Evaluation of Probabilistic Forecasts

Mathias Blicher Bjerregård and Henrik Madsen

DTU Compute, Technical University of Denmark

matbb@dtu.com

November 22, 2021

Overview

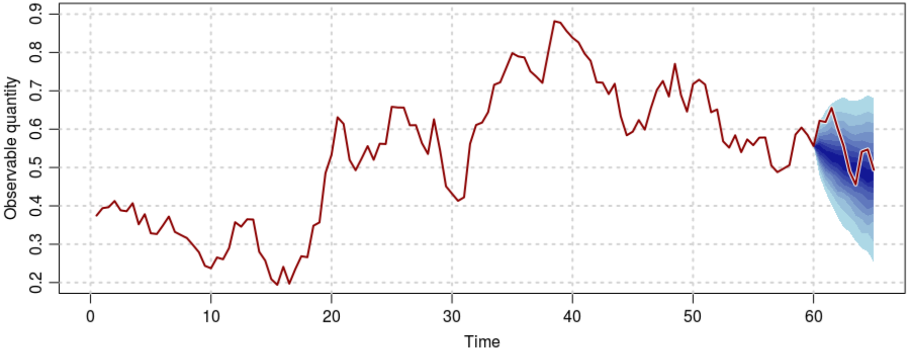
Overview

Forecasting

- *Forecasting* refers to prediction of future events in a mathematical sense.
- All forecasting of real variables comes with a stochastic component, hence suitable forecasts are *probabilistic*.
- A *forecast* of an event usually takes the form of a *point estimate*, an *interval*, a set of *quantiles* or a *probability density function* (PDF).

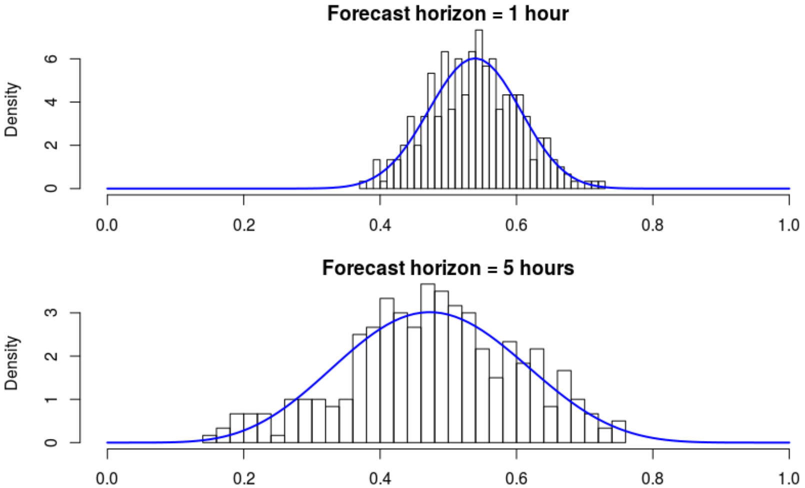
Forecasting

- A typical forecast scenario:



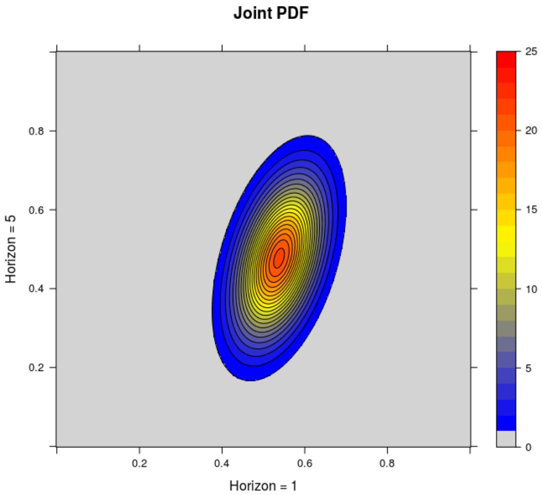
Probabilistic forecasting

- A probabilistic forecast of a variable at different horizons may be considered in terms of its marginal desnities:



Probabilistic forecasting

- \- or in terms of the joint density:



Forecast evaluation

Forecast Evaluation

- *Definition*: Either directly assessing the quality of one model or comparing a set of competing models.

- Why don't we simply apply the Mean Squared Error every time?

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

Motivation - simulation example

- Consider the conditional heteroscedastic Gaussian time series model [Gneiting, 2010]:

$$Y_t = Z_t^2$$

with,

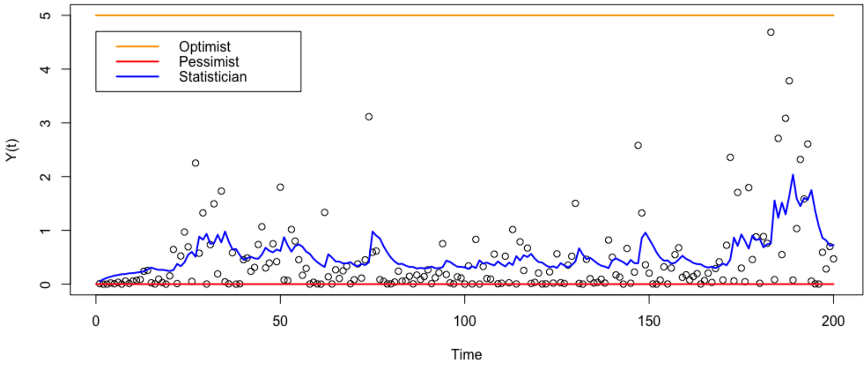
$$\begin{aligned} Z_t &\sim \mathcal{N}(0, \sigma_t^2) \\ \sigma_t^2 &= 0.20 Z_{t-1}^2 + 0.75 \sigma_{t-1}^2 + 0.05 \end{aligned}$$

- We compare three forecasting models for predicting Y_t:

Optimist:	$\hat{y}_{t\|t-1} = 5$
Pessimist:	$\hat{y}_{t\|t-1} = 0.05$
Statistician:	$\hat{y}_{t\|t-1} = \mathrm{E}[Y_t\|\sigma_t^2] = \sigma_t^2$

Motivation - simulation example

- Which model is superior?



Some commonly used scoring functions

Table 1. Some commonly used scoring functions

$S(x, y) = (x - y)^2$	squared error (SE)
$S(x, y) = \|x - y\|$	absolute error (AE)
$S(x, y) = \|(x - y)/y\|$	absolute percentage error (APE)
$S(x, y) = \|(x - y)/x\|$	relative error (RE)

Evaluation using commonly used scoring functions

Table 4. The mean error measure (1) for the three point forecasters in the simulation study, using the squared error (SE), absolute error (AE), absolute percentage error (APE) and relative error (RE) scoring functions

Forecaster	SE	AE	APE	RE
Statistician	5.07	0.97	2.58×10^5	0.97
Optimist	22.73	4.35	13.96×10^5	0.87
Pessimist	7.61	0.96	0.14×10^5	19.24

Overview

Probability integral transform

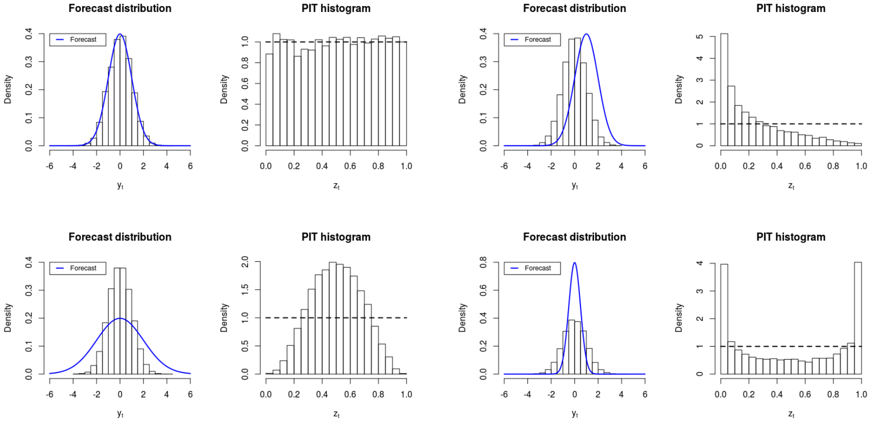
- The *probability integral transform* (PIT) z_t of an observation y_t is defined as,

$$z_t = \int_{-\infty}^{y_t} \hat{f}_t(u)\mathrm{d}u$$

- Where $\hat{f}_t$ is the forecast density. If $\hat{f}_t$ is correctly calibrated, then $z_t \sim U(0,1)$.

PIT histograms

- In practice, we inspect the histogram of $\{z_t\}$, i.e. the PIT histogram:



Overview

Scoring Rules - overview

- *Scoring rules* are summary measures intended to provide a forecaster with a form of evaluation of his og her forecasting model.

Definition (Scoring rule)

Consider a forecaster that wishes to forecast an event. If the forecaster issues the prediction f and the event x materializes, then his or her reward is $S(f, x)$ under the scoring rule S.

Note: we choose scoring rules to be *negatively* oriented. i.e. we prefer the model with the lowest score.

- When evaluating a model w.r.t. any scoring rule, the usual key measure is the average score:

$$\overline{S}(f, y) = \frac{1}{N} \sum_{i=1}^{N} S(f, y_i)$$

Logarithmic score

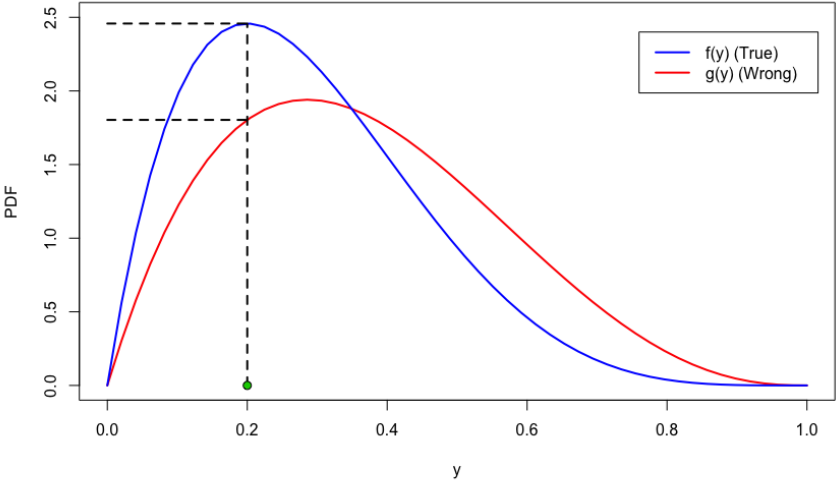
- One of the most intuitive scoring rules is the *logarithmic score* which is based on the PDF of the forecasted variable, $f(y)$, and is defined as:

$$\mathrm{LogS}(f, y) = -\log(f(y))$$

- The log score rewards observations in terms of their likelihood of occurring under the evaluated model.

Logarithmic score

- LogS visualized:



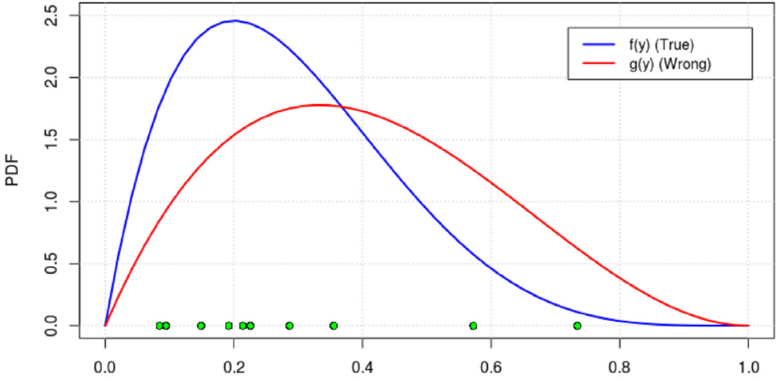
For 10 observations

- True model

$$\overline{LogS}(f, y) = -0.34$$

- Wrong model

$$\overline{LogS}(g, y) = -0.23$$



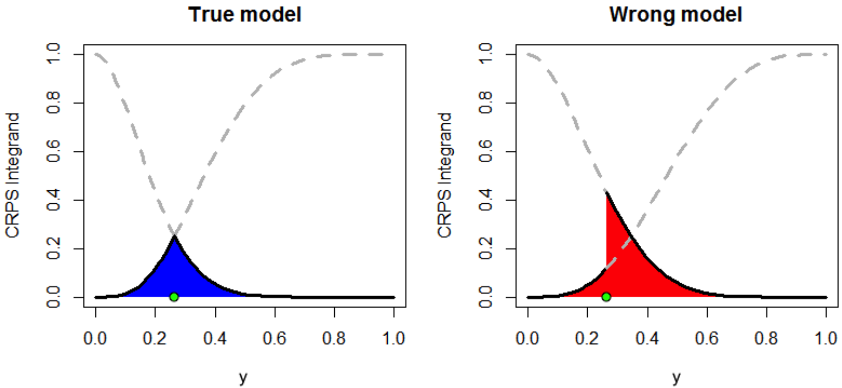
CRPS

- The *continuous ranked probability score* (CRPS) is based on the cumulative distribution function (CDF) of the forecasted variable, $F(y)$, and is defined as:

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(x) - \mathbb{I}(x \geq y))^2 \mathrm{d}x$$

- The CRPS rewards observations close to the median of the evaluated model.

- CRPS visualized:



Overview

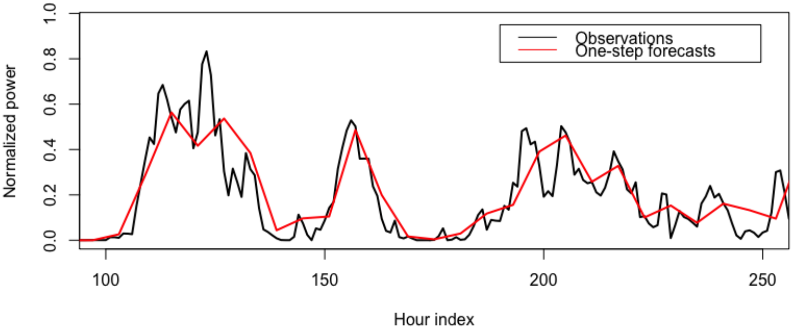
Case Study - Overview

We consider normalized wind power production at the danish Klim power plant and analyze a data set featuring,

- **Observations:** Hourly average normalized wind power production, y_t for a total of 15600 hours.
- **Predictions:** Multivariate 48-hour point forecasts, $\tilde{\boldsymbol{p}}_{\boldsymbol{t}} = (\tilde{p}_{t+1|t}, \tilde{p}_{t+2|t}, ..., \tilde{p}_{t+48|t})$ issued every 6th hour.

Case Study - Overview

- A subset of the Klim power plant data set (focus on 1-hour forecasts):



Case Study - Modelling

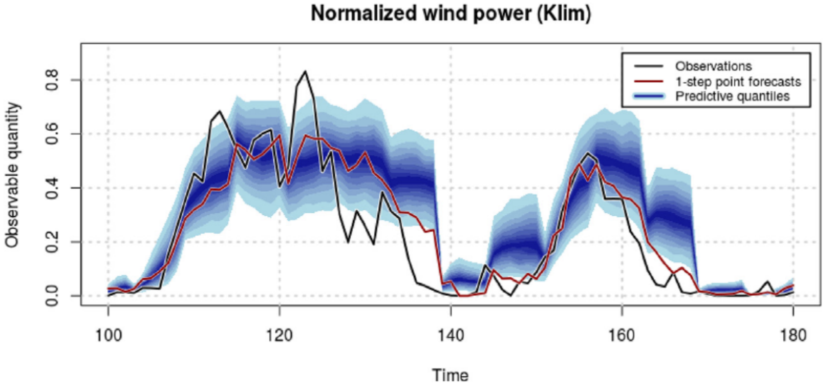
- Møller et. al. (2015) attempted to model this data by deriving and proposing the following stochastic differential equation (SDE):

$$\mathrm{d}Y_t = -\theta(Y_t - \tilde{p}_t - c\tilde{p}_t(1-\tilde{p}_t)(1-2Y_t))\mathrm{d}t + 2\sqrt{\theta\alpha\tilde{p}_t(1-\tilde{p}_t}\mathrm{d}W_t$$

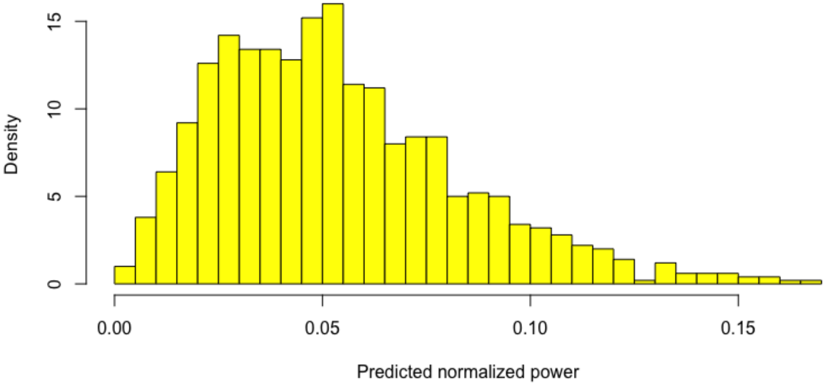
- We can use this equation to generate probabilistic forecast ensembles and subsequently estimate the forecast density $f_{t+h|t}(y_{t+h})$.
  In the following, we will be lazy and simply denote this $\hat{f}(y_t)$, where t is really $t+h|t$ and it is assumed that we consider only one specific horizon.

Case Study - Modelling

- Generated probabilistic forecasts from simulations of the SDE (6-dimensional predictions):



- An example of a forecast density for some time t and forecast horizon h obtained by simulating from the SDE model:



Case Study - Modelling

- The forecast densities generated by the SDE are double-bounded (0,1) and generally asymmetric. This suggests that the PDF may be approximated by the beta distribution:

$$\hat{f}(y_t) = \frac{1}{\mathrm{B}(\alpha,\beta)} y_t^{\alpha-1}(1-y_t)^{\beta-1}$$

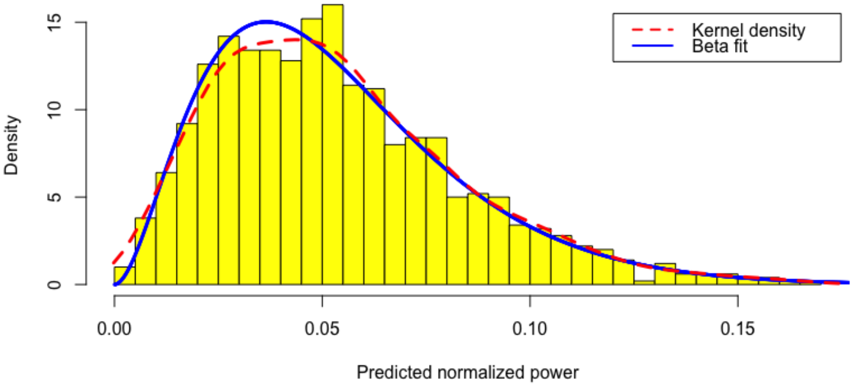
- Alternatively, we could model the forecast density with a kernel density estimate (here h denotes bandwith, not horizon):

$$\hat{f}(y_t) = \frac{1}{n}\sum_{i=1}^{N}\frac{1}{N}K\left(\frac{y_t - y_i}{h}\right)$$

- See [insert slideshowno.] for different choices of $K(\cdot)$.

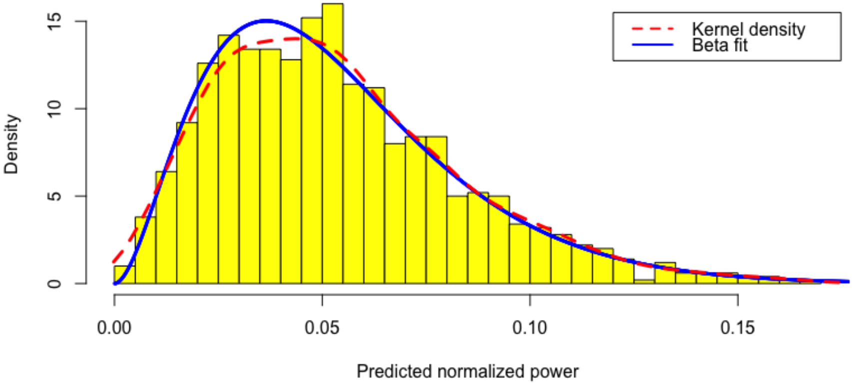
Case Study - Modelling

- The example from before with the two considered models:



How do the models compare?

- Which model should be preferred in *your* opinion?
- What do you think the *scoring rules* will conclude?



Case Study - Results

- Logarithmic scores:

Horizon	Beta	Kernel
1 hour	3.297	-1.632
2 hours	3.914	-1.084
3 hours	3.989	-0.772

- CRPS:

Horizon	Beta	Kernel
1 hour	0.040	0.039
2 hours	0.060	0.059
3 hours	0.073	0.071

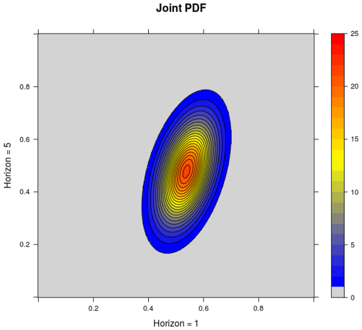
Most important lesson

- LogS penalizes very hard on observations that are unlikely under a model that may otherwise look reasonable. Is this a problem?
- CRPS finds similar models to be similar (robustness) but will more or less ignore differences in the extremes.

Overview

Why consider multivariate forecast evaluation?

- Real processes have often significant temporal or spatial correlation, which is not accounted for by univariate evaluation methods.



- Note: published research is currently limited and scattered.

Multivariate probability integral transform

- We have already defined the PIT for a univariate time series,

$$z_t = \int_{-\infty}^{y_t} \hat{f}_t(u)\mathrm{d}u$$

- A multivariate analog of this is to split the joint density $f(\boldsymbol{y}_t)$ into all possibles chains of conditional densities. For a 2-dimensional forecast, that is:

$$f(\boldsymbol{y}_t) = f(y_{2,t}|y_{1,t})f(y_{1,t})$$
$$f(\boldsymbol{y}_t) = f(y_{1,t}|y_{2,t})f(y_{2,t})$$

- If the forecast is correctly calibrated, then the PIT of each of those four univariate densities must be $U(0,1)$.

Multivariate scoring rules

- Consider the multivariate stochastic variable, $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_N)^T$. If $\boldsymbol{X}$ is a forecast of $\boldsymbol{Y}$, then we can evaluate it with the multivariate analogs of the logarithmic score and the CRPS.
- Multivariate logarithmic score:

$$\mathrm{LogS}(f_{\boldsymbol{X}}, \boldsymbol{y}) = -\log(f_{\boldsymbol{X}}(\boldsymbol{y}))$$

- Multivariate CRPS:

$$\mathrm{CRPS}(F_{\boldsymbol{X}}, \boldsymbol{y}) = \int_{-\infty}^{\infty} (F_{\boldsymbol{X}}(\boldsymbol{u}) - \mathbb{I}(\boldsymbol{u} \geq \boldsymbol{y}))^2 \mathrm{d}\boldsymbol{u}$$

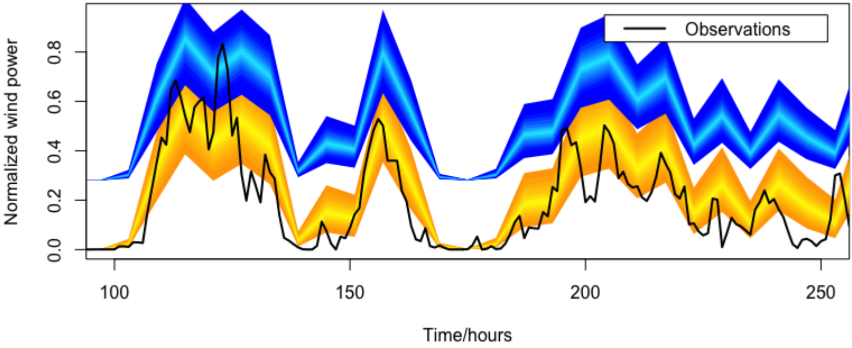
Variogram score

- Scheuerer et. al. (2015) showed that multivariate CRPS has limitations in terms of detection of misspecified correlation.
- To better address this property of a probabilistic forecast, the *variogram score* of order p was suggested,

$$\mathrm{VarS}_p(f_{\boldsymbol{X}}, \boldsymbol{y}) = \sum_{i,j=1}^{n} w_{ij}(|y_i - y_j|^p - E[|X_i - X_j|^p])^2$$

Variogram score

- The variogram score is indeed naturally sensitive to correlation structure, but it cannot detect a misspecified mean:



- Both the blue and yellow forecast density series will yield exactly the same variogram score!

Numerical implementation

How do we actually compute these scores for higher dimensions in general?

- **Log Score:** Need to estimate the PDF with a kernel density estimator (R offers `kde`, which supports 1-6 dimensional densities).
- **CRPS:** Need to estimate the CDF, which is easy, use e.g. the ECDF:

$$\hat{\boldsymbol{F}}(\boldsymbol{y}) \sum_{i=1}^{N} \mathbb{I}(\boldsymbol{x} \leq \boldsymbol{y})$$

  As a downside we need to perform multidimensional integration afterwards. One reasonable option is to use R's `vegas` which exploits importance sampling rather than usual integration methods.

- **Variogram Score:** Simple arithmetic operations.

Summary/pros and cons

- Advantages and disadvantages of the scoring rules considered:

	LogS	**CRPS**	**VarS**
Calibration	Good	Good	Useless
Correlation structure detection	Good	Useless	Good
Computation speed (non-parametric)	Okay*	Slow	Fast
Computation speed (parametric)	Fast	Fast	Fast
Scalability w.r.t dimension	Okay	Bad	Good

Table: *Computation speed is fast for 1 dimension, but slows down quickly with additional dimensions, hence categorized as 'okay'.

Recommendation

Our overall recommendation for evaluation of multivariate probabilistic forecasts is:

- Apply VarS to the full, multivariate forecast, while
- Simultaneously evaluating its marginal densities by either univariate CRPS or LogS, depending on whether the shapes of the tails are important (LogS) or not (CRPS).

Final comment

- Hence, we don't have one superior proper scoring rule that accounts for everything. We must instead decide what to put emphasis on in a given situation.

By the way

- In case you want to evaluate a quantile forecast: let I be the interval (u, l) such that this interval exactly encloses a $(1-\alpha)\cdot 100\%$ prediction interval. Then the *interval score* (IS) is defined as,

$$\mathrm{IS}(y) = (u-l) + \frac{2}{\alpha}(l-y)\mathbb{I}(y<l) + \frac{2}{\alpha}(y-u)\mathbb{I}(y>u)$$

- Long story short, narrow prediction intervals are rewarded under IS, as are observations that fall inside the prediction interval. Ezpz.

References

Bjerregård, Mathias Blicher, Møller, Jan Kloppenborg, and Madsen, Henrik (2021)
Introduction to Multivariate Probabilistic Forecast Evaluation
Energy and AI, 10.1016 (2021)

Gneiting, Tilmann, and Raftery, Adrian E.(2007)
Strictly Proper Scoring Rules, Prediction and Estimation
Journal of the American Statistical Association 102.477 (2007): 359-378.

Gneiting, Tilmann (2010)
Making and evaluating point forecasts
Journal of the American Statistical Association 106.494 (2011): 746-762.

Møller, Jan Kloppenborg, Zugno, Marco, and Madsen, Henrik (2015)
Probabilistic Forecasts of Wind Power Generation by Stochastic Differential Equation Models
Journal of Forecasting 35 (2016): 189-205.

Scheuerer, Michael, and Thomas M. Hamill (2015)
Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities
Monthly Weather Review 143.4 (2015): 1321-1334.