(Predictive) model selection

Marilena Barbieri

Università Roma Tre

O'Bayes 2019: Objective Bayes Methodology Conference
University of Warwick, 28 June - 02 July 2019

(Predictive) model selection

Marilena Barbieri

Università Roma Tre

Outline:

- Basics of Bayesian model selection
- Variable selection in linear model: the Median probability model

O'Bayes 2019: Objective Bayes Methodology Conference
University of Warwick, 28 June - 02 July 2019

General model selection: notation

Data $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ is assumed to have arisen from one of several models:

M_1: $\boldsymbol{y}$ has density $f_1(\boldsymbol{y}|\boldsymbol{\beta}_1)$

M_2: $\boldsymbol{y}$ has density $f_2(\boldsymbol{y}|\boldsymbol{\beta}_2)$

$\vdots$

M_q: $\boldsymbol{y}$ has density $f_q(\boldsymbol{y}|\boldsymbol{\beta}_q)$

$\boldsymbol{\beta}_i$ $(i = 1, \ldots, q)$ model specific parameter with dimension h_i.

Assign a prior probability $P(M_i)$ to each model.

A common choice: $P(M_i) = 1/q$

For small q, $P(M_i) = 1/q$ is sensible, although having decreasing $P(M_i)$ as model complexity increases is often reasonable.

Care is neeeded if q is huge and multiplicity is an issue.

Linear models with k variables and a common constant:

$$P(M_i) = \frac{1}{k+1} \left(\begin{array}{c} k \\ h_i \end{array} \right)^{-1}$$

which is equivalent to assigning the set of all models having the same number of parameters probability $1/(\mathrm{k}+1)$ and dividing this probability equally among them.

Under model M_i:

Prior density of β_i: $\pi_i(\boldsymbol{\beta}_i)$

Marginal density: $m_i(\boldsymbol{y}) = \int f_i(\boldsymbol{y}|\boldsymbol{\beta}_i)\pi_i(\boldsymbol{\beta}_i)d\,\boldsymbol{\beta}_i$
which measures "how likely is y under M_i"

Posterior density: $\pi_i(\boldsymbol{\beta}_i|\boldsymbol{y}) = f_i(\boldsymbol{y}|\boldsymbol{\beta}_i)\pi_i(\boldsymbol{\beta}_i)/m_i(\boldsymbol{y})$

Bayes factor of M_j to M_i:

$$B_{ji} = \frac{m_j(\boldsymbol{y})}{m_i(\boldsymbol{y})}$$

Posterior probability of M_i:

$$P(M_i|\boldsymbol{y}) = \frac{P(M_i)m_i(\boldsymbol{y})}{\sum_{j=1}^{q} P(M_j)m_j(\boldsymbol{y})} = \left[\sum_{j=1}^{q} \frac{P(M_j)}{P(M_i)} B_{ji}\right]^{-1}$$

An important note of caution

- Improper priors are problematical, because they are unnormalized:

  $$\pi_i(\beta_i) = c_i g_i(\beta_i)$$

  Can not be used for parameters not occurring in all models since c_i is arbitrary and B_{ij} is arbitrary

$$B_{ji} = \frac{\int f_j(\boldsymbol{y}|\boldsymbol{\beta}_j)\pi_j(\boldsymbol{\beta}_j)d\,\boldsymbol{\beta}_j}{\int f_i(\boldsymbol{y}|\boldsymbol{\beta}_i)\pi_i(\boldsymbol{\beta}_i)d\,\boldsymbol{\beta}_i} = \frac{c_j \int f_j(\boldsymbol{y}|\boldsymbol{\beta}_j)g_j(\boldsymbol{\beta}_j)d\,\boldsymbol{\beta}_j}{c_i \int f_i(\boldsymbol{y}|\boldsymbol{\beta}_i)g_i(\boldsymbol{\beta}_i)d\,\boldsymbol{\beta}_i}$$

- *Vague* proper priors (arbitrarily large variances) may be even worse.

Basic of Bayes prediction

For each model M_i, a future observation y^* has density $g_i(y^*|\boldsymbol{y}, \boldsymbol{\beta}_i)$.

The posterior predictive distribution of y^*, once $\boldsymbol{y}$ has been observed, is defined as

$$p_i(y^*|\boldsymbol{y}) = \int g_i(y^*|\boldsymbol{y}, \boldsymbol{\beta}_i)\, \pi_i(\boldsymbol{\beta}_i|\boldsymbol{y})\, d\boldsymbol{\beta}_i$$

A nice feature of the posterior predictive distribution is that the posterior predictive is proper, whenever the posterior distribution $\pi_i(\boldsymbol{\beta}_i|\boldsymbol{y})$ is proper.

Multiple models setting

When considering the set of parametric models $\{M_i\}$, the posterior predictive distribution, obtained by Bayesian Model Averaging, is

$$\begin{aligned} p(y^*|\boldsymbol{y}) &= \sum_i \int g_i(y^*|\boldsymbol{y}, \boldsymbol{\beta}_i)\, \pi_i(\boldsymbol{\beta}_i|\boldsymbol{y})\, P(M_i|\boldsymbol{y})\, d\boldsymbol{\beta}_i = \\ &= \sum_i P(M_i|\boldsymbol{y})\, p_i(y^*|\boldsymbol{y}) \end{aligned}$$

Basic of predictive model selection

Comparing the predictive performances of several models is often (if not always explicitly) formulated as a decision problem. The optimal predictive model is the one that minimizes the expected loss

$$\int L(m_i, a_i^*, y^*)\, p(y^*|\boldsymbol{y})\, dy^*$$

over $i = 1, 2, \ldots, q$, where m_i corresponds to selecting M_i and a_i^* denotes the optimal subsequent decision regarding a future observation y^*.

Basic of predictive model selection

The expected loss takes the form

$$-\int \log\left[p_i(y^*|\boldsymbol{y})\right]\; p(y^*|\boldsymbol{y})\, dy^*$$

for a predictive distribution of y^* with logarithmic score function and

$$\int [y^* - \hat{y}_i^*]^2\; p(y^*|\boldsymbol{y})\, dy^*$$

for point prediction with quadratic loss, where $\hat{y}_i^* = \mathbb{E}(y^*|\boldsymbol{y}, M_i)$ is the optimal prediction of a future observation y^*, given $\boldsymbol{y}$ and assuming model M_i, that is the value which minimizes

$$\int (y^* - \hat{y}^*)^2\, p_i(y^*|\boldsymbol{y})\, dy^*.$$

Motivation for the Bayesian approach

1. Ease of interpretation:
   $P(M_i|\boldsymbol{y})$ reflects real expected error rates.

2. Prior information can be incorporated, if desired:
   It is useful to separately report $\{m_i(\boldsymbol{y})\}$ (or $\{B_{ji}\}$) and $\{P(M_i)\}$. This allows computation of the $P(M_i|\boldsymbol{y})$ for any prior probabilities.

3. Consistency:
   If one of the M_i is true, then $P(M_i|\boldsymbol{y}) \to 1$ as $n \to \infty$.

Motivation for the Bayesian approach

4. Ockham's razor:
   Bayes Factors automatically seek parsimony; no adhoc penalties for model complexity are needed. (Jefferys and Berger, 1992)

5. Generality of application:
   The approach is viable for any models, regular or irregular, nested or not, large or small sample sizes, two or multiple models.

Motivation for the Bayesian approach

6. Accounting for model uncertainty:
   Selecting a single model and using it for inference ignores model uncertainty, resulting in

   inferior inferences,
   considerable overstatements of accuracy.

   The Bayesian approach incorporates this uncertainty by model averaging: if inference concerning θ (*same* meaning across models) is desired, it would be based on

$$\pi(\theta|\boldsymbol{y}) = \sum_i P(M_i|\boldsymbol{y})\, \pi_i(\theta|\boldsymbol{y}).$$

   Most frequent use is for prediction (Geisser 93, Draper 95, Raftery et al. 97, Clyde 99, Clyde and George, 04, ...)

Prediction with Normal linear models

Under the full model, the $n \times 1$ observation vector would follow

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\mathbf{X}$ is the $n \times k$ $(k < n)$ full rank design matrix of covariates, $\boldsymbol{\beta}$ is a $k \times 1$ vector of unknown coefficients and $\boldsymbol{\varepsilon}$ is $N(0, \sigma^2 I)$.
The possible submodels are

$$M_{\boldsymbol{I}} : \boldsymbol{y} = \mathbf{X}_{\boldsymbol{I}}\boldsymbol{\beta}_{\boldsymbol{I}} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{I} = (I_1, I_2, \ldots, I_k)$ is the model index, I_i being either 1 or 0 as covariate x_i is in or out of the model.

Assume that one of these models is true, and our goal is to predict a future observation at covariate values $\mathbf{x}^*$

$$y^* = \mathbf{x}^* \boldsymbol{\beta} + \varepsilon$$

using square error loss

$$(y^* - \widehat{y}^*)^2.$$

The best predictor of y^* is, via model averaging, the posterior mean of y^*

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

where $\widetilde{\boldsymbol{\beta}}_{\boldsymbol{l}}$ is the posterior mean of $\boldsymbol{\beta}$ under $M_{\boldsymbol{l}}$ (adding any needed zeroes to the β)

Selecting a single model

Often a single model $M_{\boldsymbol{l}}$ is desired for prediction, with the prediction then being $\widehat{y}^*_{\boldsymbol{l}} = \mathbf{x}^* \widetilde{\boldsymbol{\beta}}_{\boldsymbol{l}}$.

It is commonly perceived that the best model will be that with the highest $P(M_{\boldsymbol{l}}|\boldsymbol{y})$. This is true if:

- there are only two models;
- $X'X$ is diagonal, σ^2 is known, and suitable priors are used.

Selecting a single model

The best single model is the one that minimizes (where the expectation is with respect to the predictive distribution of y^* and ignoring irrelevant terms)

$$\begin{aligned}\mathbb{E}[(y^* - \widehat{y}^*_{\boldsymbol{I}})^2] &= (\widehat{y}^*_{\boldsymbol{I}} - \bar{y}^*)^2 = \\ &= \left(\widetilde{\boldsymbol{\beta}}_{\boldsymbol{I}} - \bar{\boldsymbol{\beta}}\right)' \mathbf{x}^{*\prime}\mathbf{x}^* \left(\widetilde{\boldsymbol{\beta}}_{\boldsymbol{I}} - \bar{\boldsymbol{\beta}}\right)\end{aligned}$$

and will typically depend on $\mathbf{x}^*$.

Selecting a single model

An important case is when the model will repeatedly be used to make future predictions, and covariates that will occur in the future are like those that occurred in the data, in the sense that $\mathbb{E}[(\mathbf{x}^*)'(\mathbf{x}^*)] = \mathbf{X}'\mathbf{X}$

Then the averaged squared error predictive loss for future predictions, using $M_{\boldsymbol{I}}$ is

$$\mathbb{E}^{\mathbf{x}^*}(\widehat{y}^*_{\boldsymbol{I}} - \bar{y}^*)^2 = \left(\widetilde{\boldsymbol{\beta}}_{\boldsymbol{I}} - \bar{\boldsymbol{\beta}}\right)' \mathbf{X}'\mathbf{X} \left(\widetilde{\boldsymbol{\beta}}_{\boldsymbol{I}} - \bar{\boldsymbol{\beta}}\right)$$

Posterior inclusion probabilities

The *posterior inclusion probability* for variable i is

$$p_i \equiv \sum_{\boldsymbol{l}:l_i=1} P(M_{\boldsymbol{l}}|\boldsymbol{y}),$$

that is, the overall posterior probability that variable i is in the model.
These are of interest in defining the (posterior) median probability model.

The (posterior) median probability model

If it exists, the *median probability model*, $M_{\boldsymbol{l}^*}$, is defined to be the model consisting of those variables whose posterior inclusion probability is at least $1/2$.
Formally, $\boldsymbol{l}^*$ is defined, coordinatewise, by

$$l_i^* = \begin{cases} 1, & \text{if } p_i \geq \frac{1}{2}, \\ 0, & \text{otherwise.} \end{cases}$$

Note: If computation is done by MCMC, the median probability model consists of those coordinates that were present in over half the iterations.

Existence of the median probability model

The median probability model exists when the models under consideration follow a *graphical model structure*, including:

- when any subset of variables is allowed;
- the situation in which the allowed variables consist of main effects and interactions, but a higher order interaction is allowed only if lower order interactions are included;
- a sequence of nested models, such as arises in polynomial regression and autoregressive time series.

Example: Polynomial Regression

Model M_j is

$$y = \sum_{i=0}^{j} \beta_i x^i + \varepsilon$$

(Model) j	0	1	2	3	4	5	6
$P(Mj\|y)$	≈ 0	0.06	0.22	0.29	0.38	0.05	≈ 0

(Covariate) i	0	1	2	3	4	5	6
$P(x^i$ is in model$\|y)$	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.

For nested models, the median probability model has a simpler representation as $M_{\boldsymbol{I}(j^*)}$, where j^* is such that

$$\sum_{i=0}^{j^*-1} P(M_{\boldsymbol{I}(i)}|\boldsymbol{y}) < \frac{1}{2} \quad \text{and} \quad \sum_{i=0}^{j^*} P(M_{\boldsymbol{I}(i)}|\boldsymbol{y}) \geq \frac{1}{2}.$$

In other words, one just lists the sequence of posterior model probabilities and sums them up until the sum exceeds $1/2$. The model at which the exceedance occurs is the median probability model.

Optimality theorems

Theorem 1. If

(i) the models under consideration have graphical structure,
(ii) $X'X$ is diagonal, and
(iii) the posterior mean of $\boldsymbol{\beta_I}$ is simply 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 (iii) is satisfied under any mix of

constant priors for the β_i;

independent $N(0; \sigma^2\,\lambda_i)$ priors for the β_i, with the λ_i given (objectively or subjectively specified, or estimated via empirical Bayes) and any prior for σ^2.

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 $M_{\boldsymbol{l}}$ satisfies $\widetilde{\boldsymbol{\beta}}_{\boldsymbol{l}} = b\widehat{\boldsymbol{\beta}}_{\boldsymbol{l}}$, where $b > 0$; that is, the posterior means are proportional to the least squares estimates, with the same proportionality constant across models

then the best predictive model is the median probability model.

Condition (ii) is satisfied if we use either

the objective priors for model parameters; or

g-type $\mathcal{N}_{k_{\boldsymbol{l}}}(\mathbf{0}, c\sigma^2(\mathbf{X}_{\boldsymbol{l}}'\mathbf{X}_{\boldsymbol{l}})^{-1})$ prior with the same constant $c > 0$ for each model and any prior for σ^2.

Theorems 1 and 2 essentially remain true even if there are non-orthogonal nuisance parameters (i.e., parameters common to all models) that are assigned the usual noninformative priors.

Example.

Consider Hald's regression data [Draper and Smith (1981)], consisting of $n = 13$ observations on a dependent variable y, with four (highly correlated) potential regressors: x_1, x_2, x_3, x_4. The full model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

$\varepsilon \sim N(0, \sigma^2)$, σ^2 unknown.

- All models that include the constant term are considered.
- Least squares estimates are used for parameters.
- Default posterior probabilities of each model are obtained using the Encompassing Arithmetic Intrinsic Bayes Factor (Berger and Pericchi, 1996) for linear models, together with equal prior model probabilities.

Model	$P(M_{\boldsymbol{l}}\|\boldsymbol{y})$	$R(M_{\boldsymbol{l}})$	Model	$P(M_{\boldsymbol{l}}\|\boldsymbol{y})$	$R(M_{\boldsymbol{l}})$
c	0.000003	2652.44	c,2,3	0.000229	353.72
c,1	0.000012	1207.04	c,2,4	0.000018	821.15
c,2	0.000026	854.85	c,3,4	0.003785	118.59
c,3	0.000002	1864.41	c,1,2,3	0.170990	1.21
c,4	0.000058	838.20	c,1,2,4	0.190720	0.18
c,1,2	0.275484	8.19	c,1,3,4	0.159959	1.71
c,1,3	0.000006	1174.14	c,2,3,4	0.041323	20.42
c,1,4	0.107798	29.73	c,1,2,3,4	0.049587	0.47

Computation of the posterior inclusion probabilities yields

$$p_1 = \sum_{\boldsymbol{l}:l_1=1} P(M_{\boldsymbol{l}}|\boldsymbol{y}) = 0.954556, p_2 = \sum_{\boldsymbol{l}:l_2=1} P(M_{\boldsymbol{l}}|\boldsymbol{y}) = 0.728377,$$

$$p_3 = \sum_{\boldsymbol{l}:l_3=1} P(M_{\boldsymbol{l}}|\boldsymbol{y}) = 0.425881, p_4 = \sum_{\boldsymbol{l}:l_4=1} P(M_{\boldsymbol{l}}|\boldsymbol{y}) = 0.553248.$$

Model	$P(M_l\|\boldsymbol{y})$	$R(M_l)$	Model	$P(M_l\|\boldsymbol{y})$	$R(M_l)$
c	0.000003	2652.44	c,2,3	0.000229	353.72
c,1	0.000012	1207.04	c,2,4	0.000018	821.15
c,2	0.000026	854.85	c,3,4	0.003785	118.59
c,3	0.000002	1864.41	c,1,2,3	0.170990	1.21
c,4	0.000058	838.20	c,1,2,4	0.190720	0.18
c,1,2	0.275484	8.19	c,1,3,4	0.159959	1.71
c,1,3	0.000006	1174.14	c,2,3,4	0.041323	20.42
c,1,4	0.107798	29.73	c,1,2,3,4	0.049587	0.47

- The median probability model is $\{c, x_1, x_2, x_4\}$ which clearly coincides with the optimal predictive model.
- Note that the risk of the maximum probability model $\{c, x_1, x_2\}$ is considerably higher than that of the median probability model.

R package BayesVarSel (Garcia-Donato and Forte)

- freely available at CRAN
- a number of different priors on models and parameters
- in the output: posterior inclusion probabilities, the highest probability model, the median probability model, the Bayesian model average predictor of y^* at covariates x^*

When the median probability model can fail (Merlise Clyde)
Suppose that

- the only models under consideration are M_0 with a constant term and the models M_i with the constant term and a single covariate x_i, $i = 1, \ldots, k$, with $k \geq 3$;
- all models have equal prior probability of $1/(k+1)$;
- all covariates are nearly perfectly correlated, with each other and with y.

Then the posterior probability of M_0 will be near zero, and that of each of the M_i will be approximately $1/k$.
Since these posterior inclusion probabilities are less than $1/2$, the median probability model will be M_0, which will have very poor predictive performance compared to any of the other models.

... more on this next time!

Thank you for your attention!