Forecast evaluation I

Thordis L. Thorarinsdottir
Norwegian Computing Center, Oslo, Norway
www.nr.no/~thordis

CUSO winter school 2021

Model evaluation: A fundament of the scientific process

There are three processes that are generally essential for the complete development of any branch of science, and they must be accurately applied before the subject can be considered to be satisfactorily explained. The first is the discovery of a mathematical analysis, the second is the discussion of numerous observations, and the third is a correct application of the mathematics to the observations, including a demonstration that these are in agreement.

Bigelow (MWR, 1905)

Model evaluation: A fundament of the scientific process

There are three processes that are generally essential for the complete development of any branch of science, and they must be accurately applied before the subject can be considered to be satisfactorily explained. The first is the discovery of a mathematical analysis, the second is the discussion of numerous observations, and the third is a correct application of the mathematics to the observations, including a demonstration that these are in agreement.

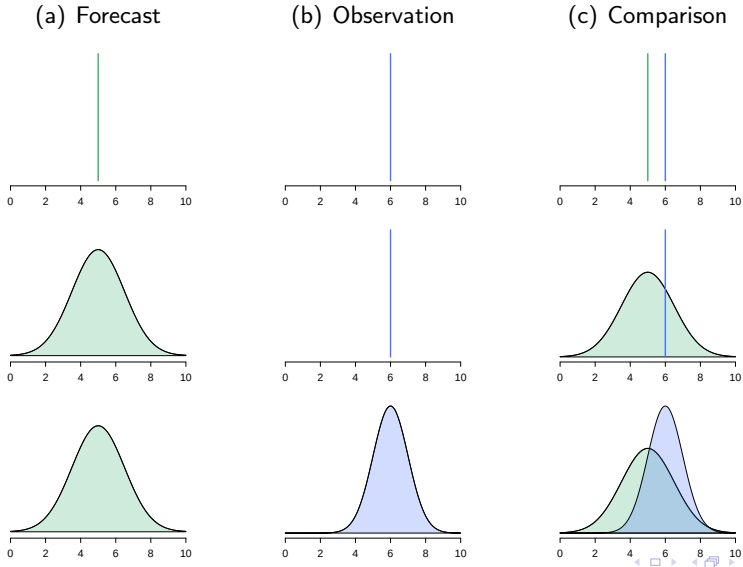
Bigelow (MWR, 1905)

Forecast evaluation: Out-of-sample model evaluation

More precisely, we should probably call them predictions...

- **Forecast**: Prediction issued **before** the predicted quantity could be determined
- **Hindcast**: Prediction issued **after** the predicted quantity could be determined
- **Projection**: Prediction conditioned on specific future boundary conditions that are intended to represent plausible, yet not necessarily probable, future scenarios

Forecast and observation classes



Outline of this lecture series

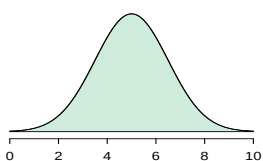
1. What constitutes a good forecast?
2. How do we evaluate the "goodness" of a forecast?

Optimally, forecasts should be probabilistic

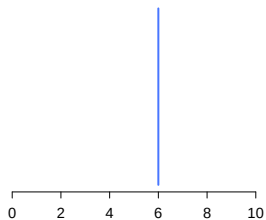
All those whose duty it is to issue regular daily forecasts know that there are times when they feel very confident and other times when they are doubtful as to coming weather. It seems to me that the condition of confidence or otherwise forms a very important part of the prediction.

Cooke (MWR, 1906)

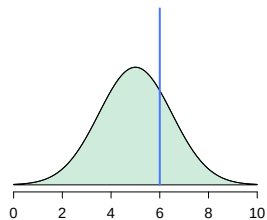
(a) Forecast



(b) Observation



(c) Comparison



What is a good probabilistic forecast?

There should be consistency between the forecaster's judgement and the forecast, there should be correspondence between the forecast and the observation, and the forecast should be informative for the user.

Murphy (WAF, 1993)

What is a good probabilistic forecast?

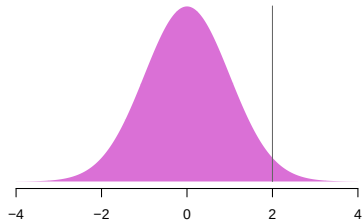
There should be consistency between the forecaster's judgement and the forecast, there should be correspondence between the forecast and the observation, and the forecast should be informative for the user.

Murphy (WAF, 1993)

We propose a diagnostic approach to the evaluation of predictive performance that is based on the paradigm of maximizing the sharpness of the predictive distribution subject to calibration.

Gneiting, Balabdaoui and Raftery (JRSSB, 2007)

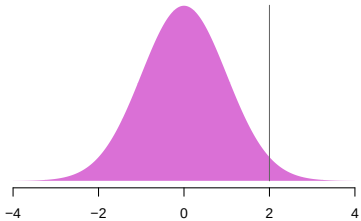
Calibration vs. sharpness



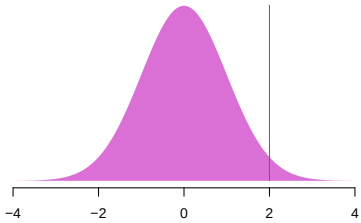
Calibration: Statistical compatibility between the forecast and the observation: An event predicted to occur with probability p should be realized with relative frequency p; joint property of the forecasts and observations

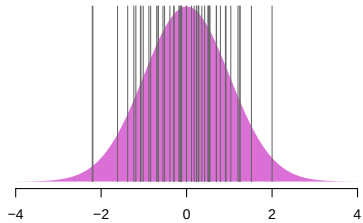
Sharpness: Information content in the forecasts; property of the forecasts only

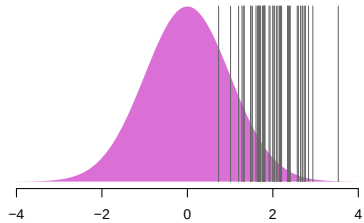
We can't assess calibration with one forecast/observation pair



We can't assess calibration with one forecast/observation pair







PIT/Rank histograms assess calibration of univariate forecasts

A probabilistic forecast for y can be

- an ensemble $E = \{x_1, \ldots, x_m\}$ of point forecasts
- a continuous distribution F

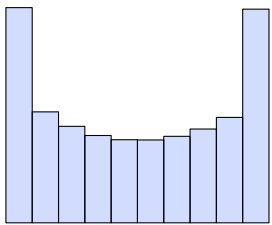
The forecast is **probabilistically calibrated** if

- the **rank** of y is uniform on $\{1, \ldots, m+1\}$
- the **probability integral transform (PIT)** fulfils $F(y) \sim \mathcal{U}([0,1])$

We assess the calibration of $E_1, \ldots, E_n$ or $F_1, \ldots, F_n$ by plotting a histogram of the ranks/PIT values.

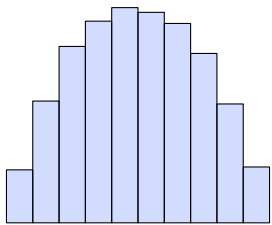
Histogram shape informs on type of miscalibration

Let the observation be $Y \sim \mathcal{N}(0, 1)$



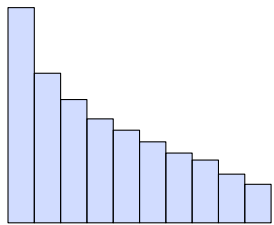
(a) Underdispersive

$F = \mathcal{N}(0, 0.5)$



(b) Overdispersive

$F = \mathcal{N}(0, 2)$



(c) Biased

$F = \mathcal{N}(0.5, 1)$

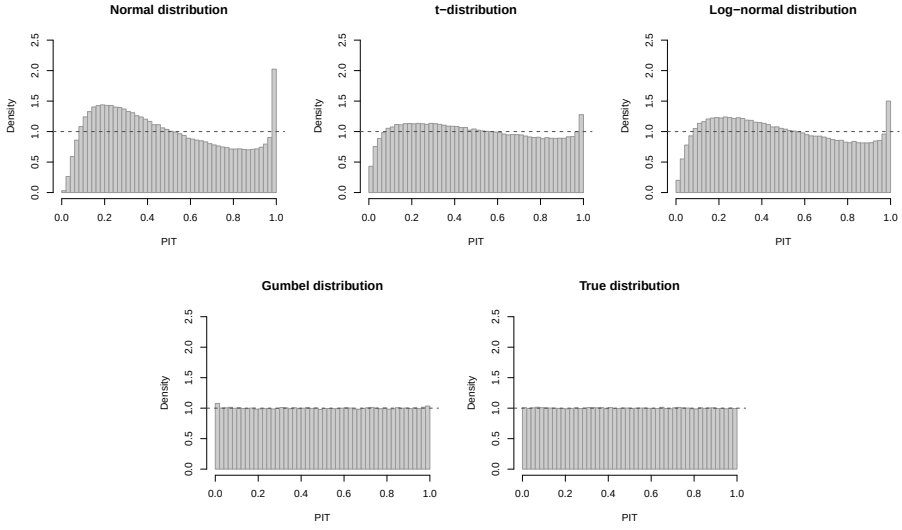
Examples we will use throughout the lecture series

Distribution $F(Y)$		$\mathbb{E}(Y)$	$\text{Var}(Y)$
Normal	$\mathcal{N}(\mu, \sigma^2)$	$\mu \sim \mathcal{N}(25, 1)$	$\sigma^2 = 9$
Gumbel	$\text{G}(\mu, \sigma)$	$\mu + \sigma \cdot \gamma \sim \mathcal{N}(25, 1)$	$\frac{\pi^2}{6}\sigma^2 = \frac{3\pi^2}{2}$

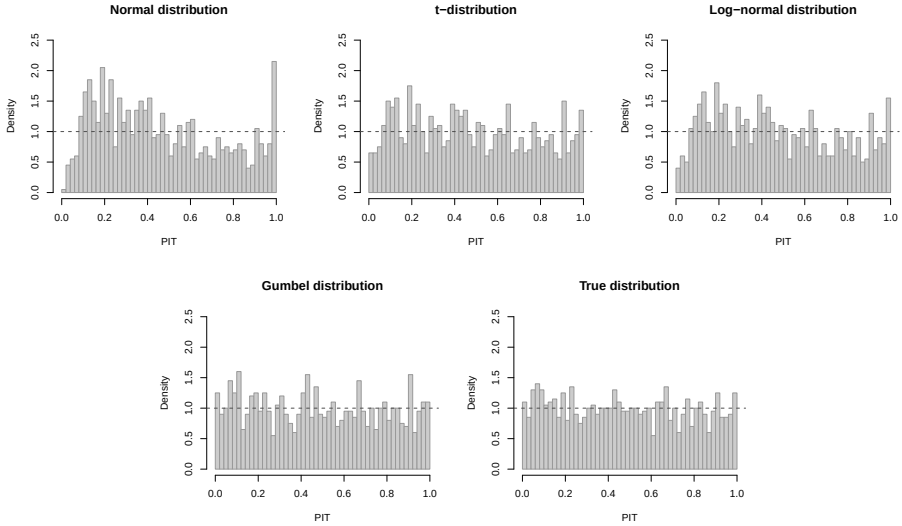
- Competing forecasts: Normal, non-central t, log-normal, Gumbel
- Each forecast is estimated based on 300 i.i.d. observations using methods of moments
- Case 1: 1 000 forecast-observation pairs
- Case 2: 1 000 000 forecast-observation pairs

Thorarinsdottir and Schuhen (2018)

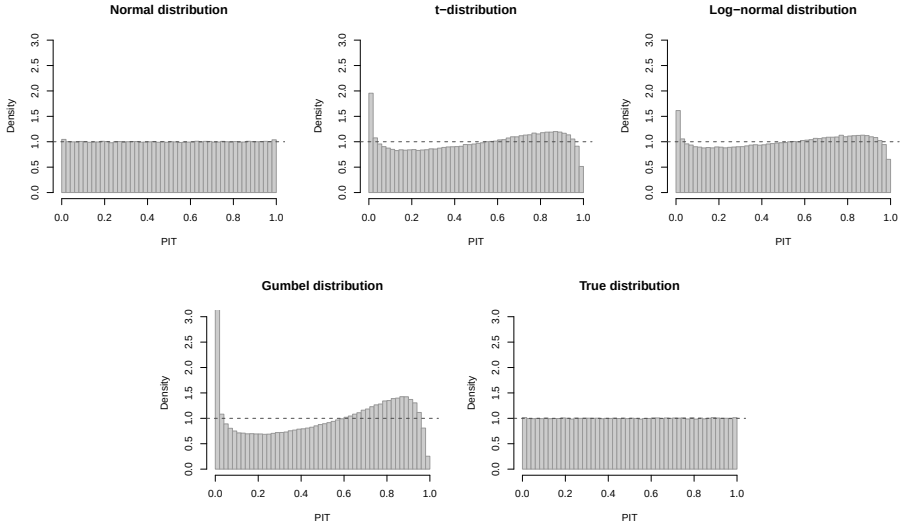
1 000 000 forecast-observation pairs, Gumbel truth



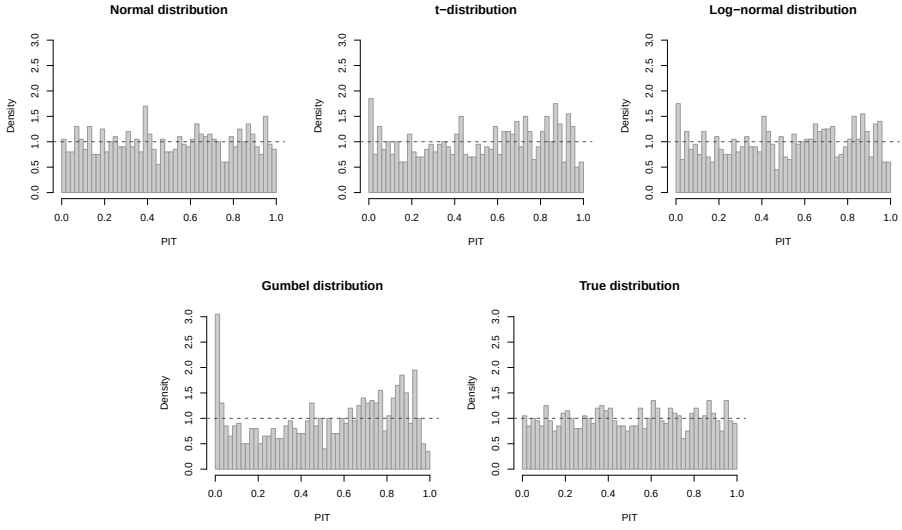
1 000 forecast-observation pairs, Gumbel truth



1 000 000 forecast-observation pairs, Normal truth



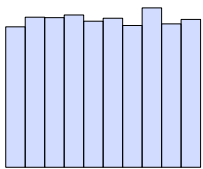
1 000 forecast-observation pairs, Normal truth



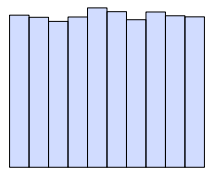
A flat histogram is necessary but not sufficient for calibration

True data distribution: $G_t = \mathcal{N}(\mu_t, 1)$ with $\mu_t \sim \mathcal{N}(0, 1)$.

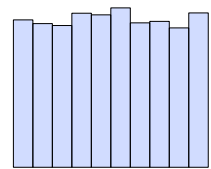
Forecaster	F_t	Parameters
Ideal	$\mathcal{N}(\mu_t, 1)$	
Marginal	$\mathcal{N}(0, 2)$	
Hamill's	$\mathcal{N}(\mu_t + \delta_t, \sigma_t^2)$	$(\delta_t, \sigma_t) \in \{(\frac{1}{2}, 1), (-\frac{1}{2}, 1), (0, \frac{169}{100})\}$



(a) Ideal

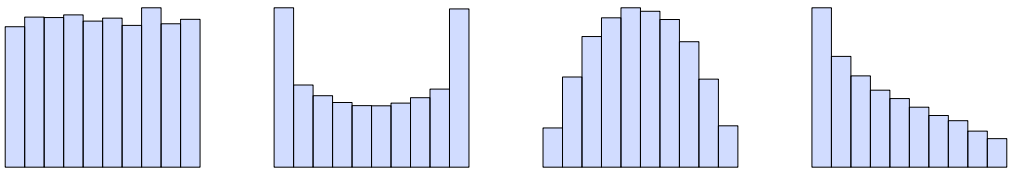


(b) Marginal



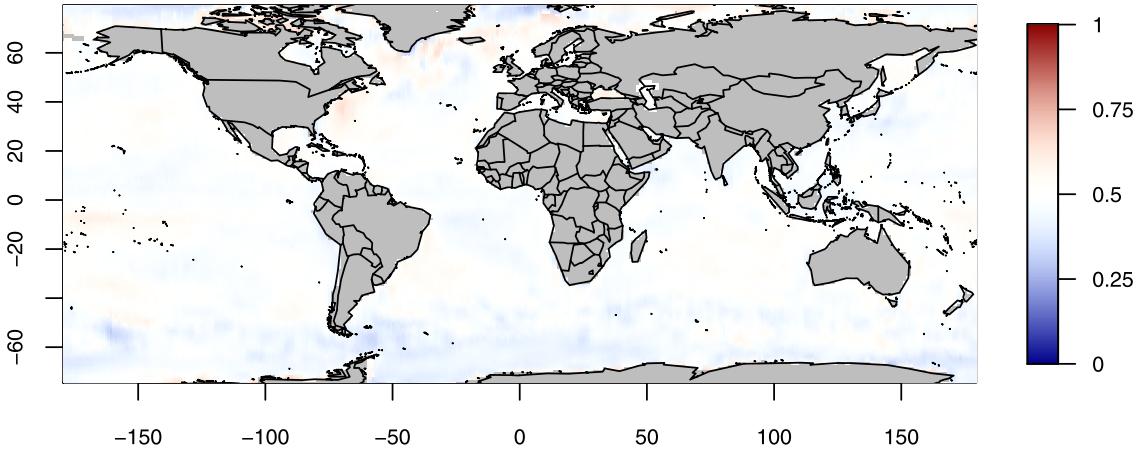
(c) Hamill's

What if we have very many histograms?



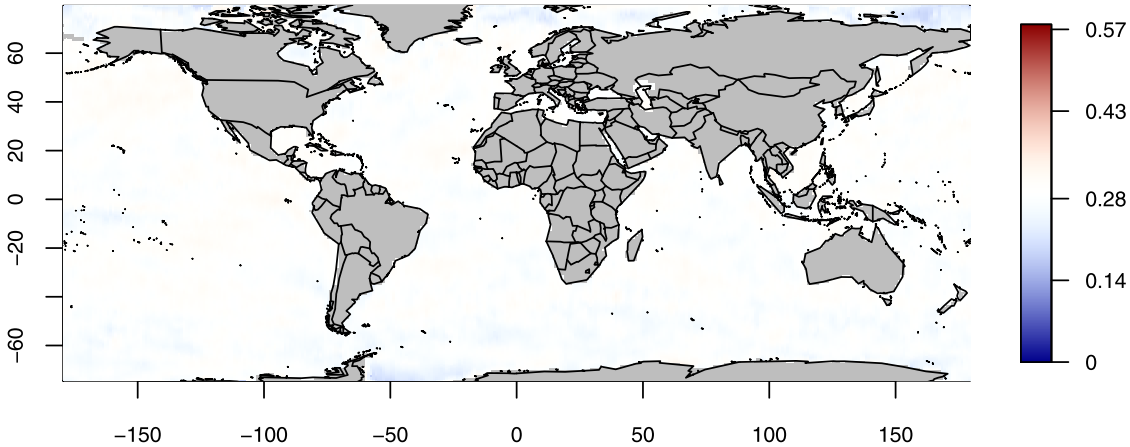
- **Calibrated** PITs are uniform on $[0, 1]$: $\mathbb{E}[F(Y)] = 0.5, \mathrm{SD}[F(Y)] \approx 0.29$
- **Underdispersive** PITs are too often close to 0 or 1: $\mathrm{SD}[F(Y)] > 0.29$
- **Overdispersive** PITs accumulate around 0.5: $\mathrm{SD}[F(Y)] < 0.29$
- **Biased** PITs have $\mathbb{E}[F(Y)] > 0.5$ or $\mathbb{E}[F(Y)] < 0.5$

Example summary PIT means

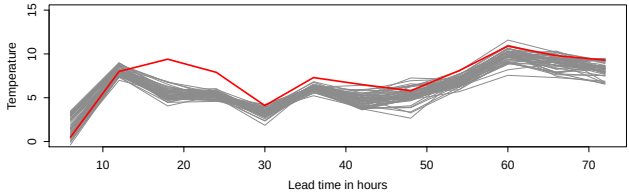


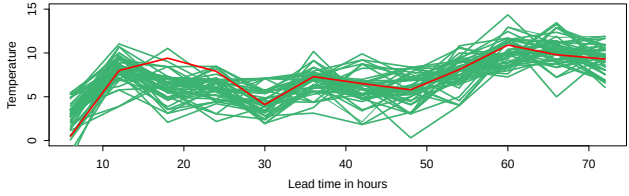
Heinrich *et al.* (JASA, 2020)

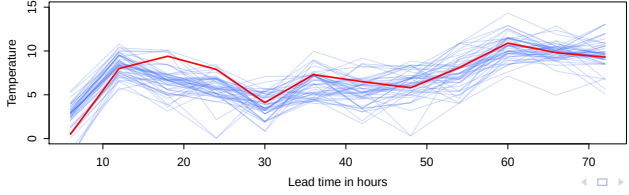
Example summary PIT standard deviations



Heinrich *et al.* (JASA, 2020)


Temperature
0
5
10
15
10
20
30
40
50
60
70
Lead time in hours


Temperature
0
5
10
15
10
20
30
40
50
60
70
Lead time in hours


Temperature
0
5
10
15
10
20
30
40
50
60
70
Lead time in hours

In higher dimensions, we lack a unique ordering

Definition (Multivariate ranks, Gneiting *et al.* (Test, 2008))

1. Apply a **pre-rank function** $\rho_S(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}_+$ to the set $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$.
2. Set the rank of $\mathbf{x}_j$ equal to the rank of $\rho_S(\mathbf{x}_j)$, with ties solved at random.

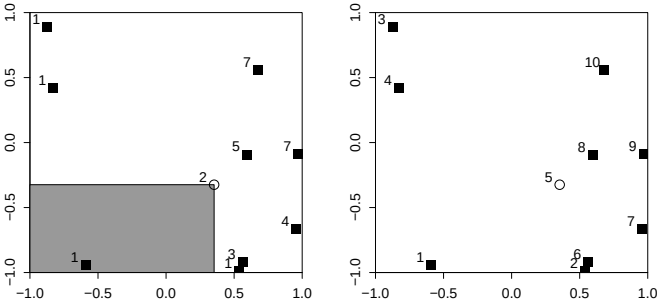
Multivariate rank histograms

The **multivariate rank histogram (MRH)** (Gneiting *et al.*, Test, 2008) uses the ordering

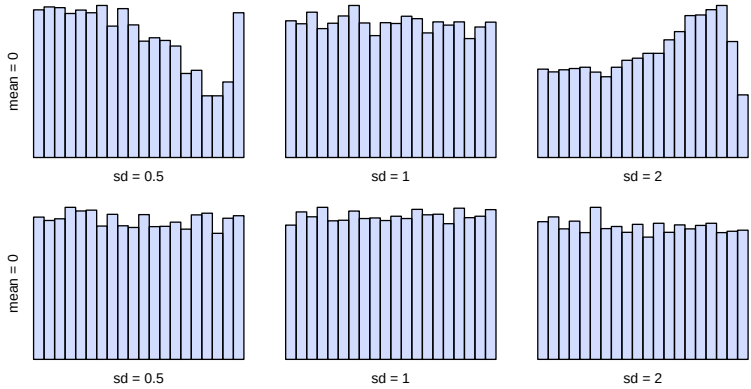
$$\mathbf{x} \preceq \mathbf{y} \text{ if and only if } x(t) \leq y(t) \text{ for all } t = 1, \ldots, T,$$

resulting in the pre-rank function

$$\rho(\mathbf{x}) = \sum_{k=0}^{m} \mathbb{1}\{\mathbf{x}_k \preceq \mathbf{x}\}.$$



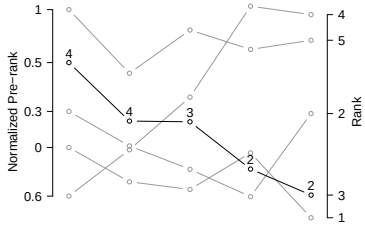
The MRH works well in low dimensions while it eventually breaks down



True distribution is standard normal in 5-dim (top) and 15-dim (bottom).

Forecast is represented by 19 curves.

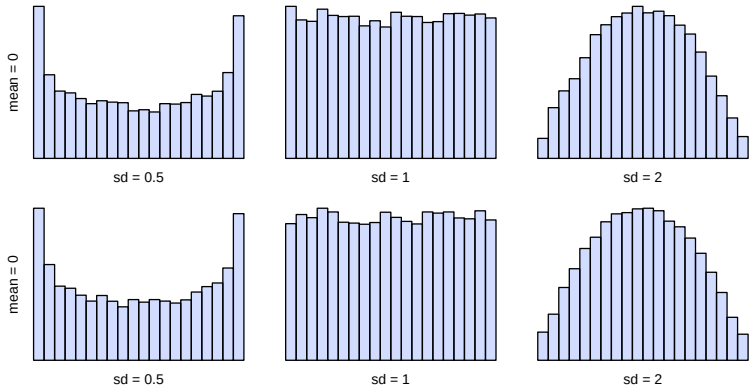
A simple alternative: Average ranking



$$\rho^{\text{avg}}(\mathbf{x}) = \frac{1}{T}\sum_{t=1}^{T}\big[\text{rank of } x(t) \text{ in } \{x_1(t),\ldots,x_m(t)\}\big].$$

(Thorarinsdottir *et al.*, JCGS, 2015)

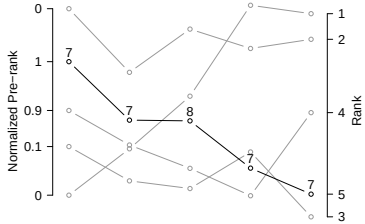
Average rank histograms apply in any dimension



True distribution is standard normal in 5-dim (top) and 15-dim (bottom).

Forecast is represented by 19 curves.

Further alternative: Centre-outwards ordering

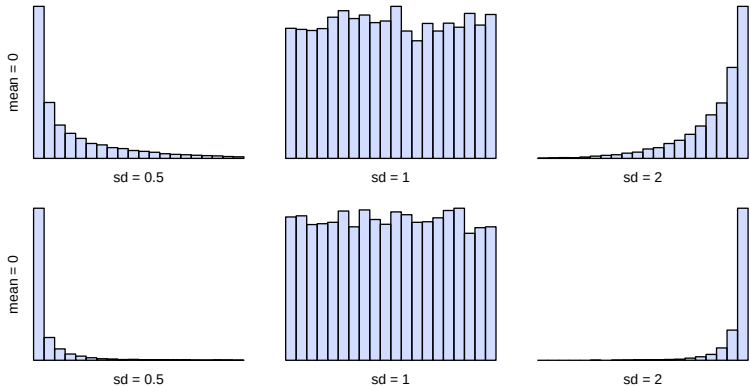


$$\rho^{\mathsf{bd}}(\mathbf{x}) = \frac{1}{T}\sum_{t=1}^{T}\sum_{1\le j<i\le m}\mathbb{1}\big\{\min\{x_j(t),x_i(t)\}\le x(t)\le\max\{x_j(t),x_i(t)\}\big\}$$

$$= \frac{1}{T}\sum_{t=1}^{T}\big[m-\mathsf{rank}\{x(t)\}\big]\big[\mathsf{rank}\{x(t)\}-1\big]+(m-1)$$

In addition to **band depth ranking** (Thorarinsdottir *et al.*, JCGS, 2015), **minimum spanning tree ranking** provides a centre-outward ordering (Smith and Hansen, MWR, 2004; Wilks, MWR, 2004).

Band depth is more sensitive, the higher the dimension



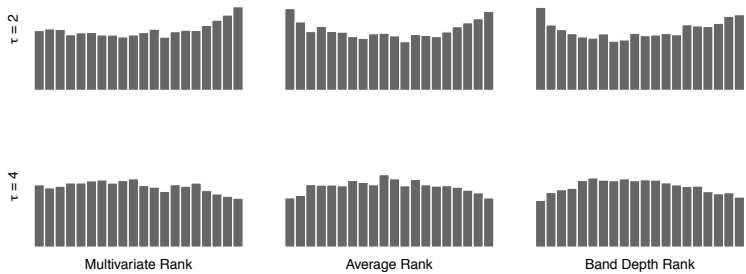
True distribution is standard normal in 5-dim (top) and 15-dim (bottom).

Forecast is represented by 19 curves.

Can averages of univariate attributes detect errors in the multivariate structure?

Let $\mathbf{Y}$ be a zero-mean Gaussian AR(1) process with

$$\mathrm{Cov}(Y(t_j), Y(t_k)) = \exp(-|t_j - t_k|/\tau), \quad \tau = 3.$$



$\mathbf{Y}$ is of dimension 5, forecast is represented by 19 curves.

Can averages of univariate attributes detect errors in the correlation structure?

Let $\mathbf{Y}$ be a zero-mean Gaussian AR(1) process with

$$\mathrm{Cov}(Y(t_j), Y(t_k)) = \exp(-|t_j - t_k|/\tau), \quad \tau = 3.$$

	Rank Mean	Rank Variance
Band depth rank		
Observation	10.7	37
Random ensemble member ($\tau = 2$)	10.5	33
Average rank		
Observation	10.5	37
Random ensemble member ($\tau = 2$)	10.5	33

Conclusions

- Calibration is a fundamental property of a "generally useful" probabilistic forecast
- Rank histograms are a simple and a convenient way of empirically assessing calibration
- The aim is to **detect miscalibration**, not **prove calibration**
- In higher dimensions, first check the marginal calibration of a method and follow by multivariate tests only if the univariate results are satisfying

"With success, you don’t always know why you succeed, but when you fail, it’s clear what you did wrong. Then you can make changes and learn." −Reinhold Messner

References

1. F Bigelow (1905): Application of mathematics in meteorology. *Monthly Weather Review*, 33, 90.
2. W E Cooke (1906): Forecasts and verifications in Western Australia. *Monthly Weather Review*, 34, 23-24.
3. A Murphy (1993): What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, 8, 281-293.
4. T Gneiting, F Balabdaoui and A Raftery (2007): Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Ser. B*, 69, 243-268.
5. T Thorarinsdottir and N Schuhen (2018): Verification: Assessment of calibration and accuracy. In *Statistical Postprocessing of Ensemble Forecasts*, pp. 155-186.
6. C Heinrich, K H Hellton, A Lenkoski and T L Thorarinsdottir (2020): Multivariate postprocessing methods for high-dimensional seasonal weather forecasts. *Journal of the American Statistical Association*, in press.
7. T Gneiting, L Stanberry, E Grimit, L Held and N Johnson (2008): Assessing probabilistic forecasts of multivariate quantities, with applications to ensemble predictions of surface winds (with discussion and rejoinder). *Test*, 17, 211-264.
8. T Thorarinsdottir, M Scheuerer and C Heinz (2016): Assessing the calibration of high-dimensional ensemble forecasts using rank histograms. *Journal of Computational and Graphical Statistics*, 25, 105-122.
9. L Smith and J Hansen (2004): Extending the limits of ensemble forecast verification with the minimum spanning tree. *Monthly Weather Review*, 132, 1522-1528.
10. D Wilks (2004): The minimum spanning tree histogram as verification tool for multidimensional ensemble forecasts. *Monthly Weather Review*, 132, 1329-1340.