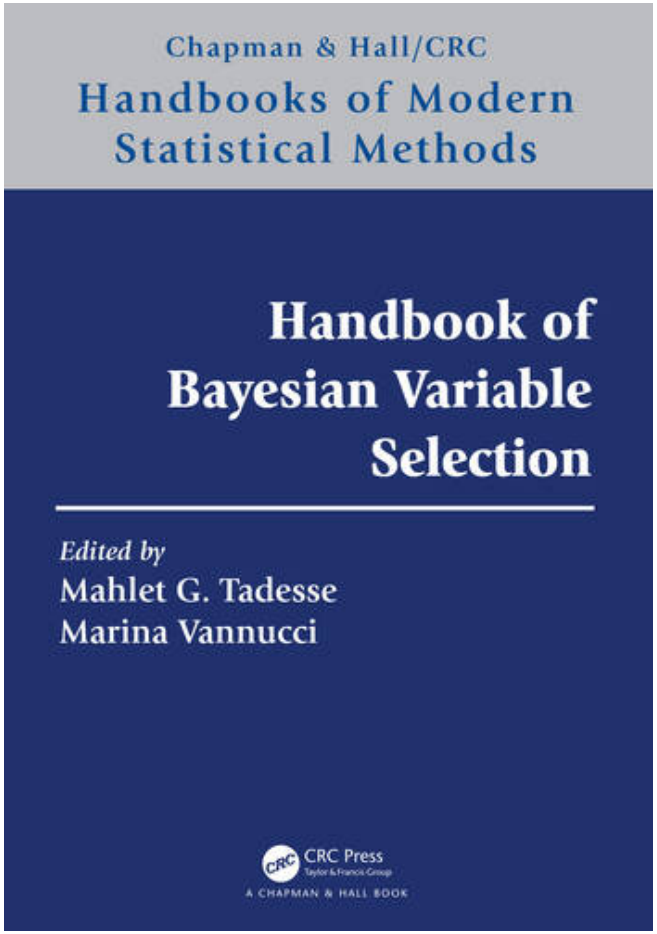
Bayesian Variable Selection - Basics

Jim Griffin

University College London

Overall structure

- Basics – Linear regression with spike-and-slab priors, prediction, summarisation
- Computation – Common computational algorithms, methods for high-dimensional regression
- GLMs and hierarchical models – Extensions needed for GLMs and use in hierarchical models


Chapman & Hall/CRC
Handbooks of Modern Statistical Methods
Handbook of Bayesian Variable Selection
Edited by
Mahlet G. Tadesse
Marina Vannucci
CRC Press
A CHAPMAN & HALL BOOK

Introduction

There are two main purposes to regression:

- Modelling relationships – we are interested in understanding the effect that a group of (explanatory) variables has on one (response) variable.
- Prediction – we are interested in predicting the value of a (response) variable for future values of the explanatory variables.

In both cases, parsimony is important.

Occam's razor: "entities must not be multiplied beyond necessity"
Betting on sparsity (Hastie et al., 2001)

Biological examples

Measurements can be made on various biological quantities, *e.g.*

- Gene expression
- Activity of proteins (Proteomics)

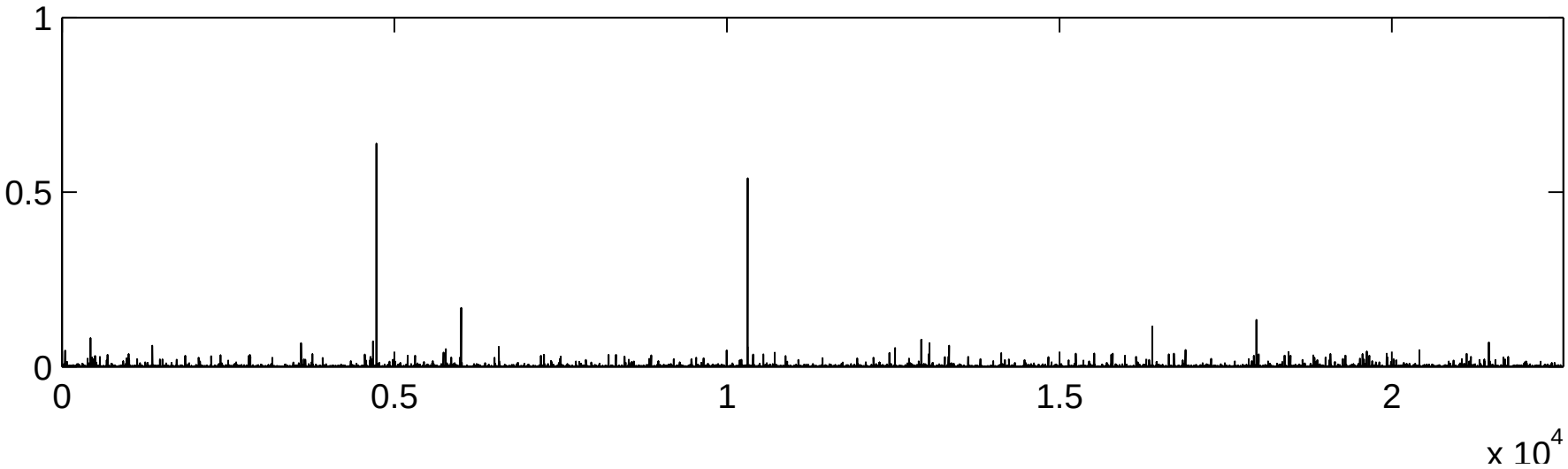
1000's (or more) measurements can be taken for a fairly low cost. This is usually linked to a phenotype (*e.g.* cholesterol levels or disease status).

The task is finding which genes or proteins explain the variation in the phenotype.

PCR Data Example

The data set includes 60 subjects from two inbred mouse populations and

- Response: phosphoenopruvate carboxykinase (PEPCK)
- Variables: gender and 22 575 gene expression measurements



Economic data

Prediction of economic variables (*e.g.* inflation) using past values of many other economic variables such as economic growth, unemployment, interest rates, ...

The potential number of variables can be very large since many past values can be considered and different methods of the same/related variables (*i.e.* different measures of inflation, different types of interest rates).

The task is to choose a subset of variables which predict inflation.

Two main approaches

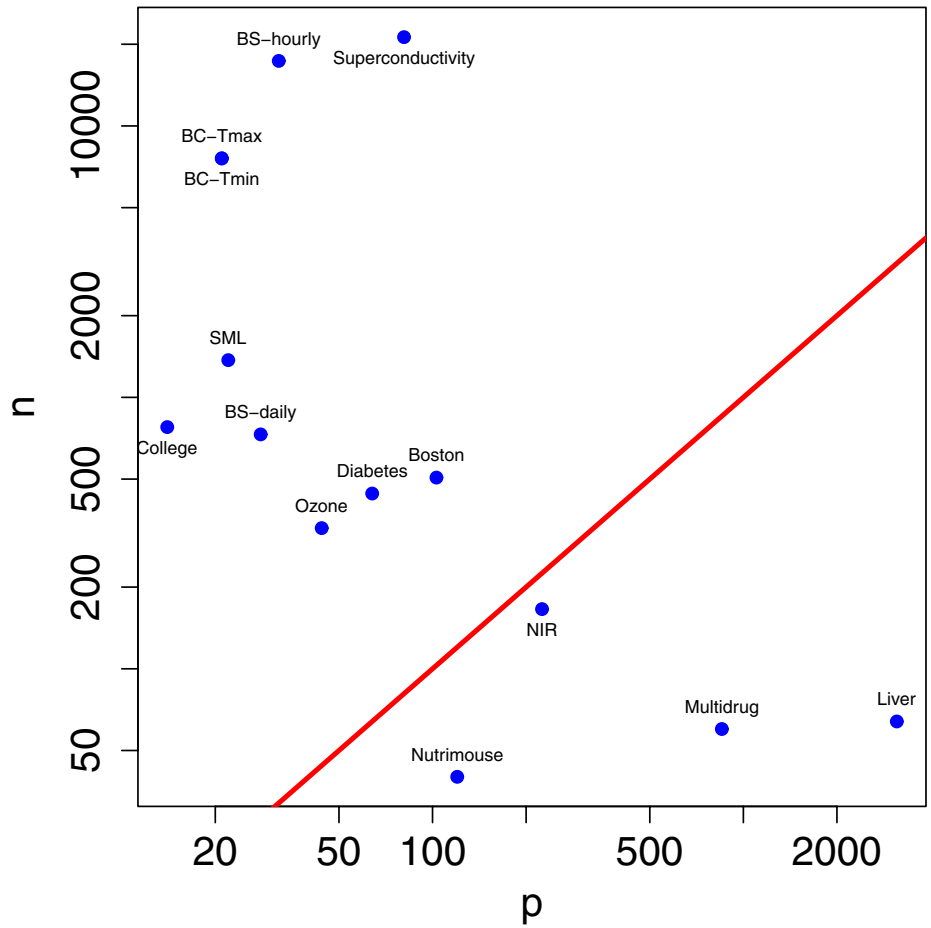
- Treat variable selection as a model selection problem with models defined by subsets of the available variables (similar to subset selection)
- Treat variable selection as a parameter estimation problem with priors that encourage "many small values" (similar to penalized maximum likelihood)

Performance of 21 variable selection methods (Porwal and Raftery, 2022) on 14 data sets

- `PointEst` is the RMSE for point estimation,
- `IntEst` is the MIS for interval estimation,
- `Inference` is the AUPRC,
- `Prediction` is the RMSE for point prediction,
- `IntPred` is the MIS for interval prediction.
- `N vars` is the average number of variables used for the task.

All metrics are standardized to equal 1 for the Jeffrey-Zellner-Siow g-prior, apart from computational time which is standardized to 1 for Lasso

The 14 data sets



	Rank	Score	PartScore	PointEst	IntEst	Inference	Prediction	IntPred	N vars	CPU time
g=sqrt(n)	1	0.974	0.982	0.978	0.927	0.999	0.968	0.996	1.294	0.949
Hyper-g	2	0.992	0.991	0.999	0.993	0.984	0.992	0.993	1.079	3.396
EB-local	3	0.993	0.996	0.995	0.978	0.995	0.998	1	1.099	0.843
JZS	4	1	1	1	1	1	1	1	1	8.835
Horseshoe	5	1.03	0.987	0.964	1.028	0.929	1.068	1.161	1.14	38.256
UIP	6	1.039	1.018	1.034	1.141	1.018	1.003	1	0.946	0.798
EB-global	7	1.073	1.029	1.024	1.238	1.035	1.026	1.042	0.876	0.909
Benchmark	8	1.15	1.111	1.072	1.394	1.189	1.072	1.021	0.719	0.742
NLP	9	1.157	1.037	1.124	1.598	0.91	1.076	1.076	2.07	254.676
LASSO	10		1.15	1.044		1.413	0.994		2.339	1
SCAD	11		1.175	1.122		1.362	1.04		1.505	7.299
BIC-BAS	12	1.21	1.214	1.144	1.206	1.088	1.41	1.201	1.227	0.936
BICREG	13	1.443	1.218	1.202	2.176	1.193	1.26	1.384	1.061	19.809
SpikeSlab	14	1.464	1.189	1.355	2.724	1.155	1.056	1.029	0.496	24.36
ElasticNet	15		1.203	1.098		1.522	0.99		3.825	60.408
MCP	16		1.221	1.148		1.417	1.099		1.227	6.725
SS Lasso	17		1.249	1.323		1.216	1.209		0.741	0.797
Lasso-1se	18		1.463	1.916		1.413	1.061		1.33	1
EMVS	19		1.501	1.703		1.508	1.291		1.026	4.634
AIC	20	3.613	3.837	6.179	4.937	1.176	4.155	1.617	2.887	1.675
g=1	21	3.859	2.256	4.016	11.153	1.194	1.557	1.373	1.66	1.194

Linear regression

We are interested in the multiple linear regression model

$$\boldsymbol{y} = \alpha \mathbf{1}_n + \boldsymbol{X}\beta + \boldsymbol{\epsilon}$$

where

- $\boldsymbol{y}$ is an $(n \times 1)$-dimensional vector of response variables
- $\boldsymbol{X}$ is an $(n \times p)$-dimensional matrix of regressors. We will assume that each column of $\boldsymbol{X}$ has been centred
- $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}_n, \sigma^2 \boldsymbol{I}_n)$ where $\boldsymbol{a}_n$ is an n-dimensional vectors of a's and $\boldsymbol{I}_n$ is a $(n \times n)$-dimensional identity matrix
- α is the intercept
- β is a $(p \times 1)$-dimensional vector of regression coefficients

Non-informative analysis

The standard non-informative prior used with a linear regression model is $p(\alpha, \boldsymbol{\beta}, \sigma^2) \propto \sigma^{-2}$.

The (conditional) posterior distribution is

$$\alpha, \boldsymbol{\beta} | \sigma^2, \boldsymbol{X}, \boldsymbol{y} \sim \mathrm{N}\left(\left(\boldsymbol{X}^{\star T} \boldsymbol{X}^{\star} \right)^{-1} \boldsymbol{X}^{\star T} \boldsymbol{y}, \sigma^2 \left(\boldsymbol{X}^{\star T} \boldsymbol{X}^{\star} \right)^{-1} \right)$$

where $\boldsymbol{X}^{\star} = (\mathbf{1}_n \;\; \boldsymbol{X})$ and

$$\sigma^{-2} | \boldsymbol{X}, \boldsymbol{y} \sim \mathrm{Ga}((n-p)/2, \mathrm{RSS}/2)$$

where $\mathrm{RSS} = \boldsymbol{y}^T \boldsymbol{y} - \boldsymbol{y}^T \boldsymbol{X}^{\star} \left(\boldsymbol{X}^{\star T} \boldsymbol{X}^{\star} \right)^{-1} \boldsymbol{X}^{\star T} \boldsymbol{y}$.

The posterior distribution is proper if $n > p$

Normal prior

The conjugate prior is $(\alpha, \beta) \sim \mathrm{N}(\boldsymbol{A}_0, \boldsymbol{B}_0\,\sigma^2)$ and $\sigma^{-2} \sim \mathrm{Ga}(c_0/2, d_0/2)$ which leads to the posterior distribution

$$\alpha, \beta | \sigma^2, \boldsymbol{X}, \boldsymbol{y} \sim \mathrm{N}\left(\boldsymbol{\mu}, \sigma^2 \left(\boldsymbol{X}^{\star T}\boldsymbol{X}^{\star} + \boldsymbol{B}_0^{-1}\right)^{-1}\right)$$

where $\boldsymbol{\mu} = \left(\boldsymbol{X}^{\star T}\boldsymbol{X}^{\star} + \boldsymbol{B}_0^{-1}\right)^{-1}\left(\boldsymbol{X}^{\star T}\boldsymbol{y} + \boldsymbol{B}_0^{-1}\boldsymbol{A}_0\right)$ and

$$\sigma^{-2} | \boldsymbol{X}, \boldsymbol{y} \sim \mathrm{Ga}((n + c_0)/2, d_1/2)$$

where $d_1 = d_0 + \boldsymbol{y}^T\boldsymbol{y} -$
$\left(\boldsymbol{X}^{\star T}\boldsymbol{y} + \boldsymbol{B}_0^{-1}\boldsymbol{A}_0\right)^T\left(\boldsymbol{X}^{\star T}\boldsymbol{X}^{\star} + \boldsymbol{B}_0^{-1}\right)^{-1}\left(\boldsymbol{X}^{\star T}\boldsymbol{y} + \boldsymbol{B}_0^{-1}\boldsymbol{A}_0\right)$

Comments

- A natural choice in the absence of substantive information is to assume that $\boldsymbol{A}_0$ is a vector of zeros. This reflects equal probability of positive and negative effects.
- The ridge regression estimator

$$\hat{\beta} = \left(\boldsymbol{X}^{\star T}\boldsymbol{X}^{\star} + \lambda \boldsymbol{I}_p\right)^{-1}\boldsymbol{X}^{\star T}\boldsymbol{y}$$

is the posterior mean if $\boldsymbol{B}_0^{-1} = \begin{pmatrix} 0 & \mathbf{0}_p^T \\ \mathbf{0}_p & \lambda \boldsymbol{I}_p \end{pmatrix}$

Bayesian variable selection (BVS) – Set-up

Define inclusion variables $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$
($\gamma_i = 1$ if the i-th variable is included and 0 otherwise)

Use the model

$$\boldsymbol{y} = \alpha \mathbf{1}_n + \boldsymbol{X}_\gamma \beta_\gamma + \boldsymbol{\epsilon}$$

where

- $\boldsymbol{X}_\gamma$ is an $(n \times p_\gamma)$-dimensional matrix formed by including the columns of $\boldsymbol{X}$ for which $\gamma_k = 1$
- β_γ is a $(p_\gamma \times 1)$-dimensional vector of regression coefficients

Bayesian variable selection

The prior distribution is extended to $\alpha, \boldsymbol{\beta}_\gamma, \sigma^2$ and $\boldsymbol{\gamma}$ where $\boldsymbol{\beta}_\gamma$ are the regression coefficients for model $\boldsymbol{\gamma}$. This is usually structured as

$$p(\alpha, \boldsymbol{\beta}_\gamma, \boldsymbol{\gamma}, \sigma^2) = p(\alpha, \boldsymbol{\beta}_\gamma | \sigma^2, \boldsymbol{\gamma})\, p(\sigma^2)\, p(\boldsymbol{\gamma})$$

This leads to a posterior distribution $p(\boldsymbol{\gamma} \mid \boldsymbol{X}, \boldsymbol{y})$, which expresses our uncertainty about $\boldsymbol{\gamma}$

Prior on the regression coefficients

In model selection problems, improper priors cannot be used for the regression coefficients but can be used for common parameters.

- The prior distribution for the intercept and observation variance is often taken to be

$$p(\alpha, \sigma^2) \propto \sigma^{-2}$$

- There are two standard priors for the regression coefficients
  - The independence prior $\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\sigma^2 \sim \mathrm{N}(\mathbf{0}_{p_{\boldsymbol{\gamma}}}, \frac{\sigma^2}{\lambda}\boldsymbol{I}_{p_{\boldsymbol{\gamma}}})$ (*i.e.* ridge)
  - The g-prior $\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\sigma^2 \sim \mathrm{N}(\mathbf{0}_{p_{\boldsymbol{\gamma}}}, \sigma^2\, g\, (\boldsymbol{X}^{\star T}\boldsymbol{X}^{\star})^{-1})$ (*i.e.* conditionally conjugate)

How can we specify the prior on γ?

In the absence of substantive information, people would often use one of the following priors

1. There are 2^p models therefore we can define $p(\gamma) = 2^{-p}$ to give a uniform weighting over models.
2. Assume that all models of size k are equally likely but that the probability of a model of size k is w_k then

$$p(\gamma) = \frac{w_{p_\gamma}}{\binom{p}{p_\gamma}}$$

   where p_γ is the number of variables included under γ.
3. Assume that $\gamma_1, \ldots, \gamma_p$ are independent and let $p(\gamma_k = 1) = \pi$

Properties of priors for the model

The prior expected model sizes are

- Prior 1: $\frac{2}{p}$
- Prior 2: $\sum_{j=1}^{\infty} j\, w_j$
- Prior 3: $p\,\pi$

In practice, people would often use Prior 3 due to its flexibility (π can be chosen to express a prior guess at the model size)

Additionally, π can be given a beta prior. Scott and Berger (2010) discuss how this can adjust for multiplicity

Posterior distribution of α, β_γ and σ^2 with independence prior

$$p(\alpha, \sigma^2) \propto \sigma^{-2}, \qquad \beta_\gamma \sim \mathsf{N}\left(\mathbf{0}_p, \frac{\sigma^2}{\lambda}\right)$$

The posterior distributions are

$$\alpha|\sigma^2, \boldsymbol{X}, \boldsymbol{y} \sim \mathsf{N}(\bar{y}, \sigma^2/n)$$

and

$$\beta_\gamma|\sigma^2, \boldsymbol{X}, \boldsymbol{y} \sim \mathsf{N}\left((\boldsymbol{X}^{\star T}\boldsymbol{X}^\star + \lambda \boldsymbol{I}_{p_\gamma})^{-1}\boldsymbol{X}^{\star T}\boldsymbol{y}, \sigma^2(\boldsymbol{X}^{\star T}\boldsymbol{X}^\star + \lambda \boldsymbol{I}_{p_\gamma})^{-1}\right)$$

and

$$\sigma^{-2} \sim \mathsf{Ga}\left(\frac{n}{2}, \frac{1}{2}\left[S_{yy} - \boldsymbol{y}^T\boldsymbol{X}^\star\left(\boldsymbol{X}^{\star T}\boldsymbol{X}^\star + \lambda \boldsymbol{I}_{p_\gamma}\right)^{-1}\boldsymbol{X}^{\star T}\boldsymbol{y}\right]\right)$$

Posterior distribution of α, β_γ and σ^2 with g-prior

$$p(\alpha, \sigma^2) \propto \sigma^{-2}, \qquad \alpha, \beta_\gamma | \sigma^2 \sim \mathrm{N}\left(\mathbf{0}_p, \sigma^2 g (\boldsymbol{X}^{\star T} \boldsymbol{X}^\star)^{-1}\right)$$

where $g > 0$ is a scalar hyperparameter

The posterior distributions are

$$\alpha | \sigma^2, \boldsymbol{X}, \boldsymbol{y} \sim \mathrm{N}(\bar{y}, \sigma^2/n)$$

$$\beta_\gamma | \sigma^2, \boldsymbol{X}, \boldsymbol{y} \sim \mathrm{N}\left(\frac{g}{1+g}\hat{\beta}_\gamma, \frac{\sigma^2 g}{1+g}\left(\boldsymbol{X}^{\star T}\boldsymbol{X}^\star\right)^{-1}\right)$$

where $\hat{\beta}_\gamma = \left(\boldsymbol{X}^{\star T}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\star T}\boldsymbol{y}$ and

$$\sigma^{-2} | \boldsymbol{X}, \boldsymbol{y} \sim \mathrm{Ga}\left(\frac{n}{2}, \frac{1}{1+g}S_{yy} + \frac{g}{1+g}\mathrm{RSS}\right)$$

The full posterior distribution

The full posterior distribution is

$$p(\alpha, \boldsymbol{\beta}_\gamma, \sigma^2, \gamma | \boldsymbol{X}, \boldsymbol{y}) = p(\alpha, \boldsymbol{\beta}_\gamma | \sigma^2, \gamma, \boldsymbol{X}, \boldsymbol{y})\, p(\sigma^2 | \gamma, \boldsymbol{X}, \boldsymbol{y})\, p(\gamma | \boldsymbol{X}, \boldsymbol{y})$$

The posterior distribution can be expressed as product of

- the posterior distribution of the regression coefficients conditional on the observational variance and model
- the posterior distribution of the observational variance conditional on the model
- the posterior distribution of the model

The marginal posterior distribution on models

The marginal posterior distribution of the models can be calculated as

$$p(\gamma|\boldsymbol{X}, \boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{X}, \gamma)\, p(\gamma)$$

where

- The marginal likelihood for model γ is $p(\boldsymbol{y}|\boldsymbol{X}, \gamma) = \int p(\boldsymbol{y}|\boldsymbol{X}, \alpha, \beta_\gamma, \sigma^2)\, p(\alpha, \beta_\gamma|\sigma^2)\, p(\sigma^2|\gamma)\, d\alpha\, d\beta_\gamma\, d\sigma^2$
- The prior probability of model γ is $p(\gamma)$

Marginal posterior distributions

- If an independence prior is used the marginal likelihood is

$$p(\boldsymbol{y}|\boldsymbol{X},\gamma) \propto \lambda^{p/2}|\boldsymbol{X}^{\star T}\boldsymbol{X}^{\star} + \lambda\,\boldsymbol{I}_p|^{-p/2} \times \left[S_{yy} - \boldsymbol{y}^T\boldsymbol{X}^{\star}(\boldsymbol{X}^{\star T}\boldsymbol{X}^{\star} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}^{\star T}\boldsymbol{y}\right]^{-(n-1)/2}$$

- If a g-prior is used the marginal likelihood is

$$p(\boldsymbol{y}|\boldsymbol{X},\gamma) \propto S_{\boldsymbol{yy}}^{-(n-1)/2} \frac{(1+g)^{(n-1-p_\gamma)/2}}{\left(1+g\left(1-R_\gamma^2\right)\right)^{(n-1)/2}}$$

where $R_\gamma^2 = 1 - \frac{\text{RSS}}{S_{\boldsymbol{yy}}}$ is the multiple correlation coefficient.

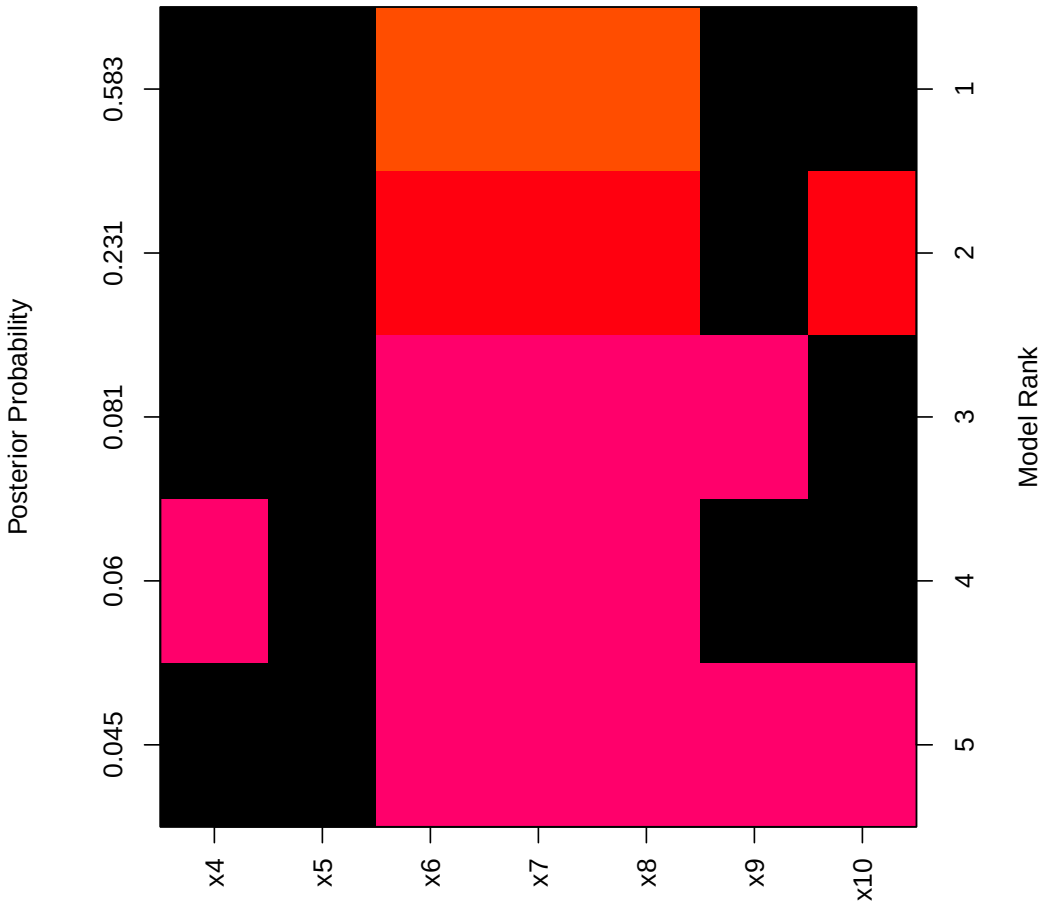
Example: Ozone data

- y – Daily maximum 1-hour-average ozone reading (ppm) at Upland, CA
- x4 – 500-millibar pressure height (m) at Vandenberg AFB
- x5 – Wind speed (mph) at LAX
- x6 – Humidity (percentage) at LAX
- x7 – Temperature (Fahrenheit degrees) at Sandburg, CA
- x8 – Inversion base height (feet) at LAX
- x9 – Pressure gradient (mm Hg) from LAX to Daggett, CA
- x10 – Visibility (miles) measured at LAX

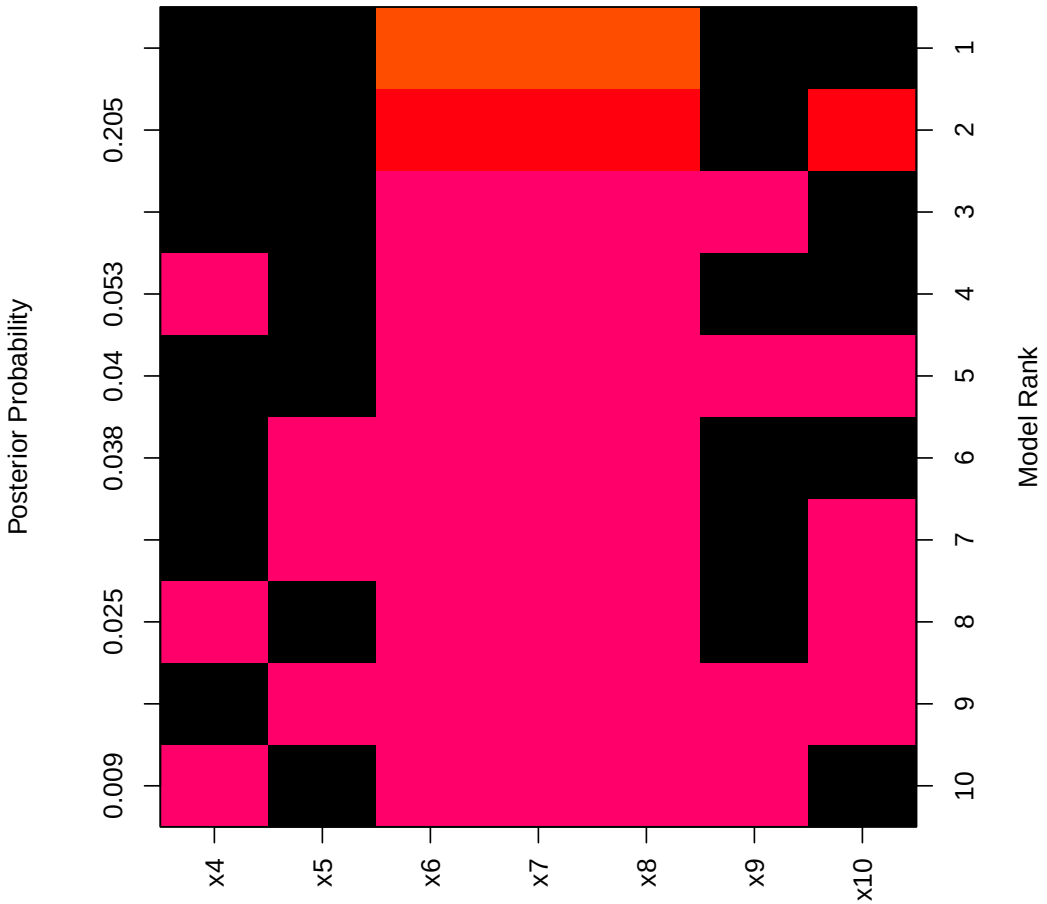
g-prior $g = n$: Top 5 models

	P(B != 0 | Y)	model 1	model 2	model 3	model 4	model 5
Intercept	1.000	1.000	1.000	1.000	1.000	1.000
x4	0.114	0.000	0.000	0.000	1.000	0.000
x5	0.105	0.000	0.000	0.000	0.000	0.000
x6	0.997	1.000	1.000	1.000	1.000	1.000
x7	0.999	1.000	1.000	1.000	1.000	1.000
x8	0.986	1.000	1.000	1.000	1.000	1.000
x9	0.154	0.000	0.000	1.000	0.000	1.000
x10	0.324	0.000	1.000	0.000	0.000	1.000
BF	NA	1.000	0.364	0.127	0.089	0.038
PostProbs	NA	0.489	0.194	0.068	0.050	0.038
R2	NA	0.686	0.692	0.688	0.687	0.693
dim	NA	4.000	5.000	5.000	5.000	6.000
logmarg	NA	93.772	92.763	91.706	91.349	90.496

g-prior $g = n$: Top 5 models



g-prior $g = n$: Top 10 models



Paradoxes

- Bartlett's Paradox As $g \to \infty$ the posterior distribution will concentrate on the null model.
  Therefore, we can't use a noninformative prior on β.
- Information Paradox As $R^2_\gamma \to 1$,

$$\frac{p(\boldsymbol{y}|\boldsymbol{X}, \gamma)}{p(\boldsymbol{y}|\boldsymbol{X}, \gamma_{null})} \to (1+g)^{(n-1-p_\gamma)/2}$$

Choosing λ

The parameter λ can be chosen

- Subjectively if we have information about the expected size of the regression effect
- A heavy-tailed prior such as a half-Cauchy distribution $p(\lambda) \propto (1+\lambda)^{-1}$
- An empirical Bayes estimate of λ can be used *i.e.* the λ that maximizes $p(\boldsymbol{y}|\boldsymbol{X},\lambda) = \sum_{\gamma \in \Gamma} p(\boldsymbol{y}|\boldsymbol{X},\gamma,\lambda)$

Choosing g

There is a large literature on choosing the parameter g.

Examples

- Unit Information prior: $g = n$ (Bayes factor is like BIC)
- Risk inflation criterion prior: $g = p^2$
- Benchmark prior: $g = \max\{n, p^2\}$
- Hannan-Quinn: $g = \log n$

Mixtures of g priors

An alternative to choosing g would be to give g a prior to a define a mixture of g priors

Examples

- Cauchy/Jeffrey-Zellner-Siow (JSZ) prior: $g \sim \text{IG}(1/2, n/2)$
- hyper-g: $p(g) \propto (1+g)^{-a/2}$
- hyper-g/n: $p(g) \propto (1+g/n)^{-a/2}$
- Robust prior: $p(g) \propto (1+g)^{-3/2}, \quad g > \frac{1+n}{p_\gamma+1} - 1$

The information paradox can be addressed if $\int_0^\infty (1+g)^{(n-1-p_\gamma)/2}\,\pi(g)\,dg = \infty$ for $p_\gamma \leq p$

Model selection consistency

- The JSZ prior is model selection consistent
- The hyper-g and robust prior is model selection consistent if the true model is not the null

Prediction consistency: the hyper-g, hyper-g/n, JSZ and robust prior are prediction consistent

Posterior prediction

Generally in Bayesian statistics, predictions are made using the posterior predictive distribution which can be calculated by integrating over all parameters.

In regression, for a new observation y_{n+1} with regressors x_{n+1}, this is

$$p(y_{n+1}|x_{n+1}) = \sum_{\gamma} \int p(y_{n+1}|x_{n+1}, \alpha, \beta_\gamma, \sigma^2, \gamma, \boldsymbol{X}, \boldsymbol{y}) \times p(\alpha, \beta_\gamma, \sigma^2, \gamma|\boldsymbol{X}, \boldsymbol{y})\, d\alpha\, d\beta_\gamma\, d\sigma^2$$

Posterior prediction and Bayesian model averaging

The posterior predictive can be expressed as

$$p(y_{n+1}|x_{n+1}) = \sum_{\gamma} p(y_{n+1}|x_{n+1}, \gamma, \boldsymbol{X}, \boldsymbol{y}) \times p(\gamma|\boldsymbol{X}, \boldsymbol{y})$$
$$= \sum_{\gamma} \text{model predictive} \times \text{model weight}$$

where

- model predictive is the posterior predictive distribution for model γ, $p(y_{n+1}|x_{n+1}, \gamma, \boldsymbol{X}, \boldsymbol{y}) = \int p(y_{n+1}|x_{n+1}, \alpha, \beta_\gamma, \sigma^2, \gamma) p(\alpha, \beta_\gamma, \sigma^2|\gamma, \boldsymbol{X}, \boldsymbol{y})\, d\alpha\, d\beta_\gamma\, d\sigma^2$
- model weight is the posterior probability of model γ, $p(\gamma|\boldsymbol{X}, \boldsymbol{y})$.

Posterior predictive expectation

The posterior predictive expectation can be expressed as

$$\begin{aligned} \mathrm{E}(y_{n+1}|x_{n+1}) &= \sum_{\gamma} \mathrm{E}(y_{n+1}|x_{n+1}, \gamma, \boldsymbol{X}, \boldsymbol{y}) \times p(\gamma|\boldsymbol{X}, \boldsymbol{y}) \\ &= \sum_{\gamma} \text{model prediction} \times \text{model weight} \end{aligned}$$

where

- model prediction is the posterior predictive expectation for model γ, $\mathrm{E}(y_{n+1}|x_{n+1}, \gamma, \boldsymbol{X}, \boldsymbol{y}) = \mathrm{E}(\alpha|x_{n+1}, \gamma, \boldsymbol{X}, \boldsymbol{y}) + \boldsymbol{X}_{\gamma}\mathrm{E}(\beta_{\gamma}|x_{n+1}, \gamma, \boldsymbol{X}, \boldsymbol{y})$
- model weight is the posterior probability of model γ, $p(\gamma|\boldsymbol{X}, \boldsymbol{y})$.

The idea of weighting predictions according to weights for different models is known as model averaging.

Using the posterior distribution of the models as weights leads to Bayesian model averaging which allows us to account for uncertainty about which model is correct.

Bayesian model averaged posterior distributions for parameters (or function of parameters) can also be constructed. For example,

$$p(\beta_i|\boldsymbol{X},\boldsymbol{y}) = \sum_{\gamma} p(\beta_i|\gamma,\boldsymbol{X},\boldsymbol{y}) \times p(\gamma|\boldsymbol{X},\boldsymbol{y})$$

which is the marginal posterior distribution of β_i.

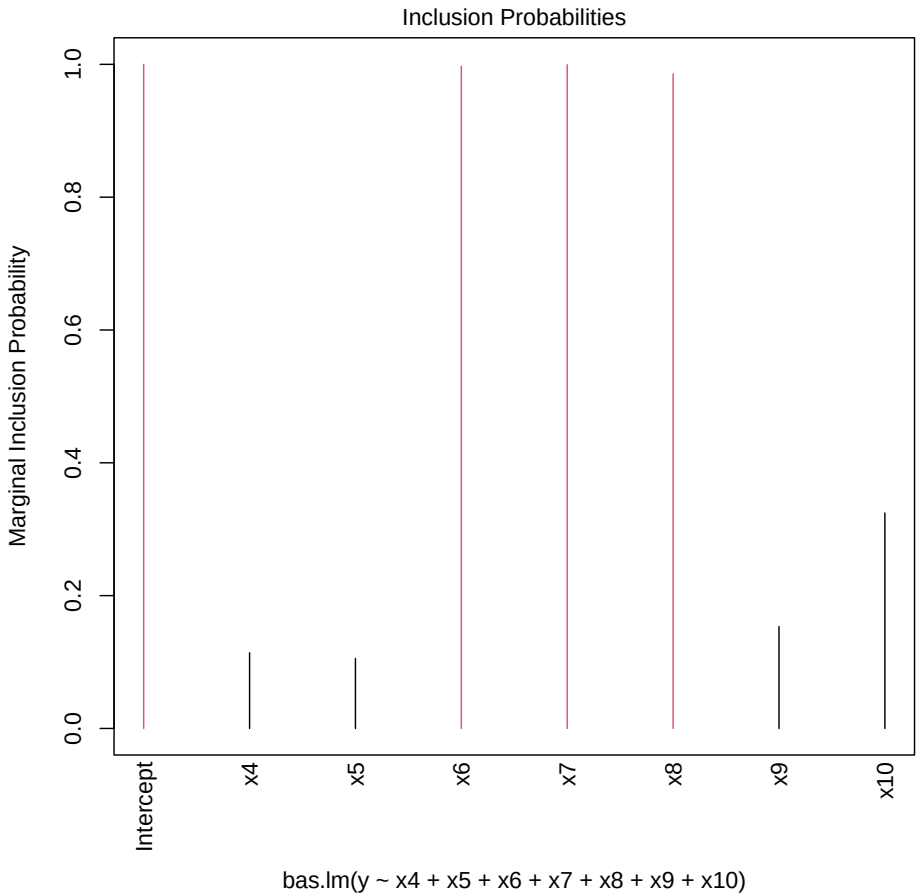
Summaries for Bayesian variable selection

To understand the relative importance of different variables, there are summaries

- Posterior inclusion probabilities (PIPs): $p(\gamma_i \mid \text{Data})$
- Maximum a posterior (MAP) model: the mode of $\gamma \mid \text{Data}$
- Median model $\hat{\gamma}$ where $\hat{\gamma}_i = \mathsf{I}(p(\gamma_i \mid \boldsymbol{X}, \boldsymbol{y}) > 0.5)$ (Barbieri and Berger, 2004)

These are summaries of importance but don't represent any relationships between variables included in models

Ozone data: g-prior $g = n$: Top 10 models



Barbieri, M. M. and Berger, J. O. (2004). Optimal predictive model selection, *The Annals of Statistics* **32**: 870–897.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning*, Springer Series in Statistics, Springer New York Inc., New York, NY, USA.

Porwal, A. and Raftery, A. E. (2022). Comparing methods for statistical inference with model uncertainty, *PNAS* **119**: e2120737119.

Scott, J. G. and Berger, J. O. (2010). Bayes and empirical Bayes multiplicity adjustment in the variable selection problem, *The Annals of Statistics* **38**: 2587–2619.