STAT 431 — Applied Bayesian Analysis — Course Notes

Model Selection Criteria

Fall 2022

You may have heard of model selection criteria like AIC or BIC ...

Several data models for the *same* data $\boldsymbol{y}$ are under consideration — possibly nested, possibly intersecting, possibly unrelated. They can differ in type and number of parameters.

Model selection criteria aim to answer the question

Which is the “best” model for the data?

Bayesians want a criterion that evaluates the *prior*, too ...

Suppose you have several candidate data models for the *same* data $\boldsymbol{y}$:

$$\mathcal{M}_1, \;\; \mathcal{M}_2, \;\; \ldots \;\; \mathcal{M}_m$$

Suppose each data model has its own prior.

Goal: Choose the “best” data model/prior combination for predicting new data (of the same kind).

Let $\mathcal{M}$ be a particular data model, parameterized by $\boldsymbol{\theta}$ (continuous), under which the data have a density

$$f(\boldsymbol{y} \mid \boldsymbol{\theta})$$

and let $\pi(\boldsymbol{\theta})$ be a prior density.

Let $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ be the resulting posterior density.

Define

$$D(\boldsymbol{y} \mid \boldsymbol{\theta}) \quad = \quad -2 \, \ln f(\boldsymbol{y} \mid \boldsymbol{\theta})$$

(Values of this are analytically computable, provided the density can be evaluated.)

Define

$$\begin{aligned}\bar{D}(\boldsymbol{y}) &= \int D(\boldsymbol{y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \boldsymbol{y})\, d\boldsymbol{\theta} \\ &= D \text{ averaged over the posterior}\end{aligned}$$

and

$$\hat{D}(\boldsymbol{y}) = D(\boldsymbol{y} \mid \hat{\boldsymbol{\theta}})$$

where $\hat{\boldsymbol{\theta}}$ is the posterior mean:

$$\hat{\boldsymbol{\theta}} = \mathrm{E}(\boldsymbol{\theta} \mid \boldsymbol{y})$$

Note: Both $\bar{D}(\boldsymbol{y})$ and $\hat{D}(\boldsymbol{y})$ can be approximated using posterior simulation.

Define

$$p_D = \bar{D}(\boldsymbol{y}) - \hat{D}(\boldsymbol{y})$$

as the **effective number of parameters**.

Note: This is *not* what a frequentist would call the "effective number of parameters." It is usually not an integer, and could possibly be negative!

The **deviance information criterion (DIC)** value (for data model $\mathcal{M}$ with prior $\pi(\boldsymbol{\theta})$) is

$$\begin{aligned} DIC &= \bar{D}(\boldsymbol{y}) + p_D \\ &= 2\,\bar{D}(\boldsymbol{y}) - \hat{D}(\boldsymbol{y}) \\ &= \hat{D}(\boldsymbol{y}) + 2p_D \end{aligned}$$

Usage: Choose the data model/prior with minimal DIC.

Remark: Motivation for DIC is similar to the frequentist-motivated AIC —

$$AIC \;=\; D(\boldsymbol{y} \mid \hat{\boldsymbol{\theta}}_{\mathsf{MLE}}) \,+\, 2p$$

where $\hat{\boldsymbol{\theta}}_{\mathsf{MLE}}$ is the maximum likelihood estimate, and p is the "true" (frequentist) number of free parameters (in $\boldsymbol{\theta}$).

- the first term penalizes lack of fit
- the second term penalizes complexity

Warning: Data models compared using DIC (or AIC or BIC) must be for *exactly* the same data $\boldsymbol{y}$.

For example, if the data are transformed, exactly the same transformation must be used for all models being compared.

Remark: Software usually assumes that $\boldsymbol{\theta}$ contains *all* parameters and hyperparameters (at all levels) of a hierarchical model.

Example: Baby Rat Weights

Model 1: Bivariate Formulation (Separate Lines)

$$\begin{aligned} Y_{ij} &= \text{weight (mass) of rat } i \text{ at time } X_j \\ &\sim \ \textit{indep} \ \mathrm{Normal}\big(\alpha_{i1} + \alpha_{i2}(X_j - \bar{X}),\ \sigma_y^2\big) \end{aligned}$$

$$\boldsymbol{\alpha}_i \ = \ \begin{bmatrix} \alpha_{i1} \\ \alpha_{i2} \end{bmatrix} \ \Bigg| \ \boldsymbol{\beta}, \boldsymbol{\Omega} \quad \sim \quad \textit{iid} \ \mathrm{Normal}(\boldsymbol{\beta}, \boldsymbol{\Omega})$$

(with appropriate priors on σ_y^2, $\boldsymbol{\beta}$, $\boldsymbol{\Omega}$)

Model 2: Univariate Formulation, Separate Lines

$$Y_{ij} \quad \sim \quad indep\ \mathrm{Normal}\big(\alpha_{i1} + \alpha_{i2}(X_j - \bar{X}),\, \sigma_y^2\big)$$

$$\left.\begin{array}{rcl} \alpha_{i1} \mid \beta_1, \sigma_{\alpha_1}^2 & \sim & \mathrm{Normal}(\beta_1, \sigma_{\alpha_1}^2) \\ \alpha_{i2} \mid \beta_2, \sigma_{\alpha_2}^2 & \sim & \mathrm{Normal}(\beta_2, \sigma_{\alpha_2}^2) \end{array}\right\} \begin{array}{c}\text{all}\\ \text{conditionally}\\ \text{independent}\end{array}$$

(with vague independent priors on σ_y^2, β_1, β_2, σ_{α_1}, σ_{α_2})

Model 3: Separate Intercepts, Common Slope

$$Y_{ij} \quad \sim \quad \textit{indep} \ \text{Normal}\big(\alpha_{i1} + \beta_2(X_j - \bar{X}),\, \sigma_y^2\big)$$

$$\alpha_{i1} \mid \beta_1, \sigma_{\alpha_1}^2 \quad \sim \quad \textit{indep} \ \text{Normal}(\beta_1, \sigma_{\alpha_1}^2)$$

Model 4: Same Line (SLR)

$$Y_{ij} \quad \sim \quad \textit{indep} \ \text{Normal}\big(\beta_1 + \beta_2(X_j - \bar{X}),\, \sigma_y^2\big)$$

R/JAGS Example 5.1:

DIC for Hierarchical Normal Regressions

Remarks:

- JAGS uses a different version of p_D suggested by Plummer (in the discussion of Spiegelhalter et al., 2002).
- As an alternative to DIC, a more recent criterion is the Watanabe-Akaike information criterion (WAIC) — see BSM, Sec. 5.5.