Bayesian measures of model complexity and fit

Lina Lin

STAT/BIOST 572

May 6, 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

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.) - see **Network Information Criterion** or **NIC**.

Effective number of parameters

Consider the following random effects model:

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

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

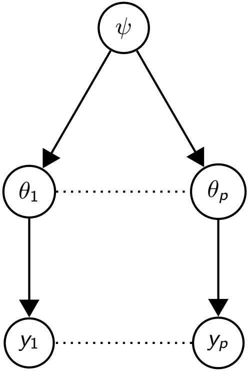
Should the effective number of parameters (or complexity) simply be **p***?*

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 $\leq p$.

The available data also influences the degree of dependency, which is consistent with the idea that **complexity should reflect the difficulty in estimation**.

Effective number of parameters

Here is a schematic representation of the random effects model (1) from the previous slide:



Parameter(s) of focus

Suppose the full probability model factorizes as:

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

This corresponds to a three-tiered hierarchical model.

From which we can construct the following marginal distributions:

$$p(y, \theta) = p(y|\theta) \int_{\Psi} p(\theta|\psi)p(\psi)d\psi = p(y|\theta)p(\theta) \ \text{ i.e. focused on } \Theta$$

OR

$$p(y, \psi) = \int_{\Theta} p(y|\theta)p(\theta|\psi)p(\psi)d\theta = p(y|\psi)p(\psi) \ \text{ i.e. focused on } \Psi$$

We assume, by default, the model to be focused on Θ for the remainder of this presentation.

A complexity measure for hierarchical models

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

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

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

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 "Bayesian deviance."

Some theoretical justification ...

If we were to assume a normal approximation to the posterior likelihood, we can expand $D(\theta)$ about the posterior mean $\bar{\theta}$ via a simple Taylor expansion as follows:

$$\begin{aligned} D(\theta) &\approx D(\bar{\theta}) + (\theta - \bar{\theta})^T \frac{\partial D}{\partial \theta}\bigg|_{\bar{\theta}} + \frac{1}{2}(\theta - \bar{\theta})^T \frac{\partial^2 D}{\partial \theta^2}\bigg|_{\bar{\theta}}(\theta - \bar{\theta}) \\ &= D(\bar{\theta}) - 2(\theta - \bar{\theta})L'_{\bar{\theta}} - (\theta - \bar{\theta})^T L''_{\bar{\theta}}(\theta - \bar{\theta}) \end{aligned} \qquad (2)$$

where $L = \log\{p(y|\theta)\}$ and L' and L'' are its first and second derivatives wrto θ respectively.

Taking posterior expectations, we achieve:

$$\begin{aligned} E_{\theta|y}\{D(\theta)\} = \overline{D(\theta)} &= D(\bar{\theta}) - E[tr\{(\theta - \bar{\theta})^T L''_{\bar{\theta}}(\theta - \bar{\theta})\}] \\ &= D(\bar{\theta}) - tr[L''_{\bar{\theta}} E\{(\theta - \bar{\theta})^T(\theta - \bar{\theta})\}] \end{aligned} \qquad (3)$$

Some theoretical justification ...

If we define $V = E\{(\theta - \bar{\theta})(\theta - \bar{\theta})^T\}$, then:

$$E_{\theta|y}\{D(\theta)\} \approx D(\bar{\theta}) + \mathrm{tr}(-L''_{\bar{\theta}} V) \tag{4}$$

which implies that, under the normal approximation to the likelihood:

$$\boxed{p_D \approx \mathrm{tr}(-L''_{\bar{\theta}} V)} \tag{5}$$

Note that: $-L''_{\bar{\theta}}$ is the observed Fisher information at $\bar{\theta}$, so p_D can be thought of as the fraction of information in the likelihood about the parameters relative to the total information.

Under negligible prior information, we obtain,

$$\boxed{p_D \approx p} \tag{6}$$

since $\bar{\theta} \approx \hat{\theta}$ and $-(\theta - \bar{\theta})^T L''_{\bar{\theta}}(\theta - \bar{\theta}) \approx \chi^2_p$.

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:

$$\boxed{p_D = tr(H) = \sum_i h_{ii}} \tag{7}$$

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.

p_D and Parameterization

p_D is **NOT** invariant to parameterization:

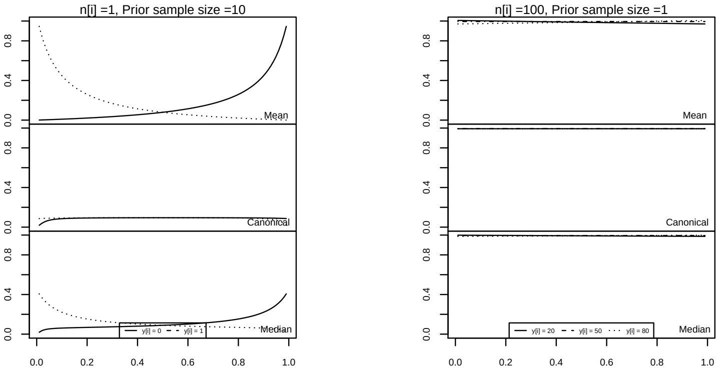
Consider a binomial likelihood $Bin(n, \mu)$ with a conjugate prior $B(a, b)$ on μ. Let $\theta = \text{logit}(\mu)$. Then, the (unstandardized) deviance is:

$$D(\mu) = -2y_i\psi(a+y) - 2(n-y)\psi(b+n-y) + 2n\psi(a+b+n)$$

and we have:

$$\begin{aligned}
\overline{D(\mu)} = \overline{D(\theta)} &= -2y\psi(a+y) - 2(n-y)\psi(b+n-y) + 2n\psi(a+b+n) \\
D(\bar{\mu}) &= -2y\log(a+y) - 2(n-y)\log(b+n-y) + 2n\log(a+b+n) \\
D(\bar{\theta}) &= -2y\psi(a+y) + 2y\psi(b+n-y) \\
&+ 2n\log[1+\exp\{\psi(a+y) - \psi(b+n-y)\}] \\
D(\mu^{med}) &= D(\theta^{med}) = -2y\log(\mu^{med}) - 2(n-y)\log(1-\mu^{med})
\end{aligned}$$

p_D and Parameterization



(a)

(b)

Figure: p_D for the binomial likelihood with conjugate prior example for different n, and prior sample sizes $(a + b)$ under the three different parameterizations

Deviance Information Criterion

Using this complexity measure, Spiegelhalter et al. propose the following criterion for comparing hierarchical models, which they term "Deviance Information Criterion,"

$$\boxed{DIC = D(\bar{\theta}) + 2p_D} \tag{8}$$

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 $p_D \approx p$ in this case.

The Derivation

DIC is an approximation of the **expected posterior loss** when adopting a particular model, assuming a logarithmic loss function: $\mathcal{L}(Y, \tilde{\theta}) = -2\log\{p(Y|\tilde{\theta})\} = D(\tilde{\theta})$

Assume that we have a replicate dataset Z derived from the same data-generating mechanism as Y, our original dataset. We favour the model that minimizes the expected loss that is suffered in predicting Z:

$$E_{z|\theta}[\mathcal{L}(Y, \tilde{\theta}(y))]$$

We can estimate this predicted loss using $\mathcal{L}(Y, \tilde{\theta}(y))$ - the loss suffered from re-predicting Y - however, this estimate is biased so we need to include an "optimism" term c:

$$\begin{aligned} E_{z|\theta}[\mathcal{L}(Y, \tilde{\theta})(y)] &= \mathcal{L}(Y, \tilde{\theta}(y)) + c_{\Theta}\{y, \theta^t, \tilde{\theta}(y)\} \\ &= D(\tilde{\theta}) + c_{\Theta}\{y, \theta^t, \tilde{\theta}(y)\} \end{aligned}$$

The Derivation

The derivation mimics that used to derive the AIC (i.e. if we were to replace $\tilde{\theta}$ with the MLE $\hat{\theta}$).

We can manipulate the above expression as:

$$\begin{aligned} c_\Theta\{y, \theta, \tilde{\theta}(y)\} &= E_{z|\theta}\{D_z(\tilde{\theta}) - D_z(\theta)\} + E_{z|\theta}\{D_z(\theta) - D(\theta)\} \\ &+ \{D(\theta) - D(\tilde{\theta})\} \end{aligned}$$

We label the three components to the sum $\mathcal{L}_1$, $\mathcal{L}_2$ and $\mathcal{L}_3$. By definition, we have

$$E_{\theta|y}[\mathcal{L}_3] = E_{\theta|y}\{D(\theta) - D(\tilde{\theta})\} = p_D$$

.

The Derivation

We perform a Taylor series expansion of $\mathcal{L}_1$ about θ (similar to proof shown a few slides back):

$$\begin{aligned}
\mathcal{L}_1 &\approx E_{z|\theta}\{-2\log[p(z|\tilde{\theta})] + 2\log[p(z|\tilde{\theta})] - 2(\tilde{\theta}-\theta)^T L'_{z,\theta} - (\tilde{\theta}-\theta)^T L''_{z,\theta}(\tilde{\theta}-\theta)\} \\
&= E_{z|\theta}\{-2(\tilde{\theta}-\theta)^T L'_{z,\theta} - (\tilde{\theta}-\theta)^T L''_{z,\theta}(\tilde{\theta}-\theta)\}
\end{aligned}$$

where $L'_{z,\theta} = \frac{\partial log[p(z|\theta)]}{\partial\theta}$ (i.e. the score function) and $L''_{z,\theta} = \frac{\partial^2 log[p(z|\theta)]}{\partial\theta^2}$. We note that $E_{z|\theta}\left\{\frac{\partial log[p(z|\theta)]}{\partial\theta}\right\} = 0$.

$$\begin{aligned}
\mathcal{L}_1 &\approx E_{z|\theta}\{-(\tilde{\theta}-\theta)^T L''_{z,\theta}(\tilde{\theta}-\theta)\} \\
&= \text{tr}\{-E_{z|\theta}[L''_{z,\theta}](\tilde{\theta}-\theta)^T(\tilde{\theta}-\theta)\} \\
&= \text{tr}\{-I(\theta)(\tilde{\theta}-\theta)(\tilde{\theta}-\theta)^T\} \\
&\approx \text{tr}\{-L''_{\tilde{\theta}}(\tilde{\theta}-\theta)(\tilde{\theta}-\theta)^T\}
\end{aligned}$$

Where the last line follows from a "good model" assumption. Taking the posterior means, we have

$$\begin{aligned}
E_{\theta|y}[\mathcal{L}_1] &= \text{tr}\{-L''_{\tilde{\theta}} E_{\theta|y}[(\tilde{\theta}-\theta)(\tilde{\theta}-\theta)^T]\} \\
&= \text{tr}\{-L''_{\tilde{\theta}} V\} \approx p_D
\end{aligned}$$

The Derivation

The $\mathcal{L}_2$ can be "ignored" because it can be shown to have a marginal expectation of 0.

$$\mathcal{L}_2 = E_{z|\theta}\{-2\log[p(z|\theta] + 2\log[p(y|\theta)]\}$$

Taking double expectations:

$$\begin{aligned} E_y E_{\theta|y}[\mathcal{L}_2] &= E_y E_{\theta|y} E_{z|\theta}\{-2\log[p(z|\theta] + 2\log[p(y|\theta)]\} \\ &= E_\theta E_{y|\theta} E_{z|\theta}\{-2\log[p(z|\theta] + 2\log[p(y|\theta)]\} \\ &= E_\theta[E_{z|\theta}\{-2\log[p(z|\theta]\} + E_{y|\theta}\{2\log[p(y|\theta)]\}] = 0 \end{aligned}$$

Putting all this together, we have

$$c_\Theta\{y, \theta, \tilde{\theta}(y)\} \approx 2p_D$$

Which implies:

$$\boxed{E_{\theta|y} E_{z|\theta}[\mathcal{L}(Y, \tilde{\theta}(y))] \approx D(\tilde{\theta}) + 2p_D = \overline{D(\theta)} + p_D} \tag{9}$$

Some Notes

- 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 - i.e. to seek out the model that minimizes information loss. It serves as a predictive approach to model selection
- It is **not** comparable to Bayes factors.

Implementation: Scottish Lip Cancer Data

We apply DIC in selecting a model for rates of lip cancer in 56 districts in Scotland (Clayton and Kaldor, 1987; Breslow and Clayton, 1993). We assume cancer counts y_i are Poisson with mean $E_i \exp(\theta_i)$ where E_i is the expected number of cases for county i, $i = 1, \ldots, 56$.

We consider the following set of candidate models:

1. **Pooled**: $\theta_i = \alpha_0$
2. **Exchangeable**: $\theta_i = \alpha_0 + \gamma_i$, γ_i exchangeable random effects
3. **Spatial**: $\theta_i = \alpha_0 + \delta_i$, δ_i spatial random effects
4. **Exchangeable + Spatial**: $\theta_i = \alpha_0 + \gamma_i + \delta_i$
5. **Saturated**: $\theta_i = \alpha_i$

Implementation: Scottish Lip Cancer Data

We placed an ...

- Improper flat prior on α_0
- Normal priors with precision parameter λ_γ on the γ_is
- Intrinsic conditional autoregressive (ICAR) prior (Besag, 1974) on the δ_is

$$\delta_i | \delta_{\setminus i} \sim N\left(\frac{1}{n_i} \sum_{j \in \mathcal{A}_i} \delta_j, \frac{1}{n_i \lambda_\delta} \right)$$

- Weakly informative $\Gamma(0.5, 0.0005)$ priors on λ_γ and λ_δ.

What Lina Did ...

- Ran two chains in `WINBUGS`; 15000 iterations each, following a burn-in period of 5000 iterations
- Attempting to replicate results in `R`; succeeded for Models 1 and 2;
- Having issues with implementing ICAR MCMC

Implementation: Scottish Lip Cancer Data

The Results

Model	p_D^{μ}	DIC^{μ}	p_D^{θ}	DIC^{θ}	p_D^{med}	DIC^{med}
Pooled	1.0	382.7	1.0	382.7	1.0	382.7
Exchangeable	42.8	103.8	43.3	104.3	43.4	104.4
Spatial	31.6	88.9	31.2	88.5	31.1	88.4
Exchangeable + Spatial	32.6	90.6	32.2	90.2	32.2	90.2
Saturated	55.9	111.9	52.9	108.9	54.7	110.7

where

- μ represents mean parameterization.
- θ represents canonical parameterization.
- *med* means taking $\tilde{\theta}$ to be posterior median of θ.

Implementation: Scottish Lip Cancer Data

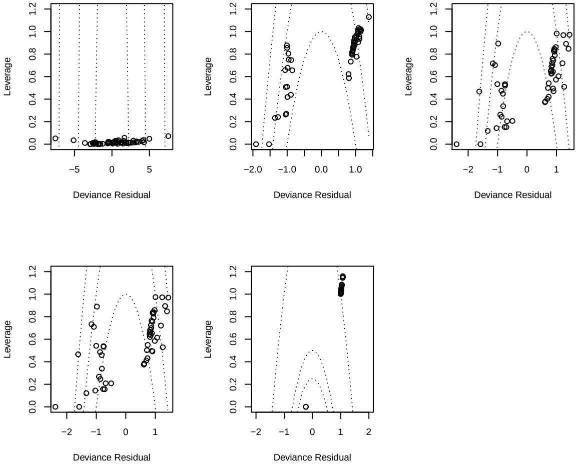


Figure: Diagnostics for the Scottish lip cancer example; residuals versus leverages for Models (1) through (5), from top left to bottom right

BPIC

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.

The form of the BPIC is quite complicated, but the penalty term reduces to $3p_D$ under similar approximations.

We note that DIC minimizes the posterior expected loss over a constrained space. If we were to repeat the proof above using $E_{z|\theta}[\mathcal{L}(Y,\theta)]$ instead of $E_{z|\theta}[\mathcal{L}(Y,\tilde{\theta}(y))]$ as the target, we can achieve $3p_D$ as the penalty term, as noted by van der Linde (2005).

Other Issues

The biggest issue (in my opinion) is that DIC is not invariant to parameterization, so you will obtain different DIC values depending on how you parameterize the model (as we noted with p_D).

We see this in the Scottish lip cancer example, although the authors also provide a more extreme demonstration in the paper (not shown).

To Do List

- Get the ICAR MCMC to work.
- Work through the last computational example in the paper (the six-cities study) - got the code to work, just haven't had time to compile or look at the results ...