Evaluating probabilistic classifiers

Reliability diagrams and score decompositions revisited

joint work with T. Dimitriadis and T. Gneiting

Alexander I. Jordan

Heidelberg Institute for Theoretical Studies
Computational Statistics

International Verification Methods Workshop Online
November 18, 2020

Reliability

Target: Binary outcome $Y \in \{0, 1\}$

- often an indicator of a threshold exceedence
- e.g., occurence of precipitation

Probabilistic Forecast $X \in [0, 1]$

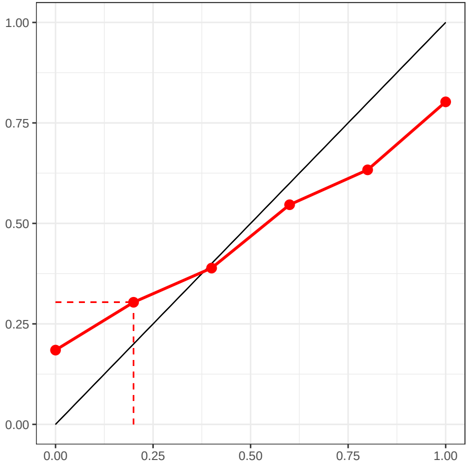
- $X = 0.25$ means we assign 25% probability to the **event** $\{Y = 1\}$

Assess **reliability** or **calibration**

- if $X = 0.25$, then 25% of the cases should be events.

Example: 6 Unique Forecast Values

Forecast	Obs. Freq.	n_j
0.0	0.18	92
0.2	0.30	79
0.4	0.39	72
0.6	0.55	86
0.8	0.63	90
1.0	0.80	81

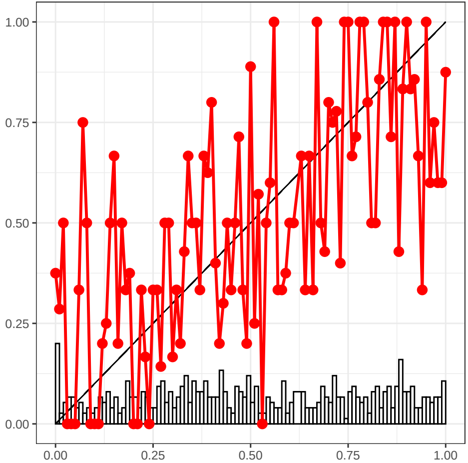


compare the red line against the diagonal.

Example: 101 Unique Forecast Values

Forecast	Obs. Freq.	n_j
0.00	0.38	8
0.01	0.29	7
0.02	0.50	2
0.03	0.00	4
0.04	0.00	5
⋮	⋮	⋮
0.99	0.60	5
1.00	0.88	8

Solution: Binning (and Counting)

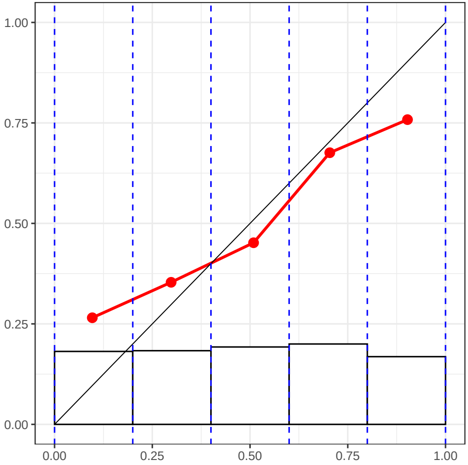


Binning and Counting

Partition $[0, 1]$ in $m \in \mathbb{N}$ bins. But:

- How many bins?
- *How* do we partition?

Common: Equidistant binning



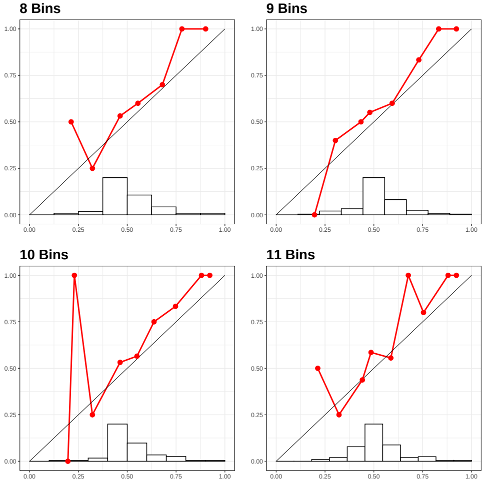
Binning and Counting: Instability

Data set with 92 observations

Location:	Niamey, Niger
Time:	July – September, 2016
Outcome:	daily occurence of precipitation
Prediction:	24-hour ahead EMOS model

Data from Vogel et al. (2020)

Quantile-based binning

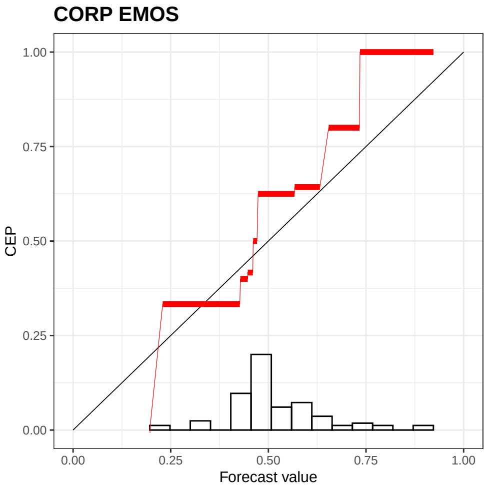


Estimate the **conditional event probability** $\mathrm{CEP}(x) = \mathbb{P}(Y = 1|X = x)$

$$\mathbb{P}(Y = 1|X = x) = \mathbb{E}(\mathbb{1}_{\{Y=1\}}|X = x)$$

isotonic (nonparametric mean) regression

- a higher x should have a higher $\mathrm{CEP}(x)$!



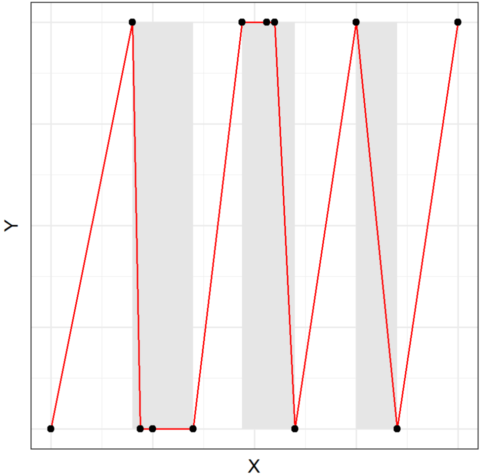
Consistent

Optimally Binned

Reproducible

PAV-Algorithm

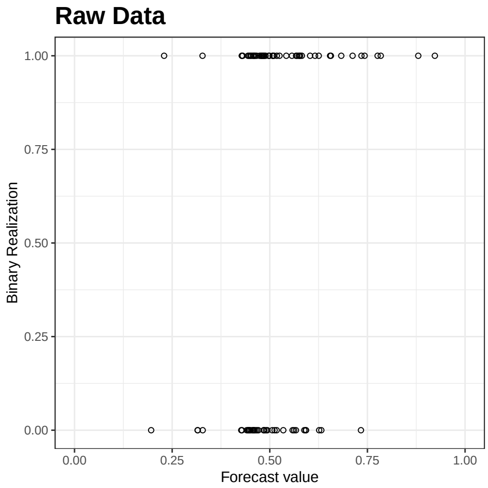
Pool-adjacent-violators algorithm



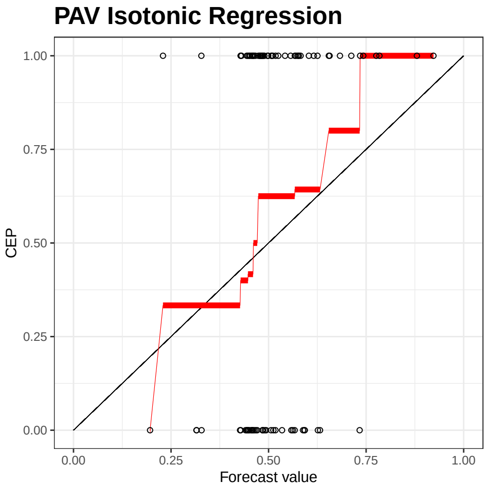
Isotonic regression finds an optimal nondecreasing free-form fit

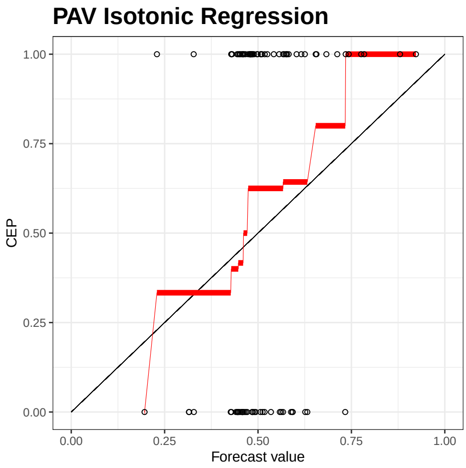
Proposed by Ayer et al. (1955)

Calibrating the predicted probabilities of supervised machine learning models (Niculescu-Mizil and Caruana 2005).


Raw Data
Binary Realization
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Forecast value

PAV

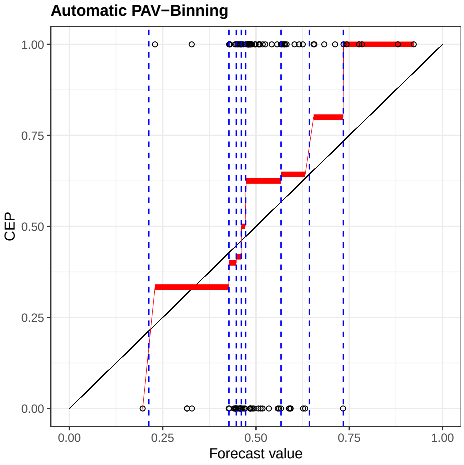

PAV Isotonic Regression
CEP
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Forecast value



Automatic

Binning

⟹



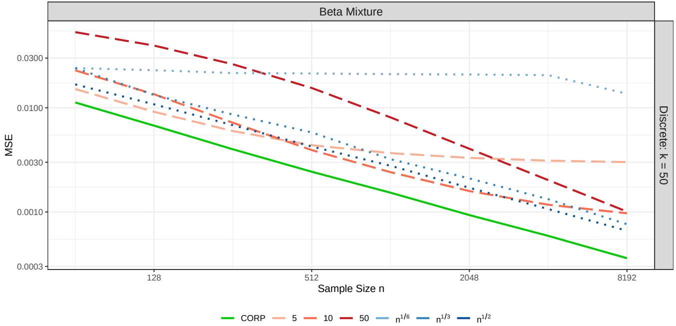
Optimal: minimizing the estimation MSE

$$\min \sum_{i=1}^{n} \left(\widehat{\mathrm{CEP}}_n(x_i) - \mathrm{CEP}(x_i) \right)^2$$

Asymptotically, choosing $\mathcal{O}(n^{1/3})$ bins is optimal.

CORP does exactly this!

CORP is also optimal in finite samples.



Further Simulation Results

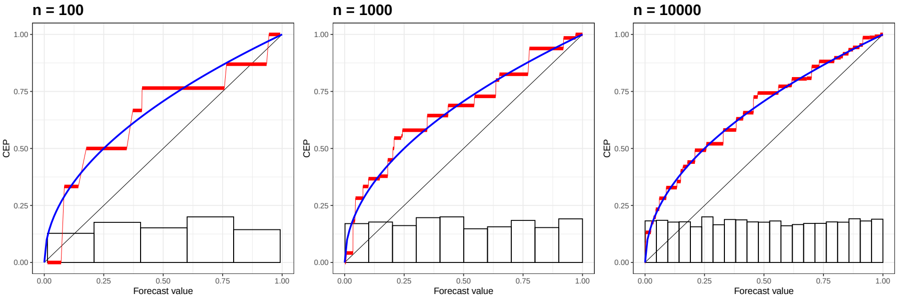
Consistent

asymptotic theory for isotonic regression (El Barmi and Mukerjee, 2005; Wright, 1981)

$$\left|\widehat{\mathrm{CEP}}_n(x) - \mathrm{CEP}(x)\right| \xrightarrow{P} 0 \qquad \forall x \in [0,1]$$

true
simulated

CORP
estimated



Uncertainty quantification

Score Decomposition

Average Brier Score:

$$\bar{S}_X = \frac{1}{n} \sum_{i=1}^{n} (x_i - y_i)^2$$

Why is forecast A better than B?

$$\bar{S}_X = \mathrm{MCB} - \mathrm{DSC} + \mathrm{UNC}$$

decomposes into

- MCB: miscalibration (reliability)
- DSC: discrimination (resolution)
- UNC: uncertainty

Decades of literature:

Murphy (1973)
Dawid (1986)
Stephenson et al. (2008)
Bröcker (2009)
Kull and Flach (2015)
Siegert (2017)
Pohle (2020)
among many others.

score of PAV-recalibrated $\hat{x}_i$

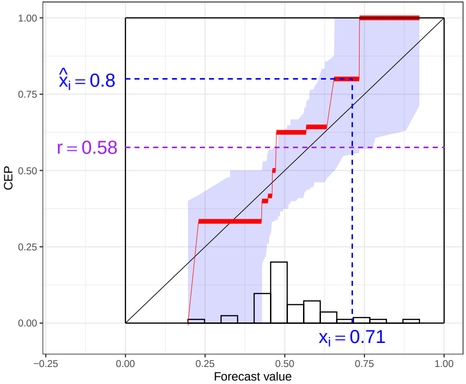
$$\bar{S}_C = \frac{1}{n}\sum_{i=1}^{n}(\hat{x}_i - y_i)^2$$

score of reference forecast $r = \frac{1}{n}\sum_{i=1}^{n} y_i$

$$\bar{S}_R = \frac{1}{n}\sum_{i=1}^{n}(r - y_i)^2$$

Adaptation from Dawid (1986) and Siegert (2017)

$$\bar{S}_X = \underbrace{(\bar{S}_X - \bar{S}_C)}_{\text{MCB}} - \underbrace{(\bar{S}_R - \bar{S}_C)}_{\text{DSC}} + \underbrace{\bar{S}_R}_{\text{UNC}} \qquad (1)$$



Theorem 1 Given any set of original forecast values and associated binary events, suppose that we apply the PAV algorithm to generate a (re)calibrated forecast, and use the marginal event frequency as reference forecast. Then, for every proper scoring rule S, the decomposition defined by Eq. [1] satisfies the following:

(a) MCB ≥ 0 with equality if the original forecast itself is calibrated.

(b) MCB > 0 if the score is strictly proper and the original forecast is not calibrated.

(c) DSC ≥ 0 with equality if the (re)calibrated forecast is constant.

(d) DSC > 0 if the score is strictly proper and the (re)calibrated forecast is not constant.

(e) The decomposition is exact.

Results

- Optimality in finite samples and asymptotically
- Stability without the need of tuning parameters
- Intuitive loss decomposition

Outlook

- Generalization to real-valued outcomes
- Blueprint for novel diagnostic and inference tools

Preprint: `https://arxiv.org/abs/2008.03033`
R package: `https://github.com/aijordan/reliabilitydiag`

References I

Bröcker, J. (2009). Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, 135(643):1512–1519.

Dawid, A. P. (1986). Probability forecasting,. In *Encyclopedia of Statistical Sciences*, volume 7, pages 210–218. Wiley-Interscience.

El Barmi, H. and Mukerjee, H. (2005). Inferences under a stochastic ordering constraint. *Journal of the American Statistical Association*, 100:252–261.

Kull, M. and Flach, P. (2015). Novel decompositions of proper scoring rules for classification: Score adjustment as precursor to calibration. In *Machine Learning and Knowledge Discovery in Databases*, pages 68–85. Springer International Publishing.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, 12:595–600.

Pohle, M.-O. (2020). The Murphy decomposition and the calibrationresolution principle: A new perspective on forecast evaluation. Preprint, `https://arxiv.org/abs/2005.01835`.

Siegert, S. (2017). Simplifying and generalising Murphy's Brier score decomposition. *Quarterly Journal of the Royal Meteorological Society*, 143:1178–1183.

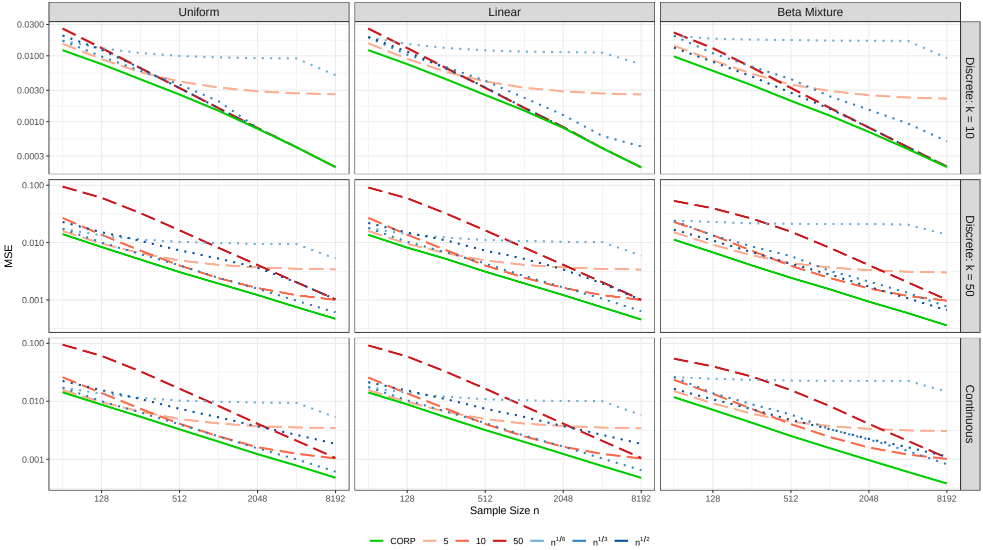
Stephenson, D. B., Coelho, C. A. S., and Jolliffe, I. T. (2008). Two extra components in the Brier score decomposition. *Weather and Forecasting*, 23:752–757.

Vogel, P., Knippertz, P., Gneiting, T., Fink, A. H., Klar, M., and Schlueter, A. (2020). Statistical forecasts for the occurrence of precipitation outperform global models over northern tropical Africa. Preprint, `https://doi.org/10.1002/essoar.10502501.1`.

Wright, F. T. (1981). The asymptotic behavior of monotone regression estimates. *Annals of Statistics*, 9:443–448.

Thanks for your attention!

Optimal Binning: Simulation Evidence



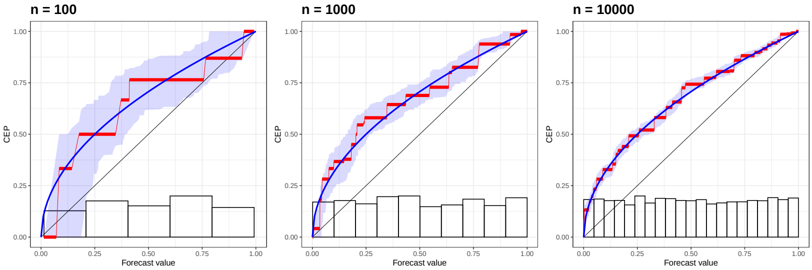
Uncertainty Quantification

Under simplifying assumptions, continuous x (Wright, 1981):

$$n^{1/3} \cdot \Sigma^{-1}(x) \cdot \left(\widehat{\mathrm{CEP}}_n(x) - \mathrm{CEP}(x)\right) \xrightarrow{d} 2\mathcal{T},$$

where $\mathcal{T}$ denotes Chernoff's distribution.

Confidence bands

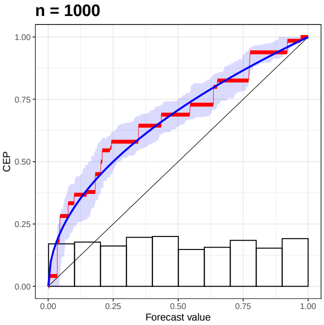


Uncertainty Quantification II

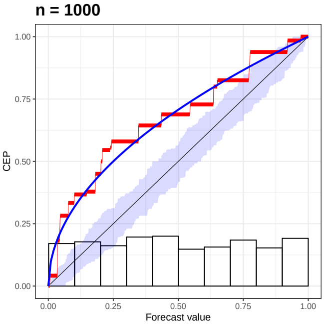
Automatic selection of:

- resampling
- *continuous* asymptotic theory
- *discrete* asymptotic theory

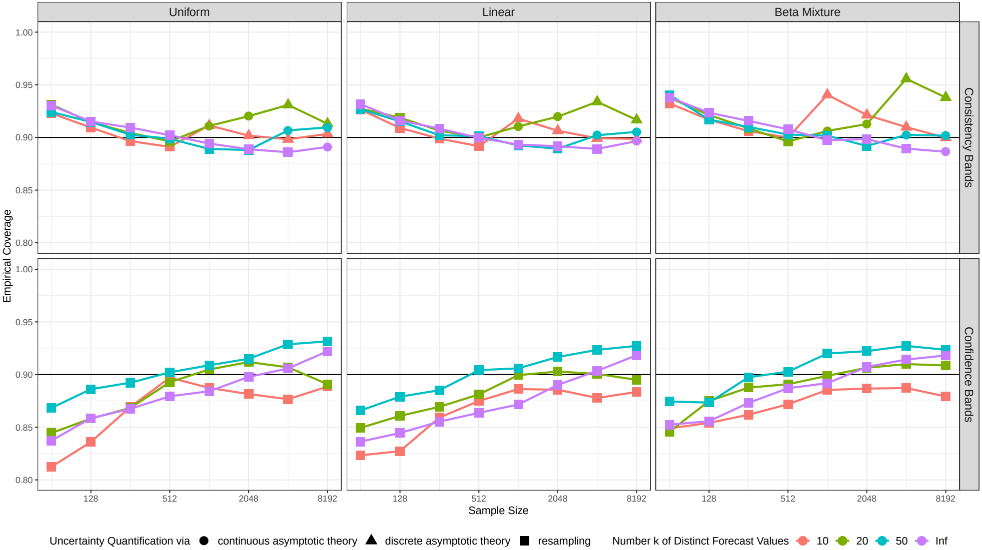
Confidence Bands



Consistency Bands



Coverage Rates: Simulation Evidence



Instability of Quantile Binning