Monte Carlo estimation techniques for model evaluation and criticism in Bayesian hierarchical models

Julia Braun Leonhard Held

University of Zurich

Reisensburg, September 2007

Outline

Introduction

One purpose of statistical modelling:
Forecasts for future observations

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

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

- Numerical value based on the predictive distribution and the true value that arised later
- Normally positively oriented, but also possible as penalty (see example 3)
- Cover both sharpness and calibration
- Proper scores: Expected value of the score is maximal if the observation is derived from the predicitive 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.

Proper scores for continuous responses

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 $f(y|\mathbf{x})$.

Proper scores for continuous responses

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

Proper scores

Logarithmic score

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

Spherical score

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

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.

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 $(0, 1)$.
- 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

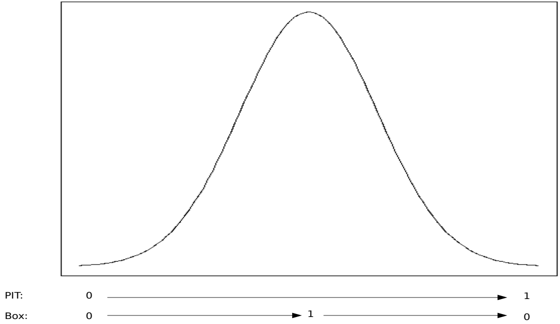
$$p_{Box} = P\{f(Y|\mathbf{x}) \leq f(y_{obs}|\mathbf{x})|\mathbf{x}\}$$

- $f(Y|\mathbf{x})$ is a function of the random variable $Y \sim f(y|\mathbf{x})$.
- Also uniformly distributed on $(0, 1)$.
- Applicable for multivariate data.

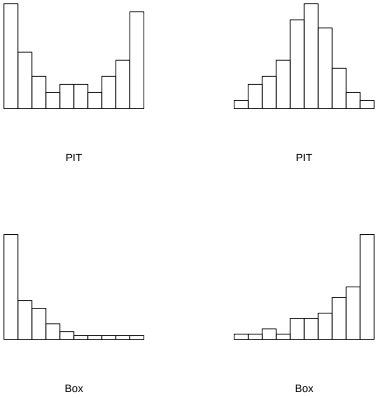
Relation

For symmetric and unimodal distributions:

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



Histograms



Calculation with MCMC methods

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

Monte-Carlo estimation

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

Estimation

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.
- Alternative calculations possible, for example all possible differences,...

PIT value

- $p_{PIT} = F(y_{obs}|\mathbf{x})$
- Estimation by evaluating $\frac{1}{N}\sum_{n=1}^{N} \mathbf{1}(y^{(n)} \leq y_{obs})$.

Estimation

For the other measures: $\hat{f}(y_{obs}|\mathbf{x})$ needed.

Logarithmic score

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

Box's p-value

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

Estimation

Spherical score

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

Toy example

Artificial data set by O'Hagan (2003):

Group	Observations						Sample mean
1	2.73	0.56	0.87	0.90	2.27	0.82	1.36
2	1.60	2.17	1.78	1.84	1.83	0.80	1.67
3	1.62	0.19	4.10	0.65	1.98	0.86	1.57
4	0.96	1.92	0.96	1.83	0.94	1.42	1.34
5	6.32	3.66	4.51	3.29	5.61	3.27	4.44

Bayesian hierarchical models

Model 1: Bayesian linear model

$$
\begin{aligned}
y_{ij}|\mu, \sigma^2 &\sim N(\mu, \sigma^2),\\
\mu &\sim N(2, 10),\\
\sigma^2 &\sim IG(10, 11).
\end{aligned}
$$

Model 2: Random intercept

$$
\begin{aligned}
y_{ij} \mid \lambda_i, \sigma^2 &\sim N(\lambda_i, \sigma^2),\\
\lambda_i \mid \mu, \tau^2 &\sim N(\mu, \tau^2),\\
\mu &\sim N(2, 10),\\
\sigma^2 &\sim IG(10, 11),\\
\tau^2 &\sim IG(10, 3).
\end{aligned}
$$

Univariate results

Mean scores:

	CRPS	ES ($\alpha = 0.5$)	LogS	SphS
Model 1	-0.73	-0.56	-1.64	0.97
Model 2	-0.38	-0.41	-1.20	1.29

P-values:

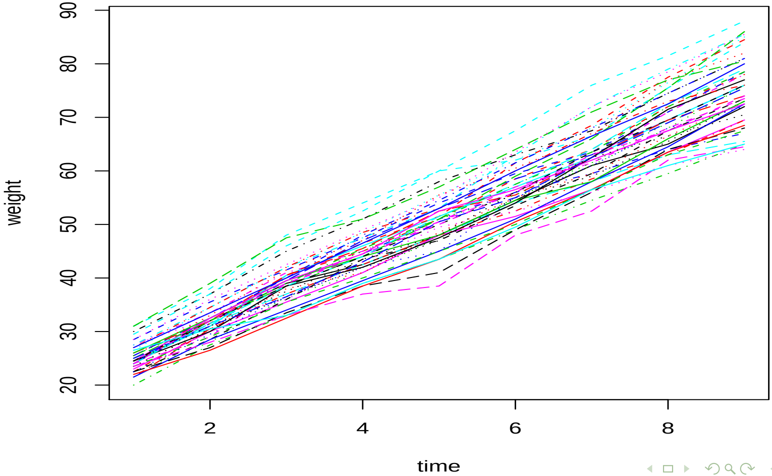
	Model 1		Model 2	
Group	PIT	Box	PIT	Box
1	0.165	0.325	0.210	0.431
2	0.163	0.316	0.154	0.318
3	0.174	0.344	0.191	0.373
4	0.289	0.575	0.420	0.850
5	0.772	0.452	0.322	0.630

Multivariate results

Multivariate:

Model	CRPS	ES ($\alpha = 0.5$)	LogS	Box
1	-1.881	-0.961	-8.766	0.447
2	-1.332	-0.811	-6.646	0.763

Pigs' weight (Diggle, 2002)



Models

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

Results

Average univariate scores:

	CRPS	ES ($\alpha = 0.5$)	LogS	SphS
Model 1	-3.753	-1.284	-20.787	0.322
Model 2	-2.093	-0.954	-3.210	0.722
Model 3	-1.099	-0.677	-2.446	0.817

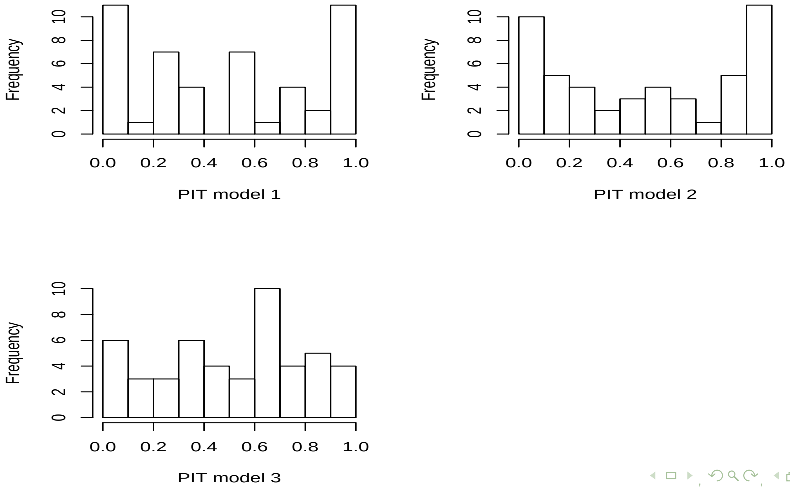
Multivariate scores:

Model	CRPS	ES ($\alpha = 0.5$)	LogS
1	-31.749	-4.03	-Inf
2	-18.57	-3.115	-151.622
3	-9.807	-2.216	-143.910

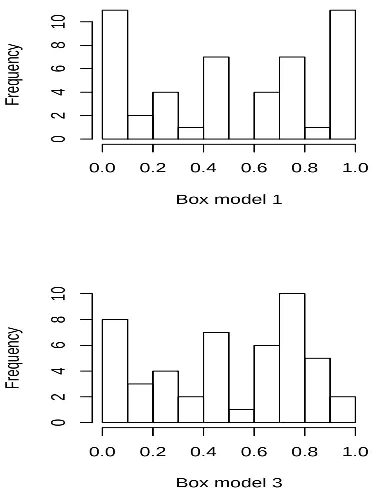
Multivariate Box's p-values:

Model 1	Model 2	Model 3
0	0	0.087

Histograms of the PIT values



Histograms of the Box's p-values



Larynx cancer in Germany

General information

- Larynx cancer data from Germany from the years 1952-2002
- Analysis of mortality counts using the age-period-cohort (APC) model
- Age groups under 30 often excluded from analysis because of low counts
- Suggestion of Baker and Bray (2005): Age-specific predictions based on full data might be more precise.
- Use of scoring rules to check this statement
- In this case: scoring rules negatively oriented

Data analysis

Age-period-cohort model

- n_{ij}: Number of persons at risk in age group i and year j
- Number of deaths in age group i and year j binomially distributed with parameters n_{ij} and π_{ij}
- Additive decomposition of the logarithmic odds η_{ij} in overall level μ, age effects θ_i, period effects ϕ_j and cohort effects ψ_k:

$$\eta_{ij} = \log\{\frac{\pi_{ij}}{1-\pi_{ij}}\} = \mu + \theta_i + \phi_j + \psi_k$$

Fitted models

Four predictive models:

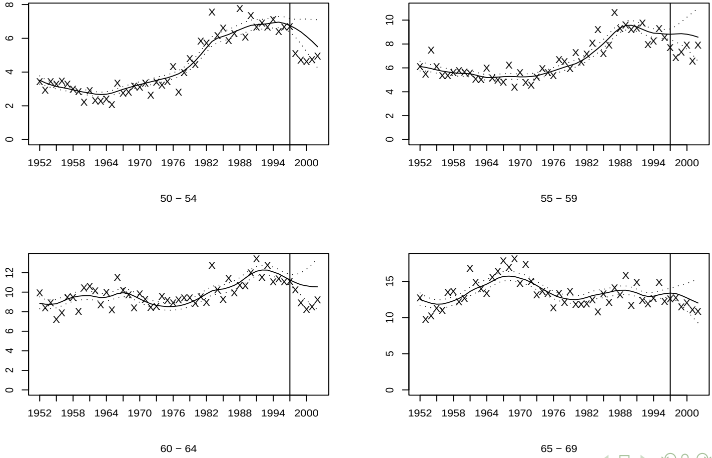
- Model 1: all age groups; overdispersion
- Model 2: all age groups; no overdispersion
- Model 3: only age groups over 30; overdispersion
- Model 4: only age groups over 30; no overdispersion

Predictions of mortality counts for 1998-2002, 12 age groups

Non-parametric smoothing priors within a hierarchical Bayesian framework

Number of deaths

Observed and fitted/predicted number of deaths per 100,000 males, based on model 4:



Scores

Scores for count data

- Logarithmic score: $\mathrm{LogS}(P, y_{obs}) = -\log p_{y_{obs}}$
- Spherical score: $\mathrm{SphS}(P, y_{obs}) = -p_{y_{obs}} / \|p\|$
- Ranked probability score:
  $\mathrm{RPS}(P, y_{obs}) = E_P|Y - y_{obs}| - \frac{1}{2} E_P|Y - Y'|$
- Additionally: Squared error score:
  $\mathrm{SqES}(P, y_{obs}) = (y_{obs} - \mu_p)^2$

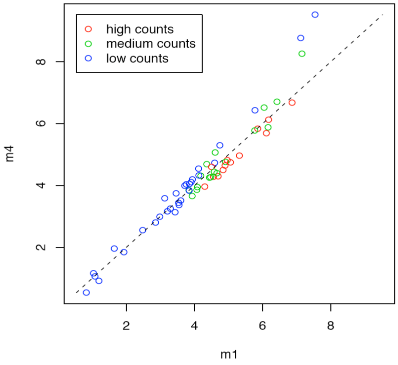
Model	age	disp	LogS	SphS	RPS	SqES
1	+	+	**4.27**	**−0.153**	14.0	852.9
2	+	−	4.35	−0.152	12.9	684.4
3	−	+	4.29	−0.152	14.2	870.0
4	−	−	4.35	−0.151	**12.2**	**564.8**

Explanation

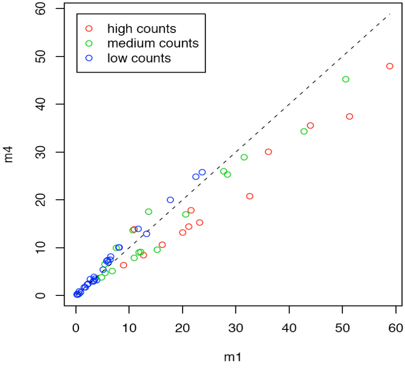
Disagreement of the scores

- LogS and SphS roughly independent of size of counts
- RPS and SqES highly dependent on the size of the counts
- Few high count cases dominate differences in the mean score.
- Better fit of model 4 in mid age groups.
- Model 1 to prefer in younger and older age groups
- As counts are especially high in mid age groups: Greater weight in the mean of RPS and SqES.

Illustrative graphic



Logarithmic score



Ranked probability score

Conclusion and Outlook

Useful methods for model comparison and criticism, but:

- computation can be time consuming,
- probably numerically instable for multivariate data,
- multivariate application needs more exploration,
- assessment of Monte Carlo error necessary,
- performance of the different scores has to be studied further.

References

Baker, A., Bray, I. (2005). Bayesian projections: What are the effects of excluding data from younger age groups? *American Journal of Epidemiology* **162**, 798-805.

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.

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.

O'Hagan, A. (2003). HSSS model criticism. *in* Green, P.J., Hjort, N.L., Richardson, E.S. (ed.), *Highly Structured Stochastic Systems*, Oxford University Press, 423-444.