Penalized loss functions for Bayesian model comparison

Josh Keller

18 April 2013
Model Selection in Bayesian Models

Bayes Factor:
- The formal solution
- Unstable with diffuse prior; undefined with improper priors

Cross-validation:
- Which model is most useful?
- Judge model by out-of-sample prediction

Posterior-Predictive approach:
- Does this model give data like what I observed?
- Simulate from posterior and compare to original data
Model Selection in Bayesian Models

Deviance Information Criterion (DIC)

\[ DIC = \overline{D} + p_D \approx \text{“Model Fit”} + \text{“Model Complexity”} \]

- Proposed by Spiegelhalter et al. (2002)
- Theoretical foundations are controversial
  - No clear generalization outside of exponential families
  - Sensitive to parameterization
- ”Experience with DIC to date suggests that it works remarkably well” –Banerjee et al. (2004)

Can we develop a formal justification for DIC?
Loss Functions for Model Selection

Suppose we have a set of data $\mathbf{Y} = (Y_1, \ldots, Y_n)$

$Y_i \sim p(\cdot | \theta) \quad \theta \sim \pi(\cdot)$

What is a suitable loss function for model comparison?

- Decision Theory suggests using *scoring rules*, which are functions of $p(\cdot)$
- Maximized when $p(\cdot)$ is true data generating density
- Deviance:

$$D(\theta) = -2 \log\{p(\mathbf{Y}|\theta)\}$$
Computing the Deviance

Ideally we have two data sets:

- Training data \( Z \)
- Test data \( Y \)
- \( p(Y|\theta, Z) = p(Y|\theta) \)

Plug-in Deviance:

\[
L^p(Y, Z) = -2 \log \left[ p\{Y|E(\theta|Z)\} \right]
\]

Expected Deviance:

\[
L^e(Y, Z) = -2 \int \log\{p(Y|\theta)\} \pi(\theta|Z) \, d\theta
\]
Penalized Loss

Typically, we only have one set of data $\mathbf{Y}$. Can we use $L(\mathbf{Y}, \mathbf{Y})$?

Yes, but we’re being optimistic by using the data to both estimate the posterior of $\theta$ and as our test data.

Add a penalty term to loss function:

$$L(\mathbf{Y}, \mathbf{Y}) + p_{opt}$$
Penalized Loss

We can split our loss function into contributions from each $Y_i$

$$L(Y, \hat{Y}) = \sum_{i=1}^{n} L(Y_i, \hat{Y})$$

Compare $L(Y_i, \hat{Y})$ to cross-validation loss $L(Y_i, \hat{Y}_{-i})$ to estimate how optimistic we are being.

$$p_{opt_i} = \mathbb{E} \left[ L(Y_i, \hat{Y}_{-i}) - L(Y_i, \hat{Y}) \bigg| \hat{Y}_{-i} \right]$$

The penalized loss function is now

$$L(Y, \hat{Y}) + \sum_{i} p_{opt_i}$$
Penalized Loss and DIC

We will see that

\[ \text{DIC} \approx L^p(\mathbf{Y}, \mathbf{Y}) + \sum_i p_{opt_i} \]

...but only when the effective number of parameters is small relative to the number of observations.

When this is not true, DIC will under-penalize complex models.
Penalized Loss and DIC in Disease Mapping

Lip Cancer in Scotland
Penalized Loss and DIC in Disease Mapping

\[ Y_i \sim \text{Poisson}(\mu_i) \quad \log(\mu_i) = \alpha_0 + \gamma_i + \delta_i + \log(E_i) \]

- \( Y_i \) – lip cancer cases in county \( i \)
- \( E_i \) – expected counts of lip cancer in county \( i \)
- \( \alpha_0 \) – fixed effect
- \( \gamma_i \) – uncorrelated random effects
- \( \delta_i \) – spatially correlated random effects

<table>
<thead>
<tr>
<th>Model</th>
<th>DIC</th>
<th>Penalized Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed Effect Only</td>
<td>1.0</td>
<td>1.1</td>
</tr>
<tr>
<td>Uncorrelated</td>
<td>43.5</td>
<td>570.5</td>
</tr>
<tr>
<td>Spatial</td>
<td>31</td>
<td>163.9</td>
</tr>
<tr>
<td>Uncorrelated + Spatial</td>
<td>31.6</td>
<td>166.4</td>
</tr>
</tbody>
</table>

*Table: Estimated penalties for model complexity for Scottish lip cancer data*
Looking Ahead

Looking more closely at $DIC \approx L^p(Y, \hat{Y}) + \sum_i p_{opt_i}$

Application to Scotland Lip Cancer Data

Can we apply penalized loss functions to other settings?
References

