Calibration of predictions

Johanna Ziegel

ETH Zurich

CUSO Summer School
7–10 September 2025

Introduction: Probabilistic predictions

- Let $Y \in \mathcal{Y}$ be an unknown future outcome.
  - Temperature tomorrow at 12:00 in Cambridge. $(Y \in \mathcal{Y} = \mathbb{R})$
  - Event of rain tomorrow in London. $(Y \in \mathcal{Y} = \{0, 1\})$
  - Default of credit card client. $(Y \in \mathcal{Y} = \{0, 1\})$
  - Amount of precipitation tomorrow in Cambridge and Oxford. $(Y \in \mathcal{Y} = \mathbb{R}^2)$
- We are interested in predictions for Y.

Introduction: Probabilistic predictions

- Let $Y \in \mathcal{Y}$ be an unknown future outcome.
  - Temperature tomorrow at 12:00 in Cambridge. $(Y \in \mathcal{Y} = \mathbb{R})$
  - Event of rain tomorrow in London. $(Y \in \mathcal{Y} = \{0, 1\})$
  - Default of credit card client. $(Y \in \mathcal{Y} = \{0, 1\})$
  - Amount of precipitation tomorrow in Cambridge and Oxford. $(Y \in \mathcal{Y} = \mathbb{R}^2)$
- We are interested in predictions for Y.

Questions that I will address:

- What is a probabilistic prediction for Y? What is a point prediction for Y? Are there other predictions for Y?

Introduction: Probabilistic predictions

- Let $Y \in \mathcal{Y}$ be an unknown future outcome.
  - Temperature tomorrow at 12:00 in Cambridge. ($Y \in \mathcal{Y} = \mathbb{R}$)
  - Event of rain tomorrow in London. ($Y \in \mathcal{Y} = \{0, 1\}$)
  - Default of credit card client. ($Y \in \mathcal{Y} = \{0, 1\}$)
  - Amount of precipitation tomorrow in Cambridge and Oxford. ($Y \in \mathcal{Y} = \mathbb{R}^2$)
- We are interested in predictions for Y.

Questions that I will address:

- What is a probabilistic prediction for Y? What is a point prediction for Y? Are there other predictions for Y?
- When is a probabilistic prediction calibrated? (Lecture 1)
- How can calibrated probabilistic predictions be constructed? (Lecture 1)

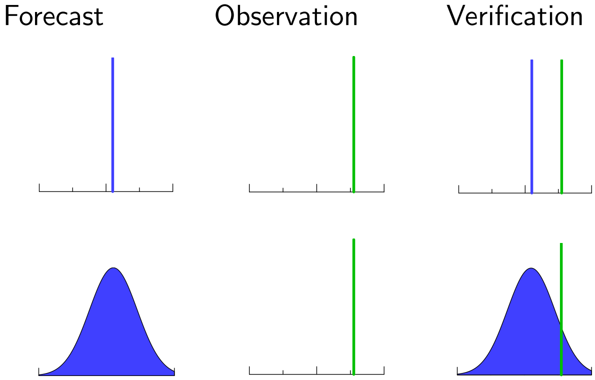
Introduction: Probabilistic predictions

- Let $Y \in \mathcal{Y}$ be an unknown future outcome.
  - Temperature tomorrow at 12:00 in Cambridge. ($Y \in \mathcal{Y} = \mathbb{R}$)
  - Event of rain tomorrow in London. ($Y \in \mathcal{Y} = \{0, 1\}$)
  - Default of credit card client. ($Y \in \mathcal{Y} = \{0, 1\}$)
  - Amount of precipitation tomorrow in Cambridge and Oxford. ($Y \in \mathcal{Y} = \mathbb{R}^2$)
- We are interested in predictions for Y.

Questions that I will address:

- What is a probabilistic prediction for Y? What is a point prediction for Y? Are there other predictions for Y?
- When is a probabilistic prediction calibrated? (Lecture 1)
- How can calibrated probabilistic predictions be constructed? (Lecture 1)
- How can we compare probabilistic predictions? With proper scoring rules. (Lecture 2)
- How should point predictions be evaluated and compared? (Lecture 3)

Probabilistic and point predictions



"Tomorrow at 12:00 temperature will be 17.5°C."

"Tomorrow at 12:00 temperature will be $\mathcal{N}(17.5, \sigma^2)$."

Today: Probabilistic predictions

- Single valued “best guess” $z \in \mathcal{Y}$ does not quantify uncertainty.

Today: Probabilistic predictions

- Single valued "best guess" $z \in \mathcal{Y}$ does not quantify uncertainty.
- Better: Quantify uncertainty of Y by a *probabilistic prediction* F.
  - F is a distribution on $\mathcal{Y}$.

Today: Probabilistic predictions

- Single valued "best guess" $z \in \mathcal{Y}$ does not quantify uncertainty.
- Better: Quantify uncertainty of Y by a *probabilistic prediction* F.
  - F is a distribution on $\mathcal{Y}$.
- If X is information available for prediction, F should approximate $\mathcal{L}(Y \mid X)$.

Today: Probabilistic predictions

- Single valued "best guess" $z \in \mathcal{Y}$ does not quantify uncertainty.
- Better: Quantify uncertainty of Y by a *probabilistic prediction* F.
  - F is a distribution on $\mathcal{Y}$.
- If X is information available for prediction, F should approximate $\mathcal{L}(Y \mid X)$.
- Other possibilities to quantify uncertainty of Y: prediction intervals, predictions of some measure of variability, ...
- Structurally, these are "point predictions" but they are often called probabilistic predictions since they quantify uncertainty to some degree.

Goal: Discuss quality criteria for probabilistic predictions

- What is calibration of probabilistic predictions?

Goal: Discuss quality criteria for probabilistic predictions

- What is calibration of probabilistic predictions?
- Are calibrated predictions good? When are predictions informative?
- How do we calibrate predictions?
- How do we compare predictions and how is this related to calibration?

Goal: Discuss quality criteria for probabilistic predictions

- What is calibration of probabilistic predictions?
- Are calibrated predictions good? When are predictions informative?
- How do we calibrate predictions?
- How do we compare predictions and how is this related to calibration?

- Forecasts are usually sequential but many concepts are easier to understand in a "hypothetical" one-period setting.
- Future outcome Y and forecast F are both random and defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Outline

Binary outcomes

- Calibration
- Discrimination ability
- Comparison

Binary outcomes: Calibration

- $Y \in \{0, 1\}$
  - Event of rain tomorrow in London. ($Y \in \{0, 1\}$)
  - Default of credit card client. ($Y \in \{0, 1\}$)

Binary outcomes: Calibration

- $Y \in \{0, 1\}$
  - Event of rain tomorrow in London. ($Y \in \{0, 1\}$)
  - Default of credit card client. ($Y \in \{0, 1\}$)
- Distribution of Y is characterised by probability of $\{Y = 1\}$: Probabilistic prediction is random variable $p \in [0, 1]$.

Binary outcomes: Calibration

- $Y \in \{0, 1\}$
  - Event of rain tomorrow in London. $(Y \in \{0, 1\})$
  - Default of credit card client. $(Y \in \{0, 1\})$
- Distribution of Y is characterised by probability of $\{Y = 1\}$: Probabilistic prediction is random variable $p \in [0, 1]$.
- Since $\mathbb{P}(Y = 1 \mid X) = \mathbb{E}(\mathbb{1}\{Y = 1\} \mid X)$,
  - p is a prediction for the conditional distribution of Y (probabilistic prediction);
  - p is a prediction for the conditional mean of Y (point prediction).

Binary outcomes: Calibration

- $Y \in \{0,1\}$
  - Event of rain tomorrow in London. ($Y \in \{0,1\}$)
  - Default of credit card client. ($Y \in \{0,1\}$)
- Distribution of Y is characterised by probability of $\{Y = 1\}$:
  Probabilistic prediction is random variable $p \in [0,1]$.
- Since $\mathbb{P}(Y = 1 \mid X) = \mathbb{E}(\mathbb{1}\{Y = 1\} \mid X)$,
  - p is a prediction for the conditional distribution of Y (probabilistic prediction);
  - p is a prediction for the conditional mean of Y (point prediction).

Definition

A probability prediction $p \in [0,1]$ for $Y \in \{0,1\}$ is *calibrated* (or *reliable*) if

$$\mathbb{P}(Y = 1 \mid p) = p.$$

Predicted probabilities should align with observed frequencies.

Example

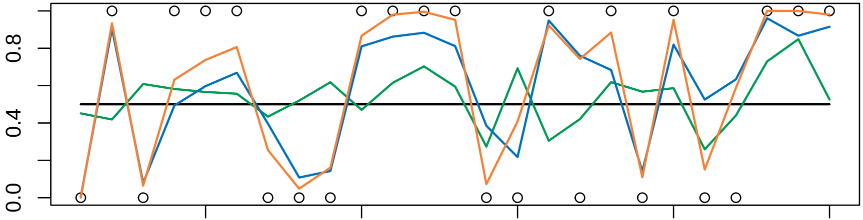
Let $X_1 \sim \mathcal{N}(0,1)$, $X_2 \sim \mathcal{N}(0,2)$ be independent, and

$$\mathbb{P}(Y = 1 \mid X_1, X_2) = \Phi(X_1 + X_2).$$

Predictions:

$$p^{(0)} = 1/2, \quad p^{(1)} = \Phi\left(\frac{X_1}{\sqrt{3}}\right), \quad p^{(2)} = \Phi\left(\frac{X_2}{\sqrt{2}}\right), \quad p^{(3)} = \Phi(X_1 + X_2).$$

- All predictions are calibrated.



Ranjan and Gneiting (2010)

Diagnostics to assess calibration: Reliability diagrams

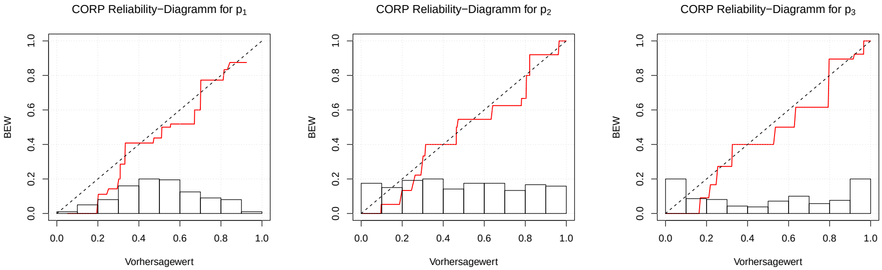
Data: $(p_1, Y_1), \ldots, (p_n, Y_n)$

Diagnostics to assess calibration: Reliability diagrams

Data: $(p_1, Y_1), \dots, (p_n, Y_n)$

Simulation example

$X_1 \sim \mathcal{N}(0,1)$, $X_2 \sim \mathcal{N}(0,2)$ independent, $\mathbb{P}(Y = 1 \mid X_1, X_2) = \Phi(X_1 + X_2)$,
$p^{(1)} = \Phi(X_1/\sqrt{3})$, $p^{(2)} = \Phi\left(X_2/\sqrt{2}\right)$, $p^{(3)} = \Phi(X_1 + X_2)$, $n = 200$.



(Dimitriadis et al., 2021)

Reliability diagrams

Forecasts and observations: $(p_1, Y_1), \ldots, (p_n, Y_n)$

Binning: Classical approach

Choose $m \in \mathbb{N}$, for example $m = 10$. Define

$$\hat{q}_j = \frac{\#\left\{i \mid Y_i = 1, \frac{j-1}{m} \le p_i \le \frac{j}{m}\right\}}{\#\left\{i \mid \frac{j-1}{m} \le p_i \le \frac{j}{m}\right\}}, \quad j = 1, \ldots, m,$$

$\hat{q}_j$ is an estimator of the conditional event probability (CEP)

$$\mathbb{P}\left(Y = 1 \mid \frac{j-1}{m} \le p \le \frac{j}{m}\right), \quad j = 1, \ldots, m.$$

For calibrated predictions, we have that

$$\frac{j-1}{m} \le \mathbb{P}\left(Y = 1 \mid \frac{j-1}{m} \le p \le \frac{j}{m}\right) \le \frac{j}{m}.$$

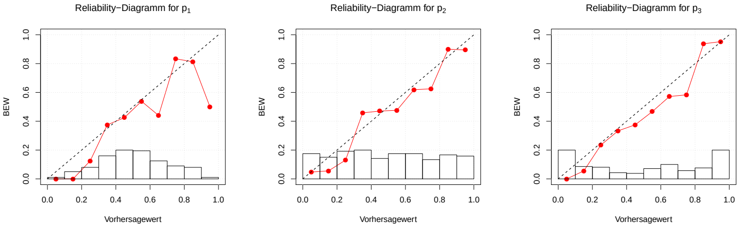
Reliability diagrams

Diagnostic tool to assess calibration

- Plot $((j-1)/2, \hat{q}_j)$, $j = 1, \ldots, m$ and joint points by a line.
- Add diagonal, that is, line from $(0,0)$ to $(1,1)$.
- Sometimes: Add histogram of $p_1, \ldots, p_n$ such that it fits.

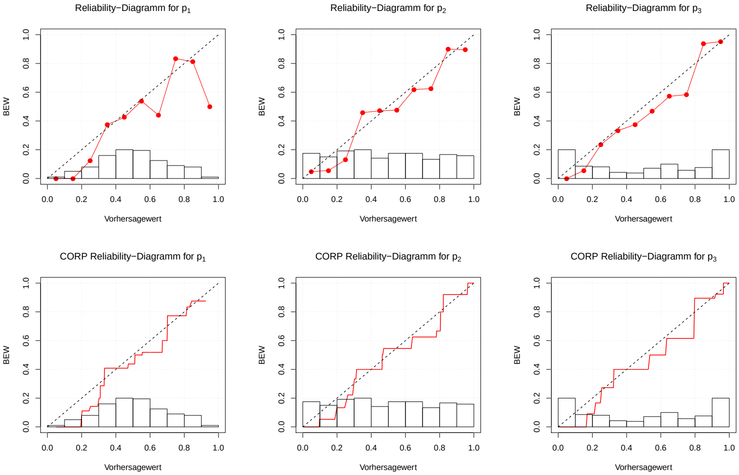
Simulation example

$X_1 \sim \mathcal{N}(0,1)$, $X_2 \sim \mathcal{N}(0,2)$ independent, $\mathbb{P}(Y=1 \mid X_1, X_2) = \Phi(X_1 + X_2)$,
$p^{(1)} = \Phi(X_1/\sqrt{3})$, $p^{(2)} = \Phi\left(X_2/\sqrt{2}\right)$, $p^{(3)} = \Phi(X_1 + X_2)$, $n = 200$.



Simulation example

$X_1 \sim \mathcal{N}(0,1)$, $X_2 \sim \mathcal{N}(0,2)$ independent, $\mathbb{P}(Y = 1 \mid X_1, X_2) = \Phi(X_1 + X_2)$,
$p^{(1)} = \Phi(X_1/\sqrt{3})$, $p^{(2)} = \Phi\left(X_2/\sqrt{2}\right)$, $p^{(3)} = \Phi(X_1 + X_2)$, $n = 200$.



Hosmer-Lemeshow Test

Goodness-of-fit test for binary regression models. Is deviation from diagonal in reliability diagram significant? (Hosmer and Lemeshow, 1980)

$$\mathcal{H}_0 = \{\mathbb{P} \mid \mathbb{P}(Y \mid p_i) = p_i, \ i = 1, \dots, n\}$$

Test statistic

$$T_{\mathrm{HL}} = \sum_{j=1}^{m} \left[\frac{(O_{1j} - E_{1j})^2}{E_{1j}} + \frac{(O_{0j} - E_{0j})^2}{E_{0j}} \right],$$

with

$$O_{1j} = \#\Big\{i \mid Y_i = 1\} \cap I_j, \quad O_{0j} = \#\Big\{i \mid Y_i = 0\} \cap I_j,$$

and

$$E_{1j} = \sum_{i \in I_j} p_i, \quad E_{0j} = \sum_{i \in I_j} (1 - p_i),$$

where $I_j = \{i \mid (j-1)/m \le p_i \le j/m\}$. Then, $T_{\mathrm{HL}} \sim \chi^2_{m-1}$ asymptotically.

Problems with Binning

- Visual impression of reliability diagram can look very different for different m, say, $m = 9, 10, 11$. Can be misleading as a diagnostic tool.
- Choice of the bins influences the test substantially: p-values of the Hosmer-Lemeshow test from 0.020 to 0.159 with six different statistical software packages (Hosmer et al., 1997).
- Reordering a data set with ties can yield p-values from 0.01 to 0.95 (Bertolini et al., 2000).

CORP Reliability Diagrams

Postulate that

$$p' \mapsto \mathbb{P}(Y = 1 \mid p = p')$$

is increasing.
For $(p_1, Y_1), \dots, (p_n, Y_n)$ with $p_1 \leq \cdots \leq p_n$ compute the isotonic regression $(q_1^{(iso)}, \dots, q_n^{(iso)})$ of Y given p. Then,

$$q_i^{(iso)} \approx \mathbb{P}(Y = 1 \mid p = p_i) \overset{\text{under calibration}}{=} p_i.$$

- Plot $(p_i, q_i^{(iso)})$, $i = 1, \dots, n$ and joint points by a line.
- Add diagonal.
- Sometimes: Add histogram of $p_1, \dots, p_n$ such that it fits.

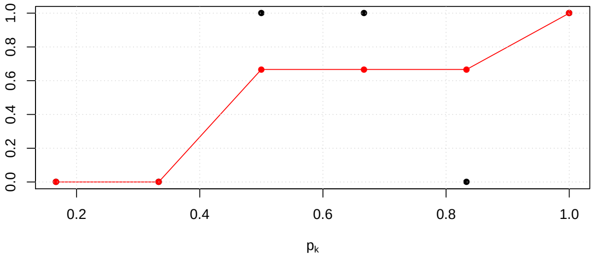
(Dimitriadis et al., 2021)

For $(p_1, Y_1), \dots, (p_n, Y_n)$ with $p_1 \leq \dots \leq p_n$, the isotonic regression $(q_1^{(iso)}, \dots, q_n^{(iso)})$ of Y given p is the solution to the optimization problem

$$\min_{\mathbf{q} \text{ isotone}} \sum_{i=1}^{n} (Y_i - q_i)^2,$$

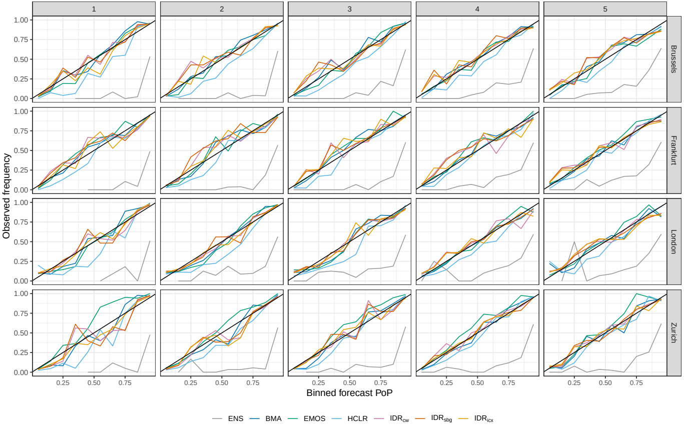
where the minimum is taken over all vectors $\mathbf{q} = (q_1, \dots, q_n) \in [0, 1]^T$ with $q_1 \leq \dots \leq q_n$.

- Solution can be computed with the Pool Adjacent Violators Algorithm (PAVA).



(van Eeden, 1958; Barlow et al., 1972)

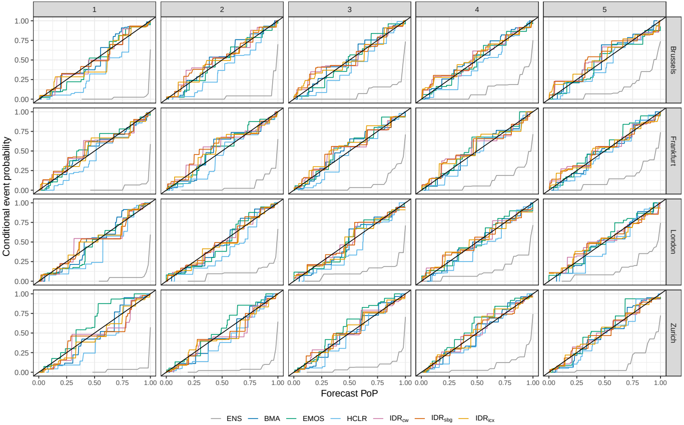
Application example



Reliability diagrams for probability of precipitation forecasts at prediction horizons of 1, 2, 3, 4 and 5 days, for the test period.

(Henzi et al., 2021b)

Application example



CORP reliability diagrams for probability of precipitation forecasts at prediction horizons of 1, 2, 3, 4 and 5 days, for the test period.

(Henzi et al., 2021b)

Summary

- Calibration assesses the statistical compatibility of forecasts and observation: Evaluation of *absolute* forecast quality.
- Calibrated predictions can be uninformative.
- How can discrimination ability of probably forecasts be assessed?

ROC curves assess discrimination ability

Use threshold $t \in [0, 1)$ to construct a hard classifier $\mathbb{1}\{p > t\}$ from the probability prediction. Define the Hit Rate (HR)

$$\mathrm{HR}(t) = \mathbb{P}(p > t \mid Y = 1),$$

and the False Alarm Rate (FAR)

$$\mathrm{FAR}(t) = \mathbb{P}(p > t \mid Y = 0).$$

- Receiver operating characteristic (ROC) curve consists of the points $(\mathrm{FAR}(t), \mathrm{HR}(t))$, $t \in [0, 1)$.

ROC curves assess discrimination ability

Use threshold $t \in [0, 1)$ to construct a hard classifier $\mathbb{1}\{p > t\}$ from the probability prediction. Define the Hit Rate (HR)

$$\mathrm{HR}(t) = \mathbb{P}(p > t \mid Y = 1),$$

and the False Alarm Rate (FAR)

$$\mathrm{FAR}(t) = \mathbb{P}(p > t \mid Y = 0).$$

- Receiver operating characteristic (ROC) curve consists of the points $(\mathrm{FAR}(t), \mathrm{HR}(t))$, $t \in [0, 1)$.

Empirical ROC curve

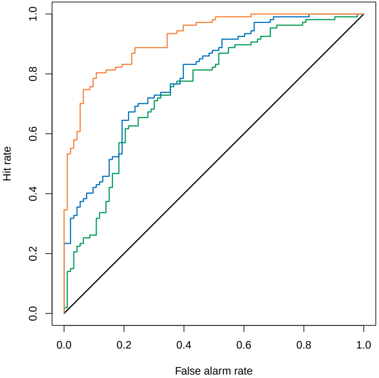
Estimate $\mathrm{HR}(t)$, $\mathrm{FAR}(t)$ by

$$\widehat{\mathrm{HR}}(t) = \frac{\sum_{i=1}^n \mathbb{1}\{Y_i = 1, p_i > t\}}{\sum_{i=1}^n \mathbb{1}\{Y_i = 1\}}, \quad \widehat{\mathrm{FAR}}(t) = \frac{\sum_{i=1}^n \mathbb{1}\{Y_i = 0, p_i > t\}}{\sum_{i=1}^n \mathbb{1}\{Y_i = 0\}},$$

respectively.

Simulation example

$X_1 \sim \mathcal{N}(0,1)$, $X_2 \sim \mathcal{N}(0,2)$ independent, $\mathbb{P}(Y=1 \mid X_1, X_2) = \Phi(X_1 + X_2)$, $p^{(0)} = 1/2$, $p^{(1)} = \Phi(X_1/\sqrt{3})$, $p^{(2)} = \Phi\left(X_2/\sqrt{2}\right)$, $p^{(3)} = \Phi(X_1 + X_2)$, $n = 200$.



- If $\mathrm{FAR}(t) = \mathrm{HR}(t)$, $t \in (0,1)$: No discrimination, ROC is on the diagonal.
- If $\exists t$ such that $\mathrm{FAR}(t) = 0, \mathrm{HR}(t) = 1$: Perfect discrimination, ROC is in upper left corner.

Properties of ROC curves

ROC curve consists of the points $(\mathrm{FAR}(t), \mathrm{HR}(t))$, $t \in [0, 1)$, where

$$\mathrm{HR}(t) = \mathbb{P}(p > t \mid Y = 1), \quad \mathrm{FAR}(t) = \mathbb{P}(p > t \mid Y = 0).$$

- ROC curve is invariant under strictly increasing transformations of forecast p: ROC ignores calibration.
- ROC discrimination ability or potential predictive ability, only!

Properties of ROC curves

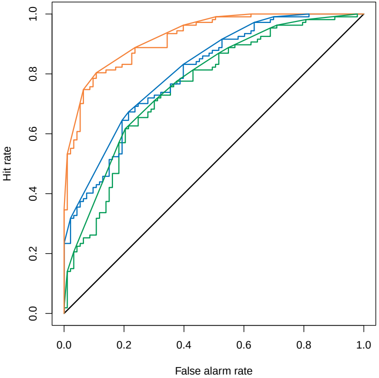
ROC curve consists of the points $(\mathrm{FAR}(t), \mathrm{HR}(t))$, $t \in [0,1)$, where

$$\mathrm{HR}(t) = \mathbb{P}(p > t \mid Y = 1), \quad \mathrm{FAR}(t) = \mathbb{P}(p > t \mid Y = 0).$$

- ROC curve is invariant under strictly increasing transformations of forecast p: ROC ignores calibration.
- ROC discrimination ability or potential predictive ability, only!
- ROC curve is concave if and only if CEP $p' \mapsto \mathbb{P}(Y = 1 \mid p = p')$ is increasing.
- Hard classifier construction only makes sense if CEP is increasing! Therefore, empirical ROC should be concave.
- Solution: Compute ROC curve from isotonically recalibrated predictions $q_i^{(iso)}$, $i = 1, \dots, n$.

Simulation example

$X_1 \sim \mathcal{N}(0,1)$, $X_2 \sim \mathcal{N}(0,2)$ independent, $\mathbb{P}(Y=1 \mid X_1, X_2) = \Phi(X_1 + X_2)$, $p^{(0)} = 1/2$, $p^{(1)} = \Phi(X_1/\sqrt{3})$, $p^{(2)} = \Phi\left(X_2/\sqrt{2}\right)$, $p^{(3)} = \Phi(X_1 + X_2)$, $n = 200$.



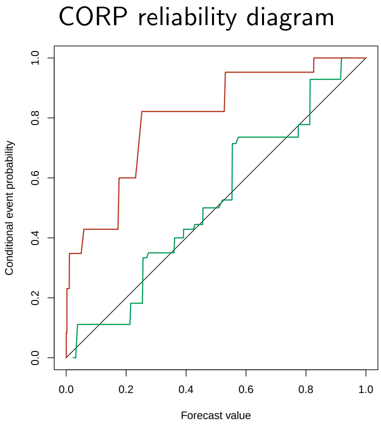
- Predictions can be differentiated with respect to discrimination ability.

(Gneiting and Vogel, 2022)

Simulation example

$X_1 \sim \mathcal{N}(0,1)$, $X_2 \sim \mathcal{N}(0,2)$ independent, $\mathbb{P}(Y = 1 \mid X_1, X_2) = \Phi(X_1 + X_2)$,
$p^{(1)} = \Phi(X_1/\sqrt{3})$, $p^{(4)} = (\Phi(X_1 + X_2))^3$, $n = 200$.

Calibration

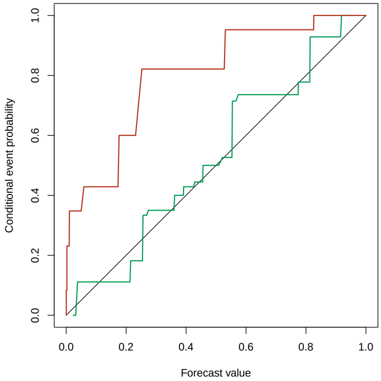


Simulation example

$X_1 \sim \mathcal{N}(0,1)$, $X_2 \sim \mathcal{N}(0,2)$ independent, $\mathbb{P}(Y=1 \mid X_1, X_2) = \Phi(X_1 + X_2)$,
$p^{(1)} = \Phi(X_1/\sqrt{3})$, $p^{(4)} = (\Phi(X_1 + X_2))^3$, $n = 200$.

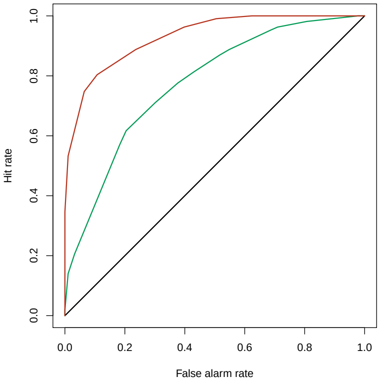
Calibration

CORP reliability diagram



Discrimination ability

ROC curve



(Dimitriadis et al., 2024)

Comparing predictions with proper scoring rules

- Need a summary measure that takes calibration and discrimination ability into account: Proper scoring rules.

Definition

A proper scoring rule is a map $S : [0,1] \times \{0,1\} \to \overline{\mathbb{R}}$ such that

$$\mathbb{E}S(p, Y) \leq \mathbb{E}S(q, Y)$$

for all $p, q \in [0,1]$ and $\mathbb{P}(Y = 1) = p$.

Give preference to forecaster with lower average score

$$\frac{1}{n}\sum_{i=1}^{n} S(p_i, Y_i).$$

Examples:

$$S(p, Y) = (p - Y)^2, \quad S(p, Y) = -Y \log p - (1 - Y)\log(1 - p).$$

Calibration

Real-valued outcomes
Multi-variate outcomes
Calibration and averaging

Real-valued outcomes

- $Y \in \mathbb{R}$.
  - Temperature tomorrow at 12:00 in Cambridge. ($Y \in \mathbb{R}$)
- Quantify uncertainty of Y by a probabilistic prediction F.
  - F is a distribution on $\mathbb{R}$ (typically specified as a CDF).

Real-valued outcomes

- $Y \in \mathbb{R}$.
  - Temperature tomorrow at 12:00 in Cambridge. ($Y \in \mathbb{R}$)
- Quantify uncertainty of Y by a probabilistic prediction F.
  - F is a distribution on $\mathbb{R}$ (typically specified as a CDF).
- If X is information available for prediction, F should approximate $\mathcal{L}(Y \mid X)$.

Real-valued outcomes

- $Y \in \mathbb{R}$.
  - Temperature tomorrow at 12:00 in Cambridge. ($Y \in \mathbb{R}$)
- Quantify uncertainty of Y by a probabilistic prediction F.
  - F is a distribution on $\mathbb{R}$ (typically specified as a CDF).
- If X is information available for prediction, F should approximate $\mathcal{L}(Y \mid X)$.

Illustration: Point and probabilistic predictions

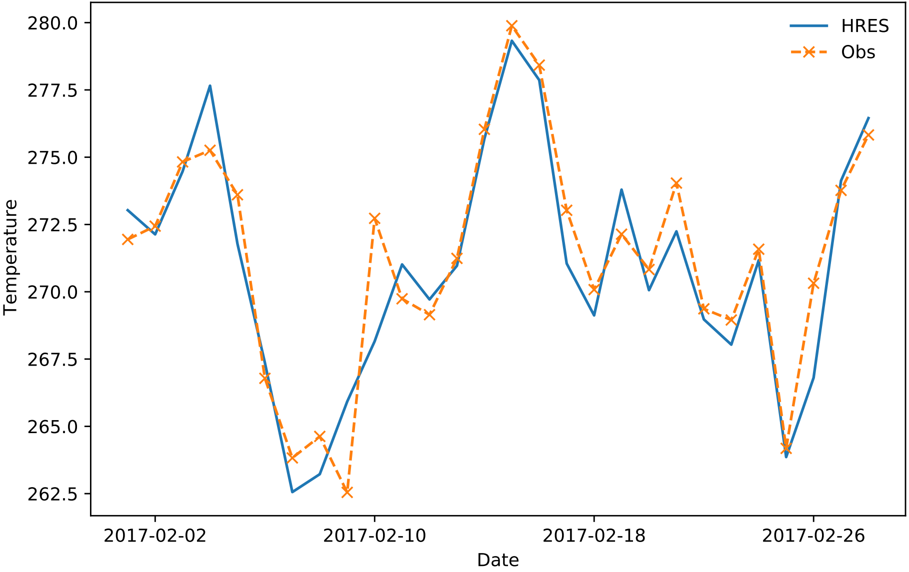
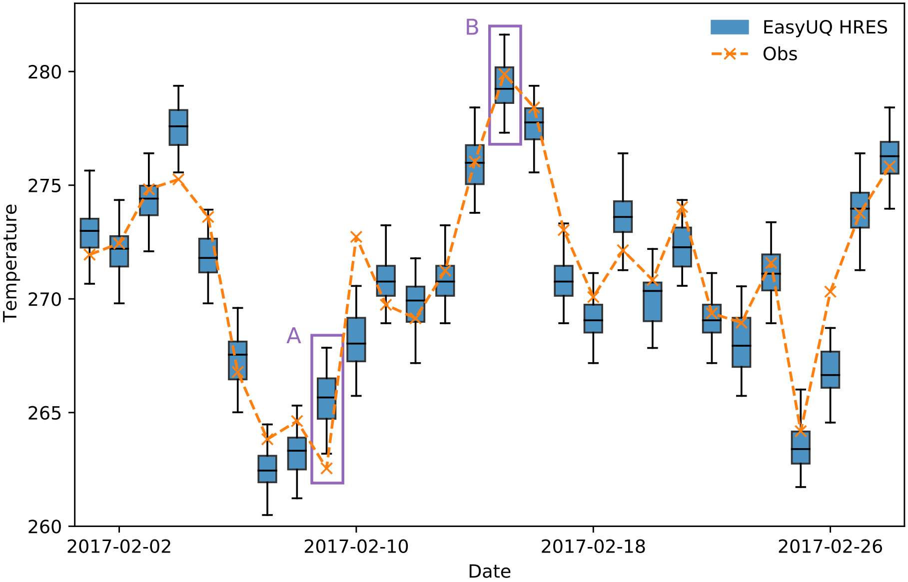
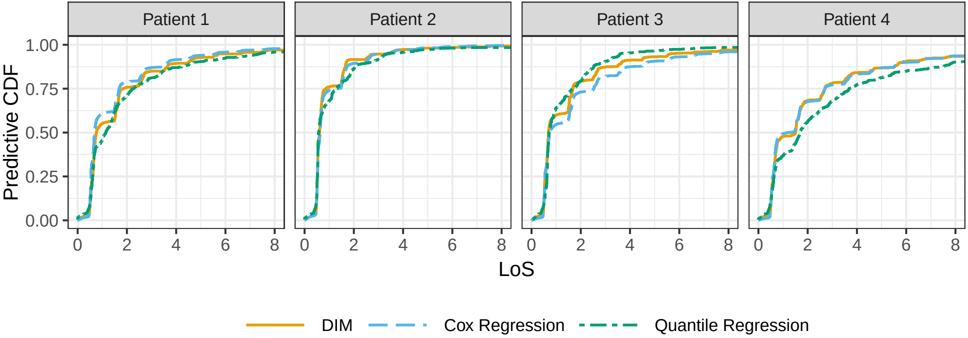


Illustration: Point and probabilistic predictions



Application example 1

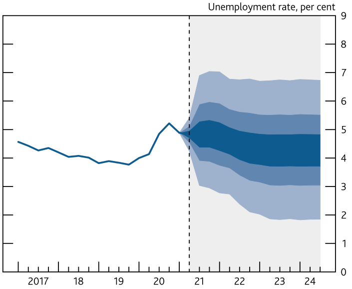
Patient length of stay in Swiss intensive care units



(Henzi et al., 2021a)

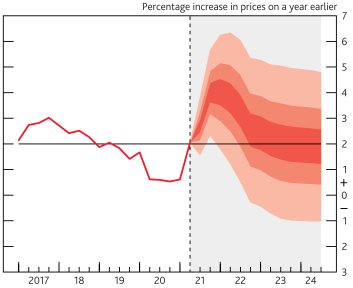
Application example 2

Chart 1.3: Unemployment projection based on market interest rate expectations, other policy measures as announced



The fan chart depicts the probability of various outcomes for LFS unemployment. It has been conditioned on the assumptions in **Table 1.A** footnote (b). The coloured bands have the same interpretation as in **Charts 1.1** and **1.2**, and portray 90% of the probability distribution. The calibration of this fan chart takes account of the likely path dependency of the economy, where, for example, it is judged that shocks to unemployment in one quarter will continue to have some effect on unemployment in successive quarters. The fan begins in 2021 Q2, a quarter earlier than for CPI inflation. That is because Q2 is a staff projection for the unemployment rate, based in part on data for April and May. The unemployment rate was 4.8% in the three months to May, and is projected to be 4.8% in Q2 as a whole. A significant proportion of this distribution lies below Bank staff's current estimate of the long-term equilibrium unemployment rate. There is therefore uncertainty about the precise calibration of this fan chart.

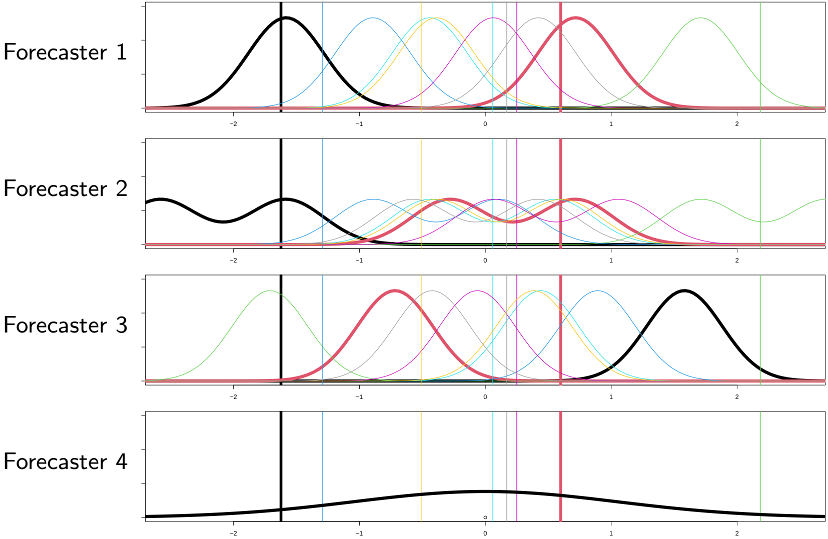
Chart 1.4: CPI inflation projection based on market interest rate expectations, other policy measures as announced



The fan chart depicts the probability of various outcomes for CPI inflation in the future. It has been conditioned on the assumptions in **Table 1.A** footnote (b). If economic circumstances identical to today's were to prevail on 100 occasions, the MPC's best collective judgement is that inflation in any particular quarter would lie within the darkest central band on only 30 of those occasions. The fan chart is constructed so that outturns of inflation are also expected to lie within each pair of the lighter red areas on 30 occasions. In any particular quarter of the forecast period, inflation is therefore expected to lie somewhere within the fans on 90 out of 100 occasions. And on the remaining 10 out of 100 occasions inflation can fall anywhere outside the red area of the fan chart. Over the forecast period, this has been depicted by the light grey background. See the box on pages 48–49 of the May 2002 *Inflation Report* for a fuller description of the fan chart and what it represents.

(Bank of England, 2021)

Calibration of probabilistic predictions



Calibration: Compatibility between forecasts and observations

Probabilities derived from predictive distributions should align with observed frequencies.

Calibration: Compatibility between forecasts and observations

Probabilities derived from predictive distributions should align with observed frequencies.

Most popular: Probabilistic calibration/“Flat PIT histogram”

$$F_i(Y_i) \sim \mathrm{UNIF}(0,1) \quad \text{for all } i$$

- $Y_i \in \mathbb{R}$, F_i predictive CDF for Y_i
- Suitable randomization if F_i is not continuous
- Closely related to validity of conformal predictive systems

Calibration: Compatibility between forecasts and observations

Probabilities derived from predictive distributions should align with observed frequencies.

Most popular: Probabilistic calibration/"Flat PIT histogram"

$$F_i(Y_i) \sim \mathrm{UNIF}(0,1) \quad \text{for all } i$$

- $Y_i \in \mathbb{R}$, F_i predictive CDF for Y_i
- Suitable randomization if F_i is not continuous
- Closely related to validity of conformal predictive systems
- **Binary outcomes**: $Y_i \in \{0,1\} : \mathbb{P}(Y_i = 1 | p_i) = p_i$
- *Many* notions of calibration, except for binary outcomes...

Why PIT values?

Motivation 1

Let X be a random variable and G a (deterministic) continuous CDF. Then,

$$G(X) \sim \mathrm{UNIF}(0,1) \quad \iff \quad X \sim G.$$

Without continuity:

$$\mathbb{P}(G(X) \leq \alpha) \leq \alpha \leq \mathbb{P}(G(X-) \leq \alpha) \quad \forall \alpha \in (0,1) \quad \iff \quad X \sim G.$$

Proposition

Let X be a random variable and G a CDF. Then,

$$\mathbb{P}(G(X) \leq \alpha) \leq \alpha \leq \mathbb{P}(G(X-) \leq \alpha) \quad \forall \alpha \in (0,1) \quad \iff \quad X \sim G.$$

Proof.

Let $Y \sim H$. If $G(z) < H(z)$ for some z, then, for $\alpha \in (G(z), H(z))$,

$$\alpha \geq \mathbb{P}(G(Y) \leq \alpha) \geq \mathbb{P}(G(Y) \leq G(z)) \geq \mathbb{P}(Y \leq z) = H(z) > \alpha \quad \lightning.$$

Proposition

Let X be a random variable and G a CDF. Then,

$$\mathbb{P}(G(X) \leq \alpha) \leq \alpha \leq \mathbb{P}(G(X-) \leq \alpha) \quad \forall \alpha \in (0,1) \quad \iff \quad X \sim G.$$

Proof.

Let $Y \sim H$. If $G(z) < H(z)$ for some z, then, for $\alpha \in (G(z), H(z))$,

$$\alpha \geq \mathbb{P}(G(Y) \leq \alpha) \geq \mathbb{P}(G(Y) \leq G(z)) \geq \mathbb{P}(Y \leq z) = H(z) > \alpha \quad ↯.$$

If $G(z) > H(z)$, then, for $\alpha \in (H(z), G(z))$,

$$\alpha \leq \mathbb{P}(G(Y-) \leq \alpha) \leq \mathbb{P}(G(Y-) < G(z)) \leq \mathbb{P}(Y \leq z) = H(z) < \alpha \quad ↯.$$

Proposition

Let X be a random variable and G a CDF. Then,

$$\mathbb{P}(G(X) \leq \alpha) \leq \alpha \leq \mathbb{P}(G(X-) \leq \alpha) \quad \forall \alpha \in (0,1) \quad \iff \quad X \sim G.$$

Proof.

Let $Y \sim H$. If $G(z) < H(z)$ for some z, then, for $\alpha \in (G(z), H(z))$,

$$\alpha \geq \mathbb{P}(G(Y) \leq \alpha) \geq \mathbb{P}(G(Y) \leq G(z)) \geq \mathbb{P}(Y \leq z) = H(z) > \alpha \quad ↯.$$

If $G(z) > H(z)$, then, for $\alpha \in (H(z), G(z))$,

$$\alpha \leq \mathbb{P}(G(Y-) \leq \alpha) \leq \mathbb{P}(G(Y-) < G(z)) \leq \mathbb{P}(Y \leq z) = H(z) < \alpha \quad ↯.$$

For reverse: Recall that $G(G^{-1}(u)) \geq u$ and $G(G^{-1}(u)-) \leq u$. Let $U \sim \mathrm{UNIF}(0,1)$, then, $G^{-1}(U) \sim G$. Therefore,

$$\mathbb{P}(G(Y) \leq \alpha) = \mathbb{P}(G(G^{-1}(U)) \leq \alpha) \leq \mathbb{P}(U \leq \alpha) = \alpha,$$
$$\mathbb{P}(G(Y-) \leq \alpha) = \mathbb{P}(G(G^{-1}(U)-) \leq \alpha) \geq \mathbb{P}(U \leq \alpha) = \alpha. \qquad \square$$

Motivation 2

Time series $(Y_t)_{t\in\mathbb{N}}$. Then,

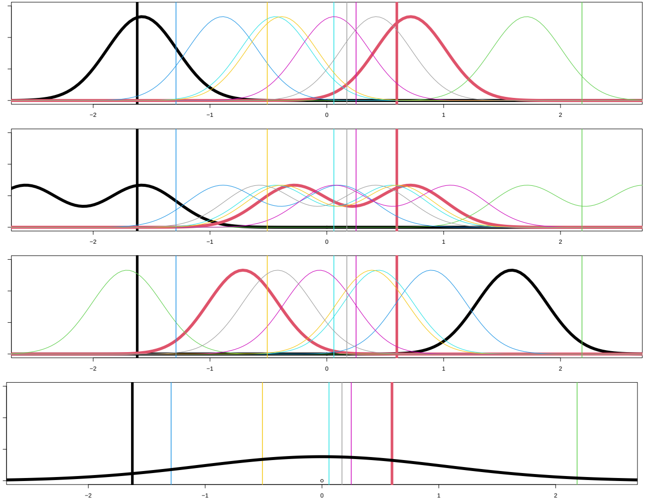
$$(F_t(Y_t))_{t\in\mathbb{N}} \overset{\text{iid}}{\sim} \text{UNIF}(0,1) \quad \Longleftrightarrow \quad F_t = \mathcal{L}(Y_t \mid Y_1, \dots, Y_{t-1}),\ t \in \mathbb{N}.$$

- Ideal prediction is equivalent to independence and uniformity of PIT values.
- Restricted to very special information set and lag 1 predictions.

(Diebold et al., 1998)

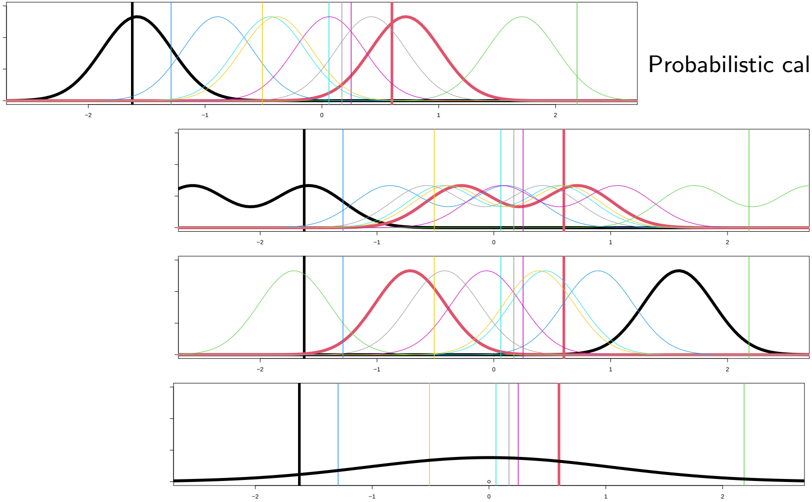
Evaluating probabilistic predictions

$$\mu \sim \mathcal{N}(0,1), \quad Y \sim \mathcal{N}(\mu, 0.09)$$



Evaluating probabilistic predictions

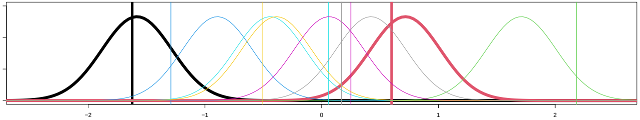
$$\mu \sim \mathcal{N}(0,1), \quad Y \sim \mathcal{N}(\mu, 0.09)$$



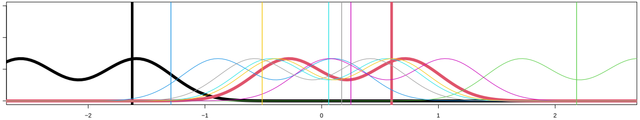
Probabilistic calibration ✓

Evaluating probabilistic predictions

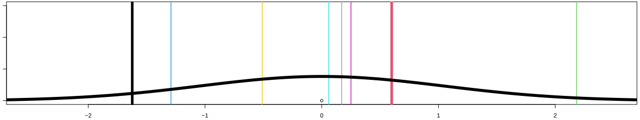
$$\mu \sim \mathcal{N}(0,1), \quad Y \sim \mathcal{N}(\mu, 0.09)$$



Probabilistic calibration ✓

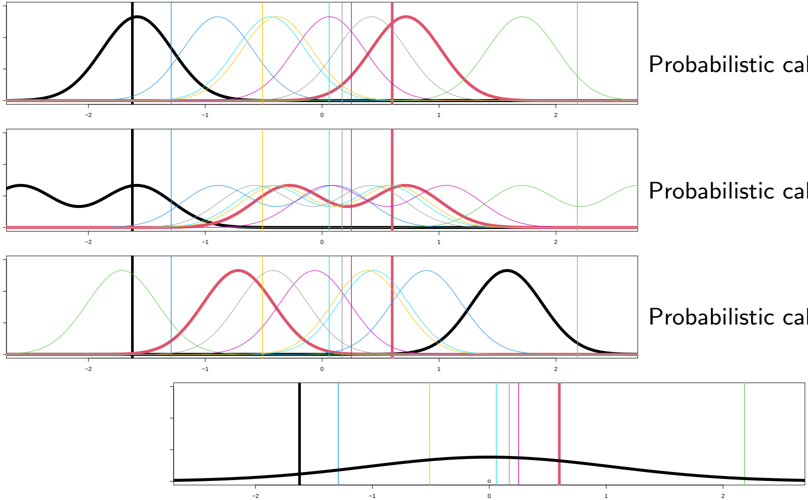


Probabilistic calibration ✓



Evaluating probabilistic predictions

$$\mu \sim \mathcal{N}(0,1), \quad Y \sim \mathcal{N}(\mu, 0.09)$$



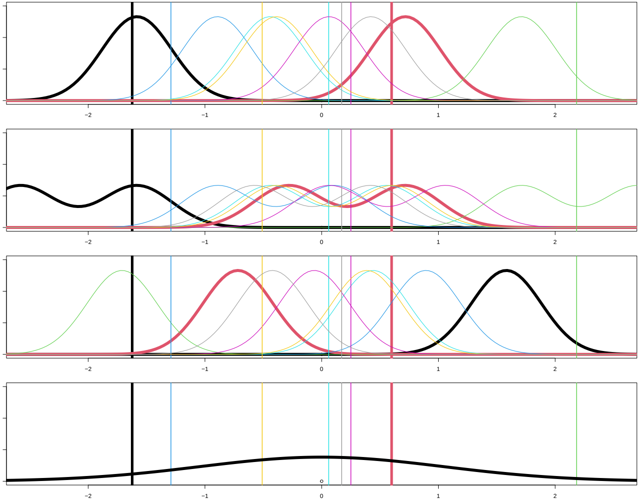
Probabilistic calibration ✓

Probabilistic calibration ✓

Probabilistic calibration ×

Evaluating probabilistic predictions

$$\mu \sim \mathcal{N}(0,1), \quad Y \sim \mathcal{N}(\mu, 0.09)$$



Probabilistic calibration ✓

Probabilistic calibration ✓

Probabilistic calibration ×

Probabilistic calibration ✓

Many notions of calibration . . .

Auto-calibration:
$\mathbb{P}(Y_i > y \mid F_i) = 1 - F_i(y)\ \forall y$
$\mathcal{L}(Y_i \mid F_i) = F_i$

$\Downarrow$

Isotonic calibration:
$\mathbb{P}(Y_i > y \mid \mathcal{A}(F_i)) = 1 - F_i(y)\ \forall y$
$\mathcal{L}(Y_i \mid \mathcal{A}(F_i)) = F_i$

$\swarrow \quad \searrow$

Threshold calibration:
$\mathbb{P}(Y_i > y \mid F_i(y)) = 1 - F_i(y)\ \forall y$

$\Downarrow$

Marginal calibration:
$\mathbb{P}(Y_i > y) = 1 - \mathbb{E}F_i(y)\ \forall y$

Quantile calibration:
$q_\alpha(Y_i \mid F_i^{-1}(\alpha)) = F_i^{-1}(\alpha)\ \forall \alpha$

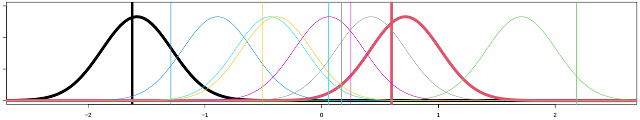
$\Downarrow$

Probabilistic calibration:
$F_i(Y_i) \sim \mathrm{UNIF}(0,1)$
$\mathbb{P}(F_i(Y_i) < \alpha) \le \alpha \le \mathbb{P}(F_i(Y_i-) \le \alpha)\ \forall \alpha$

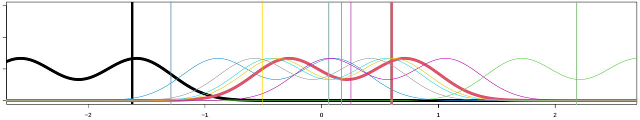
And if we want to focus on tails of F_i... (Allen et al., 2025b)

Evaluating probabilistic predictions

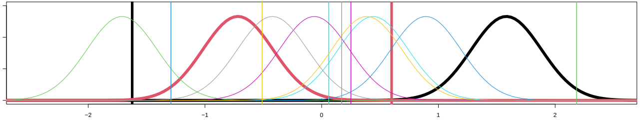
$$\mu \sim \mathcal{N}(0,1), \quad Y \sim \mathcal{N}(\mu, 0.09)$$



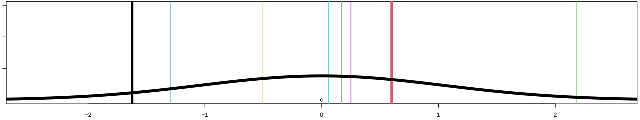
Auto-calibration ✓
Probabilistic calibration ✓
Marginal calibration ✓



Auto-calibration ×
Probabilistic calibration ✓
Marginal calibration ×



Auto-calibration ×
Probabilistic calibration ×
Marginal calibration ✓



Auto-calibration ✓
Probabilistic calibration ✓
Marginal calibration ✓

Auto-calibration

$$\mathbb{P}(Y \leq y \mid F) = F(y) \; \forall y$$
$$\mathcal{L}(Y \mid F) = F$$

Proposition

The forecast F is auto-calibrated for $Y \in \mathbb{R}$ if and only if $Z_F \sim \mathit{UNIF}(0,1)$ and $Z_F \perp\!\!\!\perp F$.

Here,

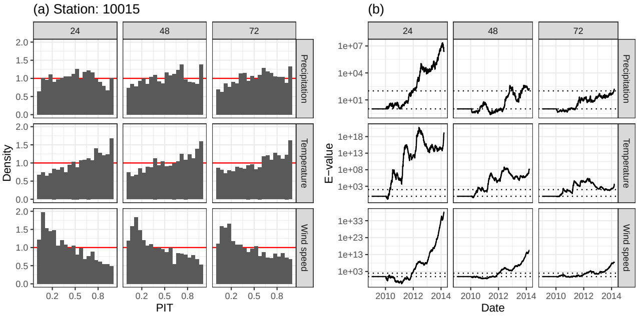
$$Z_F = F(Y-) + V(F(Y) - F(Y-)),$$

where $V \sim \mathrm{UNIF}(0,1)$ is independent of (Y, F).

(Strähl and Ziegel, 2017; Modeste, 2023)

Empirical assessment of calibration

Probabilistic calibration: Typically PIT histograms



Null hypothesis $F_i(Y_i) \sim \mathrm{UNIF}(0,1)$ for all t.

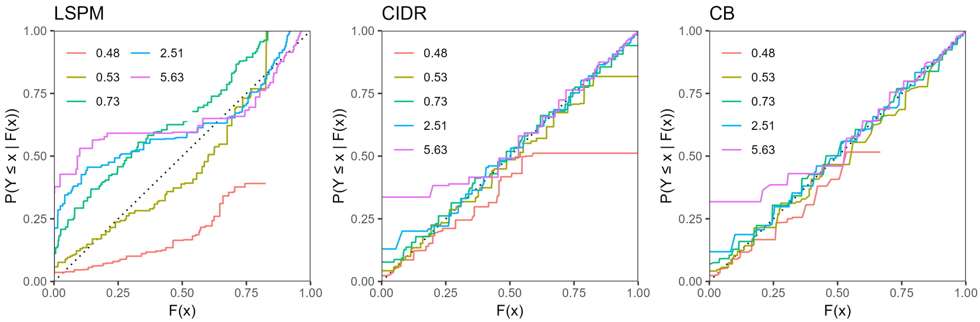
- At n: Estimate density of $(F_i(Y_i))_{i=1}^n$ by $\hat{f}_n$
- Define conditional e-value $E_{n+1} = \hat{f}_T(Y_{n+1})$
- Monitor calibration with test martingale $M_n = \prod_{i=1}^n E_i$

(Arnold et al., 2023)

Empirical assessment of calibration

Threshold calibration

- For each $z \in \mathbb{R}$, $F(z)$ is probability prediction for $\mathbb{1}\{Y \leq z\} \in \{0, 1\}$.
- Plot reliability diagrams for several thresholds.



Multi-variate outcomes

Let $\mathcal{Y} = \mathbb{R}^d$, forecast F is distribution over $\mathbb{R}^d$.

- Auto-calibration is still an applicable theoretical concept: $\mathcal{L}(Y \mid F) = F$.
- More popular: Graphical tools similar to PIT histograms
- Mainly used: Multivariate rank histograms assessing calibration of multivariate predictive distributions with finite support

Main idea

Reduce outcome Y and predictive distribution F to something univariate using map $\rho : \mathbb{R}^d \to \mathbb{R}$. Assess calibration of univariate summary.

(Gneiting et al., 2008; Ziegel and Gneiting, 2014; Thorarinsdottir et al., 2016; Allen et al., 2024)

Calibration and averaging: A word of warning

If F and G are calibrated for Y, then $(F+G)/2$ will not be calibrated for Y.

Example

Suppose that $F(Y) \sim \mathrm{UNIF}(0,1)$, $G(Y) \sim \mathrm{UNIF}(0,1)$, and $w_1 + w_2 = 1$. Then, $(w_1 F + w_2 G)(Y)$ has variance $< \frac{1}{12}$, so it cannot have distribution $\mathrm{UNIF}(0,1)$.

(Gneiting and Ranjan, 2013)

Summary

- Probabilistic predictions should be calibrated and sharp/discriminative in order to be trustworthy and informative.
- Ideally, probabilistic predictions should be auto-calibrated.
- Comparison of probabilistic predictions with proper scoring rules: Assign a real-valued score assessing calibration and discrimination ability simultaneously.
- Proper scoring rules allow to compare probabilistic predictions simultaneously with respect to calibration and discrimination ability. (Lecture 2)

Calibration and conformal prediction

Goal of conformal prediction

For data $(X_1, Y_1), \ldots, (X_n, Y_n)$ and X_{n+1}, construct

- prediction set C_{n+1} for Y_{n+1},
- predictive distribution F_{n+1} for Y_{n+1}

Goal of conformal prediction

For data $(X_1, Y_1), \ldots, (X_n, Y_n)$ and X_{n+1}, construct

- prediction set C_{n+1} for Y_{n+1},
- predictive distribution F_{n+1} for Y_{n+1}

such that

- $\mathbb{P}(Y_{n+1} \in C_{n+1}) \geq 1 - \alpha$,
- $\mathbb{P}(F_{n+1}(Y_{n+1})) \approx \mathrm{UNIF}(0, 1)$.

Goal of conformal prediction

For data $(X_1, Y_1), \ldots, (X_n, Y_n)$ and X_{n+1}, construct

- prediction set C_{n+1} for Y_{n+1},
- predictive distribution F_{n+1} for Y_{n+1}

such that

- $\mathbb{P}(Y_{n+1} \in C_{n+1}) \geq 1 - \alpha$,
- $\mathbb{P}(F_{n+1}(Y_{n+1})) \approx \mathrm{UNIF}(0, 1)$.

- Finite sample (marginal) coverage/calibration guarantees.
- Many and rapidly evolving approaches to obtain conditional coverage guarantees, understand training conditional coverage, work with multivariate data,...

What is at the heart of conformal prediction?

"In-sample calibration yields conformal calibration guarantees."

What is at the heart of conformal prediction?

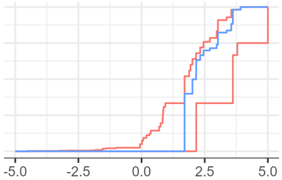
"In-sample calibration yields conformal calibration guarantees."

Predictive system

A set $\Pi \subseteq \mathbb{R} \times [0, 1]$ of the form

$$\Pi = \{(y, \tau) \mid \Pi_\ell(y) \leq \tau \leq \Pi_u(y)\}$$

with $\Pi_\ell \leq \Pi_u$ increasing, $\lim_{y \to -\infty} \Pi_\ell(y) = 0$, $\lim_{y \to \infty} \Pi_u(y) = 1$.



What is at the heart of conformal prediction?

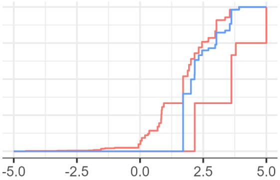
"In-sample calibration yields conformal calibration guarantees."

Predictive system

A set $\Pi \subseteq \mathbb{R} \times [0,1]$ of the form

$$\Pi = \{(y, \tau) \mid \Pi_\ell(y) \leq \tau \leq \Pi_u(y)\}$$

with $\Pi_\ell \leq \Pi_u$ increasing, $\lim_{y\to-\infty} \Pi_\ell(y) = 0$, $\lim_{y\to\infty} \Pi_u(y) = 1$.



Conformal calibration guarantee:

We can construct a predictive system that contains a calibrated CDF.

Example of in-sample calibration:

Let $w_1, \ldots, w_m \in \mathbb{R}$. Define

$$F(y) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\{w_i \leq y\}, \quad y \in \mathbb{R}.$$

Draw W uniformly at random from $w_1, \ldots, w_m$.
Then F is *in-sample* probabilistically calibrated, that is,

$$\mathbb{P}(F(W) < \alpha) \leq \alpha \leq \mathbb{P}(F(W-) \leq \alpha), \quad \alpha \in (0,1).$$

$$F(W) \approx \mathrm{UNIF}(0,1)$$

Let $W_1, \dots, W_{n+1} \in \mathbb{R}$ be exchangeable and define for $w \in \mathbb{R}$

$$F^w(y) = \frac{1}{n+1} \sum_{i=1}^{n} \mathbb{1}\{W_i \le y\} + \frac{1}{n+1} \mathbb{1}\{w \le y\}, \quad y \in \mathbb{R},$$

and

$$\Pi_\ell(y) = \inf\{F^w(y) \mid w \in \mathbb{R}\}, \quad \Pi_u(y) = \sup\{F^w(y) \mid w \in \mathbb{R}\},$$

Then,

$$\Pi_\ell(y) \le F^{W_{n+1}}(y) \le \Pi_u(y), \quad \text{and}$$

$$\mathbb{P}(F^{W_{n+1}}(W_{n+1}) < \alpha) \le \alpha \le \mathbb{P}(F^{W_{n+1}}(W_{n+1}-) \le \alpha), \quad \alpha \in (0,1).$$

Let $W_1, \dots, W_{n+1} \in \mathbb{R}$ be exchangeable and define for $w \in \mathbb{R}$

$$F^w(y) = \frac{1}{n+1} \sum_{i=1}^{n} \mathbb{1}\{W_i \le y\} + \frac{1}{n+1} \mathbb{1}\{w \le y\}, \quad y \in \mathbb{R},$$

and

$$\Pi_\ell(y) = \inf\{F^w(y) \mid w \in \mathbb{R}\}, \quad \Pi_u(y) = \sup\{F^w(y) \mid w \in \mathbb{R}\},$$

Then,

$$\Pi_\ell(y) \le F^{W_{n+1}}(y) \le \Pi_u(y), \quad \text{and}$$

$$\mathbb{P}(F^{W_{n+1}}(W_{n+1}) < \alpha) \le \alpha \le \mathbb{P}(F^{W_{n+1}}(W_{n+1}-) \le \alpha), \quad \alpha \in (0,1).$$

Proof: Conditional on empirical distribution $\hat{\mathbb{P}}_{n+1}$ of $(W_i)_{i=1}^{n+1}$, W_{n+1} is a random draw from $W_1, \dots, W_{n+1}$. By in-sample probabilistic calibration:

$$\mathbb{P}(F^{W_{n+1}}(W_{n+1}) < \alpha \mid \hat{\mathbb{P}}_{n+1}) \le \alpha \le \mathbb{P}(F^{W_{n+1}}(W_{n+1}-) \le \alpha \mid \hat{\mathbb{P}}_{n+1}) \dots$$

(Classical) conformal prediction trick

Use conformity measure $A(\hat{\mathbb{P}}, (x, y))$ to lift the one-dimensional result to general spaces $\mathcal{X} \times \mathcal{Y}$.

(Classical) conformal prediction trick

Use conformity measure $A(\hat{\mathbb{P}}, (x, y))$ to lift the one-dimensional result to general spaces $\mathcal{X} \times \mathcal{Y}$.

Let $(X_1, Y_1), \dots, (X_{n+1}, Y_{n+1}) \in \mathcal{X} \times \mathbb{R}$ be exchangeable.

- $\hat{\mathbb{P}}^y$: Empirical distribution of $(X_1, Y_1), \dots, (X_n, Y_n), (X_{n+1}, y)$ for $y \in \mathbb{R}$
- $\hat{F}^y$: Empirical CDF of

$$W_1 = A(\hat{\mathbb{P}}^y, (X_1, Y_1)), \dots, W_n = A(\hat{\mathbb{P}}^y, (X_n, Y_n)), w(y) = A(\hat{\mathbb{P}}^y, (X_{n+1}, y))$$

- $\mathbb{P}(\hat{F}^{Y_{n+1}}(w(Y_{n+1})) < \alpha) \leq \alpha \quad \leq \mathbb{P}(\hat{F}^{Y_{n+1}}(w(Y_{n+1})-) \leq \alpha)$

(Classical) conformal prediction trick

Use conformity measure $A(\hat{\mathbb{P}}, (x, y))$ to lift the one-dimensional result to general spaces $\mathcal{X} \times \mathcal{Y}$.

Let $(X_1, Y_1), \dots, (X_{n+1}, Y_{n+1}) \in \mathcal{X} \times \mathbb{R}$ be exchangeable.

- $\hat{\mathbb{P}}^y$: Empirical distribution of $(X_1, Y_1), \dots, (X_n, Y_n), (X_{n+1}, y)$ for $y \in \mathbb{R}$
- $\hat{F}^y$: Empirical CDF of

$$W_1 = A(\hat{\mathbb{P}}^y, (X_1, Y_1)), \dots, W_n = A(\hat{\mathbb{P}}^y, (X_n, Y_n)), w(y) = A(\hat{\mathbb{P}}^y, (X_{n+1}, y))$$

- $\mathbb{P}(\hat{F}^{Y_{n+1}}(w(Y_{n+1})) < \alpha) \le \alpha \quad \le \mathbb{P}(\hat{F}^{Y_{n+1}}(w(Y_{n+1})-) \le \alpha)$
- This implies $\mathbb{P}(Y_{n+1} \in C_{n+1}) \ge 1 - \alpha \quad \ge \mathbb{P}(Y_{n+1} \in C_{n+1}^-)$, where

$$C_{n+1} = \{y \in \mathbb{R} \mid \hat{F}^y(w(y)) \ge \alpha\}.$$

(Classical) conformal prediction trick

Use conformity measure $A(\hat{\mathbb{P}}, (x, y))$ to lift the one-dimensional result to general spaces $\mathcal{X} \times \mathcal{Y}$.

Let $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1}) \in \mathcal{X} \times \mathbb{R}$ be exchangeable.

- $\hat{\mathbb{P}}^y$: Empirical distribution of $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)$ for $y \in \mathbb{R}$
- $\hat{F}^y$: Empirical CDF of

$$W_1 = A(\hat{\mathbb{P}}^y, (X_1, Y_1)), \ldots, W_n = A(\hat{\mathbb{P}}^y, (X_n, Y_n)), w(y) = A(\hat{\mathbb{P}}^y, (X_{n+1}, y))$$

- $\mathbb{P}(\hat{F}^{Y_{n+1}}(w(Y_{n+1})) < \alpha) \leq \alpha \quad \leq \mathbb{P}(\hat{F}^{Y_{n+1}}(w(Y_{n+1})-) \leq \alpha)$
- This implies $\mathbb{P}(Y_{n+1} \in C_{n+1}) \geq 1 - \alpha \quad \geq \mathbb{P}(Y_{n+1} \in C_{n+1}^-)$, where

$$C_{n+1} = \{y \in \mathbb{R} \mid \hat{F}^y(w(y)) \geq \alpha\}.$$

- Predictive CDF available if $y \mapsto \hat{F}^y(w(y))$, $y \mapsto \hat{F}^y(w(y)-)$ are increasing.
  (Classical) conformal predictive system

Alternative

Use other in-sample calibrated procedures.

Auto-calibration

Let $(x_1, y_1), \dots, (x_m, y_m) \in \mathcal{X} \times \mathbb{R}$.

- Let $B_1, \dots, B_{m'}$ be a partition of $\{1, \dots, m\}$.
-
$$F_{x_k}(y) = \frac{1}{|B_i|} \sum_{j \in B_i} \mathbb{1}\{y_j \le y\}, \quad k \in B_i, y \in \mathbb{R}$$

Auto-calibration

Let $(x_1, y_1), \dots, (x_m, y_m) \in \mathcal{X} \times \mathbb{R}$.

- Let $B_1, \dots, B_{m'}$ be a partition of $\{1, \dots, m\}$.
- $$F_{x_k}(y) = \frac{1}{|B_i|} \sum_{j \in B_i} \mathbb{1}\{y_j \le y\}, \quad k \in B_i, y \in \mathbb{R}$$

  is in-sample auto-calibrated, that is,

  $$\hat{\mathbb{P}}_m(Y \le y \mid F_X) = F_X(y), \quad y \in \mathbb{R},$$

  hence, in particular, isotonically calibrated, threshold calibrated, quantile calibrated, and probabilistically calibrated.
  Here, $(X, Y) \sim \hat{\mathbb{P}}_m$, and $\hat{\mathbb{P}}_m$ is the empirical distribution of $(x_j, y_j)_{j=1}^m$.

Auto-calibration

Let $(x_1, y_1), \ldots, (x_m, y_m) \in \mathcal{X} \times \mathbb{R}$.

- Let $B_1, \ldots, B_{m'}$ be a partition of $\{1, \ldots, m\}$.
- $$F_{x_k}(y) = \frac{1}{|B_i|} \sum_{j \in B_i} \mathbb{1}\{y_j \leq y\}, \quad k \in B_i, y \in \mathbb{R}$$

  is in-sample auto-calibrated, that is,

  $$\hat{\mathbb{P}}_m(Y \leq y \mid F_X) = F_X(y), \quad y \in \mathbb{R},$$

  hence, in particular, isotonically calibrated, threshold calibrated, quantile calibrated, and probabilistically calibrated.
  Here, $(X, Y) \sim \hat{\mathbb{P}}_m$, and $\hat{\mathbb{P}}_m$ is the empirical distribution of $(x_j, y_j)_{j=1}^m$.
- We call this a *binning procedure*.

Auto-calibration

Let $(x_1, y_1), \ldots, (x_m, y_m) \in \mathcal{X} \times \mathbb{R}$.

- Let $B_1, \ldots, B_{m'}$ be a partition of $\{1, \ldots, m\}$.
- $$F_{x_k}(y) = \frac{1}{|B_i|} \sum_{j \in B_i} \mathbb{1}\{y_j \leq y\}, \quad k \in B_i, y \in \mathbb{R}$$

  is in-sample auto-calibrated, that is,

  $$\hat{\mathbb{P}}_m(Y \leq y \mid F_X) = F_X(y), \quad y \in \mathbb{R},$$

  hence, in particular, isotonically calibrated, threshold calibrated, quantile calibrated, and probabilistically calibrated.

  Here, $(X, Y) \sim \hat{\mathbb{P}}_m$, and $\hat{\mathbb{P}}_m$ is the empirical distribution of $(x_j, y_j)_{j=1}^m$.
- We call this a *binning procedure*.
- All in-sample auto-calibrated procedures are of this form.
- Choice: How is the partition constructed?

Let $(X_1, Y_1), \dots, (X_{n+1}, Y_{n+1}) \in \mathcal{X} \times \mathbb{R}$ be exchangeable.

Let Π be constructed with a binning procedure:

- Let $F^z_{X_k}$ be the binning CDF constructed with $(X_1, Y_1), \dots, (X_n, Y_n), (X_{n+1}, z)$.
- Define

$$\Pi_{\ell,X_{n+1}}(y) = \inf\{F^z_{X_{n+1}}(y) \mid z \in \mathbb{R}\}, \quad \Pi_{u,X_{n+1}}(z) = \sup\{F^z_{X_{n+1}}(y) \mid z \in \mathbb{R}\},$$

Theorem (Conformal calibration guarantee)

Predictive system contains an auto-calibrated CDF:

$$F^{Y_{n+1}}_{X_{n+1}}(y) = \mathbb{P}(Y_{n+1} \le y \mid F^{Y_{n+1}}_{X_{n+1}}), \quad y \in \mathbb{R},$$

and

$$\Pi_{\ell,X_{n+1}}(y) \le F^{Y_{n+1}}_{X_{n+1}}(y) \le \Pi_{u,X_{n+1}}(y), \quad y \in \mathbb{R}$$

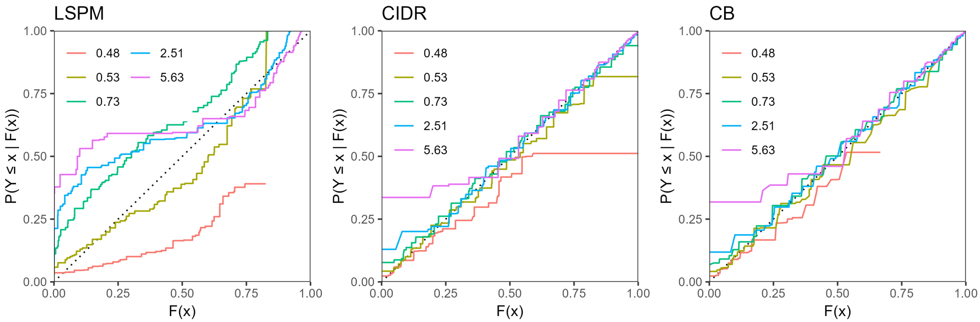
Thickness of predictive systems

- Predictive systems are only useful if they are thin.
- Classical conformal predictive systems:
  - Thickness is $1/(n+1)$.
- Auto-calibration: Binning procedures, where bins are determined only based on $X_1, \ldots, X_{n+1}$ (example: k-means clustering):
  - Thickness is $1/(\text{size of bin containing } n+1)$.

Case study: Length of stay in intensive care units

- Predictions for individual patients' length of stay in ICU's in Switzerland 24h after admission[1]

Threshold calibration



[1] Data provided by *G.-R. Kleger and Schweizerische Gesellschaft für Intensivmedizin*. Data is internal hospital data and not publicly available.

Summary

- In-sample calibration yields conformal calibration guarantees.
- Strong out-of-sample calibration guarantees are possible.
- Arguments can be extended to distribution shifts.
- Conformal binning is simple but works well.
  Only example explored so far: k-means clustering.
- Conformal IDR is a further possibility that was not presented. Allows to quantify alleatoric and epistemic uncertainty.
- Outlook: Conformal calibration guarantees for point predictions.

(Allen et al., 2025a)

References I

S. Allen, J. Ziegel, and D. Ginsbourger. Assessing the calibration of multivariate probabilistic forecasts. *Quarterly Journal of the Royal Meteorological Society*, 150:1315–1335, 2024.

S. Allen, G. Gavrilopoulos, A. Henzi, G.-R. Kleger, and J. Ziegel. In-sample calibration yields conformal calibration guarantees. *Preprint, arXiv:2503.03841*, 2025a.

S. Allen, J. Koh, J. Segers, and J. Ziegel. Tail calibration of probabilistic forecasts. *Journal of the American Statistical Association*, 2025b. To appear.

S. Arnold, A. Henzi, and J. F. Ziegel. Sequentially valid tests for forecast calibration. *Annals of Applied Statistics*, 17:1909–1935, 2023.

Bank of England. Monetary policy report, August 2021. URL `www.bankofengland.co.uk/monetary-policy-report/2021/august-2021`.

R. E. Barlow, D. J. Bartholomew, J. M. Bremner, and H. D. Brunk. *Statistical inference under order restrictions*. John Wiley & Sons Ltd., London, 1972.

G. Bertolini, R. D'Amico, D. Nardi, A. Tinazzi, and G. Apoleone. One model, several results: the paradox of the Hosmer-Lemeshow goodness-of-fit tests for the logistic regression model. *Journal of Epidemiology and Biostatistics*, 5:251–253, 2000.

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

References II

T. Dimitriadis, T. Gneiting, and A. I. Jordan. Stable reliability diagrams for probabilistic classifiers. *Proceedings of the National Academy of Sciences*, 118:e2016191118, 2021.

T. Dimitriadis, T. Gneiting, A. I. Jordan, and P. Vogel. Evaluating probabilistic classifiers: The triptych. *International Journal of Forecasting*, 40:1101–1122, 2024.

T. Gneiting and R. Ranjan. Combining predictive distributions. *Electronic Journal of Statistics*, 7:1747–1782, 2013.

T. Gneiting and P. Vogel. Receiver Operator Characteristic (ROC) curves. *Machine Learning*, 111:2147–2159, 2022.

T. Gneiting, L. I. Stanberry, E. P. Grimit, L. Held, and N. A. Johnson. Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *Test*, 17:211–235, 2008.

A. Henzi, G.-R. Kleger, M. P. Hilty, P. D. Wendel Garcia, and J. F. Ziegel. Probabilistic analysis of COVID-19 patients' individual length of stay in Swiss intensive care units. *PLOS ONE*, 16(2):e0247265, 2021a.

A. Henzi, J. F. Ziegel, and T. Gneiting. Isotonic distributional regression. *Journal of the Royal Statistical Society: Series B*, 85:963–993, 2021b.

References III

D. W. Hosmer and S. Lemeshow. Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics – Theory and Methods*, 9:1043–1069, 1980.

D. W. Hosmer, T. Hosmer, S. le Cessie, and S. Lemeshow. A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*, 16:965–980, 1997.

T. Modeste. *Évaluation et construction des prévisions probabilistes : score et calibration dans un cadre dynamique*. PhD thesis, Université Claude Bernard Lyon 1, 2023.

R. Ranjan and T. Gneiting. Combining probability forecasts. *Journal of the Royal Statistical Society: Series B*, 72:71–91, 2010.

C. Strähl and J. F. Ziegel. Cross-calibration of probabilistic forecasts. *Electronic Journal of Statistics*, 11:608–639, 2017.

T. L. Thorarinsdottir, M. Scheuerer, and C. Heinz. Assessing the calibration of high-dimensional ensemble forecasts using rank histograms. *Journal of Computational and Graphical Statistics*, 25:105–122, 2016.

C. van Eeden. *Testing and estimating ordered parameters of probability distributions*. PhD thesis, University of Amsterdam, 1958.

J. F. Ziegel and T. Gneiting. Copula calibration. *Electronic Journal of Statistics*, 8:2619–2638, 2014.