Bayesian measures of model complexity and fit

D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde (2002)

Presented by: Lina Lin

STAT/BIOST 572

May 21, 2013

Classical approaches to model selection

The goal is to select the "best approximating model", which achieves the optimal balance between "fit" and "complexity."

- **Cross-validation** is the intuitive approach, but (1) validation data is not always readily available and (2) can be computationally intensive.
- An idea that has gained much traction over recent years is the use **information criterions**.
- Ex. **Akaike's Information Criterion** (Akaike, 1973):

$$AIC = -2 \log \mathcal{L}(\hat{\theta}) + 2p$$

- Ex. **Bayesian Information Criterion** (Schwarz, 1978):

$$BIC = -2 \log \mathcal{L}(\hat{\theta}) + 2p \log n$$

where $\mathcal{L}(\hat{\theta})$ denotes the maximized likelihood, p number of free parameters and n number of observations.

A Generalized Criterion for Hierarchical Models

For practical reasons, we would like to develop a criterion for hierarchical model (e.g. random effects model) selection.

Both **AIC** and **BIC** are characterized as a sum of a "fit" term ($\mathcal{L}$) and a "complexity term" (a monotonically increasing function of p).

- While this is by no means rigorous justification, we would expect our criterion to feature these components, which implies that **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: see Wahba (1990), Moody (1992), MacKay (1995) and Ripley (1996).

Effective number of parameters (p_{eff})

Consider the following random effects model:

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

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

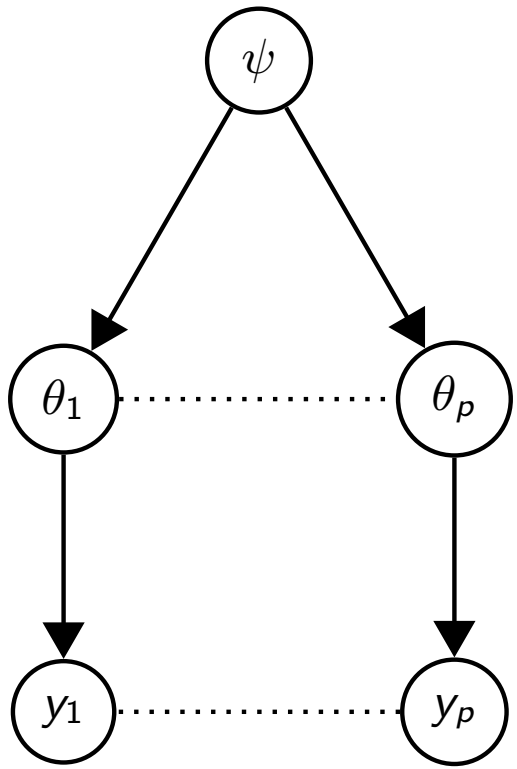
- **Question**: *Is* $p_eff = p$*?*

The presence of a prior induces **dependency** between the θ_i's, 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**. We will return to this idea later.

Effective number of parameters (p_{eff})

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



Parameter(s) of focus

The full probability model for (1) 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

OR

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

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. define 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.

We can also write p_D as:

$$p_D = \overline{D(\theta)} - D(\bar{\theta}) \tag{2}$$

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

- **Why?**

The "non-Bayesian" variant:

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

is an approximation to the complexity term found in the **Takeuchi Information Criterion (TIC)**.

p_D under normal approximation to the likelihood

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 secord-order Taylor expansion and obtain the following result:

$$\boxed{p_D \approx \text{tr}(-L''_{\bar{\theta}} V)} \tag{3}$$

where $V = E\{(\theta - \bar{\theta})(\theta - \bar{\theta})^T\}$.

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** (i.e. flat priors), we obtain,

$$\boxed{p_D \approx p} \tag{4}$$

A information-theoretic justification for p_D

The "**mutual information**" between Y and Θ is;

$$\mathcal{I}(\theta, Y) = \int p(\theta, y) \log \frac{p(\theta, y)}{p(\theta)p(y)} d\theta dy = KL(p(\theta, y), p(\theta)p(y))$$

and the symmetrized "mutual information":

$$\mathcal{J}(\theta, Y) = KL(p(\theta, y), p(\theta)p(y)) + KL(p(\theta)p(y), p(\theta, y))$$

where "KL" stands for **Kullback-Leibler divergence**, a measure of "distance" between two densities.

- "Mutual information" measures how **sensitive** the posterior distribution of θ is to the observations Y, and thus, corresponds to difficulty in estimation, which agrees with our notions on "model complexity."

- van der Linde (2005) showed that (for **conjugate exponential families**):

$$\boxed{p_D \approx \mathcal{J}(\theta_{post}, Z)} \tag{5}$$

where θ_{post} represents the posterior θ, and Z represents future observations derived from same data-generating mechanism.

A convincing example

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

$$\boxed{\begin{array}{rcl} Y & \sim & N(A_1\theta, C_1) \\ \theta & \sim & N(A_2\psi, C_2) \end{array}} \tag{6}$$

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**: how much of an effect each Y has on the overall fit.

This matches previous suggestions made by Ye (1998) for model complexity.

Deviance Information Criterion

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

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

Recall that:

$$AIC = -2 \log \mathcal{L}(\hat{\theta}) + 2p$$

- 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 $D(\bar{\theta}) \approx -2 \log \mathcal{L}(\hat{\theta})$ and $p_D \approx p$ 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 (Efron, 1986):

$$\begin{aligned} E_{z|\theta}[\mathcal{L}(Y, \tilde{\theta})(y)] &= \mathcal{L}(Y, \tilde{\theta}(y)) + c_{\Theta}\{y, \theta^t, \tilde{\theta}(y)\} \\ &= D(\tilde{\theta}) + c_{\Theta}\{y, \theta^t, \tilde{\theta}(y)\} \end{aligned} \tag{9}$$

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 (9) 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

$$(3) \quad E_{\theta|y}[\mathcal{L}_3] = E_{\theta|y}\{D(\theta) - D(\tilde{\theta})\} = p_D$$

.

The Derivation

We perform a Taylor series expansion of $\mathcal{L}_1$ about θ:

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

We note that $E_{z|\theta}\left\{\frac{\partial log[p(z|\theta)]}{\partial\theta}\right\} = 0$.

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

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.

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

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

$$\boxed{E_{\theta|y}[c_\Theta\{y, \theta, \tilde{\theta}(y)\}] \approx 2p_D} \tag{10}$$

Which implies:

$$\boxed{E_{\theta|y} E_{z|\theta}[\mathcal{L}(Y, \tilde{\theta}(y))] \approx D(\tilde{\theta}) + 2p_D = \overline{D(\theta)} + p_D} \tag{11}$$

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**: $\theta_i = \alpha_i$

We placed an improper flat prior on α_0, zero-mean normal priors with precision λ_γ on the γ_i's, an ICAR prior on the δ_i's (Besag, 1974) with precision parameter λ_δ , and weakly informative $\Gamma(0.5, 0.0005)$ on λ_γ and λ_δ.

Implementation: Scottish lip cancer data

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
	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
	42.9	104.0	43.4	104.5	43.5	104.6
Spatial	31.6	88.9	31.2	88.5	31.1	88.4
	31.7	89.9	31.2	89.5	31.1	89.3
Exchangeable + Spatial	32.6	90.6	32.2	90.2	32.2	90.2
	31.8	89.7	31.4	89.3	31.3	89.2
Saturated	55.9	111.9	52.9	108.9	54.7	110.7
	55.9	111.7	52.8	108.6	54.5	110.4

Table: Summary of calculated p_D and DIC values after running 15000 MCMC iterations following a burn-in period of 5000 iterations under the three different parameterizations: mean(μ), canonical (θ), and median(med)

Implementation: Six-cities study

We consider modelling a subset of data from the six-cities study, a longitudinal study of the health effects of air pollution (Fitmaurice and Laird, 1993).

$$\boxed{\begin{array}{lcl} Y_{ij} & \sim & Bern(p_{ij}) \\ p_{ij} & = & g^{-1}(\mu_{ij}) \\ \mu_{ij} & = & \beta_0 + \beta_1(a_{ij} - \overline{a}) + \beta_2(s_i - \overline{s}) + \beta_3(s_i a_{ij} - \overline{sa}) \end{array}} \tag{12}$$

where Y_{ij} is wheezing status (1 for yes, 0 for no) of child i at time j, s_i is smoking status of child i's mother, and a_{ij} is age of child i at time j.

The three models are:

1. $g(p_{ij}) = \log\{p_{ij}/(1 - p_{ij})\}$ logit link
2. $g(p_{ij}) = \Phi^{-1}(p_{ij})$ probit link
3. $g(p_{ij}) = \log\{-\log(1 - p_{ij})\}$ complementary log-log link

We place flat priors on $\beta_0, \ldots, \beta_3$, a normal prior on β with precision λ and a $\Gamma(0.001, 0.001)$ prior on λ.

Implementation: Six-cities study

Model	p_D^{μ}	DIC^{μ}	p_D^{θ}	DIC^{θ}
Logit	169.5	1334.3	247.2	1412.1
	168.9	1335.3	248.7	1415.1
Probit	159.0	1306.0	262.1	1409.1
	158.7	1307.3	262.7	1411.3
Complementary Log-Log	167.0	1350.8	224.1	1407.8
	167.2	1348.1	224.4	1405.3

Table: Summary of calculated p_D and DIC values after running 10000 MCMC iterations following a burn-in period of 10000 iterations under two different parameterizations: mean(μ) and canonical (θ)

.

BPIC

DIC does have the potential to select the model that over-fits since it "uses the data twice," and thus "under-penalizes" complexity.

- Ando (2007) developed a new criterion **Bayesian Predictive Information Criterion (BPIC)** which corrects for the over-fitting.

The form of BPIC is quite complicated, but its complexity term is similar to that of TIC's. The penalty term **reduces to** $3p_D$ **under similar approximations** made by Spiegelhalter et al. in the DIC derivation.

- 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).

Pros and Cons of DIC

Pros	Cons
1. Can be readily computed using MCMC. 2. Equivalent to AIC under vague priors.	1. Not invariant to parameterization. 2. Selection of $\tilde{\theta}$ arbitrary; no guidelines 3. Does not work for mixture models, or in general when posterior densities are non log-concave.

Key References

- D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde, "Bayesian measures of model complexity and fit", Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 64, no. 4, pp. 538639, 2002.
- H. Akaike, "A new look at the statistical model identification", Automatic Control, IEEE Transactions on, vol. 19, no. 6, pp. 716723, 1974.
- S. Kullback and R.A. Leibner, "On information and sufficiency", Annals of Mathematical Statistics, vol. 22, no. 1, pp. 7986, 1951.
- A. van der Linde. "DIC in variable selection", Statistical Neerlandica, vol. 59, no. 1, pp. 45-56, 2005.
- T. Ando, "Bayesian predictive information criterion for the evaluation of hierarchical bayesian and empirical bayes models", Biometrika, vol. 94, no. 2, pp. 443458, 2007.