

Evaluation and criticism of multivariate probabilistic forecasts

Julia Braun Leonhard Held

University of Zurich

Luzern, November 2007

Outline

- 1 Introduction
- 2 General principles
- 3 Measures
- 4 Conclusion and Outlook

Predictions

Major purpose of statistical modelling:
Forecasts for future observations

Examples: weather and climate forecasts, inflation report, financial risk, prediction of cancer rates or mortality,...

Forecasts try to reduce uncertainty about the future, but they will never be perfect.

⇒ Need for suitable measures of the uncertainty.

Predictive distribution

One possibility: Comparison of the point prediction and the value that later materializes

Problem: Does not take into account uncertainty. Use whole predictive distribution.

Key quantity in a Bayesian context:

Posterior predictive distribution

$$f(y|\mathbf{x}) = \int f(y|\theta, \mathbf{x})f(\theta|\mathbf{x})d\theta$$

Predictive distribution

Two main tasks:

Sharpness

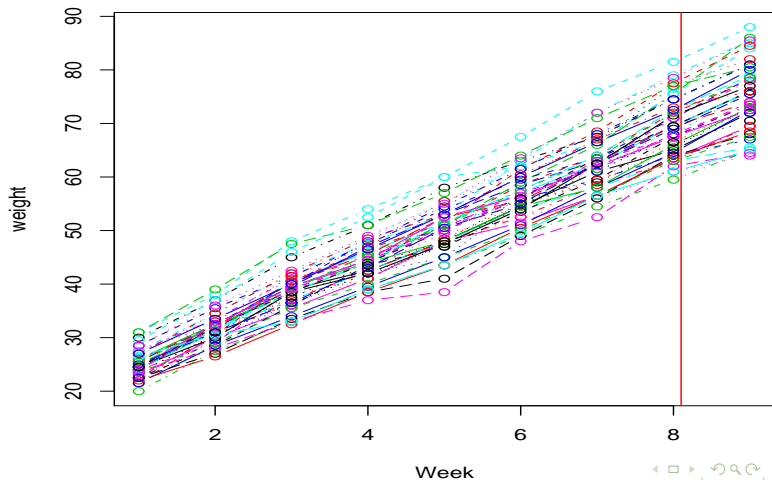
- Property of the predictions
- Refers to the concentration of the predictive distribution

Calibration

- Joint property of the predictive distribution and the real data
- Agreement of the true values and the chosen predictive distribution

Example

Weight of 48 pigs over 9 weeks (Diggle, 2002)



Models for pigs' weight

Model 1: Linear model

Model 2: Linear model with random intercept

Model 3: Linear model with random intercept and random slope

In all models: time as explanatory variable
Bayesian inference using MCMC

Quantitative assessment of probabilistic forecasts

Model evaluation

Comparing alternative models based on the predictive distribution and the true value

Model criticism

Assessing the agreement of one model with external data

Model evaluation

Scoring rules

- Sometimes also called scoring functions
- Measure for the quality of forecasts
- Numerical value based on the predictive distribution and the true value that arised later
- Normally positively oriented
- Cover both sharpness and calibration

Model evaluation

Propriety

- Proper scores: Expected value of the score is maximal if the observation is derived from the predictive distribution F .
- Strictly proper scores: Expected value has only one maximum.
- Interpretation: Proper scores do not lead the forecaster to turn away from his true belief.
- Strictly proper scores penalize such an alteration.
- The mean of proper scores is also proper.

Model criticism

- No alternative model assumptions necessary
- Helps to detect and maybe correct inappropriate models

Prequential principle (Dawid, 1984):

A measure of agreement between a predictive distribution and the real values should depend on the distribution only through the sequence of predictions.

Calculation with MCMC methods

- Calculation of many scores requires the predictive density $f(y|\mathbf{x})$.
- In most cases: predictive density unknown.
- Solution: MCMC methods
- Gibbs sampling algorithm: Sample iteratively from full conditional distributions
- Samples $\theta^{(1)}, \dots, \theta^{(N)}$ are available from posterior distribution
- For each set of model parameters $\theta^{(n)}$ we additionally draw a value for $y^{(n)}$.

Monte-Carlo estimation

$$\hat{f}(y|\mathbf{x}) = \frac{1}{N} \sum_{n=1}^N f(y|\theta^{(n)}, \mathbf{x})$$

Logarithmic score

Logarithmic score

$$\text{LogS}(Y, y_{obs}) = \log f(y_{obs}|\mathbf{x})$$

Properties

- Simple idea: Logarithmic density at the in fact observed value
- Problem: Very sensitive to outliers and extreme events
- Results for very small densities sometimes not computable
- At the same time: Not sensitive to distance! Predictive density $f(y|\mathbf{x})$ is only evaluated at y_{obs} . All other values not taken into account.

Calculation of the logarithmic score

Estimation with the estimated predictive density:

Estimated logarithmic score

$$\widehat{\text{LogS}}(Y, y_{obs}) = \log \hat{f}(y_{obs}|\mathbf{x})$$

Results for the example:

	Model 1	Model 2	Model 3
Univariate	-20.787	-3.210	-2.446
Multivariate	-Inf	-151.622	-143.910

Univariate: Mean of univariate scores for each pig

Multivariate: Calculate score for the joint distribution of all pigs

Spherical score

Spherical score

$$\text{SphS}(Y, y_{\text{obs}}) = \frac{f(y_{\text{obs}}|\mathbf{x})}{\sqrt{\int_{-\infty}^{\infty} f(y|\mathbf{x})^2 dy}}$$

Properties

- Measure of distance between the forecast and the truth.
- Problem: Computation can be difficult and time-consuming, often unfeasible in the multivariate case.

Calculation of the spherical score

Estimated spherical score

- Problem: Integral of $\hat{f}(y|\mathbf{x})^2$ in the denominator
- Numerical solution: Newton-Cotes formulas
- Samples $y^{(n)}$ serve as supporting points
- Approximation of the value of the integral between two consecutive supporting points (three different versions)
- Sum of these approximations
- Results indistinguishable for different versions of Newton-Cotes

Results for the example:

	Model 1	Model 2	Model 3
Univariate	0.322	0.722	0.817

Continuous ranked probability score (CRPS)

Continuous ranked probability score

$$\begin{aligned} CRPS(Y, y_{obs}) &= - \int_{-\infty}^{\infty} (P(Y \leq t) - \mathbf{1}(y_{obs} \leq t))^2 dt \\ &= \frac{1}{2} E|Y - Y'| - E|Y - y_{obs}|. \end{aligned}$$

where Y and Y' are independent realisations from $f(y|\mathbf{x})$.

Integral over the Brier score for binary predictions at all possible thresholds t .

Generalizations of the CRPS

Energy Score

$$ES(Y, y_{obs}) = \frac{1}{2} E|Y - Y'|^\alpha - E|Y - y_{obs}|^\alpha$$

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

Multivariate energy score

$$ES(Y, y_{obs}) = \frac{1}{2} E\|Y - Y'\|^\alpha - E\|Y - y_{obs}\|^\alpha$$

where $\|\cdot\|$ denotes the Euclidean norm.

Calculation and properties

Estimation of the energy score

- $ES(Y, y_{obs}) = \frac{1}{2} E|Y - Y'|^\alpha - E|Y - y_{obs}|^\alpha$.
- Split samples for $y^{(n)}$ in two parts $y^{(n)}$ and $y'^{(n)}$.
- As they are far enough apart, they can be seen as independent.

Comments:

- Quick calculation because $f(y|\mathbf{x})$ not needed directly
- Alternative calculations possible, for example all possible differences,...
- No recommendations concerning the choice of α

Example

Results for the example:

	Model 1	Model 2	Model 3
Univariate CRPS	-3.753	-2.093	-1.099
Univariate ES ($\alpha = 0.5$)	-1.284	-0.954	-0.677
Multivariate CRPS	-31.749	-18.57	-9.807
Multivariate ES ($\alpha = 0.5$)	-4.03	-3.115	-2.216

Tools for model criticism

Probability integral transform (PIT)

$$p_{PIT} = F(y_{obs}|\mathbf{x})$$

- F is the distribution function of the posterior predictive density.
- If F is continuous and the observation comes from F , the PIT value is uniformly distributed on $(0, 1)$.
- Estimation by evaluating $\frac{1}{N} \sum_{n=1}^N \mathbf{1}(y^{(n)} \leq y_{obs})$.
- Check: Plotting the histogram for several PIT values or testing for uniform distribution.
- Disadvantage: Only possible for univariate distributions.

Tools for model criticism

Box's predictive p-value (Box, 1980)

$$p_{\text{Box}} = P\{f(Y|\mathbf{x}) \leq f(y_{\text{obs}}|\mathbf{x})|\mathbf{x}\}$$

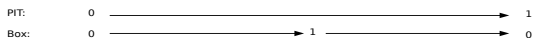
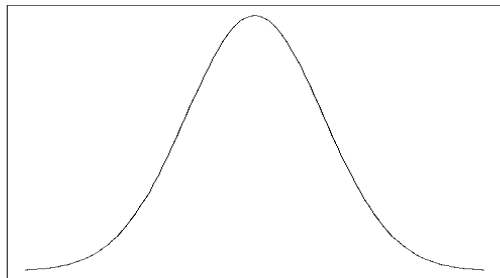
- $f(Y|\mathbf{x})$ is a function of the random variable $Y \sim f(y|\mathbf{x})$.
- Also uniformly distributed on $(0, 1)$.
- Estimation using the estimated predictive density:

$$\hat{p}_{\text{Box}} = \frac{1}{N} \sum_{n=1}^N \mathbf{1}(\hat{f}(y^{(n)}|\mathbf{x}) \leq \hat{f}(y_{\text{obs}}|\mathbf{x})).$$

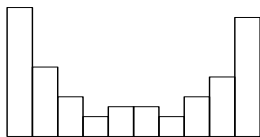
Relation

For symmetric and unimodal distributions:

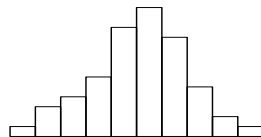
$$p_{Box} = 1 - 2|p_{PIT} - 0.5|$$



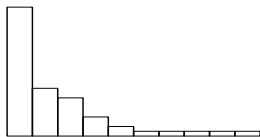
Histograms



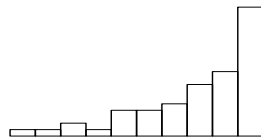
PIT



PIT

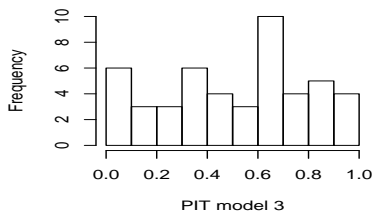
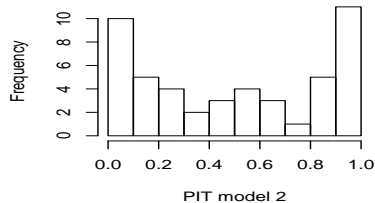
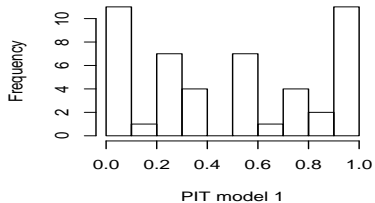


Box

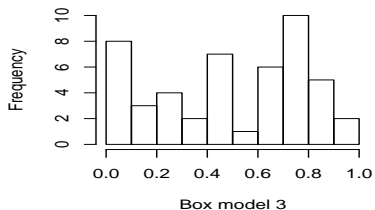
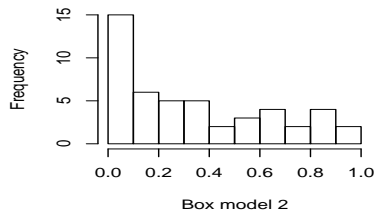
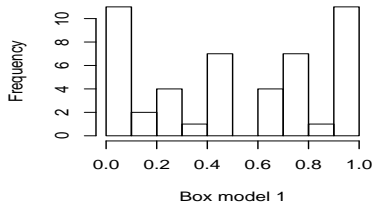


Box

Histograms of the pigs' PIT values



Histograms of the pigs' Box's p-values



Multivariate Box's p-values

- Applicable for multivariate data.
- Calculation time-consuming

Results for the example:

Model 1	Model 2	Model 3
0	0	0.087

Numerical problems?

Conclusion and Outlook

Useful methods for model comparison and criticism, but:

- no general recommendation which score to use,
- computation can be time consuming,
- probably numerically instable for multivariate data,
- comparison with analytical results from lme-predictions,
- multivariate application needs more exploration,
- assessment of Monte Carlo error necessary,
- performance of the different scores has to be studied further.

References

Box, G.E.P. (1980). Sampling and Bayes' inference in scientific modelling and robustness, *Journal of the Royal Statistical Society, Series A* **143**, 383-430.

Dawid, A.P. (1984). Statistical theory: The prequential approach, *Journal of the Royal Statistical Society, Series A* **147**, 278-292.

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

Diggle, J.P., Heagerty, P., Liang, K.Y., Zeger, S.L. (2002). *Analysis of Longitudinal Data* (second edition). Oxford University Press.

Gneiting, T., Raftery, A.F. (2007). Strictly proper scoring rules, prediction and estimation, *Journal of the American Statistical Association* **102**, 359-378.