Optimal Predictive Model Selection

Jim Berger

Institute of Statistics and Decision Sciences

Duke University

joint with Marilena Barbieri, J.K. Ghosh and Nitai Mukhopadhyay

Outline

- Linear models
- Basics of Bayesian prediction
- Optimal predictive choice of a single model: the (posterior) median probability model
- Misconceptions about BIC and AIC in prediction

Normal linear models

- Observe the $n \times 1$ vector

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}\,;$$

  where $\mathbf{X}$ is the $n \times k$ design matrix, $\boldsymbol{\beta}$ is the $k \times 1$ vector of unknown coefficients, and $\boldsymbol{\epsilon}$ is $\mathcal{N}(0, \sigma^2 I)$.

- Choose from among submodels

$$M_{\mathbf{l}} : \boldsymbol{y} = \mathbf{X}_{\mathbf{l}}\,\boldsymbol{\beta}_{\mathbf{l}} + \boldsymbol{\epsilon}\,,$$

  where $\mathbf{l} = (l_1, l_2, \ldots, l_k)$ is the model index, l_i being either 1 or 0 as covariate x_i is in or out of the model.

Basics of Bayesian prediction

- The goal is to predict a future $y^* = \mathbf{x}^* \boldsymbol{\beta} + \epsilon$, using squared error loss $(y^* - \widehat{y}^*)^2$.

- Combining the data and prior yields, for all $\mathbf{l}$,
  - $P(M_\mathbf{l} \mid \boldsymbol{y})$, the posterior probability of model $M_\mathbf{l}$;
  - $\pi^*_\mathbf{l}(\boldsymbol{\beta}_\mathbf{l}, \sigma \mid \boldsymbol{y})$, the posterior distribution of $(\boldsymbol{\beta}_\mathbf{l}, \sigma)$.

- The best predictor of y^* is, via *model averaging*,

$$\bar{\boldsymbol{y}}^* = \mathbf{x}^* \, \bar{\boldsymbol{\beta}} \equiv \mathbf{x}^* \sum_{\mathbf{l}} P(M_\mathbf{l} \mid \boldsymbol{y}) \, \widetilde{\boldsymbol{\beta}}_\mathbf{l} \, ,$$

  where $\widetilde{\boldsymbol{\beta}}_\mathbf{l}$ is the posterior mean for $\boldsymbol{\beta}$ under $M_\mathbf{l}$.

Selecting a single model

- Often a single model is desired for prediction.
- A common misperception is that the best single model is that with the largest $P(M_{\mathbf{l}} \mid \boldsymbol{y})$;
  - however, this is true if there are only two models;
  - and it is true if $\mathbf{X}'\mathbf{X}$ is diagonal, σ^2 is known, and suitable priors are used (Clyde and Parmigiani, 1996).
- The best single model will typically depend on $\mathbf{x}^*$.
- An important case is when the future covariates are like the past covariates, i.e., when $E[(\mathbf{x}^*)'\mathbf{x}^*] = \mathbf{X}'\mathbf{X}$.

The (posterior) median probability model

If it exists, the (posterior) *median probability model*, $M_{\mathbf{l}^*}$, is defined to be the model consisting of those variables which occur with posterior probability at least 1/2, when summed across models. Formally, $\mathbf{l}^*$ is defined, coordinatewise, by

$$l_i^* = 1 \text{ if } \sum_{\mathbf{l}\,:\,l_i=1} P(M_{\mathbf{l}}|\boldsymbol{y}) \geq \tfrac{1}{2}\,;\ l_i^* = 0, \text{ otherwise.}$$

Note: If a model-jumping MCMC is being used to do the Bayesian analysis, the median probability model simply consists of those coordinates that were present in over half the iterations.

Two optimality theorems

Theorem 1. If (i) $\mathbf{X}'\mathbf{X}$ is diagonal and (ii) the posterior mean of $\boldsymbol{\beta}_{\mathbf{l}}$ is found by simply taking the relevant coordinates of $\widetilde{\boldsymbol{\beta}}$ (the posterior mean in the full model), then the best predictive model is the median probability model. Condition (ii) is satisfied if either

- noninformative priors are used to determine the posterior distributions of model parameters; or
- σ^2 is known and independent normal priors are used for the β_i. *(In this case, the median probability model is often equal to the maximum probability model.)*

Theorem 2. Suppose a sequence of nested linear models is under consideration. If (i) prediction is desired at 'future covariates like the past' and (ii) the posterior mean under $M_{\mathbf{l}}$ satisfies $\widetilde{\boldsymbol{\beta}}_{\mathbf{l}} = b\,\widehat{\boldsymbol{\beta}}_{\mathbf{l}}$, where $\widehat{\boldsymbol{\beta}}_{\mathbf{l}}$ is the least squares estimate, then the best predictive model is the median probability model.

Condition (ii) is satisfied if either

- noninformative priors are used to determine the posterior distributions of model parameters; or
- g-type normal priors are used.

Example (Polynomial Regression): Model j is

$$y = \sum_{i=0}^{j} \beta_i \, x^i + \epsilon.$$

(Model) j	0	1	2	3	4	5	6
$P(M_j \mid \boldsymbol{y})$	~ 0	0.06	0.22	0.29	0.38	0.05	~ 0

(Covariate) i	0	1	2	3	4	5	6
$P(x^i \text{ is in model} \mid \boldsymbol{y})$	1	1	0.94	0.72	0.43	0.05	0

Thus M_3 is the median probability (optimal predictive) model, while M_4 is the maximum probability model.

Misconceptions about BIC and AIC

- In prediction, Bayesian selection of a single model is inferior to use of AIC. (*Actually, Bayesian selection is often better.*)
- In prediction, the chosen Bayesian model is typically too small, smaller than AIC. (*Actually, the Bayesian model is often larger, with the extra coefficients being 'shrunk' part way towards zero.*)
- BIC is a satisfactory Bayesian tool when high dimensional models are under consideration. (*This is not true and has caused the above misconceptions.*)

Definitions of BIC and AIC

- BIC (Bayes Information Criterion), Schwarz (1978)

$$BIC_{\mathbf{l}} = \max_{\theta \in \Theta_{\mathbf{l}}} \log f_{\mathbf{l}}(\mathbf{y}|\theta) - k_{\mathbf{l}} \frac{\log n}{2}$$

  $k_{\mathbf{l}}$ = dimension of the model $M_{\mathbf{l}}$ and n = sample size.

- AIC (Akaike Information Criteria), Akaike (1973)

$$AIC_{\mathbf{l}} = \max_{\theta \in \Theta_{\mathbf{l}}} \log f_{\mathbf{l}}(\mathbf{y}|\theta) - k_{\mathbf{l}}.$$

Stone's Example (Stone, 1979):

$$y_{ij} = \mu_i + \epsilon_{ij}, \;\; \epsilon_{ij} \sim N(0,1), \; i = 1, \ldots, k, \;\; j = 1, \ldots, r.$$

Compare the two models

$$M_1 : \mu_i = 0 \text{ for all } i, \quad \text{versus} \quad M_2 : \mu \in \Re^k.$$

Of interest is what happens when

- M_2 is the true model;
- The true model dimension $k \to \infty$ as the overall sample size $n = rk \to \infty$.

For Stone's example, use of BIC or AIC would lead to the model selection criterion

- $\Delta BIC = BIC_2 - BIC_1 = \frac{r}{2}\bar{\mathbf{y}}'\bar{\mathbf{y}} - \frac{k}{2}\log n$;
- $\Delta AIC = AIC_2 - AIC_1 = \frac{r}{2}\bar{\mathbf{y}}'\bar{\mathbf{y}} - k$;

here $\bar{\mathbf{y}} = (\bar{y}_1, \ldots, \bar{y}_p)'$. Stone (1979) showed that AIC is consistent as k grows with n, but BIC is not consistent.

The reason BIC **fails:** BIC is based on Laplace approximation of marginal densities of $\mathbf{y}$, but these can be highly inaccurate when k can grow with n.

A Solution: *Generalized Bayes Information Criterion*

- Based on a type of Laplace approximation that is also valid as $k \to \infty$.
- For Stone's example, this gives the model selection criterion

$$\Delta GBIC = \left(\frac{r}{2}\bar{\mathbf{y}}'\bar{\mathbf{y}} - \frac{k}{2}\log(\frac{r}{k}\,\bar{\mathbf{y}}'\bar{\mathbf{y}}) - \frac{k}{2}\right)^{+} - \frac{\log k}{2},$$

where '$+$' denotes the positive part.

- This is a consistent criterion as $k \to \infty$.
- BIC is a special case.

$\frac{r}{k}\bar{\mathbf{y}}'\bar{\mathbf{y}}$	$logBF^c$	BIC	GBIC
1.0	-3.8251	-90.129	-1.956
2.0	6.0388	-65.129	5.715
3.0	20.8203	-40.129	20.579
4.0	38.4814	-15.129	38.387
4.2	42.235	-10.129	42.167
20.0	397.369	384.871	398.151

Table 1: Log(Bayes factor) and its approximations under a Cauchy prior; k=50, r=2; Normal error

Conclusions

- For prediction, one should compute the (posterior) median probability model, as well as the maximum probability model, and compare for predictive power.
- BIC is not trustworthy unless model dimensions are small compared to the sample size. (No existing simple method is trustworthy.)
- $GBIC$ shows promise, if it can be extended to arbitrary linear models.
- The predictive model chosen by a full Bayesian analysis can be larger than, and outperform AIC.