Quantitative assessment of probabilistic forecasts with applications in epidemiology

Leonhard Held
University of Zurich

Swiss Statistics Meeting
November 17, 2006
Overview

1. Motivating problem
   Forecasting cancer rates

2. Predictive model comparison and criticism
   Goals
   Proper scoring rules
   Tools for model criticism

3. Case study

4. Discussion
Motivating problem II: Forecasting cancer rates

- Bayesian age-period-cohort models are used increasingly to project cancer incidence and mortality rates.
- Data from younger age groups (typically age < 30 years) for which rates are low are often excluded from the analysis.
- However, a recent empirical comparison (Baker and Bray, 2005) based on data from Hungary suggests that age-specific predictions based on full data are more accurate.

→ Question how to quantify the accuracy of probabilistic forecasts
→ Forecast evaluation for model comparison
Case study: Forecasting larynx cancer rates in Germany

- We will investigate if the same conclusion can be drawn for mortality data on larynx cancer from Germany, 1952-1997.
- To assess the predictive quality of the different models, we have predicted the mortality counts in the years 1998-2002.
- For all different models, we looked at the predictions in the 12 age groups with age above 30 years.
- One-step-ahead forecasts have also been calculated.
Observed (×) and fitted/predicted number of deaths (per 100,000) for larynx cancer among males. Shown are posterior means and 90% pointwise credible intervals for the mean, calculated with software BAMP.
Age-period-cohort models

- Let $n_{ij}$ be the number of persons at risk in age group $i$ and year $j$.
- We assume that the number of deaths $y_{ij}$ has a binomial distribution with parameters $n_{ij}$ and $\pi_{ij}$.
- Decompose log odds $\eta_{ij} = \log\{\pi_{ij}/(1 - \pi_{ij})\}$ additively into an overall level $\mu$, age effects $\theta_i$, period effects $\varphi_j$ and cohort effects $\psi_k$:
  \[
  \eta_{ij} = \mu + \theta_i + \varphi_j + \psi_k.
  \]
- Problem to define cohorts because age group (in 5 year steps) and period (in 1 year steps) are not on the same grid.
- Use cohort index $k = 5 \cdot (I - i) + j$ with $I$ being the number of age groups.
Bayesian age-period-cohort models

- Linear time trends for age, period or cohort effects are too restrictive.
- Period and cohort as factors cause instability of ML estimates.
- We use nonparametric smoothing priors within a hierarchical Bayesian model.
- Adjustments for overdispersion straightforward
- Inference based on Markov chain Monte Carlo (MCMC).
- Formulation is particularly useful for prediction.
2. Predictive model comparison and criticism

- Selten (1998) states:
  "Probabilistic theories [...] need to be compared with respect to their predictive success"

- Box (1980) writes:
  "Sampling theory is needed for exploration and ultimate criticism of an entertained model in the light of current data, while Bayes’ theory is needed for estimation of parameters conditional on the adequacy of the entertained model"

  and

  "In making this predictive check it is not necessary to be specific about an alternative model"
Questions

1. **Which scores** should we use to compare predictive distributions with actually observed data?
   
   For our purposes it is not enough to look at the quality of point prediction only, e.g. at the squared prediction error (squared bias)

   \[
   \text{SBS}(Y, y_{obs}) = -(\mathbb{E}(Y) - y_{obs})^2
   \]
   
   of a probabilistic forecast \( Y \sim f(y|x) \), given past data \( x \).

2. Which tools are available for **model criticism**?

3. Can we calculate/estimate them in Bayesian hierarchical models based on Monte-Carlo samples?

4. Can we apply them to **count** data and to **multivariate** forecasts?
Scoring rules (also called scoring functions) are the key measures for the evaluation of probabilistic forecasts.

- Assign a numerical score (or reward) based on the predictive density $f(y|x)$ for the unknown quantity and of that quantity's true value $y_{obs}$, that has later materialised.
- Are usually positively oriented, i.e. the larger, the better.
- Are called proper, if they do not provide any incentive to the forecaster to digress from her true belief.
- Are called strictly proper if any such digress results in a penalty, i.e. the forecaster is encouraged to quote her true belief rather than any other predictive distribution.
Logarithmic score

- The **logarithmic score** is strictly proper and defined as

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

the log predictive density ordinate at the observed value \(y_{obs}\).

- Monte-Carlo estimation is possible if at least \(f(y_{obs}|x, \theta)\) is available:

\[
\hat{f}(y_{obs}|x) = \frac{1}{J} \sum_{j=1}^{J} f(y_{obs}|x, \theta^{(j)})
\]

here \(\theta^{(j)}\) are samples from the posterior \(f(\theta|x)\).
Selten (1998) notes that

... the logarithmic scoring rule is not really recommendable. On the one hand, it is too sensitive with respect to differences between very small probabilities, on the other hand, it is sometimes not sensitive enough, in the sense that in some situations it does not matter whether the truth is near to the prediction or far from it.

⇒ Logarithmic score is not sensitive to distance.
Continuous ranked probability score

A popular strictly proper score which is less sensitive to outliers but sensitive to distance is the so-called \textbf{continuous ranked probability score}

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

which is the integral of the \textbf{Brier scores} for binary predictions at all possible thresholds $t$. 

As Gneiting and Raftery (2006) noted, the CRPS can be written as
\[
\text{CRPS}(Y, y_{\text{obs}}) = \frac{1}{2} \mathbb{E}|Y - Y'| - \mathbb{E}|Y - y_{\text{obs}}|,
\]
here \( Y \) and \( Y' \) are independent realisations from \( f(y|x) \) (see Gneiting and Raftery, 2006, for references).

- CRPS generalizes the absolute error for deterministic forecasts.
- The above equation allows for Monte-Carlo estimation of CRPS based on samples from the predictive distribution.
- Analytic forms follow for normal predictive distributions.
The continuous ranked probability score can be generalized to the energy score

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

which is strictly proper for all $\alpha \in (0, 2)$ (Gneiting and Raftery, 2006).

In the limiting case $\alpha = 2$, $\text{ES}(Y, y_{obs})$ reduces to the squared bias score

$$\text{SBS}(Y, y_{obs}) = -(E(Y) - y_{obs})^2,$$

as

$$\frac{1}{2} E(Y - Y')^2 = \text{Var}(Y).$$
Multivariate forecasts

- Sums of strictly proper scoring rules are strictly proper.
- The logarithmic score can be also used for multivariate forecasts $\mathbf{Y}$.
- Similarly, the energy (and CRPS) score can be generalized to

$$\text{ES}(\mathbf{Y}, \mathbf{y}_{\text{obs}}) = \frac{1}{2} \mathbb{E} \|\mathbf{Y} - \mathbf{Y}'\|^\alpha - \mathbb{E} \|\mathbf{Y} - \mathbf{y}_{\text{obs}}\|^\alpha,$$

here $\|\cdot\|$ denotes the Euclidean norm and, as before, $\alpha \in (0, 2)$ is required to ensure strict propriety.
Model criticism: PIT values

- For univariate forecasts, the probability integral transform (PIT) value, defined as $p = F(y_{obs} | x)$ where $F(y | x)$ is the univariate predictive distribution function, is often used.
- If the forecast is perfect and $F$ is continuous, $p$ is uniformly distributed (David, 1984), PIT values for one-step-ahead forecasts are also independent.
- In practice one usually examines the histogram of several PIT values for departure from a uniform distribution.
- If the data are counts as in our application, a correction is possible.
Corrected PIT values

- Define randomized PIT value

\[ p = U \cdot F(y_{obs} - 1|x) + (1 - U) \cdot F(y_{obs}|x) \]

where \( U \) is standard uniform.

- If the forecast is perfect, \( p \) will still be uniformly distributed (Smith, 1985).

- Corresponding histograms can be “de-randomized” by adding rectangles with between \( F(y_{obs} - 1|x) \) and \( F(y_{obs}|x) \) with height \( 1/(F(y_{obs}|x) - F(y_{obs} - 1|x)) \) (Czado et al., 2006).
Model criticism: Box’s density ordinate $p$-value

- A less-known approach, that is also applicable to multivariate forecasts, is the calculation of Box’s density ordinate $p$-value (Box, 1980).

- Let $f(y|x)$ be the (now possibly multivariate) predictive density for external data $y$, conditional on observed data $x$. Interest is in measuring the support of the predictive density for the actually observed data $y_{obs}$ by considering

$$Q(y_{obs}) = P\{f(Y|x) \leq f(y_{obs}|x)|x\},$$

where $f(Y|x)$ is a function of the random variable $Y \sim f(y|x)$.

- Box suggested to use $Q(y_{obs})$ based on the prior-predictive distribution.
Box’s density ordinate $p$-value cont.

- If $Y \sim f(y|x)$ and $f(Y|x)$ follow a continuous distribution, $Q(Y_{\text{obs}})$, viewed as a function of the random variable $Y_{\text{obs}} \sim f(y|x)$, is also uniformly distributed.

- For count data, $Q(Y_{\text{obs}})$ is only approximately uniform, but approximation should be good (better than in the univariate case).

- This allows for multivariate model checks as done with PIT values for univariate forecasts.
Monte-Carlo estimate of Box’s density ordinate $p$-value

- If $f(y|x)$ is available, Monte-Carlo estimation of $Q(y_{obs})$ is straightforward:

$$
\hat{Q}(y_{obs}) = \frac{1}{J} \sum_{j=1}^{J} 1\{f(y^{(j)}|x) \leq f(y_{obs}|x)\},
$$

if $\theta^{(j)}$ are samples from the posterior $f(\theta|x)$ (Wei and Tanner, 1990).

- If only $f(y|x, \theta)$ is available, $f(y|x)$ can be estimated as before in a preceding step (Held, 2004):

$$
\hat{f}(y_{obs}|x) = \frac{1}{J} \sum_{j=1}^{J} f(y_{obs}|x, \theta^{(j)}).
$$
3. Case study: Forecasting larynx cancer rates

- A recent empirical comparison (Baker and Bray, 2005) based on data from Hungary suggests that age-specific projections based on full data sets are more accurate for younger age groups.
- However, to quantify the accuracy of the predictions, the squared normalized prediction error \( \frac{(E(Y) - y_{\text{obs}})^2}{\text{Var}(Y)} \) was used.
- This score is not proper, as the expected score is optimized by \( \text{Var}(Y) \rightarrow \infty \).
- We will investigate if the same conclusion can be drawn for mortality data on larynx cancer from Germany based on proper scoring rules.
We have fitted four different models:

<table>
<thead>
<tr>
<th>Model</th>
<th>all age groups</th>
<th>adjustments for overdispersion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>2</td>
<td>+</td>
<td>−</td>
</tr>
<tr>
<td>3</td>
<td>−</td>
<td>+</td>
</tr>
<tr>
<td>4</td>
<td>−</td>
<td>−</td>
</tr>
</tbody>
</table>
Results

- The table below gives averaged scores over all $12 \cdot 5 = 60$ observations.

<table>
<thead>
<tr>
<th>Score</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>LogS($Y, y_{obs}$)</td>
<td>-6.36</td>
<td>-5.74</td>
<td>-6.49</td>
<td>-5.63</td>
</tr>
<tr>
<td>CRPS($Y, y_{obs}$)</td>
<td>-14.0</td>
<td>-12.8</td>
<td>-14.3</td>
<td>-12.3</td>
</tr>
<tr>
<td>SBS($Y, y_{obs}$)</td>
<td>-851.0</td>
<td>-646.2</td>
<td>-874.3</td>
<td>-564.1</td>
</tr>
<tr>
<td>$-(E(Y) - y_{obs})^2/\text{Var}(Y)$</td>
<td>-1.66</td>
<td>-2.06</td>
<td>-1.67</td>
<td>-2.16</td>
</tr>
</tbody>
</table>

- Suggestion by Baker and Bray (2005) is not confirmed.
- Allowing for overdispersion makes predictions worse.
- Squared normalized prediction error is not suitable as a scoring rule.
Multivariate CRPS

- The multivariate CRPS over all five years and all 12 age groups are $-155.6$, $-137.4$, $-158.4$, and $-129.3$ for model 1 to 4 respectively, in accordance with the results based on the mean univariate CRPS.

- We can also calculate the multivariate CRPS for each age group, which allows for a more detailed assessment of the predictive quality of the different models:
Corrected PIT histograms for one-step-ahead forecasts

Model 1

Model 2

Model 3

Model 4
Box’s density ordinate $p$-values for one-step-ahead forecasts

<table>
<thead>
<tr>
<th>model</th>
<th>1998</th>
<th>1999</th>
<th>2000</th>
<th>2001</th>
<th>2002</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.010</td>
<td>0.001</td>
<td>0.068</td>
<td>0.086</td>
<td>0.087</td>
</tr>
<tr>
<td>2</td>
<td>0.000</td>
<td>0.000</td>
<td>0.002</td>
<td>0.029</td>
<td>0.031</td>
</tr>
<tr>
<td>3</td>
<td>0.007</td>
<td>0.002</td>
<td>0.014</td>
<td>0.148</td>
<td>0.096</td>
</tr>
<tr>
<td>4</td>
<td>0.000</td>
<td>0.000</td>
<td>0.003</td>
<td>0.024</td>
<td>0.020</td>
</tr>
</tbody>
</table>

▶ All four models are discredited.
▶ Allowing for overdispersion improves $p$-values slightly.
Scores for one-step-ahead forecasts

- The table below gives averaged scores over all 60 one-step-ahead forecasts.

<table>
<thead>
<tr>
<th>Score</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>LogS($Y, y_{obs}$)</td>
<td>-5.09</td>
<td>-4.61</td>
<td>-5.12</td>
<td>-4.60</td>
</tr>
<tr>
<td>CRPS($Y, y_{obs}$)</td>
<td>-10.5</td>
<td>-10.0</td>
<td>-10.6</td>
<td>-9.8</td>
</tr>
<tr>
<td>SBS($Y, y_{obs}$)</td>
<td>-430.3</td>
<td>-347.4</td>
<td>-429.9</td>
<td>-322.8</td>
</tr>
<tr>
<td>$-(E(Y) - y_{obs})^2/\text{Var}(Y)$</td>
<td>-1.78</td>
<td>-2.43</td>
<td>-1.83</td>
<td>-2.55</td>
</tr>
</tbody>
</table>

- Results similar to five-year forecasts
4. Discussion

▶ Squared normalized prediction error is unsuitable as a scoring rule. In fact, it is a monotone function of Box’s density ordinate $p$-value and will be approximately $\chi^2_1$-distributed.

▶ Proper scoring rules and pure model criticism tools disagree in case study.

▶ Question of significance of score difference between models.

▶ More work is needed on forecast evaluation, in particular on the compatibility of univariate and multivariate scores and multivariate model criticism tools.