Kernel scores: A versatile class of proper scoring rules for evaluating probabilistic forecasts

Johanna Ziegel

University of Bern

LIKE22 – 11-14 January 2022

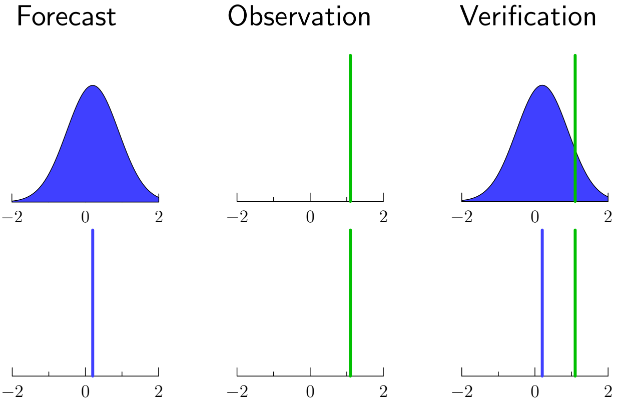
Probabilistic forecasts

- Yesterday at 21:04, the weather forecast of `meteoschweiz.admin.ch` for Bern today at 12:00 stated that

  the temperature will be -1.1°C,
  and there is a 0% chance of rain.

- Difference between these two forecasts:
  - Temperature forecast is a point forecast.
  - "Chance of rain" forecast is a probabilistic forecast.
- A probabilistic forecast for temperature could be: $\mathcal{N}(1.1, \sigma^2)$.

Forecasts for real-valued quantities



Case Study: Precipitation Forecasts

Numerical weather prediction models

- Physical model of the atmosphere is run with current (measured) inital conditions
- Initial conditions are measured with error: Several model runs with slightly perturbed inital conditions yields *ensemble of forecasts*
- Forecast ensembles are interpreted as random draws from the conditional distribution of the outcome
- Ensembles are usually biased and underdispersed: Statistical postprocessing

Bauer et al. (2015)

Case Study: Precipitation Forecasts

Data consists of

- 52-member ECMWF ensemble forecasts and associated observations of 24-hour accumulated precipitation
- prediction horizons of 1 to 5 days ahead
- from 6 January 2007 to 1 January 2017
- at weather stations on airports in London, Brussels, Zurich and Frankfurt.

Precipitation is a challenging variable:

- Mixed discrete-continuous: point mass at zero, right-skewed on $(0, \infty)$

We perform out-of-sample evaluation and comparison of different probabilistic predictions

- years 2015 and 2016 as test period
- prior years serve to provide training data

Henzi et al. (2021)

Statistical postprocessing methods

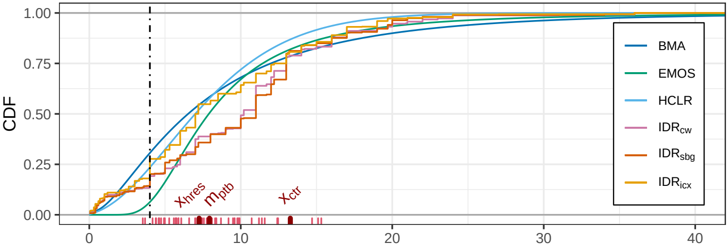
ECMWF ensemble forecast is of the form

$$x = (x_{\text{hres}}, x_{\text{ctr}}, x_1, \ldots, x_{50})$$

Compare different postprocessing/distributional regression techniques with covariate x

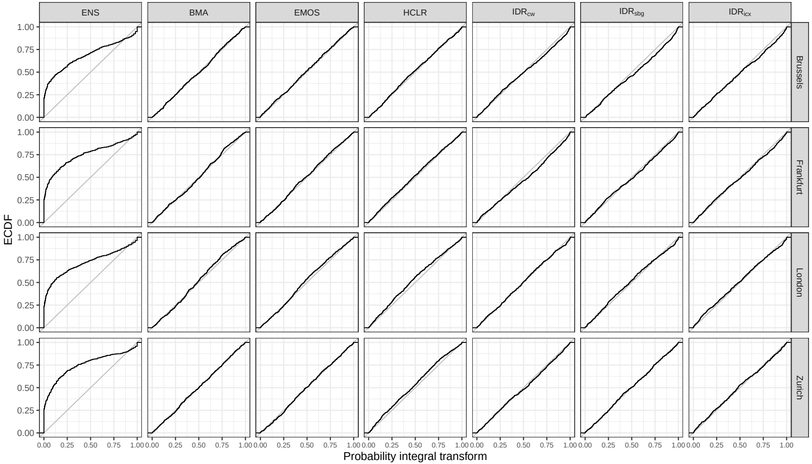
- ENS ECMWF raw ensemble forecast, i.e., the empirical distribution of the 52 ensemble members
- BMA Bayesian Model Averaging (Sloughter et al., 2007)
  - semi-parametric, based on mixtures of Bernoulli and power-transformed Gamma components
- EMOS Ensemble Model Output Statistics (Scheuerer, 2014)
  - parametric, predictive CDFs from the three-parameter family of left-censored generalized extreme value (GEV) distributions
  - location and scale parameters linked to covariates
- HCLR Heteroscedastic Censored Logistic Regression (Messner et al., 2014)
- IDR Isotonic Distributional Regression
  - non-parametric, order-constrained distributional regression
  - partial order on the covariates has to be specified

Example: Predictive CDFs for Brussels, 16 December 2015

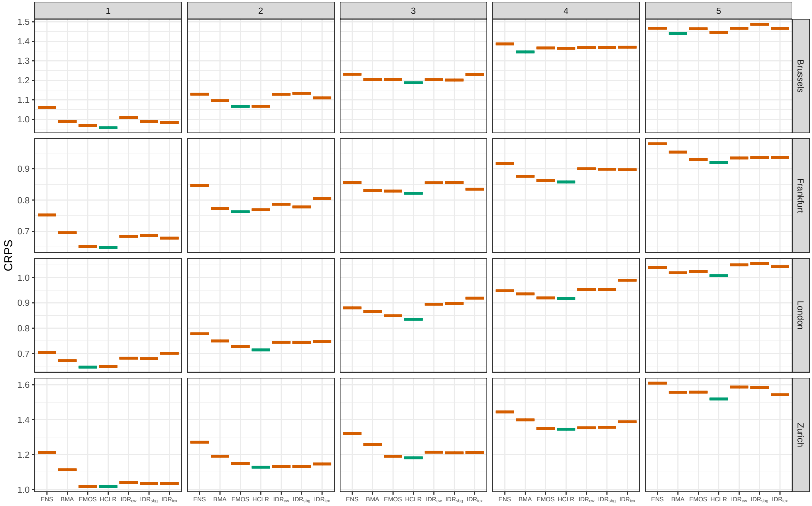


prediction horizon: two days

Calibration: Absolute forecast quality



Comparison with CRPS



Probabilistic forecasts

Let $(\Omega, \mathcal{F}, \mathbb{Q})$ be a probability space.

- The future event Y is a random element in $\mathcal{Y}$.
- Let $\mathcal{A} \subseteq \mathcal{F}$ be a sub σ-algebra. (Our information today.)
- A probabilistic forecast is a random probability measure P which is $\mathcal{A}$-measurable.
  (That is, a Markov kernel from $(\Omega, \mathcal{A})$ to $(\mathcal{Y}, \mathcal{B})$).
- Ideal forecast: $P = \mathcal{L}(Y|\mathcal{A})$.
- Goal in applications: Calibrated predictions

Guiding principle for probabilistic forecasts

"Maximize sharpness subject to calibration."

(Gneiting et al., 2007)

Proper scoring rules

Let $\mathcal{P}$ be a class of probability measures on $(\mathcal{Y}, \mathcal{B})$.

Definition

A *scoring rule* is a function $S : \mathcal{P} \times \mathcal{Y} \to \mathbb{R} \cup \{\pm\infty\}$ such that for any $P \in \mathcal{P}$, $S(P, \cdot)$ is quasi-integrable with respect to any $Q \in \mathcal{P}$. A scoring rule S is *proper* if

$$S(P, P) = \mathbb{E}_P S(P, Y) \leq \mathbb{E}_P S(Q, Y) = S(Q, P), \quad P, Q \in \mathcal{P}.$$

S is *strictly* proper if equality implies $P = Q$.

Proper scoring rules

Let $\mathcal{P}$ be a class of probability measures on $(\mathcal{Y}, \mathcal{B})$.

Definition

A *scoring rule* is a function $S : \mathcal{P} \times \mathcal{Y} \to \mathbb{R} \cup \{\pm\infty\}$ such that for any $P \in \mathcal{P}$, $S(P, \cdot)$ is quasi-integrable with respect to any $Q \in \mathcal{P}$.
A scoring rule S is *proper* if

$$S(P, P) = \mathbb{E}_P S(P, Y) \leq \mathbb{E}_P S(Q, Y) = S(Q, P), \quad P, Q \in \mathcal{P}.$$

S is *strictly* proper if equality implies $P = Q$.
Suppose that S is strictly proper.
The *entropy* associated to S is

$$G(P) = S(P, P), \quad P \in \mathcal{P}$$

and the *divergence* is

$$d(P, Q) = S(P, Q) - S(Q, Q), \quad P, Q \in \mathcal{P}.$$

Comparison of probabilistic forecasts

Available data

Sequence of (at least) two forecasts and observations

$$(P_{11}, P_{21}, Y_1), \ldots, (P_{1n}, P_{2n}, Y_n)$$

Given a proper scoring rule S, compare average realized scores:

$$\frac{1}{n}\sum_{i=1}^{n} S(P_{1i}, Y_i) \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} S(P_{2i}, Y_i)$$

Proper scoring rules

- ...assess calibration and sharpness simultaneously.
- ...are sensitive with respect to increasing information sets.

(Gneiting and Raftery, 2007)

Examples of proper scoring rules: Density forecasts

Let μ be a σ-finite measure on $\mathcal{Y}$. Specify the forecast P in terms of its density with respect to μ.

Logarithmic score

$$S(p, y) = -\log p(y)$$

Strictly proper with respect to all measures that are absolutely continuous with respect to μ.
Entropy is the Shannon entropy

$$G(p) = -\int_{\mathcal{Y}} p(y) \log p(y) \, \mathrm{d}\mu(y)$$

Divergenge is the Kullback-Leibler divergence

$$d(p, q) = \int_{\mathcal{Y}} q(y) \log \frac{q(y)}{p(y)} \, \mathrm{d}\mu(y)$$

Forecasts with finite mean

Let $\mathcal{P}$ be the class of probability measures on $\mathbb{R}$ with finite mean. Specify P in terms of its CDF F.

Continuous Ranked Probability Score (CRPS)

$$\begin{aligned} S(F, y) &= \int_{\mathbb{R}} (F(x) - \mathbb{1}\{y \leq x\})^2 \, \mathrm{d}(x) \\ &= \int_0^1 \left(\mathbb{1}\{y \leq F^{-1}(\alpha)\} - \alpha\right) \left(F^{-1}(\alpha) - y\right) \, \mathrm{d}\alpha \\ &= \mathbb{E}_F |X - y| - \frac{1}{2} \mathbb{E}_F |X - X'| \end{aligned}$$

- Allows to compare discrete, continuous and mixed discrete-continuous distributions.
- Is becoming increasingly popoular also in estimation (Gneiting et al., 2005; Hothorn et al., 2014; Gasthaus et al., 2019).

Characterization

Proper scoring rules can be characterized in terms of concave functions on $\mathcal{P}$.

Theorem (Gneiting and Raftery (2007))

A scoring rule is (strictly) proper if and only if there exists a (strictly) concave function G on $\mathcal{P}$ such that

$$S(P, y) = G(P) - \int G^*(P, y') \, dP(y') + G^*(P, y),$$

where $G^*(P, \cdot)$ is a supertangent of G at P.

- How can we generate interesting concave functions on a general outcome space $\mathcal{Y}$?

Kernel scores

Suppose we have a measurable kernel k on $\mathcal{Y}$, that is $k : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ symmetric and positive definite. Then,

$$G(P) = -\frac{1}{2} \int \int k(x, y) \, \mathrm{d}P(x) \, \mathrm{d}P(y) + \frac{1}{2} \int k(y, y) \, \mathrm{d}P(y)$$

is concave and non-negative. It has supertangent

$$G^*(P, y) = -\int k(x, y) \, \mathrm{d}P(x) + \frac{1}{2} k(y, y).$$

Definition (Kernel score)

Let $(\mathcal{Y}, \mathcal{G})$ be a measurable space and k a measurable kernel on $\mathcal{Y}$. Then,

$$S_k(P, y) = -\int_{\mathcal{Y}} k(x, y) \,\mathrm{d}P(x) + \frac{1}{2} \int_{\mathcal{Y}} \int_{\mathcal{Y}} k(x, x') \,\mathrm{d}P(x) \,\mathrm{d}P(x') + \frac{1}{2} k(y, y)$$

is the *kernel score* of k.

- Kernel scores are proper on

$$\mathcal{M}_1^k(\mathcal{Y}) = \left\{ P \mid \int_{\mathcal{Y}} \sqrt{k(y, y)} \,\mathrm{d}P(y) < \infty \right\}$$

- They have close connections to machine learning.
- They have close connections to energy statistics.

(Gneiting and Raftery, 2007; Dawid, 2007)

$$
\begin{aligned}
S_k(P, y) &= -\int_{\mathcal{Y}} k(x, y) \, \mathrm{d}P(x) \\
&\qquad + \frac{1}{2} \int_{\mathcal{Y}} \int_{\mathcal{Y}} k(x, x') \, \mathrm{d}P(x) \, \mathrm{d}P(x') + \frac{1}{2} k(y, y) \\
&= \frac{1}{2} \int_{\mathcal{Y} \times \mathcal{Y}} k \, \mathrm{d}(P - \delta_y) \otimes (P - \delta_y)
\end{aligned}
$$

Entropy

$$
G_k(P) = -\frac{1}{2} \int_{\mathcal{Y}} \int_{\mathcal{Y}} k(x, x') \, \mathrm{d}P(x) \, \mathrm{d}P(x') + \frac{1}{2} \int_{\mathcal{Y}} k(y, y) \, \mathrm{d}P(y)
$$

Divergence

$$
d_k(P, Q) = \frac{1}{2} \int_{\mathcal{Y} \times \mathcal{Y}} k \, \mathrm{d}(P - Q) \otimes (P - Q)
$$

Continuous Ranked Probability Score (CRPS)

$$S(P, y) = \mathbb{E}|X - y| - \frac{1}{2}\mathbb{E}|X - X'|,$$

where $X, X' \sim P$ and X, X' independent, is a kernel score S_k with kernel

$$k(x, y) = |x| + |y| - |x - y|.$$

Entropy

$$G_k(P) = \frac{1}{2}\mathbb{E}|X - X'|$$

Divergence

$$d_k(P, Q) = \mathbb{E}|X - Y| - \frac{1}{2}\mathbb{E}|X - X'| - \frac{1}{2}\mathbb{E}|Y - Y'|$$

where $X, X' \sim P$, $Y, Y' \sim Q$ all independent.

Propriety and connection to machine learning

Theorem

Let k be a measurable kernel with RKHS H with norm $\|\cdot\|_H$ and $\Phi\colon \mathcal{M}_1^k(\mathcal{Y}) \to H$ the kernel mean embedding defined by

$$\Phi(P) = \int_{\mathcal{Y}} k(x, \cdot)\, dP(x).$$

Then,

$$\|\Phi(P) - \Phi(Q)\|_H^2 = 2d_k(P, Q), \quad P, Q \in \mathcal{M}_1^k(\mathcal{Y}).$$

S_k is strictly proper if and only if Φ is injective.

- If k is bounded then k is called *characteristic* if Φ is injective.
- $\gamma_k(P, Q) = \|\Phi(P) - \Phi(Q)\|_H$ is the *maximum mean discrepancy* between P and Q.

(Gretton et al., 2012; Steinwart and Ziegel, 2021)

Characteristic kernels: On $\mathbb{R}^d$

Bounded continuous kernels

Radial kernels that are strictly positive definite for any d:

$$k(x, y) = \varphi(\|x - y\|),$$

where

$$\varphi(t) = \int_0^\infty \exp(-t^2 s)\, \mathrm{d}\nu(s)$$

for a measure μ with $\operatorname{supp} \mu \neq \{0\}$.

(Sriperumbudur et al., 2011)

Distance kernels

$$k(x, y) = \|x\|^\alpha + \|y\|^\alpha - \|x - y\|^\alpha,$$

for $\alpha \in (0, 2)$.

Characteristic kernels: On $\mathbb{S}^d$

Consider isotropic kernels

$$k(x, y) = \psi(\text{arccos}\langle x, y\rangle)$$

with $\psi\colon [0, \pi] \to \mathbb{R}$ continuous.

Results

- If ψ induces a positive definite kernel on $\mathbb{S}^{d+2}$ or a strictly positive definite kernel on $\mathbb{S}^{d+1}$, then it induces a characteristic kernel on $\mathbb{S}^d$ if and only if it is strictly positive definite.
- Suppose that ψ induces a positive definite kernel on $\mathbb{S}^d$ for all d. Then, ψ is strictly positive definite for all d, if and only if, it is characteristic for some d.

(Steinwart and Ziegel, 2021)

Connection to energy distance

Székely and Rizzo (2004), Baringhaus and Franz (2004) introduced the *energy distance* between two distributions P, Q on $\mathbb{R}^d$ with finite first moments

$$\mathbb{E}\|Z - W\| - \frac{1}{2}\mathbb{E}\|Z - Z'\| - \frac{1}{2}\mathbb{E}\|W - W'\|,$$

where Z, Z', W, W' are independent with $Z, Z' \sim P$, $W, W' \sim Q$.

- Energy distance is the divergence of the kernel score with kernel $k(x, y) = \|x\| + \|y\| - \|x - y\|$ called the *energy score*.
- Energy distance is a squared maximum mean discrepancy between P and Q (Sejdinovic et al., 2013).
- Energy score is a popular strictly proper scoring rule for multivariate outcomes.

Characteristic distance kernels/Metrics of strong negative type

$$k(x, y) = \|x\| + \|y\| - \|x - y\|$$

- Separable Hilbert spaces
- Separable L^p-spaces for $1 < p \leq 2$

(Linde, 1986; Lyons, 2013; Sejdinovic et al., 2013)

- Since future outcomes are uncertain, predictions should be probabilistic.
- Probabilistic predictions should be compared with proper scoring rules.
- Kernel scores provide proper scoring rules on general outcome spaces $\mathcal{Y}$ as soon as a positive definite kernel is available.
- Kernel scores can often only be computed numerically. For the CRPS, closed form expressions are available for many relevant predictive distributions (Jordan et al., 2019).

References I

L. Baringhaus and C. Franz. On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88:190–206, 2004.

P. Bauer, A. Thorpe, and G. Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525:47–55, 2015.

A. P. Dawid. The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 59:77–93, 2007.

J. Gasthaus, K. Benidis, Y. Wang, S. S. Rangapuram, D. Salinas, V. Flunkert, and T. Januschowski. Probabilistic forecasting with spline quantile function RNNs. *Proceedings of Machine Learning Research*, 89:1901–1910, 2019.

T. Gneiting. Strictly and non-strictly positive definite functions on spheres. *Bernoulli*, 19:1327–1349, 2013.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378, 2007.

T. Gneiting, A. E. Raftery, A. H. Westveld, and T. Goldman. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133:1098–1118, 2005.

References II

T. Gneiting, F. Balabdaoui, and A. E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B*, 69:243–268, 2007.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

A. Henzi, J. F. Ziegel, and T. Gneiting. Isotonic distributional regression. *Journal of the Royal Statistical Society: Series B*, 2021. To appear. Preprint available at `arXiv:1909.03725`.

T. Hothorn, T. Kneib, and P. Bühlmann. Conditional transformation models. *Journal of the Royal Statistical Society: Series B*, 76:3–27, 2014.

A. Jordan, F. Krüger, and A. Lerch. Evaluating probabilistic forecasts with scoringRules. *Journal of Statstical Software*, 90:1–37, 2019.

W. Linde. Uniqueness theorems for measures in l_r and $c_0(\omega)$. *Mathematische Annalen*, 274:617–626, 1986.

R. Lyons. Distance covariance in metric spaces. *Annals of Probability*, 41: 3284–3305, 2013.

References III

J. W. Messner, G. J. Mayr, D. S. Wilks, and A. Zeileis. Extending extended logistic regression: Extended versus separate versus ordered versus censored. *Monthly Weather Review*, 142:3003–3014, 2014.

M. Scheuerer. Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quarterly Journal of the Royal Meteorological Society*, 140:1086–1096, 2014.

I. J. Schoenberg. Metric spaces and completely monotone functions. *Ann. Math.*, 39:811–841, 1938.

I. J. Schoenberg. Positive definite functions on spheres. *Duke Math. J.*, 9: 96–108, 1942.

D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Annals of Statistics*, 41:2263–2291, 2013.

J. M. Sloughter, A. E. Raftery, T. Gneiting, and C. Fraley. Probabilistic quantitative precipitation forecasting using Baysian model averaging. *Monthly Weather Review*, 135:3209–3220, 2007.

B. K. Sriperumbudur, K. Fukumizu, and G. R. G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12:2389–2410, 2011.

References IV

I. Steinwart and J. F. Ziegel. Strictly proper kernel scores and characteristic kernels on compact spaces. *Applied and Computational Harmonic Analysis*, 51:510–542, 2021.

G. J. Székely and M. Rizzo. Testing for equal distribution in high dimension. *InterStat*, 5, November 2004.