On Prior Distributions for Bayesian Model Selection

Ioannis Ntzoufras ,
Department of Business Administration, University of the Aegean, 8 Michalon Street, Island of Chios , Greece, e-mail: ntzoufras@aegean.gr.
Petros Dellaportas
Department of Statistics, Athens University of Economics and Business, 76 Patission Street, 10434, Athens, Greece, e-mail:petros@aueb.gr.
Jonathan J. Forster
Faculty of Mathematics, University of Southampton, UK, e-mail:jjf@maths.soton.ac.uk

Contents

1. Lindley's Paradox
2. Information Criteria
3. Bayes Factors for Normal Linear Models (Posterior Penalty Specification via Adjustment of Prior Odds)
4. Discussion

1 Jeffreys-Lindley Paradox

Consider two models m_0 and m_1;

- $d(m)$ dimension of model m,
- $d(m_0) < d(m_1)$; model m_0 is simpler.

1. If sample size $n \to \infty$: $B_{10} \to 0$
   Bayes factor supports simpler models in contradiction to significance tests (Lindley, 1957, Bk).
2. If prior variance of additional parameters $\to \infty$: $B_{10} \to 0$
   (Bartlett, 1957, Bk).

(1) and/or (2) are referred in literature as

- 'Lindley's paradox' → for any case where Bayesian and significance tests result in contradictive evidence (Shafer, 1982, JASA).
- 'Bartlett' paradox → Kass and Raftery (1995, JASA)
- 'Jeffreys' paradox → Lindley (1980, *An.Stat.*), Berger and Delampady (1987, *St.Science*)
- 'Jeffreys-Lindley's paradox' → Robert (1993, *St.Sinica*).
- 'Bartlett - Lindley' paradox → Chipman *et al.* (2000, Tec.Rep.).
- For detailed discussion → Shafer (1982, JASA).

We focus on Variable Selection Problems for GLM.

Let us consider a GLM with $n \times 1$ vector of linear predictors given by

$$\boldsymbol{\eta} = \boldsymbol{X}_{(m)}\boldsymbol{\beta}_{(m)}$$

- $\boldsymbol{X}_{(m)}$ = design matrix of model m
- $\boldsymbol{\beta}_{(m)}$ = vector of parameters involved in the linear predictors.

Prior Distributions for the parameters of the linear predictor:

$$f\left(\boldsymbol{\beta}_{(m)}|m\right) \sim N\left(\boldsymbol{\mu}_{\beta_{(m)}}, \boldsymbol{\Sigma}_{(m)}\right)$$

Low Information Prior Distributions proposed in literature:

- $\boldsymbol{\mu}_{\beta_m} = \mathbf{0}$: prior centered against alternative hypothesis.
- $\boldsymbol{\Sigma}_{(m)} = c^2\boldsymbol{V}_{(m)}$ or $\boldsymbol{\Sigma}_{(m)} = c^2\boldsymbol{V}_{(m)}\sigma^2$ in regression.

The choice of $\boldsymbol{\Sigma}_{(m)}$ remains difficult. Two types of prior distributions

- Block Diagonal Covariance Matrix (independent priors)
- Non-diagonal Covariance Matrix

Normal Independent priors, $\boldsymbol{V}_{(m)} = Diagonal(v_i^2)$:

- George and McCullogh (1993, JASA) in SSVS
- Geweke (1996, *B.Stat.*): Independent truncated normal distributions in regression.

Non-diagonal Covariance Matrix

- REGRESSION: $\boldsymbol{\Sigma}_{(m)} = c^2\boldsymbol{V}_{(m)}\sigma^2$
  - $\boldsymbol{V}_{(m)}^{-1} = \boldsymbol{X}_{(m)}^T\boldsymbol{X}_{(m)} \rightarrow$ Zellner's g-priors (Zellner, 1980).
  - $c^2 \in [10, 100]$ proposed by Smith and Kohn (1996, *J.Econ.*).
  - $c^2 = n \rightarrow$ Unit Information priors (Kass and Wasserman, 1995, JASA).
  - Fernandez *et al.* (2001, *J.Econ.*) used various values for c^2; proposed $c^2 = \max\{d(m)^2, n\}$.

- Contingency tables: $\boldsymbol{\Sigma}_{(m)} = c^2\boldsymbol{V}_{(m)}$
  - Albert (1996, *Can.J.St.*): based on prior beliefs on odds ratios.
  - Dellaportas and Forster (1999,Bk) based on Knuinman and Speed (1988, *Bc*); $\boldsymbol{V}_{(m)}^{-1} = \boldsymbol{X}_{(m)}^T\boldsymbol{X}_{(m)}$, $c^2 = 2 \times \#cells$.
  - Ntzoufras *et al.* (2000, JSCS): combination of the above for SSVS.
- GLM → Raftery (1996, Bk):
  - diagonal covariance matrix and mean zero for covariates based on sample variances.
  - Nonzero mean and correlation of intercept with the rest of parameters.
  - $c^2 = 2.85^2$ based on mathematical arguments.
- Ntzoufras *et al.* (2001): Constructed 'equivalent' priors across GLM with different link function based on Taylor expansion.

- Unit Information Prior $\boldsymbol{\Sigma}_{(m)} = n(-\boldsymbol{H}_{(m)})^{-1}$ (Kass and Wasserman, 1995, JASA); $\boldsymbol{H}_{(m)}$ is the Hessian matrix.
- Kuo and Mallick (1998, Sankya): Define prior only on full model.
- Using Imaginary data to construct an informative prior: Chen *et al.* (1999, JRSSB).
- George and Foster (2000, Bk): Empirical Bayes Approach.
- Expected Posterior Prior Distributions (Perez and Berger, 2000)

Prior Distributions on Model Space

- Usual naive prior: Uniform prior on model space $\mathcal{M}$ $p(m) = 1/|\mathcal{M}|$. Informative in terms of dimension (Chipman *et al.* , 2000, Tec.Rep.).
- Alternative: Use prior on dimension (Chipman *et al.* , 2000, Tec.Rep.).
- Use Beta prior on common inclusion probability (George and McCullogh, 1997, *St.Sin.*, Kohn *et al.* , 2001 *St.Comp.*).
- Elicit imaginary data: Chen *et al.* (1999, JRSSB)
- Use Empirical Bayes Approach (George and Foster, 2000, Bk).
- Prior distribution based on Dilution of models (George, 1999, *B.Stat*).
- Our Approach: Define prior model odds using prior penalty for additional model complexity.

2 Information Criteria

Most Criteria minimize a quantity given by

$$IC_m = -2log\left(f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_m, m)\right) + d(m)F \qquad (1)$$

- $\boldsymbol{\theta}_m$: parameter vector
- $\hat{\boldsymbol{\theta}}_m$: Maximum likelihood estimates
- $d(m)$: dimension of model m.
- F: penalty for each additional parameter used in the model.

In linear regression models:

- $\boldsymbol{\theta}_m^T = [\boldsymbol{\beta}_{(m)}^T, \sigma^2]$.
- $-2log\left(f(\boldsymbol{y}|\hat{\theta}_m, m)\right) = nlog(RSS_m)$;
  RSS_m is the residual sum of squares of model m.
  $RSS_m = \boldsymbol{y}^T\boldsymbol{y} - \hat{\boldsymbol{\beta}}_{(m)}^T\boldsymbol{X}_{(m)}^T\boldsymbol{X}_{(m)}\hat{\boldsymbol{\beta}}_{(m)}$
- NOTE: if predictors are orthogonal then for $t_j^2 > F$ we remove X_j variable (George, 2000).

Special cases of (**??**) are given for different F

- AIC: $F = 2$ (Akaike, 1973, Kiado).
- BIC: $F = \log(n)$ (Schwarz,1978, *An.St.*).
- AIC$_\alpha$: $F = \alpha \in [2, 5]$ (Bhansali and Downham, 1977, Bk)
- Φ_c: $F = c\log(log(n))$ (Hannan and Quinn, 1979, JRSSB).
- SSC: $F = 3/2$ (Smith and Spiegelhalter, 1980, JRSSB).
- SC: $F = n\log[n + 2d(m)]/d(m)$ (Shibata, 1980, *An.St.*).
- RIC $F = 2p\log[d(m)]/d(m)$ (Foster and George, 1994, *An.St.*).

And there are more out there.

For pairwise comparisons of models m_0 and m_1 with $d(m_0) < d(m_1)$:

$$IC_{01} = IC_{m_0} - IC_{m_1} = -2log\left(\frac{f(\boldsymbol{y}|\hat{\theta}_0, m_0)}{f(\boldsymbol{y}|\hat{\theta}_1, m_1)}\right) - \underbrace{\{d(m_1) - d(m_0)\}F}_{\psi} \tag{2}$$

Substitute $\{d(m_1) - d(m_0)\}F$ with a more complicated function $\psi = \delta(d(m_1), d(m_0), n)$.

- $IC_{01} < 0$ supports the simpler model m_0
- $IC_{01} > 0$ supports model m_1.

Same support pattern for $-2log(PO_{01})$

- $-2log(PO_{01}) < 0$ supports the simpler model m_0
- $-2log(PO_{01}) > 0$ supports model m_1.

where PO_{01} are the posterior odds of model m_0 vs. model m_1.

3 Normal Linear Models

Consider m_0 and m_1 given by

$$\boldsymbol{y} \sim N\left(\boldsymbol{X}_{(m)}\boldsymbol{\beta}_{(m)}, \sigma^2\boldsymbol{I}_n\right),$$

- $m \in \{m_0, m_1\}$,
- n: the sample size,
- $\boldsymbol{\beta}_{(m)}$: $d(m) \times 1$ vector of model parameters,
- $\boldsymbol{X}_{(m)}$: $n \times d(m)$ design (or data) matrix of model m,
- $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$: multivariate normal distribution (mean $\boldsymbol{\mu}$, covariance matrix $\boldsymbol{\Sigma}$)
- $\boldsymbol{I}_n$: $n \times n$ identity matrix.

Conjugate prior distribution

$$f(\boldsymbol{\beta}_{(m)}|\sigma^2, m) \sim N\left(\boldsymbol{\mu}_{\beta_{(m)}}, \boldsymbol{\Sigma}_{(m)}\sigma^2\right)$$

and the improper prior for the residual variance $f(\sigma^2) \propto \sigma^{-2}$.

Posterior odds:

$$PO_{01} = \left(\frac{|\tilde{\boldsymbol{\Sigma}}_{(m_1)}|}{|\boldsymbol{\Sigma}_{(m_1)}|}\right)^{-1/2}\left(\frac{|\tilde{\boldsymbol{\Sigma}}_{(m_0)}|}{|\boldsymbol{\Sigma}_{(m_0)}|}\right)^{1/2} \times \times\left(\frac{SS_{m_0}}{SS_{m_1}}\right)^{-n/2}\frac{f(m_0)}{f(m_1)} \tag{3}$$

SS_m = posterior sum of squares

$$SS_m = \boldsymbol{y}^T\boldsymbol{y} - \tilde{\boldsymbol{\beta}}_{(m)}^T\tilde{\boldsymbol{\Sigma}}_{(m)}^{-1}\tilde{\boldsymbol{\beta}}_{(m)} + \boldsymbol{\mu}_{\beta_{(m)}}^T\boldsymbol{\Sigma}_{(m)}^{-1}\boldsymbol{\mu}_{\beta_{(m)}},$$

$$\tilde{\boldsymbol{\beta}}_{(m)} = \tilde{\boldsymbol{\Sigma}}_{(m)}\left(\boldsymbol{X}_{(m)}^T\boldsymbol{X}_{(m)}\hat{\boldsymbol{\beta}}_{(m)} + \boldsymbol{\Sigma}_{(m)}^{-1}\boldsymbol{\mu}_{\beta_{(m)}}\right),$$

$$\tilde{\boldsymbol{\Sigma}}_{(m)}^{-1} = \boldsymbol{X}_{(m)}^T\boldsymbol{X}_{(m)} + \boldsymbol{\Sigma}_{(m)}^{-1},$$

$\tilde{\boldsymbol{\beta}}_{(m)}$ = posterior mean of $\boldsymbol{\beta}_{(m)}$
$\tilde{\boldsymbol{\Sigma}}_{(m)}$= posterior cov.matrix of $\boldsymbol{\beta}_{(m)}$.

Alternative expression of SS_m (Atkinson, 1978, Bk):

$$SS_m = RSS_m + \left(\hat{\boldsymbol{\beta}}_{(m)} - \boldsymbol{\mu}_{\beta_{(m)}}\right)^T\left[\left(\boldsymbol{X}_{(m)}^T\boldsymbol{X}_{(m)}\right)^{-1} + c^2\boldsymbol{V}_{(m)}\right]^{-1}\left(\hat{\boldsymbol{\beta}}_{(m)} - \boldsymbol{\mu}_{\beta_{(m)}}\right)$$

RSS_m residual sum of squares.
SS_m = [Data Goodness of fit] + [Distance of MLE - Prior Mean].

Consider two normal linear models m_0 and m_1 then

$$-2log(PO_{01}) = n\log\left(\frac{SS_{m_0}}{SS_{m_1}}\right) - \psi$$

- SS_{m_0}/SS_{m_1}: Goodness of fit Measure
- ψ: (Posterior) Penalty parameter which controls the weight of the parsimony principle.

$$\begin{aligned}\psi &= -\log\left(\frac{|\boldsymbol{\Sigma}_{(m_1)}|^{-1}}{|\boldsymbol{\Sigma}_{(m_0)}|^{-1}}\right) + \log\left(\frac{|\tilde{\boldsymbol{\Sigma}}_{(m_1)}|^{-1}}{|\tilde{\boldsymbol{\Sigma}}_{(m_0)}|^{-1}}\right) - 2\log\left(\frac{f(m_1)}{f(m_0)}\right)\\ &= +\log\left(\frac{|\tilde{\boldsymbol{\Sigma}}_{(m_1)}|^{-1}}{|\boldsymbol{\Sigma}_{(m_1)}|^{-1}}\right) - \log\left(\frac{|\tilde{\boldsymbol{\Sigma}}_{(m_0)}|^{-1}}{|\boldsymbol{\Sigma}_{(m_0)}|^{-1}}\right) - 2\log\left(\frac{f(m_1)}{f(m_0)}\right).\end{aligned}$$

For large n or $\boldsymbol{\Sigma}_{(m)} = c^2\boldsymbol{V}_{(m)}$, c^2 large then $SS_m \approx RSS_m$.
THE KEY for penalty specification: PRIOR MODEL ODDS.

Alternative Interpretation

- Define as ψ_0 the prior penalty defined by:

$$-2\log\left(\frac{f(m_0)}{f(m_1)}\right) = -\psi_0$$

  $\psi_0 > 0$ prior penalty imposed for using models with more parameters.
- $|\tilde{\boldsymbol{\Sigma}}_{(m)}|^{-1}$ volume of posterior information (from data and prior)
- $|\boldsymbol{\Sigma}_{(m)}|^{-1}$ volume of prior information
- $|\tilde{\boldsymbol{\Sigma}}_{(m)}|^{-1}/|\boldsymbol{\Sigma}_{(m)}|^{-1}$ is measure of new knowledge we gain from the data.

- 1st Interpretation:
  Posterior Penalty =
  + Prior Penalty
  + Difference of Posterior information between models m_1 and m_0 [A]
  − Difference of Prior information between models m_1 and m_0 [B]

$$\psi = \underbrace{-2\log\left(\frac{f(m_1)}{f(m_0)}\right)}_{Prior Penalty} + \underbrace{\log\left(\frac{|\tilde{\boldsymbol{\Sigma}}_{(m_1)}|^{-1}}{|\tilde{\boldsymbol{\Sigma}}_{(m_0)}|^{-1}}\right)}_{A} - \underbrace{\log\left(\frac{|\boldsymbol{\Sigma}_{(m_1)}|^{-1}}{|\boldsymbol{\Sigma}_{(m_0)}|^{-1}}\right)}_{B}$$

- 2nd Interpretation:
  Posterior Penalty =
  + Prior Penalty
  + Difference between Posterior and Prior information for model m_1 [C]
  − Difference between Posterior and Prior information for model m_0 [D] .

$$\psi = \underbrace{-2\log\left(\frac{f(m_1)}{f(m_0)}\right)}_{Prior Penalty} + \underbrace{\log\left(\frac{|\tilde{\boldsymbol{\Sigma}}_{(m_1)}|^{-1}}{|\boldsymbol{\Sigma}_{(m_1)}|^{-1}}\right)}_{C} - \underbrace{\log\left(\frac{|\tilde{\boldsymbol{\Sigma}}_{(m_0)}|^{-1}}{|\boldsymbol{\Sigma}_{(m_0)}|^{-1}}\right)}_{D}.$$

Let us consider m_0 and m_1 differ only in $\boldsymbol{X}_j$ term. And a low information prior distribution.

Extreme Cases:

$\boldsymbol{X}_j$ Orthogonal to $\boldsymbol{X}_{(m_0)}$ →

$$\psi \approx \psi_0 - \log\left(\frac{|\boldsymbol{\Sigma}_{(m_1)}|^{-1}}{|\boldsymbol{\Sigma}_{(m_0)}|^{-1}}\right) + \log|\boldsymbol{X}_j^T\boldsymbol{X}_j|$$

$\boldsymbol{X}_j$ collinear to $\boldsymbol{X}_{(m_0)} \rightarrow \psi \approx \psi_0 - \log\left(\frac{|\boldsymbol{\Sigma}_{(m_1)}|^{-1}}{|\boldsymbol{\Sigma}_{(m_0)}|^{-1}}\right)$

POSSIBLE SOLUTION 1: Zellner's g-Priors

- Prior model odds =1 → Prior Penalty $\psi_0 = 0$
- $\boldsymbol{\Sigma}_{(m)} = c^2\left(\boldsymbol{X}_{(m)}^T\boldsymbol{X}_{(m)}\right)^{-1}$; see Zellner (1980, *Bayesian Statistics*), Smith and Kohn (1996, *J.Econ.*) and Fernandez *et al.* (2001, *J.Econ.*).
- Result: Posterior Penalty $\psi = \{d(m_1) - d(m_0)\}log(c^2)$.
- Special Case: For $c^2 = n$: Unit information priors (Kass and Wasserman, 1995, JASA).
- Problem: Informative priors on $\boldsymbol{\beta}_{(m)}$ influence: posterior distribution of $\boldsymbol{\beta}_{(m)}$, goodness of fit and the penalty imposed. Also this prior involves data in other GLM.

POSSIBLE SOLUTION 2

- Use Prior Penalty $\psi_0 \neq 0$. Possible 'good' choice

$$\psi_0 = \{d(m_1) - d(m_0)\}F + \log\left(\frac{|\boldsymbol{\Sigma}_{(m_1)}|^{-1}}{|\boldsymbol{\Sigma}_{(m_0)}|^{-1}}\right) - \log\left(\frac{|\tilde{\boldsymbol{\Sigma}}_{(m_1)}|^{-1}}{|\tilde{\boldsymbol{\Sigma}}_{(m_0)}|^{-1}}\right)$$

  prior model odds →

$$\frac{f(m_1)}{f(m_0)} = \left\{e^{-F/2}\right\}^{d(m_1)-d(m_0)}\left(\frac{|\boldsymbol{\Sigma}_{(m_1)}|^{-1}}{|\boldsymbol{\Sigma}_{(m_0)}|^{-1}}\right)^{1/2}\left(\frac{|\tilde{\boldsymbol{\Sigma}}_{(m_1)}|^{-1}}{|\tilde{\boldsymbol{\Sigma}}_{(m_0)}|^{-1}}\right)^{-1/2}$$

- Posterior Penalty $\psi = \{d(m_1) - d(m_0)\}F$.
- Now we may use informative or non-informative priors on $\boldsymbol{\beta}_{(m)}$ which will influence GOF but not the (posterior) penalty ψ.
- Problem: How do we interpret the above prior penalty?

- F = Posterior Penalty for each additional parameter.
- $\log\left(\frac{|\boldsymbol{\Sigma}_{(m_1)}|^{-1}}{|\boldsymbol{\Sigma}_{(m_0)}|^{-1}}\right)$: Additional prior penalty imposed for additional information added via prior distribution of model parameters.
- $\log\left(\frac{|\tilde{\boldsymbol{\Sigma}}_{(m_1)}|^{-1}}{|\boldsymbol{\Sigma}_{(m_0)}|^{-1}}\right)$: Reduced prior penalty imposed for information added via the use of data and prior.
  If the additional term is collinear $\rightarrow 0$
  If the additional term is orthogonal$\rightarrow -\log(\boldsymbol{X}_j^T\boldsymbol{X}_j)$

4 Discussion

- The specification of Prior distributions is Important for Bayesian Model Selection
- Why not express our beliefs for models via prior penalties?
- Divide model selection procedure in:
  (a) Estimation (prior of $\boldsymbol{\beta}_{(m)}$)
  (b) Model selection (penalize to support parsimony principle).