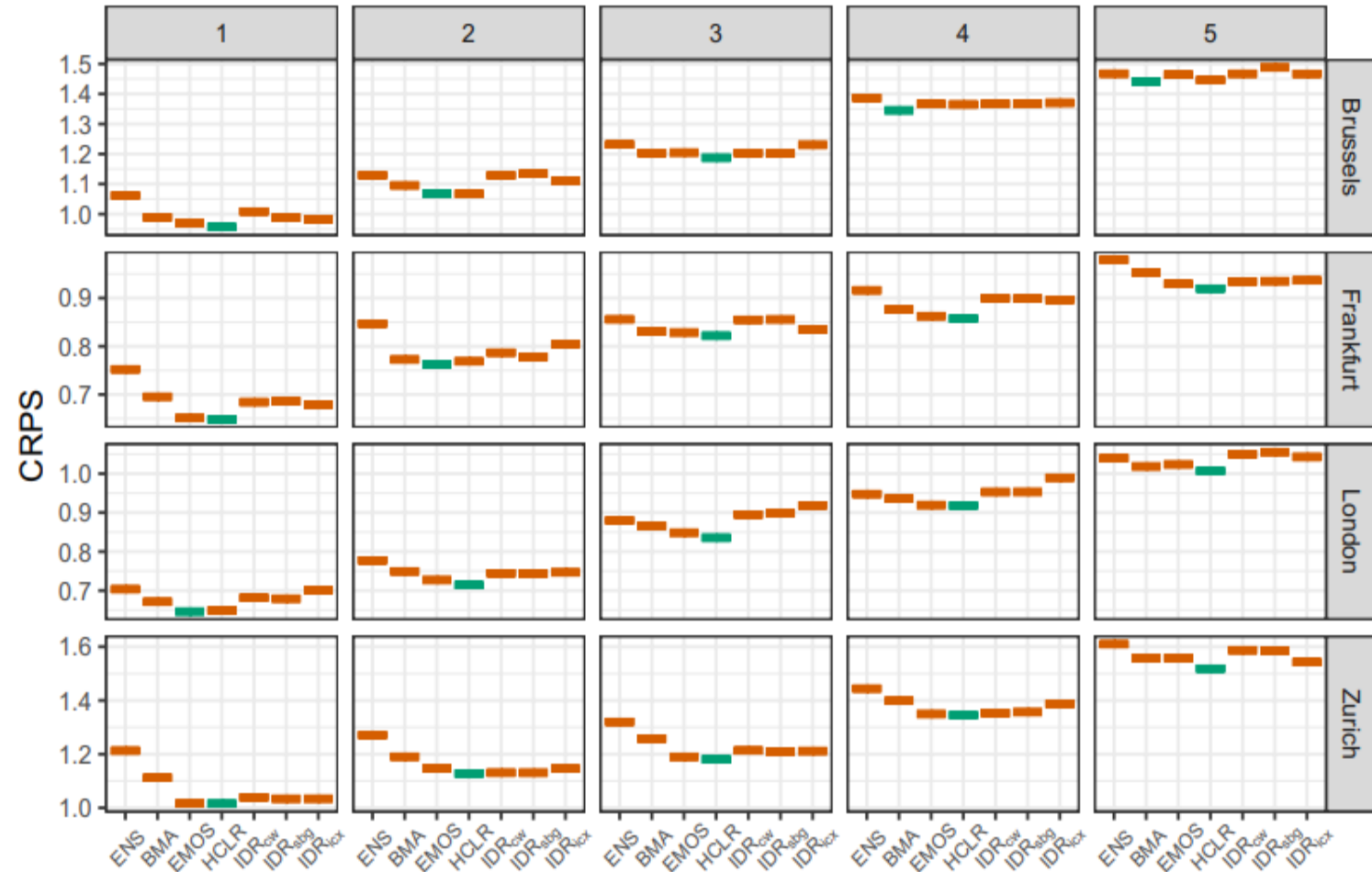
- The marks at the bottom refer to the covariates and the mean of the perturbed members
- The observation at 4mm is indicated by the real line
- Under all four methods the probability of no precipitation (point mass at 0) is extremely small
- The BMA, EMOS and HCLR CDFs are smooth and supported on the positive half-axis
- The CDFs of the IDR variants are piecewise constant with jump points at observed values in the training period

# Results

We now use the mean CRPS over the test period as an overall measure of out-of-sample predictive performance.

The figure shows the CRPS of the raw and post-processed forecasts for all stations and lead times.

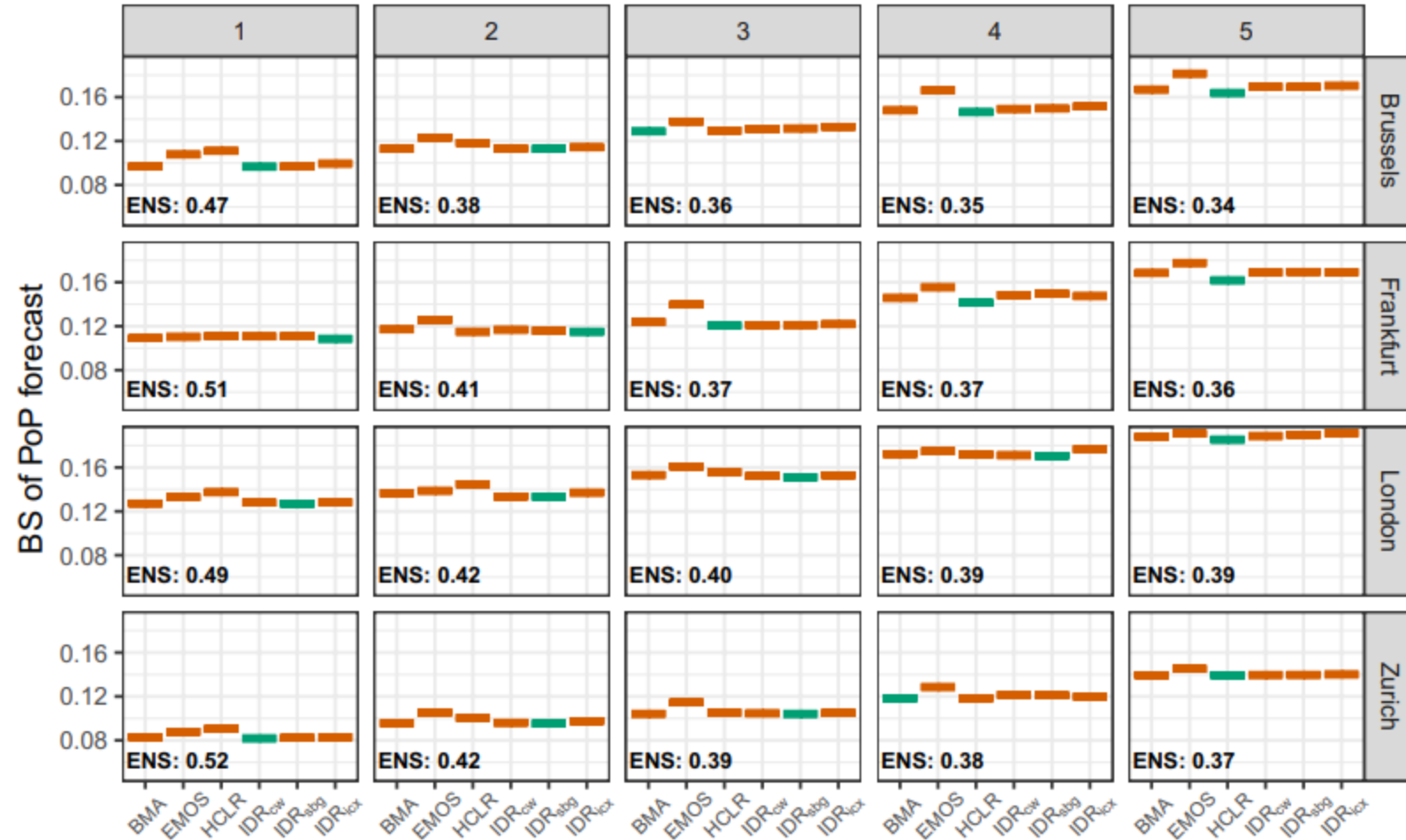
- HCLR performs best in terms of the CRPS
- the IDR variants still show scores of a similar magnitude and outperform BMA in many instances
- All three IDR variants show a PIT-distribution close to uniform
- And so do BMA, EMOS and HCLR, as opposed to the raw ensemble which is underdispersed.



# Results

We now evaluate probability of precipitation forecasts by means of the Brier score.

- As opposed to the raw ensemble forecast, all distributional regression methods yield reliable probability forecasts
- BMA,  $IDR_{cw}$ ,  $IDR_{sbg}$  and  $IDR_{icx}$  separate the estimation of the point mass at zero, and of the distribution for positive accumulations, and perform ahead of EMOS
- HCLR is outperformed by BMA and the IDR variants at lead times of one or two days but achieves a lower Brier score at the longest lead time of 5 days



# Results

- IDR tends to outperform EMOS and HCLR for probability of precipitation forecasts (Brier score), but not for precipitation accumulations (CRPS).
- This is attributed to the fact that parametric techniques are capable of extrapolating beyond the range of the training responses, whereas IDR is not: the highest precipitation amount judged feasible by IDR equals the largest observation in the training set.
- IDR does not use information about the spread of the raw ensemble. This is inconsequential for the forecast of the probability of precipitation but it may impede forecasts of precipitation accumulations.
- In the implementation, the simple subbagging method used in  $IDR_{\text{sbg}}$  reduced the computation time by up to one half.
- The results underscore the suitability of IDR as a benchmark technique in probabilistic forecasting problems. Despite being generic and fully automated, IDR performs competitively relative to techniques designed specifically for the purpose.

# Discussion

- IDR takes advantage of the partial order relations within the covariate space.
- It can be implemented through PAV algorithm (a generalization).
- It is fully automated and provides for a unified treatment of different types of response variables once the partial order and the training set have been identified.
- IDR relies only on information provided by order constraint and the choice of the partial order is crucial prior to the analysis.
- There is evidence that IDR is robust under misspecifications of the partial order: IDR has guaranteed in-sample threshold calibration (Universality Theorem) and therefore satisfies a minimal requirement for reliable probabilistic forecasts under any partial order.

# Discussion

- A limitation of IDR in its present form is that we only consider the usual stochastic order on the space  $\mathcal{P}$  of the conditional distributions. Hence, IDR is unable to distinguish situations where the conditional distributions agree in location but differ in spread, shape or other regards.
- Another direction of further research would be to extend IDR to multivariate response variables so that it allows for simultaneous post-processing of forecasts for several variables.