# Part 2: Penalized loss functions for Bayesian model comparison

Martyn Plummer, *Biostatistics* (2008)

Josh Keller

7 May 2013

# Setting: Penalized Loss Functions

Goal: Develop a formal justification for DIC
Have a measure of Bayesian model fit

Approach: Use the deviance as loss function

$$D(\theta) = -2\log\{p(y|\theta)\}$$

Estimators: Plug-in Deviance

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

Expected Deviance

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

Note: Need to add optimism penalty for using data twice

$$L(Y,Y) + p_{opt} \quad \text{where} \quad p_{opt_i} = \mathrm{E}[L(Y_i,\mathbf{Y}_{-i}) - L(Y_i,\mathbf{Y})|\mathbf{Y}_{-i}]$$

# Connection to DIC

Recall DIC (Speigelhalter *et al.*, 2002):

$$DIC = \overline{D} + p_D = \overline{D} + [\overline{D} - D(\bar{\theta})]$$

In the current setting,

$$\overline{D} = \overline{D(\theta)} = L^e(Y, Y) \quad \text{and} \quad D(\bar{\theta}) = L^p(Y, Y)$$

So

$$DIC = L^e(Y, Y) + p_D$$

$$p_D = L^e(Y, Y) - L^p(Y, Y)$$

Optimism penalties are missing!

Plummer is not the only person to notice this. For example, this motivated Ando (2007) to develop BPIC

# A Normal Example

The hierarchical linear model of Lindley and Smith (1972):

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

Let $V = \mathrm{Var}(\theta|\mathbf{Y})$ and break $\mathbf{Y}$ into conditionally independent subvectors $Y_1, \ldots, Y_n$.

For plug-in deviance $L^p$,

$$\begin{aligned} p_{opt_i} &= \mathrm{Tr}(C_{1i}^{-1}A_{1i}VA_{1i}^T) + \mathrm{Tr}(C_{1i}^{-1}A_{1i}V_{-i}A_{1i}^T) \\ &= \mathrm{Tr}(H_i) + \mathrm{Tr}((I - H_i)^{-1}H_i) \end{aligned}$$

where $H_i$ is the $i$th block of the hat matrix $H$.

$$p_{opt} = \sum_i p_{opt_i} = \mathrm{Tr(H)} + \sum_i \mathrm{Tr}((I - H_i)^{-1}H_i)$$

# A Normal Example

Spiegelhalter *et al.* showed $p_D = \mathrm{Tr}(H)$, so

$$p_{opt} = p_D + \sum_i \mathrm{Tr}((I - H_i)^{-1} H_i)$$

This gives an expression for the penalized plug-in deviance

$$L^p(Y, Y) + p_{opt} = \overline{D} + \sum_i \mathrm{Tr}((I - H_i)^{-1} H_i).$$

For scalar outcomes, $\mathrm{Tr}((I - H_i)^{-1} H_i) = \frac{p_{D_i}}{1 - p_{D_i}}$. If the dimension of $\theta$ is fixed, then $\sum_i \frac{p_{D_i}}{1 - p_{D_i}} = p_D + O\left(\frac{1}{n}\right)$, so

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

# ANOVA Example

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

Consider the ANOVA model:

$$\begin{aligned} Y_i|\theta_i &\sim N(\theta_i, \tau_i^{-1}) \\ \theta_i|\psi &\sim N(\psi, \lambda^{-1}) \end{aligned}$$

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

Letting $\rho_i = \tau_i/(\lambda + \tau_i)$ be the intraclass correlation,

$$p_{D_i} = \rho_i + \frac{\rho_i(1-\rho_i)}{\sum_{j=1}^n \rho_j}.$$

# ANOVA Example

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

ANOVA model $\to$ pooled model with mean $\psi$

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

Conceptually, $\mathbf{Y}_{-i}$ contains more information about mean of $Y_i$,

# ANOVA Example

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

ANOVA model $\to$ fixed effects model with individual means

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

Conceptually, $\mathbf{Y}_{-i}$ 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.

# $L^p(Y, Y)$ in Exponential Families

In an exponential family, the log likelihood is given by

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

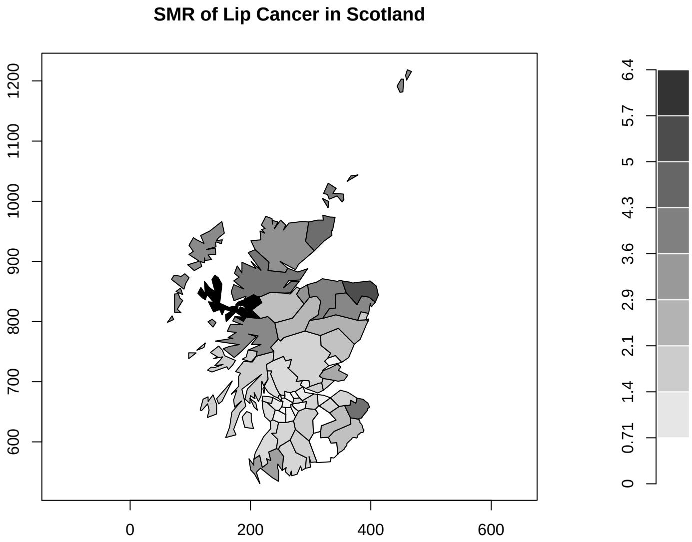
With some work, we can show that

$$p_{opt_i} = 2\phi^{-1}\mathrm{Cov}(\theta_i, \mu_i|\mathbf{Y}_{-i}) - p_{D_i}(\mathbf{Y}_{-i}) + \mathrm{E}[p_{D_i}(\mathbf{Y})|\mathbf{Y}_{-i}],$$

where $\mu_i = E[Y_i|\theta_i]$. We can then estimate $\mathrm{E}[p_{D_i}(\mathbf{Y})|\mathbf{Y}_{-i}]$ by $p_{D_i}(\mathbf{Y})$ and get an estimator for the penalized plug in deviance:

$$L^p(Y, Y) + \hat{p}_{opt} = \overline{D} + 2\phi^{-1}\sum_{i=1}^{n}\mathrm{Cov}(\theta_i, \mu_i|\mathbf{Y}_{-i}) - p_{D_i}(\mathbf{Y}_{-i}).$$

# 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
$\delta_i$ – spatially correlated 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

Computing $\hat{p}_{opt}$ 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 $\hat{p}_{opt}$ exactly, but can use the approximation $\hat{p}_{opt} \approx \sum_i p_{D_i}/(1 - p_{D_i})$.

# Lip Cancer Data

Results from Lip Cancer models:

| Model | $\overline{D}$ | $p_D$ | DIC | $\hat{p}_{opt}$ | $L^p + \hat{p}_{opt}$ |
|---|---|---|---|---|---|
| Fixed Effect Only | 589.7 | 0.99 | 590.7 | 1.0 | 590.7 |
| Uncorrelated | 269.1 | 43.3 | 312.4 | 572.5 | 841.6 |
| Spatial | 266.3 | 31.0 | 297.3 | 163.9 | 430.2 |
| Uncorrelated + Spatial | 265.9 | 31.6 | 297.5 | 166.4 | 432.3 |

- For all but the simplest model, $p_D$ does not well approximate $p_{opt}$
- DIC is under-penalizing the more complex models

# Summary

What we've seen:

- Plug-in deviance $L^p$ can be used to assess model fit
- Require a penalty term $p_{opt}$ to be added to $L^p$
- $p_{opt}$ has exact form in linear models and approximate form in exponential families
- When $p_D \ll n$, DIC can be a good approximation to $L^p + p_{opt}$
- In spatial settings, *DIC* under-penalizes complex models

What's left:

- $L^e$ in non-exponential families
- Mixture distribution example

# References

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.