Bayesian measures of model complexity and fit

Paper by **Spiegelhalter et al. (2002)**
Presented by: **Lina Lin**

STAT/BIOST 572

April 16, 2013

Classical approaches to model selection

If candidate models are nested, we can use the **likelihood ratio** test.

Cross validation (CV) is also an option, but (1) validation data is not always readily available and (2) can be computationally intensive.

Otherwise, the general approach is to reduce each model we are considering to a single number according to a **well-justified criterion**. There are two that we are already familiar with ...

- **Akaike's Information Criterion**:

$$AIC = -2 \log \mathcal{L} + 2p$$

- **Bayesian Information Criterion**:

$$BIC = -2 \log \mathcal{L} + p \log n$$

where $\mathcal{L}$ denotes likelihood, p number of parameters (or complexity), and n number of data points.

Classical approaches to model selection

One must be careful with how one uses these criterions, as they are motivated by different philosophies.

There are many more ... (i.e. {rest of the alphabet}-IC); however, they behave "similarly" in that their target model achieves a some form of **balance between measure of fit and complexity**.

To develop a criterion for hierarchical model (e.g. random effects model) selection, **we need to first define a complexity measure for hierarchical models**. The most ambitious attempts so far have been made in smoothing and neural network literature (Moody (1992), etc.).

Effective number of parameters

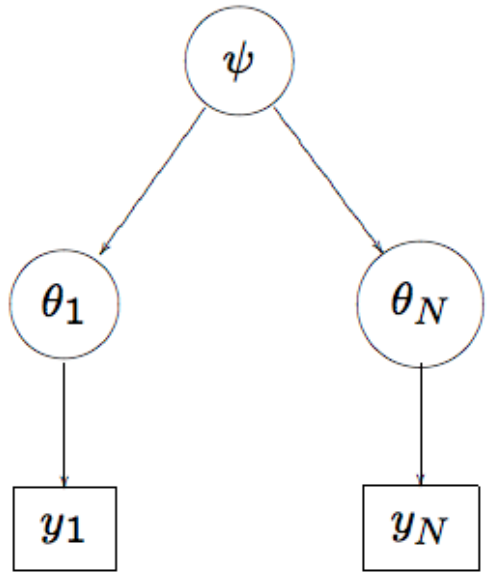
Consider the following random effects model:

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

for $i = 1, \ldots, p$.

Should the effective number of parameters (or complexity) simply be p?

The presence of a prior induces **dependency** between the θ_is, which reduces the dimensionality of the model, so the actual complexity of the model is $\leq p$.



Parameter(s) of focus

Suppose the full probability model factorizes as:

$$p(y, \theta, \psi) = p(y|\theta)p(\theta|\psi)p(\psi)$$

From which we can construct the following marginal distributions:

$$p(y, \theta) = p(y|\theta) \int_{\Psi} p(\theta|\psi)p(\psi)d\psi = p(y|\theta)p(\theta)$$

(i.e focused on θ)

$$p(y, \psi) = \int_{\Theta} p(y|\theta)p(\theta|\psi)p(\psi)d\theta = p(y|\psi)p(\psi)$$

(i.e. focused on ψ).

Depending on the level (or focus), we can argue that two variants of the model are of different complexities. For the rest of this presentation, we take Θ to be our parameter(s) of focus.

A complexity measure for hierarchical models

Define:

$$d_\Theta\{y, \theta, \theta(\tilde{y})\} = -2\log\{p(y|\theta)\} + 2\log[p\{y|\tilde{\theta}(y)\}]$$

Spiegelhalter et al. defines the complexity of the focused model to be:

$$p_D\{y, \Theta, \tilde{\theta}(y)\} = E_{\theta|y}[d_\Theta\{y, \theta, \theta(\tilde{y})\}] = E_{\theta|y}[-2\log\{p(y|\theta)\}] + 2\log[p\{y|\tilde{\theta}(y)\}]$$

where $\tilde{\theta}$ is often selected to be the posterior mean of θ.

We can also write p_D as:

$$p_D = \overline{D(\theta)} - D(\bar{\theta})$$

where $D(\theta) = -2\log\{p(y|\theta)\} + 2\log\{f(y)\}$, which we refer to as the "Bayesian deviance."

Some theoretical justification ...

If we were to assume a normal approximation to the posterior likelihood, and expand $D(\theta)$ about the posterior mean $\bar{\theta}$, then we can show, through a series of (fun!) calculations that

$$E_{\theta|y}\{D(\theta)\} \approx D(\bar{\theta}) + \text{tr}(-L''_{\bar{\theta}} V)$$

which implies that:

$$p_D \approx \text{tr}(-L''_{\bar{\theta}} V)$$

where $L''_{\bar{\theta}} = \frac{\partial^2 \log p(y|\theta)}{\partial \theta^2}$ so $-L''_{\bar{\theta}}$ is the observed Fisher information at $\bar{\theta}$, and $V = E[(\theta - \bar{\theta})(\theta - \bar{\theta})^T]$.

If $\bar{\theta} = \hat{\theta}$ (i.e. under negligible prior information), we obtain,

$$p_D \approx p$$

A convincing example

Consider the general hierarchical model described by Lindley and Smith (1972). Suppose that:

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

Then, through a series of (tedious) calculations, we can show that:

$$p_D = tr(H) = \sum_i h_{ii}$$

where H is the hat matrix (i.e. projection matrix). In other words, the effective number of parameters is the sum of the individual leverages.

Deviance Information Criterion

Based on this complexity measure, Spiegelhalter et al. propose the following criterion for comparing hierarchical models, which they term **"Deviance Information Criterion (DIC)"**,

$$DIC = D(\bar{\theta}) + 2p_D$$

Recall that:

$$AIC = -2 \log \mathcal{L} + 2p$$

We can think of DIC as a **"generalized" version of AIC**. In fact, when working with flat priors, DIC serves as a decent approximation of AIC, since $\bar{\theta} \approx \hat{\theta} \implies p_D \approx p$.

Deviance Information Criterion

DIC is an approximation of the **expected posterior loss** when adopting a particular model, assuming a logarithmic loss function.

One appealing aspect of using DIC is that it can be readily calculated using **Markov Chain Monte Carlo (MCMC)** methods.

Because DIC behaves like AIC, DIC should be motivated by the same reasons we apply AIC (as opposed to say, BIC) - i.e. to seek out the model that minimizes information loss. Thus, it is a **predictive** approach in that it chooses the model that would generate good predictions but is not necessarily the true model.

Deviance Information Criterion

DIC does have the potential to select the model that over-fits since it "uses the data twice."

Ando (2007) developed a new criterion BPIC which corrects for the over-fitting by adjusting for the asymptotic bias of the posterior mean of the log-likelihood as an estimator for its expected log-likelihood.

It has also been noted that DIC is **not** invariant to parameterization, so you will obtain different DIC values depending on how you parameterize the model.