Penalized loss functions for Bayesian model comparison

Martyn Plummer, *Biostatistics* (2008)

Josh Keller
Biost 572 Presentation

4 June 2013

Model Selection in Bayesian Models

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)

$$DIC = \overline{D} + p_D = \text{Deviance} + \text{“Effective number of parameters”}$$

- 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 $\mathbf{Z}$
  - Test data $\mathbf{Y}$
- Deviance as loss-function: $D(\boldsymbol{\theta}) = -2\log\{p(\mathbf{Y}|\boldsymbol{\theta})\}$
- Estimators:

  Plug-in Deviance

$$L^p(Y,Z) = -2\log[p\{Y|\overline{\theta}(Z)\}]$$

  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 $\mathbf{Y}$. Can we use $L(\mathbf{Y}, \mathbf{Y})$?

Yes, but we're being optimistic by using the data to both estimate the posterior of $\boldsymbol{\theta}$ and as our test data

Add an optimism penalty term to loss function:

$$\widetilde{L}(\mathbf{Y}, \mathbf{Y}) = L(\mathbf{Y}, \mathbf{Y}) + p_{opt}$$

Penalized Loss

We can split our loss function into contributions from each Y_i

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

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

$$p_{opt_i} = \mathrm{E}\left[L(Y_i, \mathbf{Y}_{-i}) - L(Y_i, \mathbf{Y}) \middle| \mathbf{Y}_{-i} \right]$$

The penalized loss function is now

$$\widetilde{L}(\mathbf{Y}, \mathbf{Y}) = L(\mathbf{Y}, \mathbf{Y}) + \sum_{i} p_{opt_i}$$

Note: $\mathrm{E}[\widetilde{L}(Y_i, \mathbf{Y}) | \mathbf{Y}_{-i}] = \mathrm{E}[L(Y_i, \mathbf{Y}_{-i}) | \mathbf{Y}_{-i}]$.

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

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

$$\mathbf{Y}|\theta \sim N(A_1\theta, C_1)$$
$$\theta|\psi \sim N(A_2\psi, C_2)$$

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 $H = C_1^{-1}A_1\mathrm{Var}(\theta|\mathbf{Y})A_1^T$.

If the dimension of θ is fixed, then

$$\widetilde{L}^p(Y, Y) = L^p(Y, Y) + p_{opt} = DIC + O\left(\frac{1}{n}\right).$$

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

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

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 τ_i and a flat prior on ψ.

Two cases:

- $\lambda \to \infty$
- $\lambda \to 0$

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

Case 1: $\lambda \to \infty$

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

$$
\begin{aligned}
p_D &\to 1 \\
DIC &\to \sum_i \tau_i (Y_i - \overline{Y})^2 + 2 \\
\widetilde{L}^p(Y, Y) &\to \sum_i \tau_i (Y_i - \overline{Y})^2 + 2
\end{aligned}
$$

Intuition: $\mathbf{Y}_{-i}$ contains much information about mean of Y_i,

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

Case 2: $\lambda \to 0$

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

$$
\begin{aligned}
p_D &\to n \\
DIC &\to 2n \\
\widetilde{L}^p(Y, Y) &\to \infty
\end{aligned}
$$

Intuition: $\mathbf{Y}_{-i}$ 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.

$\widetilde{L}^p(Y, Y)$ in Exponential Families

Consider an exponential family distribution, with density

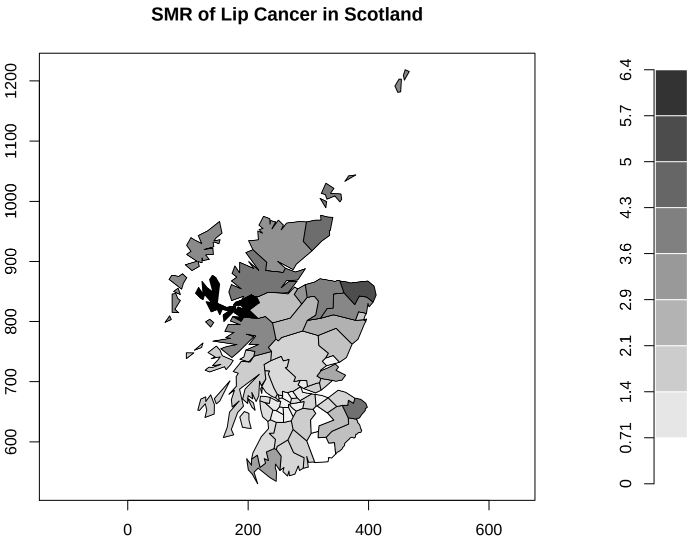
$$p(Y_i|\theta_i)\} = \exp\{[y_i\theta_i - b(\theta_i)]/\phi\}c(y_i, \phi)$$

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

$$\begin{aligned}\widetilde{L}^p(Y, Y) &= D(\overline{\boldsymbol{\theta}}) + \sum_{i=1}^{n} \mathrm{E}[p_{D_i}(\mathbf{Y})|\mathbf{Y}_{-i}] + 2\phi^{-1}\mathrm{Cov}(\theta_i, \mu_i|\mathbf{Y}_{-i}) - p_{D_i}(\mathbf{Y}_{-i}) \\ &\approx \overline{D} + \sum_{i=1}^{n} \left[2\phi^{-1}\mathrm{Cov}(\theta_i, \mu_i|\mathbf{Y}_{-i}) - p_{D_i}(\mathbf{Y}_{-i})\right], \\ &:= \overline{D} + r_{opt}.\end{aligned}$$

Recall $DIC = \overline{D} + p_D$. Let's compare p_D and r_{opt}.

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
α_0 – fixed effect
γ_i – uncorrelated random effects
δ_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 α. Gamma(0.5, 0.0005) priors on precisions for γ_i and δ_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 $\boldsymbol{\theta}|\mathbf{Y}$

Lip Cancer Data

Results from Lip Cancer models:

Model	p_D	r_{opt}	A1	A2
Fixed Effect Only	1.0	1.1	1.0	
Uncorrelated	43.4	570.8	294.7	568.2
Spatial	30.9	162.5	150.0	151.6
Uncorrelated + Spatial	30.8	165.0	110.9	153.0

- 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.

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

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

$\Rightarrow$ Now consider $\widetilde{L}^e(\mathbf{Y}, \mathbf{Y})$.

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

$$J_i(\boldsymbol{\theta}, \boldsymbol{\theta}') = J\Big(p(Y_i|\boldsymbol{\theta}), p(Y_i|\boldsymbol{\theta}')\Big)$$

Then the optimism for expected deviance is

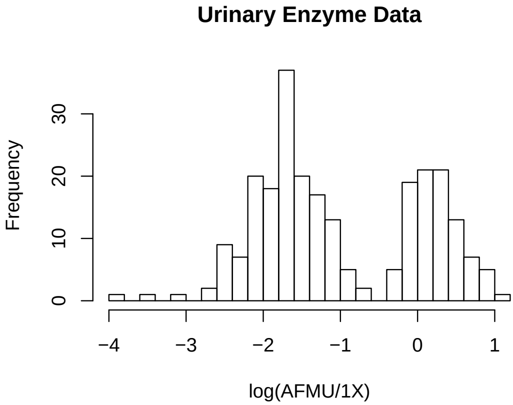
$$p_{opt_i} = \int\int J_i(\boldsymbol{\theta}, \boldsymbol{\theta}') p(\boldsymbol{\theta}|\mathbf{Y}_{-i}) p(\boldsymbol{\theta}'|\mathbf{Y}_{-i}) d\boldsymbol{\theta}' d\boldsymbol{\theta}$$

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

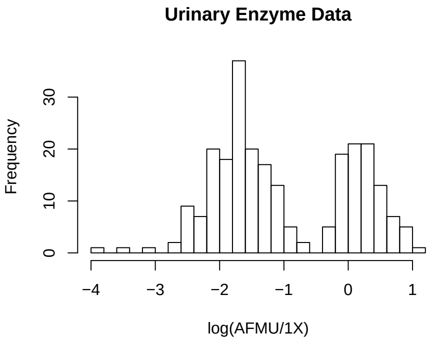
$$p(y_i|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \sum_{g=1}^{G} \pi_g \phi\left(\frac{Y_i - \mu_g}{\sigma_g}\right)$$

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

Requires two simultaneous MCMC runs

Mixture Example Results

# of Comps	L^e	$\hat{p}_{opt}$	$\widetilde{L}^e$
1	720.5	3.9	724.4
2	596.1	9.2	605.3
3	587.3	12.9	600.3
4	586.7	13.3	600.0
5	586.5	13.1	599.7



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 $\widetilde{L}^p$ and $\widetilde{L}^e$, *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 $\neq$ Appropriate to use

References

Lindley, D. and A. Smith (1972). Bayes Estimates for the Linear Model. *JRSSB* **34**, 141.

Plummer, M. (2008) Penalized loss functions for Bayesian model comparison. *Biostatistics*, **9**, 523-539.

Spiegelhalter, D., Best, N., Carlin, B., and van der Linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). *JRSSB* **64**, 583-639.