# Nonparametric Predictive Inference - Introduction

Frank Coolen

IIT Hyderabad

16 December 2015

# Objective Inference?

Jon Williamson (2004): Two norms for (precise) Objective (Bayesian) Inference

- **Empirical:** An agent's knowledge of the world should constrain her degrees of belief. Thus if one knows that a coin is symmetrical and has yielded heads roughly half the time, then one's degree of belief that it will yield heads on the next throw should be roughly $1/2$.
- **Logical:** An agent's degrees of belief should also be fixed by her lack of knowledge of the world. If the agent knows nothing about an experiment except that it has two possible outcomes, then she should award degree of belief $1/2$ to each outcome.

# Objective Inference Reformulated

Perhaps we can interpret these norms, loosely, as follows:

- **'Empirical':** Objective inferences should not disagree with empirical evidence.
- **'Logical':** If one has no information suggesting that one possible outcome is more likely than another, then this should be reflected by identical uncertainty quantifications for these outcomes.

# Hill's assumption $A_{(n)}$ (Hill, 1968)

- $X_1, \ldots, X_n, X_{n+1}$ are real-valued and exchangeable random quantities
- $x_1 < x_2 < \ldots < x_n$ are the ordered observed values of $X_1, \ldots, X_n$ (and let $x_0 = -\infty$ and $x_{n+1} = \infty$)
- For $X_{n+1}$, $A_{(n)}$ is given by

$$P(X_{n+1} \in I_j = (x_{j-1}, x_j)) = \frac{1}{n+1} \ , \quad j = 1, \ldots, n+1$$

Durham University

# Nonparametric predictive inference (NPI)

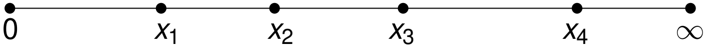
- NPI is based on Hill's assumption $A_{(n)}$
- Let $\mathcal{B}$ be the Borel $\sigma$-field over $\mathbb{R}$. For any element $B \in \mathcal{B}$, lower probability $\underline{P}(.)$ and upper probability $\overline{P}(.)$ for the event $X_{n+1} \in B$, based on the intervals $I_j = (x_{j-1}, x_j)$ ($j = 1, 2, \ldots, n+1$) created by $n$ real-valued non-tied observations, and the assumption $A_{(n)}$, are

$$\underline{P}(X_{n+1} \in B) = \frac{1}{n+1} |\{j : I_j \subseteq B\}|$$
$$\overline{P}(X_{n+1} \in B) = \frac{1}{n+1} |\{j : I_j \cap B \neq \emptyset\}|$$

Durham University

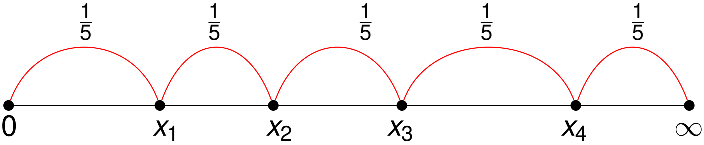
# Illustration example

$n = 4$



# Illustration example

$n = 4$

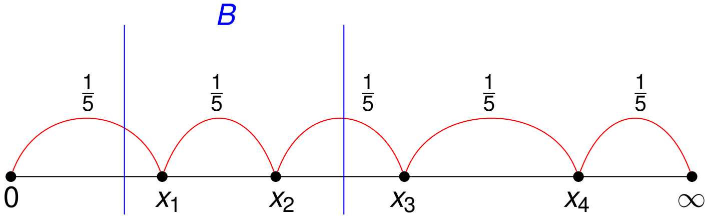



$P(X_5 \in (0, x_1)) = \frac{1}{5}$ $\qquad$ $P(X_5 \in (x_4, \infty)) = \frac{1}{5}$

$P(X_5 \in (x_i, x_{i+1})) = \frac{1}{5}$ , $\quad i = 1, 2, 3$

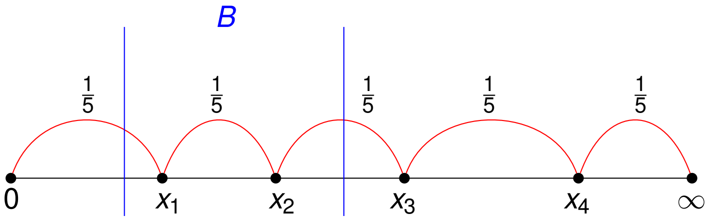
# Illustration example

$n = 4$



# Illustration example

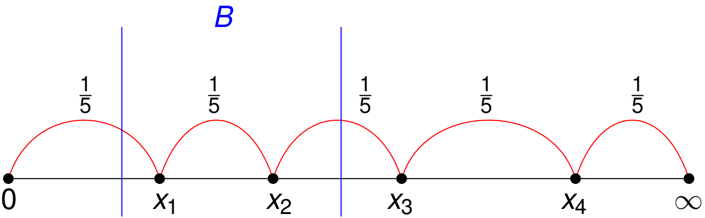
$n = 4$




$\underline{P}[X_5 \in B] = \frac{1}{5}$ $\overline{P}[X_5 \in B] = \frac{3}{5}$

# Illustration example

$n = 4$




$\underline{P}[X_5 \in B] = \frac{1}{5}$ $\overline{P}[X_5 \in B] = \frac{3}{5}$

$Imprecision = \overline{P} - \underline{P} = \frac{2}{5} = 0.4$

# Comparing two independent groups

Data from two independent groups $X$ and $Y$:

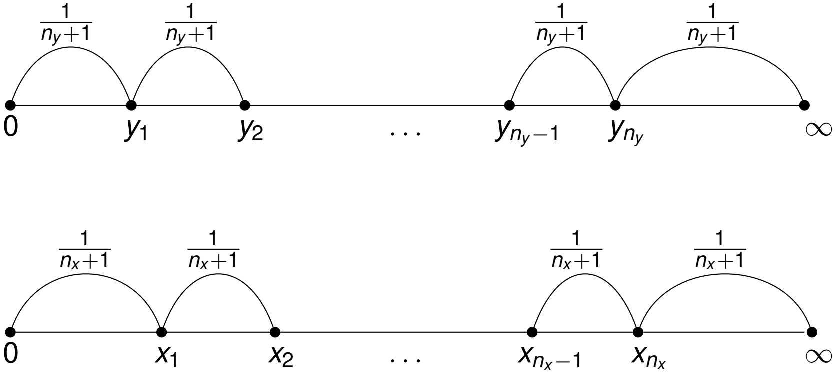
$$x_1 < x_2 < \ldots < x_{n_x} \qquad \text{and} \qquad y_1 < y_2 < \ldots < y_{n_y}$$

The classical methods test $H_0 : F_X = F_Y$.

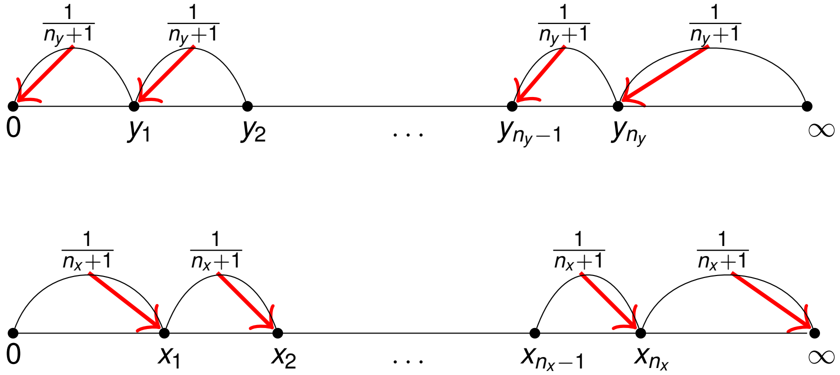
For complete data, Coolen (1996) introduced NPI to compare two independent groups depending on $A_{(n)}$ . This is given via the lower and upper probabilities

$$\underline{P}(X_{n_x+1} < Y_{n_y+1}) \qquad \overline{P}(X_{n_x+1} < Y_{n_y+1})$$

Durham University

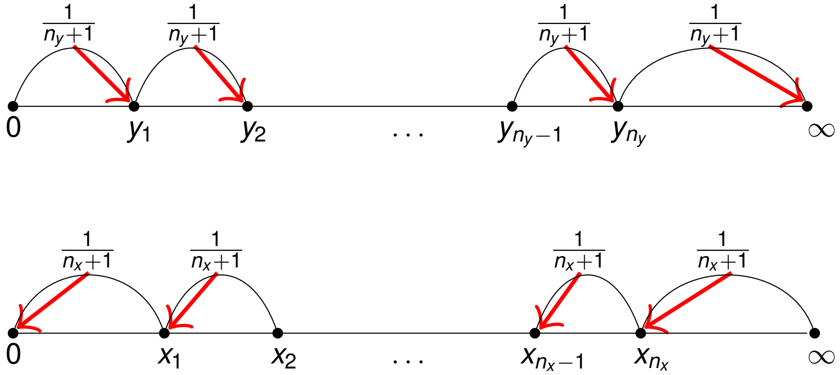

$\frac{1}{n_y+1}$ $\frac{1}{n_y+1}$ $\frac{1}{n_y+1}$ $\frac{1}{n_y+1}$
$0$ $y_1$ $y_2$ $\dots$ $y_{n_y-1}$ $y_{n_y}$ $\infty$
$\frac{1}{n_x+1}$ $\frac{1}{n_x+1}$ $\frac{1}{n_x+1}$ $\frac{1}{n_x+1}$
$0$ $x_1$ $x_2$ $\dots$ $x_{n_x-1}$ $x_{n_x}$ $\infty$

## Lower Probability, $\underline{P}(X_{n_x+1} < Y_{n_y+1})$




Durham University

# Upper Probability, $\overline{P}(X_{n_x+1} < Y_{n_y+1})$




Durham University

# Example

We use data on birthweights for 12 male and 12 female babies as presented by Dobson (1983).

| | | | | | | |
|---|---|---|---|---|---|---|
| Male ($X$) | 2625 | 2628 | 2795 | 2847 | 2925 | 2968 |
| | 2975 | 3163 | 3176 | 3292 | 3421 | 3473 |
| Female ($Y$) | 2412 | 2539 | 2729 | 2754 | 2817 | 2875 |
| | 2935 | 2991 | 3126 | 3210 | 3231 | 3317 |

$$\underline{P}(X_{13} > Y_{13}) = \frac{86}{169} = 0.509$$

$$\overline{P}(X_{13} > Y_{13}) = \frac{111}{169} = 0.657.$$

Durham University

# NPI for $m$ future observations

- We are interested in $m \geq 1$ future observations, $X_{n+i}$ for $i = 1, \ldots, m$.
- We link the data and future observations via Hill's assumption $A_{(n)}$, actually via $A_{(n+m-1)}$ (which implies $A_{(n+k)}$ for all $k = 0, 1, \ldots, m-2$).
- Let $S_j = \#\{X_{n+i} \in I_j,\ i = 1, \ldots, m\}$, then inferences about these $m$ future observations, assuming $A_{(n+m-1)}$, can be based on the following probabilities, for any $(s_1, \ldots, s_{n+1})$ with non-negative integers $s_j$ with $\sum_{j=1}^{n+1} s_j = m$

$$P(\bigcap_{j=1}^{n+1} \{S_j = s_j\}) = \binom{n+m}{n}^{-1}$$

Durham University

# Reproducibility of tests

Will a statistical test, when the experiment is repeated under the same circumstances, give the same overall result (e.g. reject a null-hypothesis or not)?

This is a topic of much confusion in (classical) statistics, particularly also in the literature in a range of application areas. One reason for confusion may be misunderstanding of a p-value.

This problem seems, quite obviously, to have a predictive nature!

*PhD thesis Sulafah Bin Himd, 2014*
Also introduced NPI-Bootstrap, and also used this for test reproducibility

# NPI-RP for the one-sample signed-rank test

$H_0 : X_1, \ldots, X_n$ symmetrically distributed around median $\theta$.

$$W = \sum_{X_i > \theta} \text{Rank}(|X_i - \theta|)$$

Reject $H_0$ in favour of $H_1$ : median $> \theta$ iff $W \geq W_\alpha$, the $100(1-\alpha)$ percentile of the null-distribution for $W$.

Without loss of generality: set $\theta = 0$.

NPI considers future observations $X_{n+1}, ..., X_{2n}$. Given real test results $x_{(1)} < ... < x_{(n)}$, there are $\binom{2n}{n}$ equally likely possible orderings of the future observations among the real test results.

*For each specific ordering*, we calculate the minimum and maximum possible test statistic values, $\underline{W}^f$ and $\overline{W}^f$.

If original data led to rejection of $H_0$, as $W \geq W_\alpha$, then $\underline{RP}$ is the proportion of all $\binom{2n}{n}$ orderings with $\underline{W}^f \geq W_\alpha$ and $\overline{RP}$ the proportion with $\overline{W}^f \geq W_\alpha$.

$\underline{W}^f$ and $\overline{W}^f$ can be calculated without the need to order the $n$ future observations.

For a specific ordering, let $S_j$ be the number of the $n$ future observations in interval $(x_{(j-1)}, x_{(j)})$ (with $x_{(0)} = -\infty, x_{(n+1)} = \infty$).

Durham University

To calculate $\underline{W}^f$, all $S_j$ future observations in $(x_{(j-1)}, x_{(j)})$ are put at ('just to the right of') $x_{(j-1)}$.

Order the absolute data and $-\infty$, with ranks $j = 1, \ldots, n+1$. Let $x_{|j|}$ denote the $j$-th ordered value if positive, $x_{-|j|}$ if negative ($x_{-|n+1|} = -\infty$).

For $j = 1, \ldots, n+1$, Let $T_j$ be the number of future observations, in the specific ordering considered, that are put at $x_{|j|}$, and $T_{-j}$ the number of such future observations that are put at $x_{-|j|}$. This means that $T_j = S_l$ with $x_{(l-1)} = x_{|j|} > 0$ and $T_{-j} = S_l$ with $x_{(l-1)} = x_{-|j|} < 0$.

$$\underline{W}^f = \sum_{j>0} T_j \left[ \frac{(T_j + 1)}{2} + \sum_{|i|<j} T_i \right] \tag{1}$$

$\overline{W}^f$ is similarly derived, with all $S_j$ future observations in $(x_{(j-1)}, x_{(j)})$ put at ('just to the left of') $x_{(j)}$.

# Example signed-rank test

| sign-ranked data | $W$ | $\underline{RP}$ | $\overline{RP}$ |
|---|---|---|---|
| 1,2,3,4,5,6 | 21 | 0.5 | 1 |
| -1,2,3,4,5,6 | 20 | 0.364 | 0.773 |
| -2,1,3,4,5,6 | 19 | 0.326 | 0.712 |
| -3,1,2,4,5,6 | 18 | 0.364 | 0.718 |
| -2,-1,3,4,5,6 | 18 | 0.5 | 0.788 |
| -4,1,2,3,5,6 | 17 | 0.429 | 0.750 |
| -3,-1,2,4,5,6 | 17 | 0.538 | 0.810 |
| -3,-2,-1,4,5,6 | 15 | 0.728 | 0.902 |
| -6,1,2,3,4,5 | 15 | 0.494 | 0.773 |
| -6,-3,-1,2,4,5 | 11 | 0.805 | 0.935 |
| -6,-5,-4,-3,-2,-1 | 0 | 0.992 | 1 |

**Table:** NPI-RP for signed-rank test with $H_1$: median$> 0$, $n = 6$, $\alpha = 0.05$, $W_{0.05} = 19$.

Durham University

# Diagnostic tests

One has real-valued measurements for two groups, say healthy and diseased people, and wants to determine an optimal threshold for classification. For example:

$X$: 'healthy', $n_x = 2$ observations (underlined)

$Y$: 'disease', $n_y = 14$ observations

140, <u>150</u>, 180, 185, 188, 190, 203, 204, <u>205</u>, 230, 260, 280, 300, 305, 330, 344

Aim: find optimal threshold $c$ such that '$X \leq c < Y$'

Durham University

A popular nonparametric method considers the ROC curve and determines optimal $c$ by maximising the empirical Youden's index

$$J_e(c) = TPF_e(c) - FPF_e(c) = \frac{\sum_{i=1}^{n_x} 1[x_i \leq c]}{n_x} + \frac{\sum_{j=1}^{n_y} 1[y_j > c]}{n_y} - 1$$

We can consider this explicitly as a predictive problem.

Durham University

Consider $m \geq 1$ future healthy people ($X$ group) and also $m$ future diseased people ($Y$ group).

Using threshold $c$:

$C_c^X(m)$: number of correct diagnoses for $m$ future healthy people

$C_c^Y(m)$: number of correct diagnoses for $m$ future diseased people

Durham University

One possibility is to consider the lower and upper expected values

$$\begin{aligned}
\underline{E}(C_c^X(1)) + \underline{E}(C_c^Y(1)) &= \frac{\sum_{i=1}^{n_x} 1[x_i \leq c]}{n_x+1} + \frac{\sum_{j=1}^{n_y} 1[y_j > c]}{n_y+1} \\
\overline{E}(C_c^X(1)) + \overline{E}(C_c^Y(1)) &= \frac{\sum_{i=1}^{n_x} 1[x_i \leq c]}{n_x+1} + \frac{\sum_{j=1}^{n_y} 1[y_j > c]}{n_y+1} \\
&\quad + \frac{1}{n_x+1} + \frac{1}{n_y+1}
\end{aligned}$$

Maximising these gives the same optimal threshold $c$

Using $m > 1$ for these criteria leads to exactly the same $c$

Durham University

140, <u>150</u>, $c(NPI)$, 180, 185, 188, 190, 203, 204, <u>205</u>, $c(YI)$, 230, 260, 280, 300, 305, 330, 344

Classification of the actual data:

Youden: both $X$ correct, 7 of the 14 $Y$ correct

NPI (Expectation): 1 of the 2 $X$ correct, 13 of the 14 $Y$ correct

*Note:* Most practical examples no difference, and always identical if $n_x = n_y$

But we can easily consider more exciting criteria, e.g. for $\alpha, \beta \in [0, 1]$ we can maximise

$$\underline{P}(C_c^X(m_x) \geq \alpha m_x, C_c^Y(m_y) \geq \beta m_y)$$

or

$$\overline{P}(C_c^X(m_x) \geq \alpha m_x, C_c^Y(m_y) \geq \beta m_y)$$

Use of $\alpha, \beta$ reflects importance to get diagnoses right for specific groups, so related to use of utilities (*possibly more intuitive*?)

Durham University

$X : n_x = 14$; $Y : n_y = 18$ (underlined)

120,$\underline{130}$,135,155,157,159,162,$\underline{166}$,168,$\underline{172}$,185,$\underline{187}$,
188,$\underline{189}$,191,$\underline{194}$,$\underline{199}$,200,$\underline{207}$,220,$\underline{227}$,230,$\underline{231}$,$\underline{240}$,
$\underline{242}$,$\underline{244}$,$\underline{250}$,$\underline{255}$,270,$\underline{277}$,$\underline{280}$,$\underline{282}$

Optimal values for $c$:

Empirical Youden's index gives $c \in (191, 194)$

$\alpha = 0.5, \beta = 0.6$ gives same interval for lower and upper probabilities for most values of $m$ considered, but for large $m$, 100 and 150, the lower probability gives the same but the upper probability gives interval $(188, 189)$.

Of course, $\alpha$ smaller and $\beta$ larger moves optimal $c$ to the left, and $\alpha$ larger and $\beta$ smaller moves it to the right.

Different values of $m$ can also have some (usually) minor effect on optimal interval, and lower and upper probabilities often lead to the same interval for $c$ but not always.

It may also be important to choose $m_x$ future people from $X$ and $m_y$ from $Y$ with $m_x \neq m_y$, straightforward to implement.

Durham University

# Regression

We consider the basic regression model

$$y_i = \alpha + \beta x_i + \epsilon_i$$

Assume that the $\epsilon_i$ are exchangeable, the $x_i$ are not random

Use the standard criterion to fit the line: minimum sum of squares of the residuals

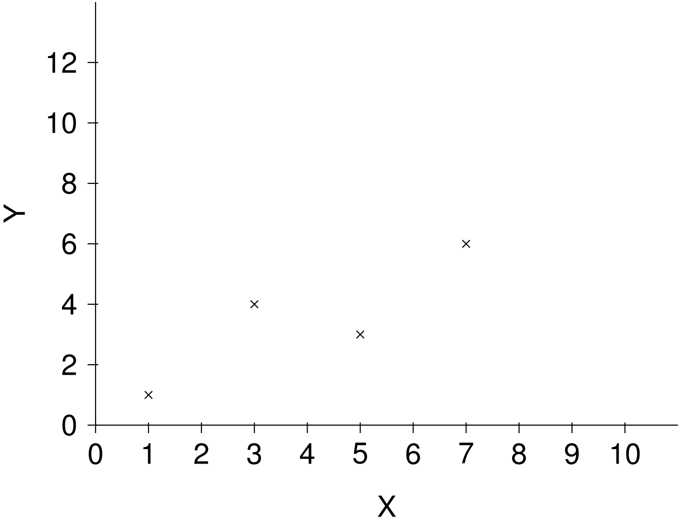
How can we use NPI?

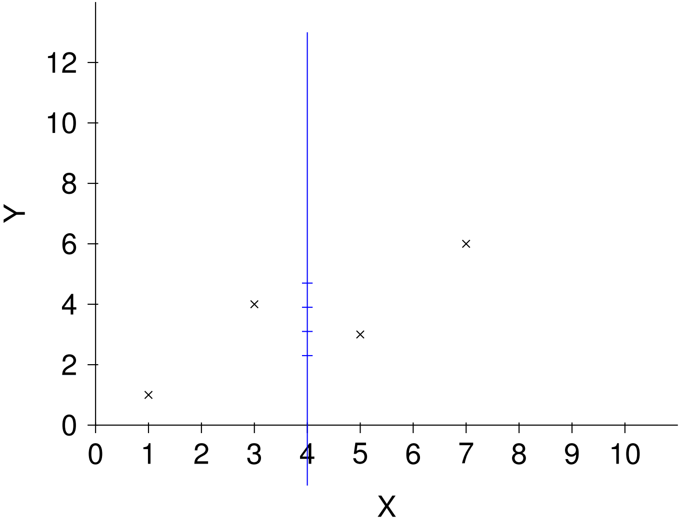
# Example

4 observations $(x_i, y_i)$: $(1, 1), (3, 4), (5, 3), (7, 6)$

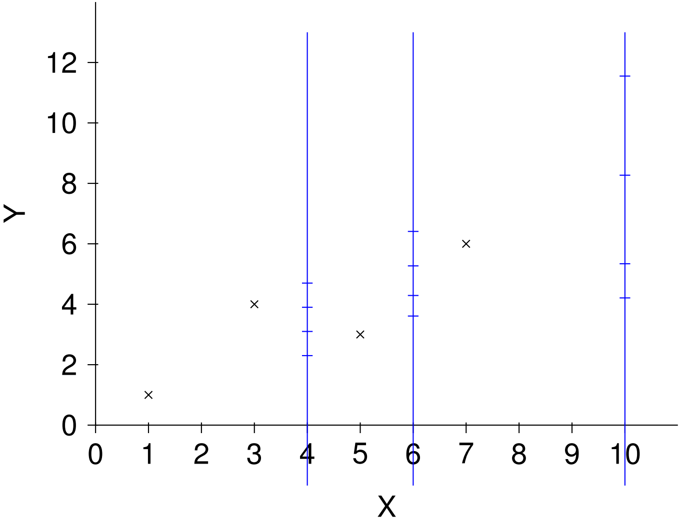
Goal: predict $y$-values corresponding to $x = 4$, $x = 6$ and $x = 10$, thereafter for all values of $x$

Durham University


Y
X
0
2
4
6
8
10
12
1
3
5
7
9



Y
X
0
2
4
6
8
10
12
0
1
2
3
4
5
6
7
8
9
10



Y
12
10
8
6
4
2
0
0 1 2 3 4 5 6 7 8 9 10
X

# Algorithm

1. Range for Prediction:

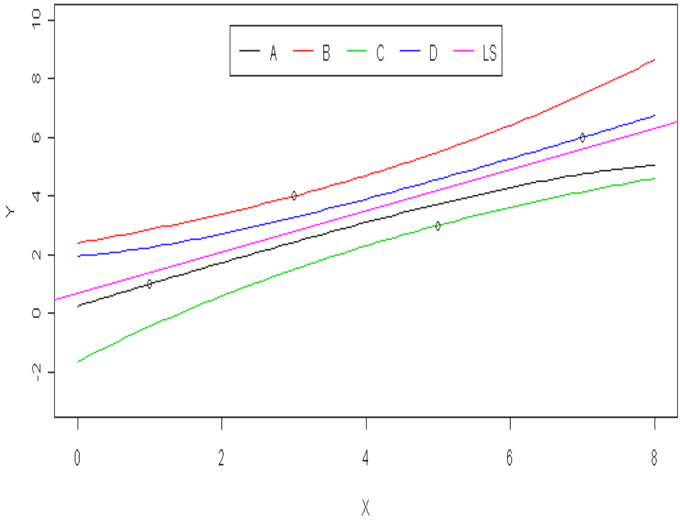
$$RP = \{x_{n+1} | (x_{n+1} - \bar{x})(\bar{x} - x_i) < \sum_{j=1}^{n} (x_j - \bar{x})^2\} \quad \text{for all} \quad i = 1, \ldots, n$$

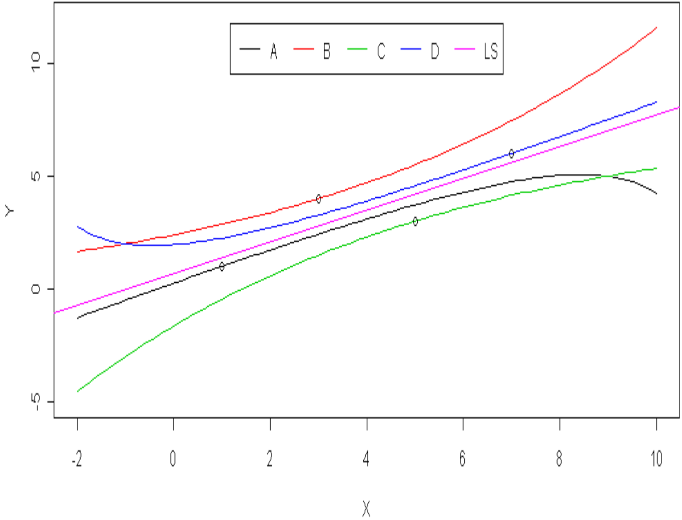
2. For $x_{n+1} \in RP$, calculate, for $i = 1, \ldots, n$,

$$\tilde{y}_i = \frac{\left[\sum_{j=1}^{n} (x_j - \bar{x}) y_j\right] (x_{n+1} - x_i) + y_i \sum_{j=1}^{n+1} (x_j - \bar{x})^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2 - (x_{n+1} - \bar{x})(\bar{x} - x_i)}$$

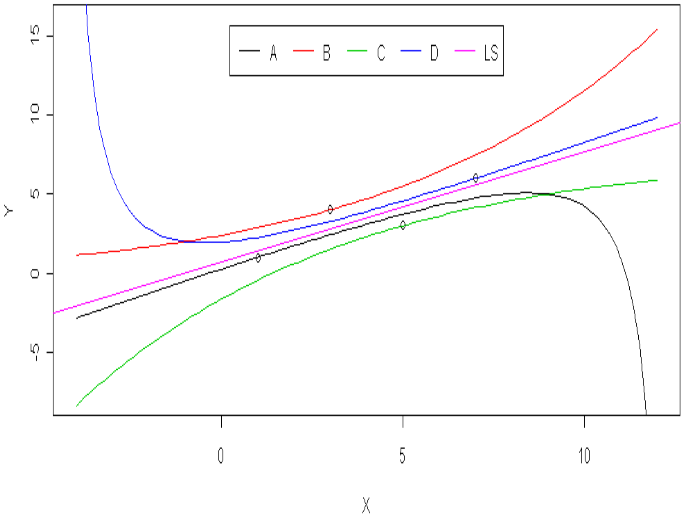
3. Ordered values $\tilde{y}_{(1)} \leq \tilde{y}_{(2)} \leq \ldots \leq \tilde{y}_{(n)}$.
4. NPI prediction for $Y_{n+1}$ corresponding to $x_{n+1} \in RP$ gives, for $j = 1, \ldots, n+1$ and with $\tilde{y}_{(0)} = -\infty$ and $\tilde{y}_{(n+1)} = \infty$,

$$P(Y_{n+1} \in (\tilde{y}_{(j-1)}, \tilde{y}_{(j)})) = \frac{1}{n+1}$$


A
B
C
D
LS
10
8
6
4
2
0
-2
Y
0
2
4
6
8
X




Durham University




Durham University

This method can be used for any parametric model of the form $y = g(x) + \epsilon$ (with real-valued $y$) with the $A_{(n)}$ assumption for the $\epsilon$'s, and with any loss function.

This method is closely related to conformal prediction!

# Some further results

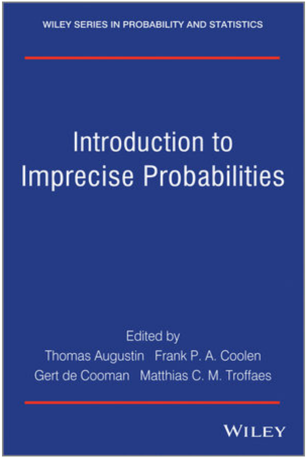
- NPI has been presented for other kinds of data, including Bernoulli, multinomial, and right-censored data
- A start has been made on research towards NPI for multivariate data
- NPI has been presented for many problems in Statistics, Reliability, Risk and OR
- NPI is never in disagreement with inferences based on empirical probabilities, so one could call NPI ‘objective’
- NPI has helped us to get better understanding of foundations of statistics with imprecise probabilities

Durham University

# Challenges

- Develop further methodology for data with covariates and multivariate data
- A wide range of topics (e.g. general censoring) for which we have a good idea how to do them but not enough time...
- Applications!

Durham University




*www.npi-statistics.com*

*maths.dur.ac.uk/stats/people/fc/fc.html*

frank.coolen@durham.ac.uk

Durham University