Penalized loss functions for Bayesian model comparison

Martyn Plummer, *Biostatistics* (2008)

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 $\mathbf{Y} = (Y_1, \ldots, Y_n)$
$Y_i \sim p(\cdot|\boldsymbol{\theta}) \quad \boldsymbol{\theta} \sim \pi(\cdot)$

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:
$$D(\boldsymbol{\theta}) = -2\log\{p(\mathbf{Y}|\boldsymbol{\theta})\}$$

Computing the Deviance

Ideally we have two data sets:

- Training data $\mathbf{Z}$
- Test data $\mathbf{Y}$
- $p(\mathbf{Y}|\boldsymbol{\theta}, \mathbf{Z}) = p(\mathbf{Y}|\boldsymbol{\theta})$

Plug-in Deviance:

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

Expected Deviance:

$$L^e(\mathbf{Y}, \mathbf{Z}) = -2 \int \log\{p(\mathbf{Y}|\boldsymbol{\theta})\} \pi(\boldsymbol{\theta}|\mathbf{Z})\, d\boldsymbol{\theta}$$

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 $\boldsymbol{\theta}$ 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

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

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

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

Penalized Loss and DIC

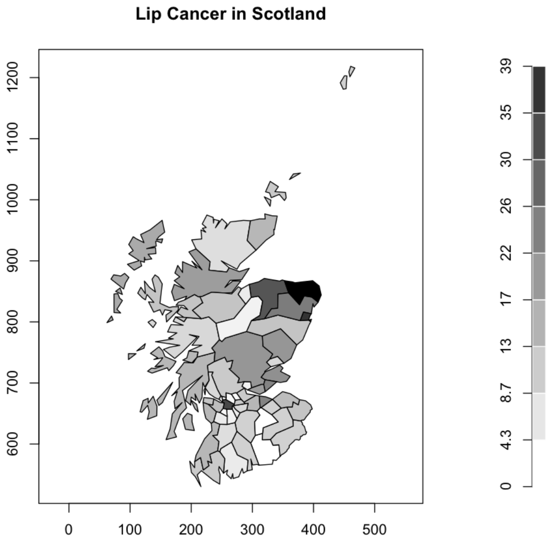
We will see that

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

...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



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

α_0 – fixed effect

γ_i – uncorrelated random effects

δ_i – spatially correlated random effects

Model	DIC	Penalized Loss
Fixed Effect Only	1.0	1.1
Uncorrelated	43.5	570.5
Spatial	31	163.9
Uncorrelated + Spatial	31.6	166.4

Table: Estimated penalties for model complexity for Scottish lip cancer data

Looking Ahead

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

Application to Scotland Lip Cancer Data

Can we apply penalized loss functions to other settings?

References

Banerjee, S., Carlin, B., and Gelfand, A. (2004) Hierarchical Modeling and ANalysis for Spatial Data. Boca Raton, FL: Chapman & Hall/CRC.

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.