Penalized loss functions for Bayesian model comparison

Josh Keller
Biost 572 Presentation

4 June 2013
Bayes Factor:
- The formal solution
- Unstable with diffuse prior; undefined with improper priors

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

Cross-validation:
- Which model is most useful?
- Judge model by out-of-sample prediction
Model Selection in Bayesian Models

Deviance Information Criterion (DIC)

\[ DIC = \bar{D} + p_D = \text{Deviance} + \text{“Effective number of parameters”} \]

- Proposed by Spiegelhalter et al. (2002)
- Theoretical foundations are controversial
  - No clear generalization outside of exponential families
  - Doesn’t work for mixture distributions
  - Sensitive to parameterization

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

Plummer’s (2008) approach:

- Use cross-validation argument
- Estimate out-of-sample model fit
  - Training data $Z$
  - Test data $Y$
- Deviance as loss-function: $D(\theta) = -2 \log \{ p(Y|\theta) \}$
- Estimators:
  - Plug-in Deviance
    \[ L^p(Y, Z) = -2 \log \{ p(Y|\bar{\theta}(Z)) \} \]
  - Expected Deviance
    \[ L^e(Y, Z) = -2 \int \log \{ p(Y|\theta) \} p(\theta|Z) \, d\theta \]
- Similar to theoretical argument for AIC
Penalized Loss

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

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

Add an optimism penalty term to loss function:

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

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

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

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

$$p_{opt_i} = \mathbb{E}[L(Y_i, Y_{-i}) - L(Y_i, Y)|Y_{-i}]$$

The penalized loss function is now

$$\tilde{L}(Y, Y) = L(Y, Y) + \sum_i p_{opt_i}$$

Note: $\mathbb{E}[\tilde{L}(Y_i, Y)|Y_{-i}] = \mathbb{E}[L(Y_i, Y_{-i})|Y_{-i}]$. 

DIC as an approximation to $\tilde{L}^p$

Consider the hierarchical linear model of Lindley and Smith (1972):

$$\begin{align*}
Y|\theta & \sim N(A_1\theta, C_1) \\
\theta|\psi & \sim N(A_2\psi, C_2)
\end{align*}$$

with $A_1, A_2, C_1, C_2$ known matrices.

We can write the optimism penalty $p_{opt}$ in terms of the entries in the hat matrix $H = C_1^{-1}A_1\text{Var}(\theta|Y)A_1^T$.

If the dimension of $\theta$ is fixed, then

$$\tilde{L}^p(Y, Y) = L^p(Y, Y) + p_{opt} = DIC + O\left(\frac{1}{n}\right).$$
DIC as an approximation to $\tilde{L}^p$

But what if dimension of $\theta \rightarrow \infty$?

Consider a simplified hierarchical model:

$Y_i|\theta_i \sim N(\theta_i, \tau_i^{-1})$

$\theta_i|\psi \sim N(\psi, \lambda^{-1})$

with fixed precisions $\tau_i$ and a flat prior on $\psi$.

Two cases:

$\lambda \rightarrow \infty$

$\lambda \rightarrow 0$
DIC as an approximation to $\tilde{L}^p$

**Case 1:** $\lambda \to \infty$

Hierarchical model $\to$ pooled model with mean $\psi$ for all $Y_i$

\[
p_D \to 1
\]

\[
DIC \to \sum_i \tau_i (Y_i - \overline{Y})^2 + 2
\]

\[
\tilde{L}^p(Y, \overline{Y}) \to \sum_i \tau_i (Y_i - \overline{Y})^2 + 2
\]

Intuition: $\overline{Y_{-i}}$ contains much information about mean of $Y_i$, 
DIC as an approximation to $\tilde{L}^p$

**Case 2:** $\lambda \to 0$

Hierarchical model $\to$ fixed effects model with different mean for each $Y_i$

\[
p_D \to n \\
DIC \to 2n \\
\tilde{L}^p(Y, Y) \to \infty
\]

Intuition: $Y_{-i}$ contains no information about mean of $Y_i$.

So when $p_D \ll n$, DIC is a good approximation to penalized plug-in deviance. But when $p_D/n$ is large, then DIC is not a good approximation.
\[ \tilde{L}^p(Y, Y) \] in Exponential Families

Consider an exponential family distribution, with density

\[ p(Y_i|\theta_i) = \exp\{[y_i\theta_i - b(\theta_i)]/\phi\}c(y_i, \phi) \]

Let \( \mu_i = E[y_i|\theta_i] \). With some work, we can show that

\[ \tilde{L}^p(Y, Y) = D(\bar{\theta}) + \sum_{i=1}^{n} E[p_{D_i}(Y)|Y_{-i}] + 2\phi^{-1}\text{Cov}(\theta_i, \mu_i|Y_{-i}) - p_{D_i}(Y_{-i}) \]

\[ \approx \bar{D} + \sum_{i=1}^{n} \left[ 2\phi^{-1}\text{Cov}(\theta_i, \mu_i|Y_{-i}) - p_{D_i}(Y_{-i}) \right] \]

\[ := \bar{D} + r_{opt}. \]

Recall \( DIC = \bar{D} + p_D \). Let's compare \( p_D \) and \( r_{opt} \).
Lip cancer in Scotland

SMR of Lip Cancer in Scotland
Models for Lip cancer data

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

- \( Y_i \) – 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
- \( \delta_i \) – spatial (ICAR) random effects

Four models:
1. Fixed Effect only
2. Uncorrelated random effects
3. Spatial random effects
4. Uncorrelated and spatial random effects
Implementation

Posterior samples of the parameters are computed using MCMC.

Improper flat prior on $\alpha$. Gamma(0.5, 0.0005) priors on precisions for $\gamma_i$ and $\delta_i$.

Computing $r_{opt}$ requires $n = 56$ MCMC runs (leaving one observation out each time), which is feasible in this case, but not practical in general.

Here we compute $r_{opt}$ directly, and using two approximations that require only one chain:

A1: $\hat{r}_{opt} \approx \sum_i p_{D_i} / (1 - p_{D_i})$.
A2: Make replicate random effect draws from $\theta | Y$. 

## Lip Cancer Data

Results from Lip Cancer models:

<table>
<thead>
<tr>
<th>Model</th>
<th>$p_D$</th>
<th>$r_{opt}$</th>
<th>A1</th>
<th>A2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed Effect Only</td>
<td>1.0</td>
<td>1.1</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>Uncorrelated</td>
<td>43.4</td>
<td>570.8</td>
<td>294.7</td>
<td>568.2</td>
</tr>
<tr>
<td>Spatial</td>
<td>30.9</td>
<td>162.5</td>
<td>150.0</td>
<td>151.6</td>
</tr>
<tr>
<td>Uncorrelated + Spatial</td>
<td>30.8</td>
<td>165.0</td>
<td>110.9</td>
<td>153.0</td>
</tr>
</tbody>
</table>

- For all but the simplest model, $p_D$ does not well approximate $r_{opt}$
- DIC is under-penalizing the more complex models
Penalized loss for Mixture Distributions

- Lack of formalization outside of exponential families, specifically mixture distributions, was a limiting aspect of DIC.

- $\tilde{L}^p$ can be difficult to compute outside of exponential families

- Both use $\bar{\theta}$, which is problematic for mixtures

$\Rightarrow$ Now consider $\tilde{L}^e(Y, \mathcal{Y})$. 
Penalized loss for Mixture Distributions

Let \( J(p, q) = KL(p, q) + KL(q, p) \) be the undirected divergence between distributions \( p \) and \( q \).

Define
\[
J_i(\theta, \theta') = J\left(p(Y_i|\theta), p(Y_i|\theta')\right)
\]

Then the optimism for expected deviance is
\[
p_{opt_i} = \int \int J_i(\theta, \theta') p(\theta|Y_{-i}) p(\theta'|Y_{-i}) d\theta' d\theta
\]

Estimate \( p_{opt_i} \) using MCMC with two parallel chains.

Instead of running \( 2n \) chains with an observation left out, just run 2 chains on full data and use importance sampling to make draws.
Mixture Example

- Ratio of two urinary metabolites after administration of caffeine
- Originally from Richardson and Green (1997)
Mixture Example

\[ p(y_i | \pi, \mu, \sigma) = \sum_{g=1}^{G} \pi_g \phi \left( \frac{Y_i - \mu_g}{\sigma_g} \right) \]

- \( G \in \{1, 2, 3, 4, 5\} \)
- \( \pi \sim \text{Dirichlet}(5, \ldots, 5) \)
- \( \mu_g \sim N\left(\frac{1}{2}(Y_{(1)} + Y_{(n)}), R^2\right) \)
- \( \sigma_g^{-2} \sim \text{Gamma}(2, \beta) \)
- \( \beta \sim \text{Gamma}(0.2, 10/R^2) \)
- \( R = Y_{(n)} - Y_{(1)} \)

Requires two simultaneous MCMC runs
Mixture Example Results

<table>
<thead>
<tr>
<th># of Comps</th>
<th>$L^e$</th>
<th>$\hat{\rho}_{opt}$</th>
<th>$\tilde{L}^e$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>720.5</td>
<td>3.9</td>
<td>724.4</td>
</tr>
<tr>
<td>2</td>
<td>596.1</td>
<td>9.2</td>
<td>605.3</td>
</tr>
<tr>
<td>3</td>
<td>587.3</td>
<td>12.9</td>
<td>600.3</td>
</tr>
<tr>
<td>4</td>
<td>586.7</td>
<td>13.3</td>
<td>600.0</td>
</tr>
<tr>
<td>5</td>
<td>586.5</td>
<td>13.1</td>
<td>599.7</td>
</tr>
</tbody>
</table>

Urinary Enzyme Data

```
log(AFMU/1X)
Frequency
-4 -3 -2 -1 0 1
0 10 20 30
```
Conclusions and Critiques

- Establishes penalized deviance as a theoretically valid model comparison approach
- Provides theoretical argument for DIC as an approximation to penalized deviance
- Demonstrates situations in which DIC is a bad approximation
- Doesn’t solve the parameterization problem with the plug-in deviance
- Not clear that $\tilde{L}^p$ and $\tilde{L}^e$, as implemented, are practical
  - Requires either $n$ MCMC runs or uses an approximation
  - For plug-in deviance, approximations for $p_{opt}$ are better than DIC but aren’t always good
  - For expected deviance, $\hat{p}_{opt}$ is easily obtained in JAGS, but the importance sampling approximation may not always be valid

Easily obtained via software ≠ Appropriate to use
References

