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

Julia Braun    Leonhard Held

University of Zurich

Reisensburg, September 2007
Outline

1. Introduction
2. Model evaluation and model criticism
3. Calculation with MCMC methods
4. Examples
5. Conclusion and Outlook
Introduction

One purpose of statistical modelling:
Forecasts for future observations

Key quantity in a Bayesian context:

Posterior predictive distribution

\[ f(y|x) = \int f(y|\theta, x) f(\theta|x) d\theta \]
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 arose 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 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.
Proper scores for continuous responses

Continuous ranked probability score

\[ CRPS(Y, y_{obs}) = -\int_{-\infty}^{\infty} (P(Y \leq t) - 1(y_{obs} \leq t))^2 dt \]

\[ = \frac{1}{2} E|Y - Y'| - E|Y - y_{obs}|. \]

where \( Y \) and \( Y' \) are independent realisations from \( f(y|x) \).
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**

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

**Spherical score**

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

\[ p_{PIT} = F(y_{obs} | x) \]

- \( F \) 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

PIT PIT
Box Box
Calculation with MCMC methods

- In most cases: predictive density $f(y|x)$ 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 additionally draw a value for $y^{(n)}$.

Monte-Carlo estimation

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

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

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

**Logarithmic score**

\[
\hat{\text{LogS}}(Y, y_{obs}) = \log \hat{f}(y_{obs}|x)
\]

**Box’s p-value**

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

Spherical score

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

- Problem: Integral of \( \hat{f}(y|x)^2 \) 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):

<table>
<thead>
<tr>
<th>Group</th>
<th>Observations</th>
<th>Sample mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.73 0.56 0.87 0.90 2.27</td>
<td>0.82 1.36</td>
</tr>
<tr>
<td>2</td>
<td>1.60 2.17 1.78 1.84 1.83</td>
<td>0.80 1.67</td>
</tr>
<tr>
<td>3</td>
<td>1.62 0.19 4.10 0.65 1.98</td>
<td>0.86 1.57</td>
</tr>
<tr>
<td>4</td>
<td>0.96 1.92 0.96 1.83 0.94</td>
<td>1.42 1.34</td>
</tr>
<tr>
<td>5</td>
<td>6.32 3.66 4.51 3.29 5.61</td>
<td>3.27 4.44</td>
</tr>
</tbody>
</table>
Bayesian hierarchical models

Model 1: Bayesian linear model

\[ y_{ij} | \mu, \sigma^2 \sim N(\mu, \sigma^2), \]
\[ \mu \sim N(2, 10), \]
\[ \sigma^2 \sim IG(10, 11). \]

Model 2: Random intercept

\[ y_{ij} | \lambda_i, \sigma^2 \sim N(\lambda_i, \sigma^2), \]
\[ \lambda_i | \mu, \tau^2 \sim N(\mu, \tau^2), \]
\[ \mu \sim N(2, 10), \]
\[ \sigma^2 \sim IG(10, 11), \]
\[ \tau^2 \sim IG(10, 3). \]
Univariate results

Mean scores:

<table>
<thead>
<tr>
<th></th>
<th>CRPS</th>
<th>ES ($\alpha = 0.5$)</th>
<th>LogS</th>
<th>SphS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>−0.73</td>
<td>−0.56</td>
<td>−1.64</td>
<td>0.97</td>
</tr>
<tr>
<td>Model 2</td>
<td>−0.38</td>
<td>−0.41</td>
<td>−1.20</td>
<td>1.29</td>
</tr>
</tbody>
</table>

P-values:

<table>
<thead>
<tr>
<th>Group</th>
<th>Model 1</th>
<th></th>
<th>Model 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PIT</td>
<td>Box</td>
<td>PIT</td>
<td>Box</td>
</tr>
<tr>
<td>1</td>
<td>0.165</td>
<td>0.325</td>
<td>0.210</td>
<td>0.431</td>
</tr>
<tr>
<td>2</td>
<td>0.163</td>
<td>0.316</td>
<td>0.154</td>
<td>0.318</td>
</tr>
<tr>
<td>3</td>
<td>0.174</td>
<td>0.344</td>
<td>0.191</td>
<td>0.373</td>
</tr>
<tr>
<td>4</td>
<td>0.289</td>
<td>0.575</td>
<td>0.420</td>
<td>0.850</td>
</tr>
<tr>
<td>5</td>
<td>0.772</td>
<td>0.452</td>
<td>0.322</td>
<td>0.630</td>
</tr>
</tbody>
</table>
Multivariate results

Multivariate:

<table>
<thead>
<tr>
<th>Model</th>
<th>CRPS</th>
<th>ES (α = 0.5)</th>
<th>LogS</th>
<th>Box</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>−1.881</td>
<td>−0.961</td>
<td>−8.766</td>
<td>0.447</td>
</tr>
<tr>
<td>2</td>
<td>−1.332</td>
<td>−0.811</td>
<td>−6.646</td>
<td>0.763</td>
</tr>
</tbody>
</table>
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:

<table>
<thead>
<tr>
<th></th>
<th>CRPS</th>
<th>ES ($\alpha = 0.5$)</th>
<th>LogS</th>
<th>SphS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>-3.753</td>
<td>-1.284</td>
<td>-20.787</td>
<td>0.322</td>
</tr>
<tr>
<td>Model 2</td>
<td>-2.093</td>
<td>-0.954</td>
<td>-3.210</td>
<td>0.722</td>
</tr>
<tr>
<td>Model 3</td>
<td>-1.099</td>
<td>-0.677</td>
<td>-2.446</td>
<td>0.817</td>
</tr>
</tbody>
</table>

Multivariate scores:

<table>
<thead>
<tr>
<th>Model</th>
<th>CRPS</th>
<th>ES ($\alpha = 0.5$)</th>
<th>LogS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-31.749</td>
<td>-4.03</td>
<td>-Inf</td>
</tr>
<tr>
<td>2</td>
<td>-18.57</td>
<td>-3.115</td>
<td>-151.622</td>
</tr>
<tr>
<td>3</td>
<td>-9.807</td>
<td>-2.216</td>
<td>-143.910</td>
</tr>
</tbody>
</table>

Multivariate Box’s p-values:

<table>
<thead>
<tr>
<th>Model</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0.087</td>
</tr>
</tbody>
</table>
Histograms of the PIT values

PIT model 1

PIT model 2

PIT model 3

Julia Braun, Leonhard Held
University of Zurich

Monte Carlo estimation techniques for model evaluation and criticism in Bayesian hierarchical models
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 $\pi_{ij}$
- Additive decomposition of the logarithmic odds $\eta_{ij}$ in overall level $\mu$, age effects $\theta_i$, period effects $\phi_j$ and cohort effects $\psi_k$:

\[
\eta_{ij} = \log\left\{ \frac{\pi_{ij}}{1-\pi_{ij}} \right\} = \mu + \theta_i + \phi_j + \psi_k
\]
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 for count data

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

<table>
<thead>
<tr>
<th>Model</th>
<th>age</th>
<th>disp</th>
<th>LogS</th>
<th>SphS</th>
<th>RPS</th>
<th>SqES</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>+</td>
<td>+</td>
<td>4.27</td>
<td>-0.153</td>
<td>14.0</td>
<td>852.9</td>
</tr>
<tr>
<td>2</td>
<td>+</td>
<td>-</td>
<td>4.35</td>
<td>-0.152</td>
<td>12.9</td>
<td>684.4</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>+</td>
<td>4.29</td>
<td>-0.152</td>
<td>14.2</td>
<td>870.0</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>-</td>
<td>4.35</td>
<td>-0.151</td>
<td>12.2</td>
<td>564.8</td>
</tr>
</tbody>
</table>

Julia Braun, Leonhard Held
University of Zurich
Monte Carlo estimation techniques for model evaluation and criticism in Bayesian hierarchical models
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

Julia Braun, Leonhard Held
Monte Carlo estimation techniques for model evaluation and criticism in Bayesian hierarchical models
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


