Lecture 5: Conventional Model Selection Priors

Susie Bayarri

University of Valencia

CBMS Conference on Model Uncertainty and Multiplicity
July 23-28, 2012

Outline

- The general linear model and Orthogonalization
- Historical Conventional priors: g-priors and Zellner-Siow priors
- Desiderata for choice of model priors
- Variable selection in linear models
- A proposed new prior
- Extensions of conventional priors

I. The General Linear Model and Orthogonality

The General Linear Model

Notation: with $Y_i = X_{i1}\beta_1 + \ldots + X_{ik}\beta_k + \epsilon_i$ we have

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$\boldsymbol{Y}^t = (Y_1, \ldots, Y_n) \rightsquigarrow$ dependent variables

$\boldsymbol{\beta}^t = (\beta_1, \ldots, \beta_k) \rightsquigarrow$ regression coefficients

$\boldsymbol{X}_{[n\times k]}$ of rank $k \rightsquigarrow$ independent var. (design matrix)

$\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$

Hence $\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$

Often $X_{i1} = 1 \rightsquigarrow \beta_1$ called 'intercept' and denoted by α

Likelihood: for observed $\boldsymbol{y}, \boldsymbol{X}$

$$f(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2) = \frac{1}{(\sigma^2 2\pi)^{n/2}} \exp\{-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^t(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\}$$

$$= \frac{1}{(\sigma^2 2\pi)^{n/2}} \exp\{-\frac{1}{2\sigma^2}\left[\nu s^2 + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^t \boldsymbol{X}^t \boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right]\}$$

$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y} \rightsquigarrow$ O.L.S. estimator and MLE

$\nu = n - k \rightsquigarrow$ degrees of freedom

$s^2 = \frac{1}{\nu}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^t(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \rightsquigarrow$ residual s.o.s

Properties

$\hat{\boldsymbol{\beta}}$ sufficient for $\boldsymbol{\beta}$ given σ^2, $(\hat{\boldsymbol{\beta}}, s^2)$ sufficient for $(\boldsymbol{\beta}, \sigma^2)$

$\hat{\boldsymbol{\beta}} \sim N_k(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^t\boldsymbol{X})^{-1})$, $\nu s^2 \perp \hat{\boldsymbol{\beta}}$ and $\frac{\nu s^2}{\sigma^2} \sim \chi^2_{(\nu)}$

Orthogonal parameters

Important concept in ‘conventional priors’ derivations:

- $\boldsymbol{Y} \mid \boldsymbol{\theta} \sim f(\boldsymbol{y} \mid \boldsymbol{\theta})$
- $\boldsymbol{J}(\boldsymbol{\theta}) \rightsquigarrow$ (expected) Fisher Information matrix :

$$\boldsymbol{J}(\boldsymbol{\theta}) = E\Big(- \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log f(\boldsymbol{y} \mid \boldsymbol{\theta})\Big).$$

- If $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$, partition $\boldsymbol{J}(\boldsymbol{\theta})$ as

$$\boldsymbol{J}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{J}^{\alpha}(\boldsymbol{\theta}) & \boldsymbol{J}^{\alpha,\beta}(\boldsymbol{\theta}) \\ \boldsymbol{J}^{\alpha,\beta}(\boldsymbol{\theta}) & \boldsymbol{J}^{\beta}(\boldsymbol{\theta}) \end{pmatrix}$$

- *Definition (Jeffreys):*

  Parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are (globally) orthogonal if $\boldsymbol{J}^{\alpha,\beta}(\boldsymbol{\theta}) = \mathbf{0}$, $\forall \boldsymbol{\theta}$.

‘similar’, ‘common’ parameters

- Model-specific parameters do not have, in general, the same meaning and should not in general be identified the same.

- Ignoring this principle produces erroneous prior assessments:

$$\begin{aligned} M_1 &: \quad f_1(\boldsymbol{y} \mid \boldsymbol{\alpha}) = f(\boldsymbol{y} \mid \boldsymbol{\alpha}, \mathbf{0}) \\ M_2 &: \quad f_2(\boldsymbol{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = f(\boldsymbol{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}). \end{aligned}$$

- It is tempting (and frequent) to assess a unique prior under M_2, $\pi_2(\boldsymbol{\alpha}, \boldsymbol{\beta})$ (probably assuming independence) and then *deduce* $\pi_1(\boldsymbol{\alpha})$, either by
  - conditioning (may not be invariant under transformations)
  - marginalizing (McCulloch y Rossi 92, Verdinelli and Wasserman 93)

- This is usually erroneous, since $\boldsymbol{\alpha}$ has different meanings in M_1 and M_2

Equally identified, common parameters

- For truly subjective assessments, if priors under both models are similar, then similar ‘meaning’ and same ’names’ seems fine.
- Common ’Objective assessments’ must rely on different criteria.
- A rigorous criteria is when ‘common’ parameters have an invariant structure (Berger, Pericchi and Vasarhski), and then use of right Haar prior is justified.
- A less justifiable, but frequent practice is the following

  ‘Conventional’ assumption: “If under $M_2 : f_2(\boldsymbol{y} \mid \boldsymbol{\alpha}_2, \boldsymbol{\beta})$, $\boldsymbol{\alpha}_2$ and $\boldsymbol{\beta}$ are orthogonal, then $\boldsymbol{\alpha}_1$ in M_1 and $\boldsymbol{\alpha}_2$ in M_2 can be identified the same ($\boldsymbol{\alpha}$, say) and then

$$\pi_1(\boldsymbol{\alpha}), \qquad \pi_2(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \pi_1(\boldsymbol{\alpha})\pi(\boldsymbol{\beta} \mid \boldsymbol{\alpha}),$$

  is a suitable structure for prior assessment”

But there does not seem a serious justification for this practice: Berger and Pericchi (1996) write:

“That use of orthogonal parameters overcomes this difficulty is a belief in the statistical folklore and is undoubtedly true in certain asymptotic senses, but we have not seen a clear Bayesian argument as to why this should be so. The other problems with orthogonalization are (i) it is frequently extremely difficult or impossible to find orthogonal parameters, and (ii) orthogonal parameters typically have no intuitive meaning, and so models expressed in terms of subsets of orthogonal parameters often have no meaning. Nevertheless, the use of orthogonal parameters, when possible, appears to be a quite effective tool. Jeffreys (1961) provides a number of convincing examples. For a modern successful use of the idea, see Clyde and Parmigiani (1995)”.

see also Clyde and George (2004?), and others ...

- However, these difficulties are much milder (and often absent) in the linear models scenario
- A different difficulty is that orthogonal parameters do not imply same conventional priors under both models. It is true under certain assumptions (met in the Normal scenario) for Jeffreys priors:

Result 1. *Let $\pi_1^N(\boldsymbol{\alpha}_1)$ and $\pi_2^N(\boldsymbol{\alpha}_2, \boldsymbol{\beta})$ be Jeffreys priors, with α_2 and $\boldsymbol{\beta}$ orthogonal. If $\boldsymbol{J}^{\alpha_2}(\boldsymbol{\alpha}_2, \boldsymbol{\beta}) = G(\boldsymbol{\alpha}_2)$, then*

$$\pi_2^N(\boldsymbol{\alpha}_2, \boldsymbol{\beta}) = \pi_1^N(\boldsymbol{\alpha}_2)|\boldsymbol{J}^{\beta}(\boldsymbol{\alpha}_2, \boldsymbol{\beta})|^{1/2}.$$

- The condition $\boldsymbol{J}^{\alpha_2}(\boldsymbol{\alpha}_2, \boldsymbol{\beta}) = G(\boldsymbol{\alpha}_2)$ does not hold in general (although it does in the Normal case); note that $|\boldsymbol{J}^{\beta}(\boldsymbol{\alpha}_2, \boldsymbol{\beta})|^{1/2}$ is typically improper and so cannot be used.
- In spite of all this, the "Conventional Assumption" is typically used in the Linear Model after reparameterization to achieve orthogonality.

Example: let

$$f(\boldsymbol{y} \mid \alpha, \beta, \sigma) = N_n(\boldsymbol{y} \mid \alpha\, \mathbf{1}_n + \beta \boldsymbol{X}, \sigma^2 \boldsymbol{I}_n)$$

with $\boldsymbol{X}' = (x_1, \ldots, x_n)$. To choose between models:

$$\begin{aligned} M_1 &: \quad f_1(\boldsymbol{y} \mid \alpha_1, \sigma_1) = f(\boldsymbol{y} \mid \alpha_1, 0, \sigma_1) \\ M_2 &: \quad f_2(\boldsymbol{y} \mid \alpha_2, \beta, \sigma_2) = f(\boldsymbol{y} \mid \alpha_2, \beta, \sigma_2). \end{aligned}$$

Fisher information matrix for (α, β, σ) is:

$$\boldsymbol{J}(\alpha, \beta, \sigma) = \frac{1}{\sigma^4} \begin{pmatrix} n & \sum x_i & 0 \\ \sum x_i & \sum x_i^2 & 0 \\ 0 & 0 & 2n\sigma^2 \end{pmatrix}.$$

σ is orthogonal to (α, β) but α is not orthogonal to β. We re-parameterize by considering $\boldsymbol{Z} = \boldsymbol{X} - \bar{x}\mathbf{1}_n$ and $(\gamma, \beta, \sigma) = g(\alpha, \beta, \sigma) = (\alpha + \beta\bar{x}, \beta, \sigma)$ resulting in the re-parameterized model:

$$f^o(\boldsymbol{y} \mid \gamma, \beta, \sigma) = N_n(\boldsymbol{y} \mid \gamma\, \mathbf{1}_n + \beta \boldsymbol{Z}, \sigma^2 \boldsymbol{I}_n),$$

with Fisher information matrix

$$\boldsymbol{J}^o(\gamma, \beta, \sigma) = \frac{1}{\sigma^4} \left(\begin{array}{ccc} n & 0 & 0 \\ 0 & \sum z_i^2 & 0 \\ 0 & 0 & 2n\sigma^2 \end{array} \right),$$

so that β and (γ, σ) are orthogonal.

This suggests using $\pi_1^o(\gamma, \sigma) = \pi_2^o(\gamma, \sigma) = 1/\sigma$ in both models (with $\pi_2^o(\beta \mid \gamma, \sigma)$ to be determined by another method). Note that these priors transform back to the original parameterization as

$$\pi_1(\alpha_1, \sigma_1) = \pi_1^o(\alpha_1, \sigma_1) \text{ and}$$

$$\begin{aligned}\pi_2(\alpha_2, \beta, \sigma_2) &= \pi_2^o(\alpha_2 + \beta\bar{x}, \beta, \sigma_2) \begin{vmatrix} 1 & \bar{x} & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix} \\ &= \pi_2^o(\alpha_2 + \beta\bar{x}, \beta, \sigma_2)\,.\end{aligned}$$

II. Historical conventional priors: g-Priors and Zellner-Siow Priors

(traditional) *Conventional* arguments for π_i

(Jeffreys, Zellner-siow, ...)

Test $M_0 : N_n(\boldsymbol{y} \mid \boldsymbol{X}_0\boldsymbol{\beta}_0, \sigma^2\boldsymbol{I}_n)$ vs $M_i : N_n(\boldsymbol{y} \mid \boldsymbol{X}_0\boldsymbol{\beta}_0 + \boldsymbol{X}_i\boldsymbol{\beta}_i, \sigma^2\boldsymbol{I}_n)$

For "old" parameters $(\boldsymbol{\beta}_0, \sigma)$, $\pi(\boldsymbol{\beta}_0, \sigma)$

- Orthogonalize $\boldsymbol{\beta}_i$ and $(\boldsymbol{\beta}_0, \sigma)$ (in Fisher sense)
- Intuitively argue that in this case, $(\boldsymbol{\beta}_0, \sigma)$ could be taken to have similar meaning in all models
- Arguing that then a common prior distribution $\pi(\boldsymbol{\beta}_0, \sigma)$ could be taken under each of the models.
- Because of small impact of the common prior, Jeffreys argued for the objective estimation prior (with a common arbitrary constant that cancels out in BF's

Intuitively sensible but ad-hoc arguments; not formally justified.

For "new" parameters $\boldsymbol{\beta}_i$: $\pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma)$ when comparing model M_i vs M_0, should be

- proper to avoid the indeterminacy of BF
- symmetric around the null $\boldsymbol{\beta}_0 = 0$
- scaled by σ, and oriented like the likelihood
- with no moments (flat tails result in information consistency)
- Jeffreys: For n $=$ 1 the Bayes factor should be one (since a single observation allows no discrimination between the two models).

J-Z-S proposal Jeffreys (for normal mean) Zellner-Siow (for variable selection) argue that the simplest prior with the above requirements is

(the improper $\frac{1}{\sigma}$ for 'common' parameters) × (a specific Cauchy prior for 'model specific' parameters)

Zellner g-priors

- A conjugate prior proposed by Zellner (1986) often used for model selection
- Let $M : \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$
- g-prior: $\pi(\sigma^2) = \frac{1}{\sigma^2} \quad \pi(\boldsymbol{\beta} \mid \sigma^2) = N_n(\boldsymbol{\beta} \mid \boldsymbol{0}, g\, \sigma^2\, (\boldsymbol{X}^t \boldsymbol{X})^{-1})$ Conjugate/objective prior with $\boldsymbol{m}_0 = \boldsymbol{0}$ and $\boldsymbol{V}_0 = g\,(\boldsymbol{X}^t \boldsymbol{X})^{-1}$
- some choices for g:
  - g fixed, typically at $g = n$ (since $g\,(\boldsymbol{X}^t \boldsymbol{X})^{-1}$ then 'stabilizes')
  - g estimated via empirical-Bayes ($\hat{g} = (\frac{\hat{\boldsymbol{\beta}}^t (\boldsymbol{X}^t \boldsymbol{X}) \hat{\boldsymbol{\beta}}}{k s^2} - 1)^+$).
- predictive distribution is closed form:

$$m(\boldsymbol{y}) = \frac{\Gamma(n/2)}{2\,\pi^{n/2} (1+g)^{k/2}} \left(\boldsymbol{y}^t \boldsymbol{y} - \frac{g}{1+g}\, \boldsymbol{y}^t \boldsymbol{X} (\boldsymbol{X}^t \boldsymbol{X})^{-1} \boldsymbol{X}^t \boldsymbol{y} \right)^{-n/2}$$

But g-priors have undesirable features when used for model selection

Assume that, in the previous scenario, we want to test

$$M_0 : \boldsymbol{\beta} = 0 \quad \text{vs} \quad M_1 : \boldsymbol{\beta} \neq 0$$

It can be shown that as $\hat{\boldsymbol{\beta}} \to \infty$ (that is, overwhelming evidence against M_0), $B_{01} \to (1+g)^{(k-n)/2}$ a non-zero constant

This was the main reason that motivated Jeffreys to use the Cauchy, later generalized by Zellner-Siow ; those priors produce BF for $M_0 \to 0$ as evidence against $M_0 \to \infty$

This however is only serious for small n (compare to parameter dimension) and often one needs fast computations of marginals

Zellner-Siow basic proposals (orthogonal case)

- to choose between

$$M_1 : f_1(\boldsymbol{y} \mid \boldsymbol{\beta}_1, \sigma) = N_n(\boldsymbol{y} \mid \boldsymbol{X}_1\boldsymbol{\beta}_1, \sigma^2\boldsymbol{I}_n)$$

$$M_2 : f_2(\boldsymbol{y} \mid \boldsymbol{\beta}_1, \boldsymbol{\beta}_e, \sigma) = N_n(\boldsymbol{y} \mid \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_e\boldsymbol{\beta}_e, \sigma^2\boldsymbol{I}_n),$$

  with $\boldsymbol{X}_1 : n \times k_1$, $\boldsymbol{X}_e : n \times k_e$ full rank

- alternatively, to test: $H_1 : \boldsymbol{\beta}_e = \boldsymbol{0}$ versus $H_2 : \boldsymbol{\beta}_e \neq \boldsymbol{0}$

- Assume $\boldsymbol{X}_1^t\boldsymbol{X}_e = \boldsymbol{0} \rightsquigarrow$ ‘common’ parameters $\boldsymbol{\beta}_1, \sigma$ are orthogonal to ‘new’ parameters $\boldsymbol{\beta}_e$ in M_2

- ZS (1980, 1984) prior π_i under $M_i,\ i = 1, 2$:
  - common' parameters have $same$ improper prior:

$$\pi_1(\boldsymbol{\beta}_1, \sigma) = \pi_2(\boldsymbol{\beta}_1, \sigma) = \sigma^{-1},$$

  - conditionally on the 'common', the non-common $\boldsymbol{\beta}_e$ has a (proper) Cauchy prior:

$$\pi_2(\boldsymbol{\beta}_e \mid \boldsymbol{\beta}_1, \sigma) = Ca_{k_e}(\boldsymbol{\beta}_e \mid \mathbf{0}, n\sigma^2(\boldsymbol{X}_e^t \boldsymbol{X}_e)^{-1})$$

intuitive arguments for $\pi_i(\boldsymbol{\beta}_1, \sigma)$

(i) Common orthogonal parameters have same meaning across models $\leadsto$ can be given the *same* prior

(ii) Bayes factor not very sensitive to the (common) prior used for common orthogonal parameters

(Jeffreys, 1961; Kass and Vaidyanathan, 1992)

(i) and (ii) $\leadsto$ fine to assess *same improper* prior for common (orthogonal) parameters

Note: arbitrary constants cancel out in Bayes factor

For a rigorous argument based on invariance see Berger, Pericchi y Varshavsky (1998).

'Non common' $\boldsymbol{\beta}_e$ can NOT have an improper prior

intuitive arguments for $\pi_2(\boldsymbol{\beta}_e \mid \sigma)$

ZS Cauchy prior in the same spirit as Jeffreys proposal:

(i) centered (spiked) and symmetric around the simpler model $(\mathbf{0})$ in this case

(ii) has no moments

(some advantages of priors with no moments for model selection are reviewed in Liang et al. 2007)

(iii) 'right' scale $\rightsquigarrow$ oriented like the likelihood and wider

(so the prior does not 'wash-out' the likelihood, which is very easy in high dimensional problems)

(iv) it is the $IGa(.5, .5)$ scale mixture of the g-prior with $g = n$

Crucial: ZS prior is invariant to scale changes in the $\boldsymbol{X}_i$.

ZS-priors: non orthogonal case

When $\boldsymbol{X}_1^t \boldsymbol{X}_e \neq \mathbf{0} \rightsquigarrow$ reparameterize the original model

$$N_n(\boldsymbol{y} \mid \boldsymbol{X}_1 \boldsymbol{\beta}_1 + \boldsymbol{X}_e \boldsymbol{\beta}_e, \sigma^2 \boldsymbol{I}_n)$$

to orthogonality resulting in the reparameterized model:

$$N_n(\boldsymbol{y} \mid \boldsymbol{X}_1 \gamma + \boldsymbol{V} \boldsymbol{\beta}_e,\, \sigma^2 \boldsymbol{I}_n)$$

with new parameters:

$$(\boldsymbol{\gamma}, \boldsymbol{\beta}_e, \sigma) = g(\boldsymbol{\beta}_1, \boldsymbol{\beta}_e, \sigma) = (\boldsymbol{\beta}_1 + (\boldsymbol{X}_1^t \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^t \boldsymbol{X}_e \boldsymbol{\beta}_e,\, \boldsymbol{\beta}_e,\, \sigma)$$

and new design matrix for the ‘extra’ parameter:

$$\boldsymbol{V} = (\boldsymbol{I}_n - \boldsymbol{P}_1)\boldsymbol{X}_e, \quad \boldsymbol{P}_1 = \boldsymbol{X}_1(\boldsymbol{X}_1^t \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^t$$

- The original model selection problem can be equivalently formulated as that of choosing between:

$$M_1^0 : f_1^0(\boldsymbol{y} \mid \boldsymbol{\beta}_1, \sigma) = N_n(\boldsymbol{y} \mid \boldsymbol{X}_1\boldsymbol{\gamma},\, \sigma^2\boldsymbol{I}_n)$$
$$M_2^0 : f_2^0(\boldsymbol{y} \mid \boldsymbol{\beta}_1, \boldsymbol{\beta}_e, \sigma) = N_n(\boldsymbol{y} \mid \boldsymbol{X}_1\boldsymbol{\gamma} + \boldsymbol{V}\boldsymbol{\beta}_e,\, \sigma^2\boldsymbol{I}_n),$$

same 'names' $(\boldsymbol{\gamma}, \sigma)$ are used in both models since $(\boldsymbol{\gamma}, \sigma) \perp \boldsymbol{\beta}_e$

- Using JZS priors proposals for the orthogonal case and transforming back, gives JZS priors in the original formulation:

$$\pi_1(\boldsymbol{\beta}_1, \sigma) = \ \pi_2(\boldsymbol{\beta}_1, \sigma) = \sigma^{-1},$$
$$\pi_2(\boldsymbol{\beta}_e \mid \boldsymbol{\beta}_1, \sigma) = Ca_{k_e}(\boldsymbol{\beta}_e \mid \mathbf{0}, n\sigma^2(\boldsymbol{V}^t\boldsymbol{V})^{-1}),$$

- In variable selection in linear models, it is common practice to assume (without lost of generality) that
  - Regressors measured as deviations from their sample means $(\mathbf{1}^t \boldsymbol{X}_i = 0, \quad i = 1, e)$
  - Orthogonal reparameterization, so that $\boldsymbol{X}_1^t \boldsymbol{X}_e = \mathbf{0}$
- We have explicitly provided the required reparameterization, but when implementing JZS priors (Bayes factors), there is no need to worry about orthogonality since these priors are valid for both orthogonal and non-orthogonal situations
- JZS priors have desirable properties (see Berger and Pericchi, 2001, and next slides). In particular JZS priors are consistent whereas other default Bayesian methods are inconsistent (see Berger, Ghosh and Mukhopadhyay, 2003)

Z-S priors for multiple model selection

Suppose all linear sub-models are under consideration. Z-S priors are defined for comparing one model to another, so two obvious choices:

- Use the Z-S Bayes factors of each model to the full model; but this does not correspond to an actual Bayesian analysis.
- Use the Z-S Bayes factors of each model to the simplest model; this does correspond to an actual Bayesian analysis. When the simplest model is the intercept only model with $\boldsymbol{X}_1 = \mathbf{1}_n$, the resulting prior for the models is

$$\pi_0(\boldsymbol{\beta}_1, \sigma) = \pi_1(\boldsymbol{\beta}_1, \sigma) = \sigma^{-1},$$

$$\pi_e(\boldsymbol{\beta}_1, \boldsymbol{\beta}_e, \sigma) = \sigma^{-1} Ca_{k_e}(\boldsymbol{\beta}_e \mid \mathbf{0}, n\sigma^2(\boldsymbol{V}^t\boldsymbol{V})^{-1}),$$

where $\boldsymbol{V} = (\boldsymbol{I}_n - \boldsymbol{X}_1(\boldsymbol{X}_1^t\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^t)\boldsymbol{X}_e = (\boldsymbol{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^t)\boldsymbol{X}_e$.

Computation for Z-S priors

Let $\boldsymbol{X}_2 = (\boldsymbol{X}_1,\ \boldsymbol{X}_e)$; SSE_i, residual sums of squares:

$$SSE_i = \boldsymbol{y}^t(\boldsymbol{I}_n - \boldsymbol{P}_i)\boldsymbol{y}, \quad \boldsymbol{P}_i = \boldsymbol{X}_i(\boldsymbol{X}_i^t\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i^t, \quad i = 1, 2,$$

Writing the Cauchy prior as a scale mixture of normal priors, the Z-S Bayes factor is can be expressed as the one-dimensional integral:

$$\begin{aligned} B_{21} &= \int \left(1 + t\,n\,\frac{SSE_2}{SSE_1}\right)^{-(n-k_1)/2} (1+t\,n)^{(n-k)/2}\, IGa(t \mid \frac{1}{2}, \frac{1}{2})\, dt \\ &= \int \left(1 + t\,\frac{SSE_2}{SSE_1}\right)^{-(n-k_1)/2} (1+t)^{(n-k)/2}\, IGa(t \mid \frac{1}{2}, \frac{n}{2})\, dt\,. \end{aligned}$$

NOTE: This is valid whether or not the problem has beenn orthogonalized, and whether or not the matrices are full-rank (later)

Liang et al. (2007) $\rightsquigarrow$ very good Laplace approximation to B_{21}:

$$B_{21} \approx \sqrt{2\pi}\,\tilde{v}\left(1 + n\,\hat{t}\,\frac{SSE_2}{SSE_1}\right)^{-(n-k_1)/2} (1+n\,\hat{t})^{(n-k)/2}\,IGa(\hat{t} \mid \frac{1}{2}, \frac{1}{2}),$$

where $\hat{t}$ is the (real) positive solution of the cubic equation:

$$t^3\,R\left(k_1 - k - 3\right) + t^2\big(-k + n - 3 + \,R\,(k_1 - 3)\big) + t\Big(n - 3 + n\,R\Big) + n = 0,$$

where $R = SS2/SS1$, and

$$\tilde{v} = \Big(-\frac{d^2}{dt^2}\log L(t) IGa(t \mid \frac{1}{2}, \frac{1}{2})\Big|_{t=\hat{t}}\Big)^{-1/2}$$

Paulo (2003) shows, through an extensive simulation study, the accuracy of the approximation

.

III. Desiderata for Choice of Model Priors[a]

.

[a] **based on a recent paper with J. Berger, A. Forte, and G. Garcia-Donato**

Foundations of Objective Bayesian Model Selection

- there have been many efforts (over more than 30 years) to develop ‘objective model selection priors,’
- several methodologies have been proposed to derive these
  - the conventional priors (Jeffreys 1961; Zellner and Siow 1980)
  - the Intrinsic priors (Berger and Pericchi 1996; Moreno *et al.* 1998; O’Hagan 1997),
  - the Expected posterior priors (Pérez and Berger 2002),
  - the Integral priors (Cano *et al.* 2008),
  - the Divergence based priors (Bayarri and García-Donato 2008) ....

- no single criterion has emerged as dominant in defining objective prior distributions
- For the most part, these proposals have started with a good idea, used it to develop the priors, and then studied the behavior of the priors.
- The conventional priors of Jeffreys and Z-S are perhaps the most successful and widely used. They differ from the previous strategy in that
  - They first list a serie of desiderata that the objective priors for the problem should have
  - See if they can find a prior matching these desiderata

Our main goal (inspired by Jeffreys)

> Compiling, formally formulating and extending the various criteria that have been deemed essential for model selection priors and seeing if these criteria can essentially determine the priors

Notation for general model selection

- observe a vector $\boldsymbol{y} \sim f(\boldsymbol{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})$ of size n
- data can come from one of the N models:

$$
\begin{aligned}
M_0 &: f_0(\boldsymbol{y} \mid \boldsymbol{\alpha}) = f(\boldsymbol{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}_0) \\
M_i &: f_1(\boldsymbol{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = f(\boldsymbol{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}), \quad i = 1, 2, \ldots, N-1
\end{aligned}
$$

  $\boldsymbol{\alpha}$ (the ‘common’ parameter) is of dimension k_0,
  $\boldsymbol{\beta}_i$ (model specific parameters) have dimension k_i.

- priors:
  - under the null M_0: $\pi_0(\boldsymbol{\alpha})$
  - under M_i: $\quad \pi_i(\boldsymbol{\alpha}, \boldsymbol{\beta}_i) = \pi_i(\boldsymbol{\alpha})\, \pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\alpha})$.
-

Desiderata in Prior Selection

- Jeffreys' desiderata (and extensions) are intuitively sensible but mainly ad-hoc arguments: difficult to use
- We formalize the most general and compelling of the various criteria that have been used in the literature [a] , provide formal formulations and suggest a new criterion
- We have organized the criteria in four blocks:
  - I. Basic criteria,
  - II. Consistency criteria,
  - III. Predictive matching criteria,
  - IV. Invariance criteria.

[a] some few modern references relevant to development of such criteria: Fernández *et al.* (2001); Berger and Pericchi (2001); Berger *et al.* (2003); Liang *et al.* (2008); Moreno *et al.* (2009); Casella *et al.* (2009)

I. Basic criteria

Priors for the non-common parameters $\boldsymbol{\beta}_i$

- should be proper, otherwise B_{i0} is ill-defined
- cannot be arbitrarily vague, since the arbitrary scale of vagueness appears as a multiplicative term in the Bayes factor, again rendering the Bayes factor arbitrary

This elemental desideratum is reflected in the following criterion

Criterion 1 - Basic

The conditional priors $\pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\alpha})$ must be proper (integrating to one) and cannot be arbitrarily vague

II. Consistency criteria

we consider two primary consistency criteria, plus a natural third

The first criterium establishes that the Bayes procedure will select the right model with enough data

> **Criterion 2 - Model selection consistency -**
> *If data $\boldsymbol{y}$ have been generated by M_i, then the posterior probability of M_i should converge to 1 as the sample size $n \to \infty$.*

This is an obvious criterium to require, and indeed model selection consistency is generally satisfied [a] (but not always!) so it is not usually very useful to characterize priors.

[a] (see e.g. O'Hagan 1994)

The second criterium stablishes that, for any fixed sample size, if the evidence in favor of M_i (and against M_0) grows to ∞, then the Bayes factor B_{i0} should also grow to ∞

Criterion 3 - Information consistency:
For any model M_i, and sample size n, if the likelihood ratio arising from comparing M_i to M_0 goes to ∞, then B_{i0} should $\to \infty$.

In normal linear models the criteria is equivalently formulated in terms of the F (or t) statistics growing to infinity

Jeffreys required (for the normal mean testing) $B_{i0} \to \infty$ as $\bar{y} \to \infty$, and let him to recommend the Cauchy prior instead of the normal

A third type of consistency is the formalization of Berger & Pericchi[a] requirement that a model selection procedure should correspond, at least approximately (or asymptotically) to a *genuine* Bayes procedure (in particular, with a fixed prior, independent of the data, and of n)

Criterion 4 - Intrinsic prior consistency:

Consider $\pi_i(\boldsymbol{\beta}_i \mid n, \boldsymbol{\beta}_0, \sigma)$ *for sample size* n. *Then, as* $n \to \infty$, $\pi_i(\boldsymbol{\beta}_i \mid n, \boldsymbol{\beta}_0, \sigma)$ *should converge to a proper, fixed, 'intrinsic' prior* $\pi_i^I(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma)$

If there is such a limiting prior it is called an *intrinsic prior*

.

[a] see B&P 2001 for extensive discussion and previous references.

III. Predictive matching criteria

- Perhaps the most crucial aspect of objective model selection priors is that they be appropriately 'matched' across models of different dimensions
- Jeffreys argued that with 'minimal sample size' one could not discriminate between the two hypotheses and the Bayes factor should then be 1. Argument formalized in Berger & Pericchi 2001

Criterion 5 - Predictive matching:

For samples $\boldsymbol{y}^*$ of 'minimal size', in comparing M_i with M_j, one should have model selection priors such that $m_i(\boldsymbol{y}^*)$ and $m_j(\boldsymbol{y}^*)$ are close. Optimal, but not always possible, is exact predictive matching: $m_i(\boldsymbol{y}^*) = m_j(\boldsymbol{y}^*)$.

.

- In Berger and Pericchi (2001), minimal sample size n^* was defined as the smallest sample size for which
$$0 < {m_i}^{N}(\boldsymbol{y}^*) < \infty$$
for *all* models M_i when *objective estimation* priors π_i^N under Mi are used

- BP01 minimal sample size typically equals the number of observations needed for all parameters to be identifiable.

- For model selection minimal sample size should be defined relative to the model selection priors being utilized. We propose general definition

Definition: A Minimal training sample $\boldsymbol{y}_i^*$ for $\{M_i, \pi_i\}$ is a sample of minimal size $n_i^* \geq 1$ such that $0 < m_i(\boldsymbol{y}^*) < \infty$

Some consequences

- Because of the *Basic criteria* this new n^* is smaller than the B&P01 n^*; the predictive matching criteria becomes a weaker condition.
- In problems with more than 2 competing models (e.g variable selection) the concept of minimal size is not so sensitive to the dimension of the largest model.

- Exact predictive matching is usually understood to imply that $m_i(\boldsymbol{y}^*) = m_j(\boldsymbol{y}^*)$ for *all* entertained models, and with the *same* MTS size n^*

- The new (weaker) definition allows entertaining partial comparisons resulting in exact predictive matching among interesting subsets of models

- We highlight two types of this kind of 'partial' exact predictive matching which are of particular relevance for the variable selection problem

Definition: Null predictive matching The model specific priors are *null predictive matching* if all pairs $\{M_i, \pi_i\}$ and $\{M_0, \pi_0\}$ are exact predictive matching for all minimal training samples $\boldsymbol{y}_i^*$ for $\{M_i, \pi_i\}$.

- This concept formalizes the idea that data of minimal size not allow one to distinguish between the null and alternative models
- NOTE: This null matching is entertained for possibly *different* n_i*, so whereas all models can be matched to the null, it might be that no other two models of differing dimensions are matched.
- A very interesting concept of 'exact' (but partial) predictive matching occurs when only models of the same complexity are required to be matched

Definition: Dimensional predictive matching The model selection priors are *dimensional predictive matching* if each of the model/prior pairs $\{M_i, \pi_i\}$ and $\{M_j, \pi_j\}$ of the same complexity/dimension (i.e. $k_i = k_j$) are exact predictive matching for all minimal training samples $\boldsymbol{y}_i^*$ for models of that dimension.

IV. Invariance criteria

- Invariance has played an important role in objective Bayes methods.
- It says that if the problem is invariant under some transformations, the objective priors should leave them invariant.
- We introduce two invariant criteria: the first one is rather obvious, and the second is a new criterium with important consequences

  Criterion 6 - Measurement invariance:
  The units of measurement used for the observations or model parameters should not affect the Bayesian answers.

The second refers to a much more powerful, but special type of invariance:

Criterion 7: Group invariance criterion (new):

If all models are invariant under a group of transformations G_0, then the conditional distributions, $\pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\alpha})$, should be chosen in such a way that the conditional marginal distributions

$$f_i(\boldsymbol{y} \mid \boldsymbol{\alpha}) = \int f_i(\boldsymbol{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}_i)\, \pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\alpha})\, d\boldsymbol{\beta}_i$$

are also invariant under G_0.

- Indeed, the $\pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\alpha})$ could hardly be called objective model selection priors if they eliminated an invariance structure that was possessed by all of the original models.

- *Note:* If it exists, G_0 is the relevant group of transformations for the null model M_0.

- Otherwise stated we ask $\pi(\boldsymbol{\beta}_i \mid \boldsymbol{\alpha})$ to be in 'agreement' with the structure of the null model
- This can also be viewed as a formalization of Jeffreys' requirement that the prior for a non-null parameter should be "centered at the simple model."

A crucial implication of the Group Invariance Criteria, which could be formulated as part of the criteria is the following:

> Criterion 7, Group invariance criterion (cont.): prior for the common parameters $\boldsymbol{\alpha}$ in each model should be assigned the right-Haar prior corresponding to the group of transformations.

- Indeed, with the group invariance criterion, the problem becomes that of selecting among the models:

$$f_0(\boldsymbol{y} \mid \boldsymbol{\alpha}),\; f_i(\boldsymbol{y} \mid \boldsymbol{\alpha}),\quad i = 1, \ldots, N$$

  with *same dimension* and *common invariance structure*.

- In this situation choosing $\pi_i(\boldsymbol{\alpha}) = \pi^H(\boldsymbol{\alpha})$ where $\pi^H(\cdot)$ is the right-Haar density of G_0 guarantees (under commonly satisfied conditions) exact predictive matching (Berger *et al*, 1998)

- Thus, for invariant models, the combination of the Group invariance criterion and (exact) Predictive matching criterion allows complete specification of the prior for $\boldsymbol{\alpha}$ in all models.

- Most surprisingly, this argument does NOT require orthogonality of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}_i$

Example: testing a normal standard deviation

- Let $\boldsymbol{y}$ be an iid sample of a normal population with μ, σ unknown
- To test: $H_0 : \sigma = \sigma_0, \qquad$ vs $H_1 : \sigma \neq \sigma_0$
- Need to assess objective priors
  $\pi_0(\mu)$ and $\pi_1(\mu, \sigma) = \pi_1(\sigma \mid \mu)\pi_1(\mu)$

implications of several criteria

- Basic: $\pi_1(\sigma \mid \mu)$ must be a proper and not arbitrarily vague
- Invariance: M_0 and M_1 are invariant under the group
  $G_0 = \{g \in \mathcal{R}\}$ with $g(\boldsymbol{y}) = \boldsymbol{y} + g\mathbf{1}_n$

  Result: $\pi_1(\sigma \mid \mu)$ satisfies the invariance criterion *if and only if* $\pi_1(\sigma \mid \mu) = h(\sigma)$

- Predictive matching: The minimal sample size for $\pi_1(\mu, \sigma) = h(\sigma)\pi_1(\mu)$ is $n^* = 1$.

  Result: The priors $\pi_0(\mu) = \pi_1(\mu) = \pi^H(\mu)$ where $\pi^H(\mu) = 1$ (right Haar measure for G_0) are *exact* predictive matching.

- Consistency: For fixed n, $\Lambda_{10} \to \infty$ if and only if $n \geq 2$ and either $S \to \infty$ or $S \to 0$.

  Result: Under these conditions, B_{10} also $\to \infty$ if and only if $\int_0^\infty \sigma^{1/2}\, h(\sigma)\, d\sigma = \infty$.

Notice: this is a stronger requirement than having no moments and is not met, for instance, by the conjugate prior.

In summary:

The class of priors meeting the criteria satisfy

$$\{(\pi_0, \pi_1) : \ \pi_0(\mu) = 1, \ \ \pi_1(\mu, \sigma) = h(\sigma)\}$$

with

$$\int_0^\infty h(\sigma)\, d\sigma = 1, \ \int_0^\infty \sigma^{-1} h(\sigma)\, d\sigma = \infty.$$

Example: testing a Gamma shape parameter

- Let $\boldsymbol{y}$ be an iid sample from a Gamma density with mean μ and shape parameter α : $f(y \mid \alpha) \propto y^{\alpha-1} \, e^{(-\alpha \, y)/\mu}$
- To test: $H_0 : \alpha = \alpha_0, \qquad H_1 : \alpha \neq \alpha_0,$
- Need to assess priors $\pi_0(\mu)$ and $\pi_1(\mu, \alpha) = \pi_1(\alpha \mid \mu)\pi_1(\mu)$

implications of several criteria

- Basic: $\pi_1(\alpha \mid \mu)$ must be a proper and not arbitrarily vague
- Invariance: M_0 and M_1 are invariant under the group $G_0 = \{g \in (0, \infty)\}$ with $g(\boldsymbol{y}) = g\boldsymbol{y}$ and $g(\boldsymbol{y}) = \boldsymbol{y} + g\mathbf{1}_n$

  Result: $\pi_1(\alpha \mid \mu)$ satisfies the invariance criterion *if and only if* $\pi_1(\alpha \mid \mu) = h(\alpha)$.

- Predictive matching: The minimal sample size for $\pi_1(\mu, \alpha) = h(\alpha)\pi_1(\mu)$ is $n^* = 1$.

  The priors $\pi_0(\mu) = \pi^H(\mu)$ and $\pi_1(\mu) = \pi^H(\mu)$ where $\pi^H(\mu) = 1/\mu$ (ie the right-Haar measure for G_0), are *exact* predictive matching.

- Consistency: For fixed n, $\Lambda_{10} \to \infty$ if and only if $n \geq 2$ and either $\hat{\alpha} \to \infty$ or $\hat{\alpha} \to 0$.

  Under these conditions, B_{10} also $\to \infty$ if and only if $\int_1^\infty \alpha^{1/2}\, h(\alpha)\, d\alpha = \infty$.

Notice: this is a stronger requirement than having no moments.

In summary:

The class of priors meeting the criteria satisfy

$$\{(\pi_0, \pi_1) : \;\; \pi_0(\mu) = 1, \;\; \pi_1(\mu, \alpha) = h(\alpha)\}$$

with

$$\int_0^\infty h(\alpha)\, d\alpha = 1, \; \int_1^\infty \sqrt{\alpha}\, h(\alpha)\, d\alpha = \infty.$$

IV. Variable Selection in Linear Models

Variable selection as model selection

- Observations $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^t$ explained by, at most, $k_0 + p$ covariates

- Simplest model contains k_0 pre-chosen covariates:

$$M_0 : \boldsymbol{Y} \sim N_n(\boldsymbol{X}_0\boldsymbol{\beta}_0, \sigma^2\boldsymbol{I})$$

- The other $2^p - 1$ models correspond to additionally adding each of the $2^p - 1$ non-null subsets of the remaining p covariates:

$$M_i : \boldsymbol{Y} \sim N_n(\boldsymbol{X}_0\boldsymbol{\beta}_0 + \boldsymbol{X}_i\boldsymbol{\beta}_i, \sigma^2\boldsymbol{I})$$

- $(\boldsymbol{\beta}_0, \sigma^2)$ ‘occur’ in all models (*common parameters*), whereas $\boldsymbol{\beta}_i$ do not, $i = 1, \ldots, 2^p - 1$

Objective Bayes model selection

- Is based on posterior probabilities for each model:

$$P(M_i \mid \boldsymbol{y}) = \frac{m_i(\boldsymbol{y})P(M_i)}{\sum_{l=0}^{2^p} m_l(\boldsymbol{y})P(M_l)} = \left[1 + \sum_{l \neq i} \pi_{li} B_{li}\right]^{-1}$$

- π_{li} is prior odds $Pr(M_l)/Pr(M_i)$
- B_{li} is Bayes factor $m_l(\boldsymbol{y})/m_i(\boldsymbol{y})$

$$m_i(\boldsymbol{y}) = \int f_i(\boldsymbol{y} \mid \boldsymbol{\beta}_i,\, \boldsymbol{\beta}_0,\, \sigma)\, \pi_i(\boldsymbol{\beta}_i,\, \boldsymbol{\beta}_0,\, \sigma)\, d(\boldsymbol{\beta}_i,\, \boldsymbol{\beta}_0,\, \sigma),$$

- Derive the $2^p - 1$ Bayes factors B_{0i} and compute all pairwise Bayes Factors as $B_{li} = B_{l0}B_{0i} = B_{0l}^{-1}B_{0i}$

The Zellner-Siow Priors

In the variable selection problem Zellner-Siow's prior

$$\pi_i^{ZS}(\boldsymbol{\beta}_i, \boldsymbol{\beta}_0, \sigma) = \frac{1}{\sigma}\, Cauchy_{k_i}(\boldsymbol{\beta}_i \mid 0, n\sigma^2 (\boldsymbol{V}_i^t \boldsymbol{V}_i)^{-1})$$

with $\boldsymbol{V}_i = (\boldsymbol{I}_n - \boldsymbol{X}_0(\boldsymbol{X}_0^t \boldsymbol{X}_0)^{-1}\boldsymbol{X}_0^t)\boldsymbol{X}_i$ satisfies

- *Basic criterion:* The conditional priors $\pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma)$ are proper and are not arbitrarily vague.
- *Model selection consistency* and *Information consistency* are established in Liang et. al. (2008).
- *Intrinsic prior consistency* holds if $\lim_{n\to\infty} n^{-1}\, \boldsymbol{V}_l^t \boldsymbol{V}_l = \boldsymbol{\Lambda}_l$ for positive definite $\boldsymbol{\Lambda}_l$. This would trivially happen if either there is a fixed design with replicates, or when the covariates arise randomly from a fixed distribution having second moments.

- *Predictive matching* occurs for samples of size $k_0 + 1$. Surprisingly, it is also the case that $m_i(\boldsymbol{y}^*) = m_0(\boldsymbol{y}^*)$ for all samples of size $k_0 + k_i$, and this is only true if the conditional covariance matrix in the prior is proportional to $(\boldsymbol{V}_i^t \boldsymbol{V}_i)^{-1}$.

- *Measurement invariance* is easily seen to be satisfied with the choice $(\boldsymbol{V}_i^t \boldsymbol{V}_i)^{-1}$ for the conditional covariance matrix.

- *Group invariance*, with respect to the location-scale group, holds for the $\pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma)$, and the common right-Haar prior $\pi(\boldsymbol{\beta}_0, \sigma) = 1/\sigma$ is used for all models.

The only negative feature of the Zellner-Siow prior is it does not lead to closed form answers, though only one-dimensional integrals are required.

The Robust prior

We now characterize which classes of priors satisfy the different desideratum and the implications that they have in the choice of priors. We call our ultimate choice the The Robust Prior

- The chosen flat-tailed prior is a generalization of proposals by Strawderman (1971, 1973) and Berger (1976, 1980, 1985) in the context of minimax and robust Bayes estimation

- Our proposal for $\pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0,\, \sigma)$, the **robust** prior is

$$\pi_i^R(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma) = \int_0^1 N_{k_i}(\boldsymbol{\beta}_i \mid \mathbf{0},\, (\lambda^{-1}\rho_i\,(b\,+n)\,-\,b)\,\boldsymbol{\Sigma}_i)\,a\,\lambda^{a-1}\,d\lambda,$$

where $\boldsymbol{\Sigma}_i = \mathsf{Cov}[\hat{\boldsymbol{\beta}}_i] = \sigma^2(\boldsymbol{V}_i^t\boldsymbol{V}_i)^{-1}$.

- a, b, ρ_i are chosen to achieve optimal properties for model selection

a scale mixture(s) of normals

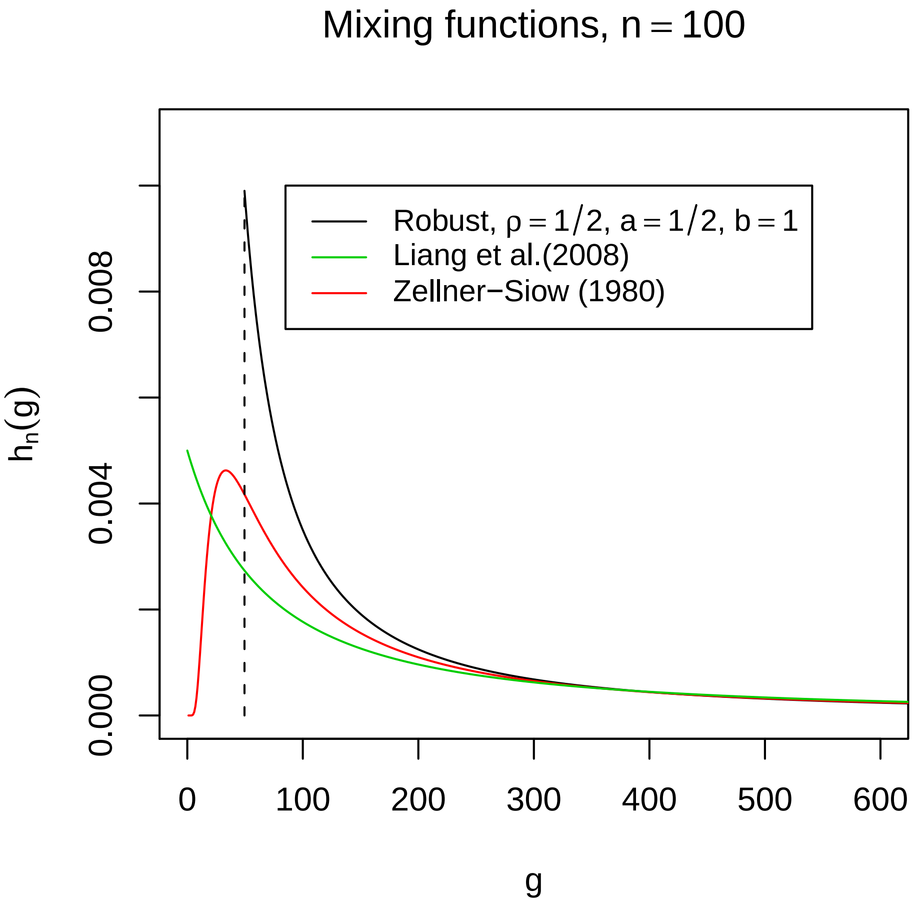
as many other conventional priors for model selection, our robust prior can be expressed as an scale mixture of normals:

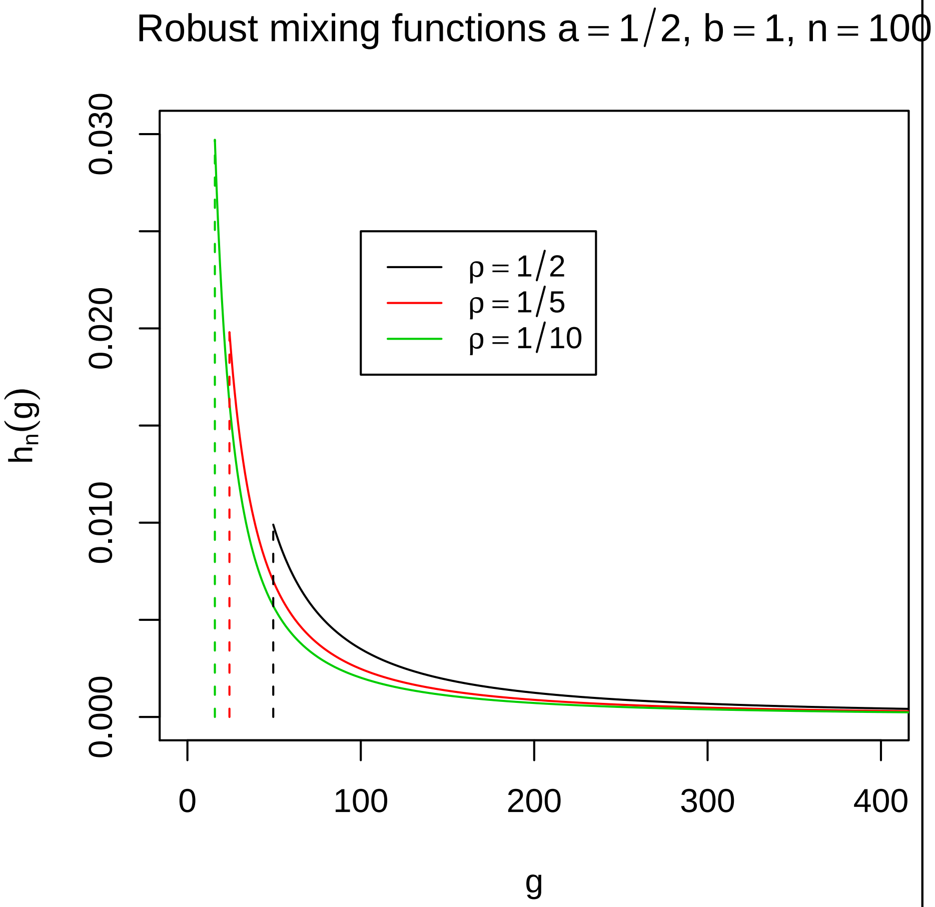
$$\pi_i^R(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma) = \int_0^\infty N_{k_i}(\boldsymbol{\beta}_i \mid \mathbf{0}, g\,\sigma^2\,(\boldsymbol{V}_i^t \boldsymbol{V}_i)^{-1})\, p_n(g)\, dg$$

- For the robust:

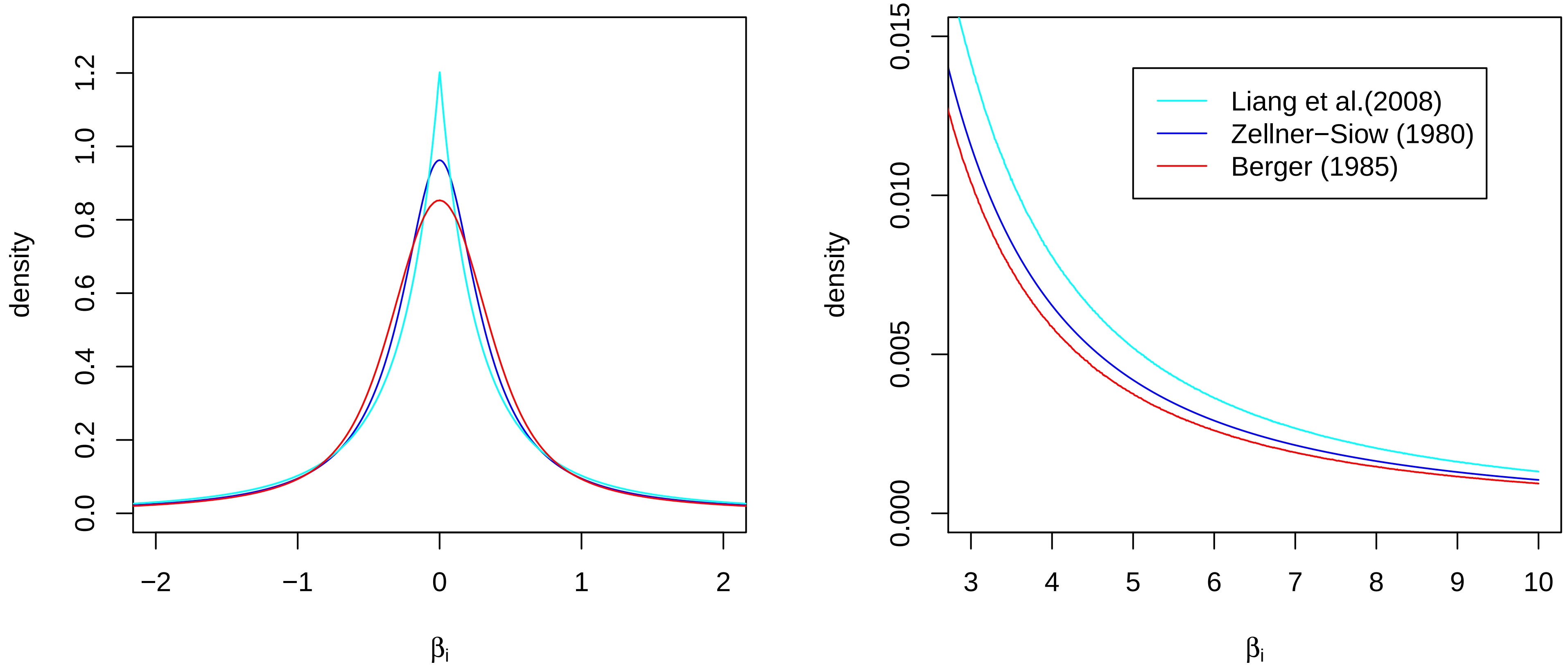
$$p_n^R(g) = \begin{cases} a(\rho_i(b+n))^a (g+b)^{-(a+1)} & g > \rho_i(b+n) - b \\ 0 & \text{otherwise} \end{cases}$$

- For Zellner $h_n^{ZS}(g) = Ga^{-1}(g \mid \frac{1}{2}, \frac{n}{2}), \quad g \geq 0$
- The *consistent* choice in Liang et al (2008) is a particular case of the Robust prior for $a = 1/2$, $b = n$, and $\rho_i = 1/2$.
- Berger (1985)'s robust prior for estimation has $a = 1/2$, $b = 1$, and $\rho_i = (k_i + 1)/(k_i + 3)$


Mixing functions, n = 100
$h_n(g)$
0.000
0.004
0.008
0
100
200
300
400
500
600
g
Robust, ρ = 1/2, a = 1/2, b = 1
Liang et al.(2008)
Zellner−Siow (1980)


Robust mixing functions a = 1/2, b = 1, n = 100
$h_n(g)$
0.000
0.010
0.020
0.030
0
100
200
300
400
g
ρ = 1/2
ρ = 1/5
ρ = 1/10

They have Student t tails, and for $a \leq 1/2$, no moments



The Robust priors:

$$\pi_i^R(\boldsymbol{\beta}_0, \boldsymbol{\beta}_i, \sigma) = \sigma^{-1} \times \int_0^\infty N_{k_i}(\boldsymbol{\beta}_i \mid \mathbf{0}, g\,\boldsymbol{\Sigma}_i)\, p_i^R(g)\, dg,$$

- Also satisfy all the desiderata.
- Yield closed form Bayes factors

$$B_{i0} = \left[\frac{n+1}{k_i+k_0}\right]^{-\frac{k_i}{2}} \frac{Q_{i0}^{-(n-k_0)/2}}{k_i+1}\, {}_2F_1\Big[\frac{k_i+1}{2}; \frac{n-k_0}{2}; \frac{k_i+3}{2}; \frac{(1-Q_{i0}^{-1})(k_i+k_0)}{(1+n)}\Big],$$

where ${}_2F_1$ is the standard hypergeometric function and

$$Q_{i0} = SSE_i/SSE_0$$

is the ratio of the sum of squared errors of models M_i and M_0.

- Adjust for the fact that n is not always the *effective sample size* for the parameters of the models.

An invariant argument for “common” parameters

we show when the (conditional) marginal likelihood, $m_i^R(\boldsymbol{y} \mid \boldsymbol{\beta}_0,\, \sigma)$ is invariant, and hence the right Haar prior produces well defined BF’s:

Theorem: The likelihood $m_i(\boldsymbol{y} \mid \boldsymbol{\beta}_0,\, \sigma)$ for $(\boldsymbol{\beta}_0,\, \sigma)$ under model M_i for $i = 1, \ldots 2^p - 1$ derived by integrating out $\boldsymbol{\beta}_i$ with any prior of the form: (iif)

$$\pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0,\, \sigma) = \sigma^{-k_i} f_i(\frac{\boldsymbol{\beta}_i}{\sigma}),$$

is invariant under the group of transformations

$G = \{g_{c,b} : g_{c,b}(\boldsymbol{y}) = c\,\boldsymbol{y} + \boldsymbol{X}_0\boldsymbol{b};\ \boldsymbol{b} \in \Re^{k_0};\ c > 0\}$

m_0 is also invariant under G

Corolary: The right Haar measure for this group, namely $1/\sigma$, can then be used as priors for these parameters, under each of the models, when deriving Bayes factors for model selection.

Important remarks

- We have justified use of the default independent Jeffreys prior for use as prior for 'common' parameters in a large class of variable selection problems

- Justification does not rely on ad-hoc arguments: as a result of invariance, $\pi(\boldsymbol{\beta}_0, \sigma) = 1/\sigma$ produce well defined Bayes factors

- σ needs to be a scale parameter in π_i, hence we have also formally justified the common (but not formally justified) practice of scaling the prior on $\boldsymbol{\beta}_i$ by σ, the scale of the model. Note that it also is centered at 0 and do not depend on β_0

- The Robust Prior satisfy the Group invariance criterion

Arguments for priors on the “new” parameters

From now on, we consider an important particular case of priors satisfying the invariance criteria, that is, of priors of the form

$$\pi_i(\boldsymbol{\beta}_0, \boldsymbol{\beta}_i, \sigma) = \sigma^{-1-k_i}\, h_i(\frac{\boldsymbol{\beta}_i}{\sigma})$$

namely, the scale mixture of normals: (Basic criterium)

$$\pi_i(\boldsymbol{\beta}_0, \boldsymbol{\beta}_i, \sigma) = \sigma^{-1} \times \int_0^\infty N_{k_i}(\boldsymbol{\beta}_i \mid \mathbf{0}, g\,\sigma^2\, \boldsymbol{\Sigma}_i)\; p_i(g)\, dg,$$

The robust prior being a particular case for

- $\boldsymbol{\Sigma}_i = Cov(\widehat{\boldsymbol{\beta}}_i) = (\boldsymbol{V}_i^t \boldsymbol{V}_i)^{-1}$
- $p_i(g) = p_i^R(g)$

argument for the particular form of $h_i(\cdot)$

- It is a very rich, ‘natural’ class of densities symmetric about 0
- only scale mixtures of normals seem to have any possibility of yielding Bayes factors that have 'easy' expressions (not necessarily close-form, like ZS)

argument for the form of the mixing density

The mixture density $p_i^R(g)$ encompasses virtually all [a] of the mixtures that have been found which can lead to closed form expressions for Bayes factors (Zellner-Siow priors have a different mixing) .

[a] Maruyama and George, 2008, being an interesting exception

arguments for the choice of the conditional scale matrix

- A standard argument: The Measurement Invariance criterion

  if $\Sigma_i = (\boldsymbol{V}_i^t \boldsymbol{V}_i)^{-1}$, Bayes factors will be unaffected by changes in the units of measurement of either $\boldsymbol{y}$ or the model parameters

  But there are many other choices with this property

- A quite surprising Predictive matching result:

  RESULT: *Scale mixture of normals with this $\boldsymbol{\Sigma}_i$ are null predictive matching and dimensional predictive matching for samples of size $k_0 + k_i$, and no choice of the conditional scale matrix other than $(\boldsymbol{V}_i^t \boldsymbol{V}_i)^{-1}$ (or a multiple) can achieve this predictive matching*

Selecting hyperparameters of the robust prior

- Recap: to meet previous desiderata, we are considering priors of the form:

$$\pi_i(\boldsymbol{\beta}_0, \boldsymbol{\beta}_i, \sigma) = \sigma^{-1} \times \int_0^\infty N_{k_i}(\boldsymbol{\beta}_i \mid \mathbf{0}, g\,\sigma^2\,(\boldsymbol{V}_i^t \boldsymbol{V}_i)^{-1})\; p_i(g)\,dg,$$

- our proposal, The robust prior is the particular case for $p_i(g) = p_i^R(g) = a(\rho_i(b+n))^a (g+b)^{-(a+1)}$ for $g > \rho_i(b+n) - b$

- Need to pick 'good' values for the hyperparameters a, b and ρ_i

- The values for the hyperparameters that will be recommended are $a = 1/2$, $b = 1$ and $\rho_i = (k_i + k_0)^{-1}$. The arguments justifying this specific recommendation follow.

Implications of the consistency criteria

1.- Model selection consistency

- Recall: as $n \to \infty$, probability of correct model $\to 1$
- A key assumption: models are asymptotically differentiated

$$\lim_{n\to\infty} \frac{\boldsymbol{\beta}_i^t \boldsymbol{V}_i^t (\boldsymbol{I} - \boldsymbol{P}_j) \boldsymbol{V}_i \boldsymbol{\beta}_i}{n} = b_j \in (0, \infty) \qquad \text{Fernández et al. 01}$$

where $\boldsymbol{P}_j = \boldsymbol{V}_j (\boldsymbol{V}_j^t \boldsymbol{V}_j)^{-1} \boldsymbol{V}_j^t$ and M_i is true model

- Model selection consistency results if $p_i(g)$ are proper densities such that $\lim_{n\to\infty} \int_0^\infty (1+g)^{-k_i/2}\, p_i(g)\, dg = 0\,,$

Robust Priors Result : *If* $\lim_{n\to\infty} \rho_i\,(b+n) = \infty$, *then the Conventional Robust Bayes factors are MS consistent*

2.- Intrinsic prior consistency

- Recall: asymptotically, MS procedures should be Bayesian, corresponding to a fixed, intrinsic prior
- another key assumption related to 'differentiated models' above: $\lim_{n\to\infty} \frac{1}{n} \boldsymbol{V}_l^t \boldsymbol{V}_l = \boldsymbol{\Xi}_l$, for some positive definite matrix $\boldsymbol{\Xi}_l$
- This would trivially happen if either there is a fixed design with replicates, or when the covariates arise randomly from a fixed distribution having second moments.

Robust prior result:*if 'key assumption' holds, a and ρ_i do not depend on n, and $\frac{b}{n} \to c$, then the conditional robust prior converges to a fixed intrinsic prior*

3.- Information consistency

Recall: for a fixed n, as the support in the data for M_i grows to ∞, B_{01} should go to 0

Robust prior result: *If $\rho_i \geq b/(b+n)$, The Bayes factor, B_{i0}^R is information consistent if and only if $n \geq k_i + k_0 + 2a$*

Consequence : If $\boldsymbol{0 < a \leq 1/2}$, then all $2^p - 1$ Bayes factors B_{i0}^R are information consistent for $n \geq p + k_0 + 1$.

Summary: The three consistency criteria are satisfied by the robust prior if a and ρ_i do not depend on n, $\lim_{n\to\infty} \frac{b}{n} = c \geq 0$, $\lim_{n\to\infty} \rho_i\,(b+n) = \infty$, and $n \geq k_i + k_0 + 2a$

Close-form Bayes factors

Remarkably, the Conventional Robust priors produce marginal likelihoods (and hence Bayes factors) in closed form, which is particularly simple for $\boldsymbol{b = 1}$

Result : For $b = 1$, the Conventional Robust Bayes factors are:

$$B^R_{i0} = Q_{i0}^{-\frac{n-k_0}{2}} \frac{2a}{k_i + 2a} \left(\rho_i \, (n+1)\right)^{-\frac{k_i}{2}} \, \mathsf{H}G_{i0},$$

where $\mathsf{H}G_{i0}$ is the hypergeometric function of one variable:

$$\mathsf{H}G_{i0} = \ {}_2F_1\left[a + \frac{k_i}{2}; \frac{n-k_0}{2}; a + 1 + \frac{k_i}{2}; \frac{1 - Q_{i0}^{-1}}{\rho_i \, (1+n)}\right].$$

and $Q_{i0} = SSE_i/SSE_0$ is ratio of residual sums of squares

- Thus computation with robust priors with $\boldsymbol{b=1}$ is remarkably simpler than with other flat-tailed priors
  - BF's for robust priors with $b \neq 1$ while still in close form (in terms of the Apell function), are considerably more complex
  - BF's for student priors can not be expressed in closed form
- This was a main motivation for its use in robust Bayesian analysis
- This is a very appealing characteristic specifically for problems with huge model spaces

A specific proposal:

$$\boldsymbol{a = 1/2, \quad b = 1, \quad \rho_i = (k_0 + k_i)^{-1}}$$

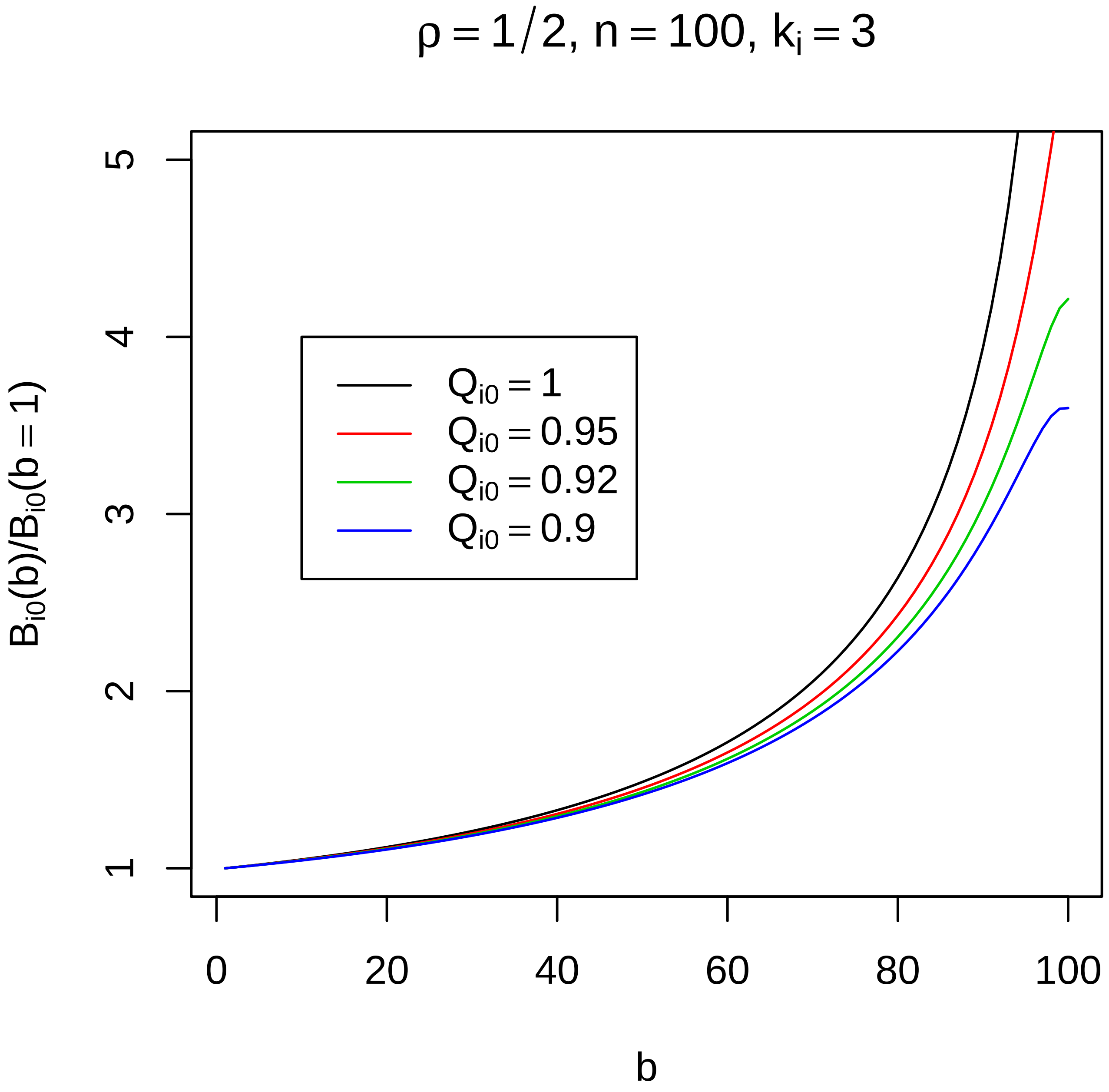
We choose specific values for the hyperparameters a, b, ρ_i so resulting procedure has desirable properties

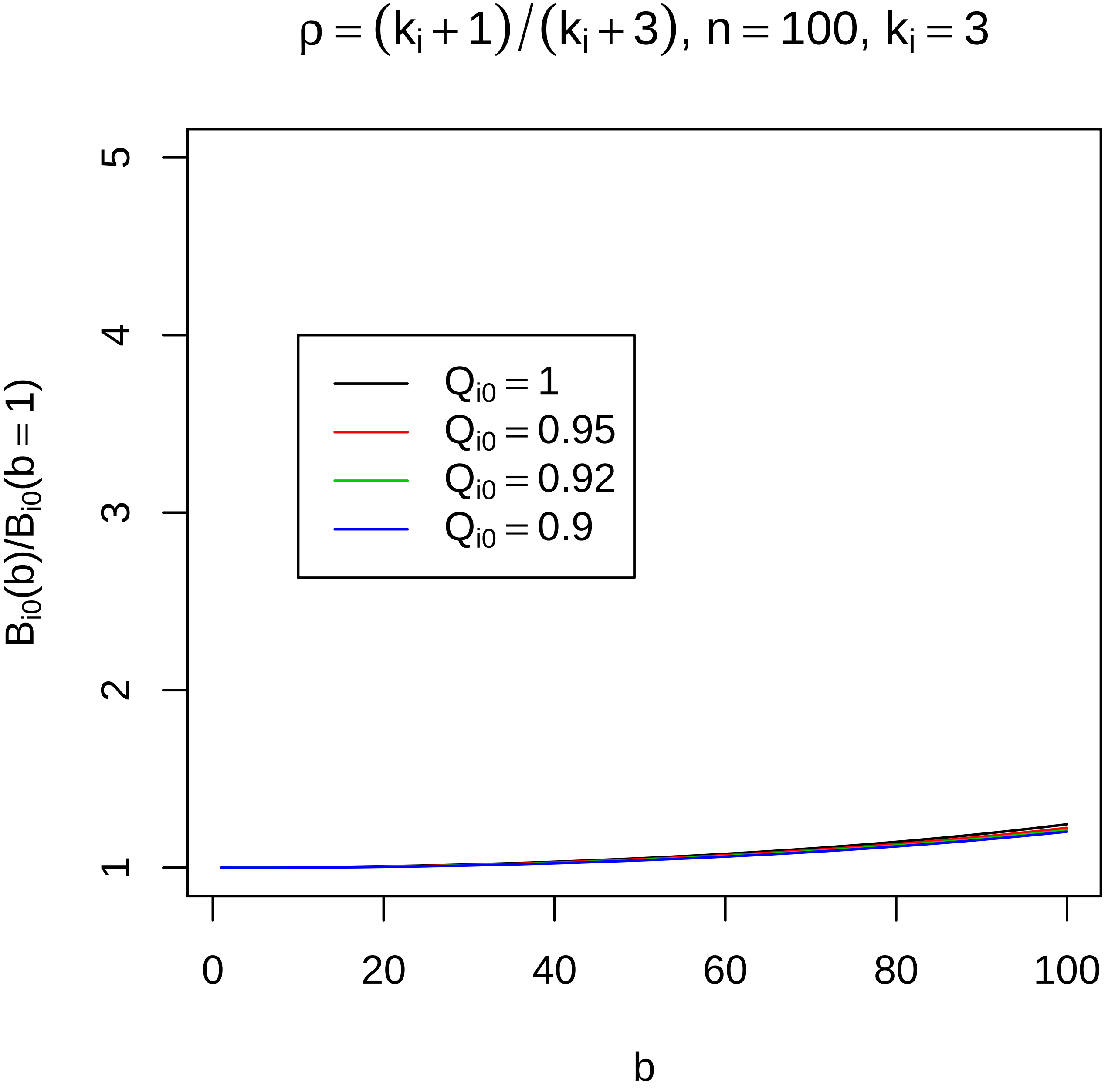
Choosing a $= 1/2$: behavior on the tails

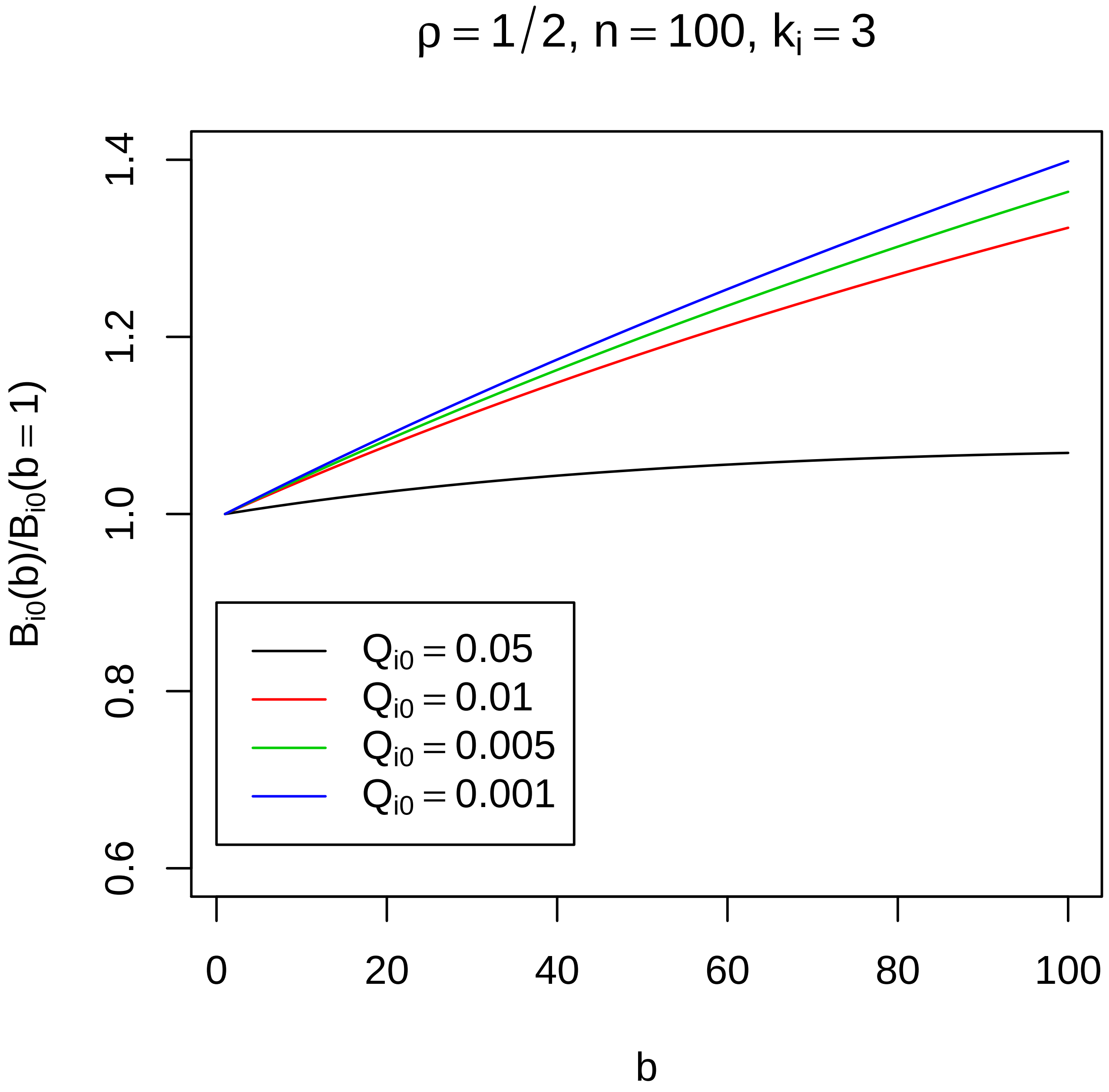
- gives $\pi_i^R(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma)$ with no moments
- results in information consistency of all Bayes factors for the minimal (frequentist) sample size $n \geq p + k_0 + 1$
- it is the largest value of a with this property $\rightsquigarrow$ avoids too conservative a procedure
- This is the choice in Berger(80) and Liang et al. (08); also, the resulting π_i^R has Cauchy tails (as in Jeffreys, ZS)

Choosing b = 1: computational convenience

- It gives (by far) the simplest computation of Bayes factors (Apell function is more complex and much less available in software)
- Bayes factors are, in general, considerably robust to choice of b (and the intrinsic prior does not depend on b)
- Significant sensitivity only can occur **only** when **both**
  - data is *extremely* compatible with M_0
  - b is *very* close to n
- and even this only occurs for some values ρ_i:
  - Liang et al. 08'choices does have this behavior, but the mixing density then has an unbounded spike at 0
  - Berger 85's choice produces remarkably insensitive BF's for all choices of b (even for b close to n)


$\rho = 1/2,\ n = 100,\ k_i = 3$
$B_{i0}(b)/B_{i0}(b = 1)$
b
0
20
40
60
80
100
1
2
3
4
5
$Q_{i0} = 1$
$Q_{i0} = 0.95$
$Q_{i0} = 0.92$
$Q_{i0} = 0.9$


$\rho = (k_i + 1)/(k_i + 3)$, $n = 100$, $k_i = 3$
$B_{i0}(b)/B_{i0}(b = 1)$
1
2
3
4
5
0
20
40
60
80
100
b
$Q_{i0} = 1$
$Q_{i0} = 0.95$
$Q_{i0} = 0.92$
$Q_{i0} = 0.9$


$\rho = 1/2$, $n = 100$, $k_i = 3$
$B_{i0}(b)/B_{i0}(b = 1)$
1.4
1.2
1.0
0.8
0.6
0
20
40
60
80
100
b
$Q_{i0} = 0.05$
$Q_{i0} = 0.01$
$Q_{i0} = 0.005$
$Q_{i0} = 0.001$

Choosing $\rho_i = (1 + k_0 + k_i)^{-1}]$

- ρ_i can be quite influential, so it needs to be carefully chosen.
- Restrictions so far:
  - propriety of π_i^R for $a = .5, b = 1$ requires $\rho_i \geq (1+n)^{-1}$
  - model selection consistency requires $\lim_{n\to\infty} \rho_i(1+n) = \infty$
  - existence of intrinsic prior $\rightsquigarrow$ ρ_i should not depend on n
- Note: For existence of the robust prior and marginal likelihoods: $n \geq 1/(k_0 + k_1)$
- Choose ρ_i so that previous conditions are satisfied for all such n:
  - ρ_i must be a constant (independent of n)
  - $\rho_i \geq 1/(1 + k_0 + k_i)$
- Desiderata met by all such ρ_i, further arguments required

Argument 1: more on predictive matching

- since $B_{i0} = 1$ for a sample of size $n = k_i + k_0$ it seems reasonable to require that a single extra observation would not be able to strongly discriminate between the models

- To quantify intuition, we look for BF's that are close to 1 for data being as supportive as possible of M_0 and sample size $n = k_0 + k_i + 1$ (also Ghost & Samantha 02, Spiegelhalter & Smith 82)

- This is achieved by choosing ρ_i to be as small as is reasonable. The choice $\rho_i = 1/(k_0 + k_i + 1)$ is the minimum value of ρ_i and is a candidate, and so is $\rho_i = 1/(k_0 + k_i)$

Argument 2: more on the intrinsic prior

- With previous choices, the intrinsic prior can be written as

$$\pi_i(\boldsymbol{\beta}_0, \boldsymbol{\beta}_i, \sigma) = \sigma^{-1} \times \int_0^{\infty} N_{k_i}(\boldsymbol{\beta}_i \mid \mathbf{0}, \tilde{g}\, \rho_i\, \boldsymbol{\Xi}^{-1})\, p_i(\tilde{g})\, d\tilde{g}\,,$$

  where $\tilde{g} = g^*/\rho_i$ and $p_i(\tilde{g}) = (1/2)(\tilde{g})^{-3/2} 1_{\{\tilde{g}>1\}}$

- in the intrinsic prior approximation to the robust prior, ρ_i can be interpreted as simply a scale factor to the conditional covariance matrix

- previous suggestions related to ‘unit information priors’ scenarios of this type suggests the overall choice $\rho_i = 1/(k_0 + k_i)$ which is obviously very close to earlier suggested $1/(k_0 + k_i + 1)$.

VI. Extensions of Conventional Bayes Factors

Extending Conventional Bayes factors

JZS proposals directly apply to

- $\boldsymbol{X} = (\boldsymbol{X}_1,\ \boldsymbol{X}_e)$ is full rank
- reduced model defined by $\boldsymbol{\beta}_e = \mathbf{0}$
- $\boldsymbol{\beta}_e$ does not need to be orthogonal to $(\boldsymbol{\beta}_1, \sigma)$.

Extension to

- $\boldsymbol{X} = (\boldsymbol{X}_1,\ \boldsymbol{X}_e)$ not necessarily of full rank
- reduced model defined by $\boldsymbol{C}^T\boldsymbol{\beta} = \mathbf{0}$

is conceptually very simple $\rightsquigarrow$ reparameterize to problems with known solutions

$\boldsymbol{X}$ full (non full) rank $\rightsquigarrow$ regression (ANOVA) models

Regression models

- $\boldsymbol{X}$: $n \times k$ full rank
- to choose between

$$M_1 : f_1(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma) = \{N_n(\boldsymbol{y} \mid \boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}_n) : \ \boldsymbol{C}^{\,t}\boldsymbol{\beta} = \boldsymbol{0}\}$$
$$M_2 : f_2(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma) = N_n(\boldsymbol{y} \mid \boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}_n)$$

- Alternatively, to test:

$$H_1 : \boldsymbol{C}^{\,t}\boldsymbol{\beta} = \boldsymbol{0}, \quad \text{vs} \quad H_2 : \boldsymbol{C}^{\,t}\boldsymbol{\beta} \neq \boldsymbol{0}.$$

(w.l.o.g. $\boldsymbol{C}$: $k \times k_e$, $k_e \leq k$ can be assumed full rank)

reparameterizing to JZS situation

- Let $\boldsymbol{A} : k \times (k - k_e)$ be *any* matrix so that $\boldsymbol{R}^t = (\boldsymbol{A},\ \boldsymbol{C})$ non singular (and w.l.g. $|det\boldsymbol{R}| = 1$)
- partition $\boldsymbol{R}^{-1} : k \times k$ as $\boldsymbol{R}^{-1} = (\boldsymbol{S},\ \boldsymbol{T})$
  $(\boldsymbol{S} : k \times (k - k_e)$ and $\boldsymbol{T} : k \times k_e; \quad k_1 = k - k_e\ )$
- define $\boldsymbol{X}_e = \boldsymbol{X}\boldsymbol{T}$, $\boldsymbol{X}_1 = \boldsymbol{X}\boldsymbol{S}$
  $(\boldsymbol{X}_1 : n \times k_1$ and $\boldsymbol{X}_e : n \times k_e; \quad k_1 = k - k_e\ )$
- reparameterization:
  for M_1: $(\boldsymbol{\beta}_1, \sigma) = g_1(\boldsymbol{\beta}, \sigma) = (\boldsymbol{A}^t\boldsymbol{\beta}, \sigma)$
  for M_2: $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_e, \sigma) = g_2(\boldsymbol{\beta}, \sigma) = (\boldsymbol{A}^t\boldsymbol{\beta}, \boldsymbol{C}^t\boldsymbol{\beta}, \sigma)$

- In the original formulation there is no "$\boldsymbol{\beta}_e$" nor "$\boldsymbol{\beta}_1$", as in the covariable selection problem. Nevertheless, in the proposed reparameterization, $\boldsymbol{C}^t\boldsymbol{\beta}$ of dimension k_e, plays the role of $\boldsymbol{\beta}_e$ (the parameter of interest) and $\boldsymbol{A}^t\boldsymbol{\beta}$ of dimension $k - k_e$ plays the role of $\boldsymbol{\beta}_1$ (the nuisance or common parameter).
- Previous model selection problem equivalent to:

$$M_1^* : f_1^*(\boldsymbol{y} \mid \boldsymbol{\beta}_1, \sigma) = N_n(\boldsymbol{y} \mid \boldsymbol{X}_1\boldsymbol{\beta}_1, \sigma^2\boldsymbol{I}_n)$$
$$M_2^* : f_2^*(\boldsymbol{y} \mid \boldsymbol{\beta}_1, \boldsymbol{\beta}_e, \sigma) = N_n(\boldsymbol{y} \mid \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_e\boldsymbol{\beta}_e, \sigma^2\boldsymbol{I}_n).$$

Interestingly, the conventional Bayes factor can be shown not to depend on the arbitrary $\boldsymbol{A}$

ANOVA models

- $\tilde{\boldsymbol{X}}$: $n \times k$ is of rank r, with $r < k$
- to choose between

$$M_1 : f_1(\boldsymbol{y} \mid \tilde{\boldsymbol{\beta}}, \sigma) = \{N_n(\boldsymbol{y} \mid \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}, \sigma^2 \boldsymbol{I}_n) : \; \tilde{\boldsymbol{C}}^t \tilde{\boldsymbol{\beta}} = \boldsymbol{0}\}$$

$$M_2 : f_2(\boldsymbol{y} \mid \tilde{\boldsymbol{\beta}}, \sigma) = N_n(\boldsymbol{y} \mid \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}, \sigma^2 \boldsymbol{I}_n),$$

- Alternatively, to test:

$$H_1 : \tilde{\boldsymbol{C}}^t \tilde{\boldsymbol{\beta}} = \boldsymbol{0} \quad \text{vs} \quad H_2 : \tilde{\boldsymbol{C}}^t \tilde{\boldsymbol{\beta}} \neq \boldsymbol{0}$$

(w.l.o.g. $\tilde{\boldsymbol{C}}^t$: $k_e \times k$, $k_e \leq k$ can be assumed of rank k_e)

The problem here is overparameterized, but often full rank parameterization exists:

reparameterizing to full rank

Result: if hypothesis $\tilde{\boldsymbol{C}}^t\tilde{\boldsymbol{\beta}} = \boldsymbol{0}$ is *testable* $\rightsquigarrow$ a (non unique) full rank parameterization exists and the conventional Bayes factor is independent of the reparameterization

$\tilde{\boldsymbol{C}}^t\tilde{\boldsymbol{\beta}} = \boldsymbol{0}$ is a testable hypothesis if $\tilde{\boldsymbol{C}}^t\boldsymbol{G}\tilde{\boldsymbol{X}}^t\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{C}}^t$, where $\boldsymbol{G}$ is a generalized inverse of $\tilde{\boldsymbol{X}}^t\tilde{\boldsymbol{X}}$ (Rencher, 2000; Ravishanker and Dey, 2002))

$\tilde{\boldsymbol{C}}^t\tilde{\boldsymbol{\beta}} = \boldsymbol{0}$ is testable $\leftrightarrow \exists\ \boldsymbol{X}_{n\times r}$, $\boldsymbol{E}_{r\times k}$, $\boldsymbol{C}_{r\times k_e}$, s.t.:

(i) $\boldsymbol{X}$ and $\boldsymbol{E}$ of full rank r,

(ii) $\boldsymbol{X}\boldsymbol{E} = \tilde{\boldsymbol{X}}$, and

(iii) $\boldsymbol{C}^t\boldsymbol{E} = \tilde{\boldsymbol{C}}^t$.

For a testable null $\rightsquigarrow$ alternative full rank formulation

$$M_1^* : f_1^*(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma) = \{N_n(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n) : \boldsymbol{C}^t\boldsymbol{\beta} = \boldsymbol{0}\},$$

$$M_2^* : f_2^*(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma) = N_n(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n).$$

Note that M_1^* and M_2^* are full rank models.

Although from the straight derivation B_{21} seems to depend on (arbitrary) choice of $\boldsymbol{X}, \boldsymbol{C}, \boldsymbol{E}$, the conventional Bayes factor can be shown to be independent of the reparameterization

Conventional Bayes factors

For ALL previous situations, CBF can be expressed as

$$B_{21} = \int \left(1 + t\,n\,\frac{SSE_f}{SSE_r}\right)^{-(n-r+k_e)/2} (1 + t\,n)^{(n-r)/2}\, IGa(t \mid \frac{1}{2}, \frac{1}{2})\, dt$$

SSE_f and SSE_r are residual sums of squares for the *original* full (M_2) and restricted (M_1) models

If only BF are required, no need to explicitly reparameterize

Conventional BF easy to evaluate: numerically, by MC or by efficient Laplace approximation

Conventional Prior Distributions

- Goal $\rightsquigarrow$ explicitly derive priors producing previous conventional Bayes factors
- motivation $\rightsquigarrow$ judge adequacy of derived Bayes factors studying the corresponding priors (not always done in objective model selection)
- general procedure is simple: we know the conventional prior $\pi_i^*(\boldsymbol{\nu}_i)$ for the convenient reparameterization $f_i^*(\boldsymbol{y} \mid \boldsymbol{\nu}_i)$ of the original problem $f_i(\boldsymbol{y} \mid \boldsymbol{\theta}_i)$, with $\boldsymbol{\nu}_i = g_i(\boldsymbol{\theta}_i)$, for $i = 1, 2 \rightsquigarrow$ derive $\pi_i(\boldsymbol{\theta}_i)$ from $\pi_i^*(\boldsymbol{\nu}_i)$

- if g_i is 1-1 $\rightsquigarrow$ usual transformation rule:

  $$\pi_i(\boldsymbol{\theta}_i) = \pi_i^*(g_i(\boldsymbol{\theta}_i))\,\big|det\,\mathcal{J}_i(\boldsymbol{\theta}_i)\big|,$$

  where $\mathcal{J}_i$ is jacobian matrix of transformation g_i

- Not always the case, i.e. when $\dim(\boldsymbol{\nu}_i) < \dim(\boldsymbol{\theta}_i)$ (ANOVA) $\rightsquigarrow$ derive π_i such that the predictive distributions in both the original and reparameterized models are equal, that is:

  $$\pi_i(\boldsymbol{\theta}_i):\ \int f_i^*(\boldsymbol{y} \mid \boldsymbol{\nu}_i)\pi_i^*(\boldsymbol{\nu}_i)d\boldsymbol{\nu}_i = \int f_i(\boldsymbol{y} \mid \boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i$$

  also $\rightsquigarrow$ B_{12} not affected by reparameterization.

  Note: $\pi_i(\boldsymbol{\theta}_i)$ does not have to be unique.

- Interestingly $\leadsto$ conventional prior distributions (CPD) closely related to the *Partially Informative Normal (PIN) Distributions* (Ibrahim and Laud, 94; Sun, Tsutakawa and Speckman 99; Speckman and Sun (SS) 03)

- It is possible to generalize PIN's and use scale mixtures of resulting GPIN's to provide a unified, convenient way to derive CPD's (Bayarri and García-Donato, 2004).

- the term "Partially Informative" nicely reflects essence of CPD's, which typically have, in the *convenient* reparameterization, improper distributions for the 'common' parameters, and proper (conditional) distributions for the parameters not occurring in the restricted model.

- *Note:* for nulls of the form $C^t\beta = 0$, 'common parameters' might not be obviously recognized in the original parameterization

- We do not pursue the PIN connection here

Regression models

Recall that the (full rank) selection model problem:

$$M_1 : \boldsymbol{Y} \sim \{N_n(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n) : \ \boldsymbol{C}^t \boldsymbol{\beta} = \boldsymbol{0}\}$$

$$M_2 : \boldsymbol{Y} \sim N_n(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n),$$

can be reparameterized as:

$$M_1^* : \boldsymbol{Y} \sim N_n(\boldsymbol{y} \mid \ \boldsymbol{X}_1 \boldsymbol{\beta}_1, \sigma^2 \boldsymbol{I}_n)$$

$$M_2^* : \boldsymbol{Y} \sim N_n(\boldsymbol{y} \mid \boldsymbol{X}_1 \boldsymbol{\beta}_1 + \boldsymbol{X}_e \boldsymbol{\beta}_e, \sigma^2 \boldsymbol{I}_n)$$

with $\boldsymbol{A}$ arbitrary $(\boldsymbol{R}^t = (\boldsymbol{A}, \boldsymbol{C})$ non singular$)$ and

. for M_1: $(\boldsymbol{\beta}_1, \sigma) = g_1(\boldsymbol{\beta}, \sigma) = (\boldsymbol{A}^t \boldsymbol{\beta}, \sigma)$

. for M_2: $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_e, \sigma) = g_2(\boldsymbol{\beta}, \sigma) = (\boldsymbol{A}^t \boldsymbol{\beta}, \boldsymbol{C}^t \boldsymbol{\beta}, \sigma)$

Conventional Prior Distributions

The CPD's in the original parameterization are:

$$\pi_1(\boldsymbol{\beta}, \sigma) = \sigma^{-1} \, 1_{k_e}(\boldsymbol{C}^t\boldsymbol{\beta} = \mathbf{0}),$$

$$\pi_2(\boldsymbol{\beta}, \sigma) = \sigma^{-1} Ca_{k_e}(\boldsymbol{C}^t\boldsymbol{\beta} \mid \mathbf{0}, (\frac{\boldsymbol{V}^t\boldsymbol{V}}{n\sigma^2})^{-1})$$

where

$$\boldsymbol{V} = (\boldsymbol{I}_n - \boldsymbol{P}_1)\boldsymbol{X}_e, \qquad \boldsymbol{X}_1 = \boldsymbol{X}\boldsymbol{S}, \quad \boldsymbol{X}_e = \boldsymbol{X}\boldsymbol{T},$$

$(\boldsymbol{S} \, , \, \boldsymbol{T})$ the inverse of $(\boldsymbol{A} \, , \, \boldsymbol{C})$

Here, $\boldsymbol{C}^t\boldsymbol{\beta}$ (dimension k_e), plays the role of $\boldsymbol{\beta}_e$ (parameter of interest); $\boldsymbol{A}^t\boldsymbol{\beta}$ (dimension $k_1 = k - k_e$) the role of $\boldsymbol{\beta}_1$ (the nuisance or common parameter).

the CPD's distributions depend on the arbitrary matrix $\boldsymbol{A}$. However, the Conventional Bayes Factor does not, as shown previously

ANOVA models

Let $\tilde{\boldsymbol{X}} : n \times k$ of rank $r < k$ and $\tilde{\boldsymbol{C}}^t \tilde{\boldsymbol{\beta}} = \mathbf{0}$ testable. Then the model selection problem:

$$M_1 : \boldsymbol{Y} \sim \{N_n(\boldsymbol{y} \mid \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}, \sigma^2 \boldsymbol{I}_n) : \tilde{\boldsymbol{C}}^t \tilde{\boldsymbol{\beta}} = \mathbf{0}\}$$

$$M_2 : \boldsymbol{Y} \sim N_n(\boldsymbol{y} \mid \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}, \sigma^2 \boldsymbol{I}_n)$$

can be reparameterized as the full rank problem :

$$M_1^* : \boldsymbol{Y} \sim \{N_n(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n) : \boldsymbol{C}^t \boldsymbol{\beta} = \mathbf{0}\}$$

$$M_2^* : \boldsymbol{Y} \sim N_n(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n),$$

where

$\boldsymbol{X} : n \times r,\ \boldsymbol{E} : r \times k$ full rank,

$(\boldsymbol{\beta}, \sigma) = g(\tilde{\boldsymbol{\beta}}, \sigma) = (\boldsymbol{E}\tilde{\boldsymbol{\beta}}, \sigma)$

$\boldsymbol{E}$ is arbitrary ($\boldsymbol{X}\boldsymbol{E} = \tilde{\boldsymbol{X}}$, and $\boldsymbol{C}^t\boldsymbol{E} = \tilde{\boldsymbol{C}}^t$)

With previous matrices $\boldsymbol{X}, \boldsymbol{E}$ and $\boldsymbol{C}$,

- take $\boldsymbol{A}$ s.t. $\boldsymbol{R}^t = (\boldsymbol{A}\ ,\ \boldsymbol{C})$ non singular and $\boldsymbol{R}^{-1} = (\boldsymbol{S}\ ,\ \boldsymbol{T})$. Let $\boldsymbol{X}_e = \boldsymbol{X}\boldsymbol{T}$, $\boldsymbol{X}_1 = \boldsymbol{X}\boldsymbol{S}$
- let $\boldsymbol{Q}_2 : k \times (k - r)$ be any matrix such that $\boldsymbol{Q} = (\boldsymbol{E}^t\ ,\ \boldsymbol{Q}_2)$ is non singular.

Conventional Prior Distributions

Alternatively, we can write:

$$\pi_1(\tilde{\boldsymbol{\beta}}, \sigma) = \sigma^{-1} \, 1_{k_e}(\tilde{\boldsymbol{C}}^t \tilde{\boldsymbol{\beta}} = \mathbf{0}) \; h^1_{k-r}(\boldsymbol{Q}_2^t \, \tilde{\boldsymbol{\beta}})$$

$$\pi_2(\tilde{\boldsymbol{\beta}}, \sigma) = \sigma^{-1} \, h^2_{k-r}(\boldsymbol{Q}_2^t \tilde{\boldsymbol{\beta}}) \, Ca_{k_e}(\tilde{\boldsymbol{C}}^t \tilde{\boldsymbol{\beta}} \mid \mathbf{0}, \big(\frac{\boldsymbol{V}^t \boldsymbol{V}}{n\sigma^2}\big)^{-1})$$

Now $\tilde{\boldsymbol{C}}^t \tilde{\boldsymbol{\beta}}$ (dimension k_e), is the 'parameter of interest' while $\boldsymbol{Q}_2^t \tilde{\boldsymbol{\beta}}$ (dimension $k - r$) is that part of the nuisance parameters which overparameterizes the problem (in the proposed parameterization, the likelihood does not depend on $\boldsymbol{Q}_2^t \tilde{\boldsymbol{\beta}}$).
h^i_m is an arbitrary (proper) density in $\mathcal{R}^m$, $i = 1, 2$

Change point problem

For $\boldsymbol{\beta}_i$ of dimension k, and $\boldsymbol{X}_i$ of full rank, let

$$\boldsymbol{Y}_a = \boldsymbol{X}_a \boldsymbol{\beta}_a + \boldsymbol{\epsilon}_a, \qquad \boldsymbol{\epsilon}_a \sim N_{n_a}(\mathbf{0}, \sigma^2 \boldsymbol{I}_{n_a}),$$

$$\boldsymbol{Y}_b = \boldsymbol{X}_b \boldsymbol{\beta}_b + \boldsymbol{\epsilon}_b, \qquad \boldsymbol{\epsilon}_b \sim N_{n_b}(\mathbf{0}, \sigma^2 \boldsymbol{I}_{n_b}),$$

The *change-point problem* is the testing:

$$H_1 : \boldsymbol{\beta}_a = \boldsymbol{\beta}_b \quad \text{vs} \quad H_2 : \boldsymbol{\beta}_a \neq \boldsymbol{\beta}_b.$$

Frequentist solutions usually based on the F statistic (Chow's Test).

Moreno, Torres and Casella (2002) derived intrinsic priors (heteroscedaticity)

The change point problem can be expressed as the following model selection problem:

$$M_1 : \boldsymbol{Y} \sim \{N_n(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n) : \boldsymbol{C}^t \boldsymbol{\beta} = \boldsymbol{0}\}$$
$$M_2 : \boldsymbol{Y} \sim N_n(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n),$$

where

$$\boldsymbol{y}^t = (\boldsymbol{y}_a^t, \boldsymbol{y}_b^t), \qquad \boldsymbol{X} = \boldsymbol{X}_a \oplus \boldsymbol{X}_b : n \times 2k \ \text{ (full rank)}$$
$$\boldsymbol{\beta}^t = (\boldsymbol{\beta}_a^t, \boldsymbol{\beta}_b^t)), \qquad \boldsymbol{C}^t = (\boldsymbol{I}_k, \ -\boldsymbol{I}_k) : k \times 2k.$$

Conventional Bayes factor

given by the usual one-dimensional integral:

$$B_{21} = \int \left(1 + n\,t\,\frac{SSE_f}{SSE_r}\right)^{-(n-k)/2} (1+n\,t)^{(n-2k)/2}\, IGa(t \mid \frac{1}{2}, \frac{1}{2})\, dt,$$

where $SSE_f = \boldsymbol{y}^t(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t)\boldsymbol{y}$ and
$SSE_r = SSE_f + \boldsymbol{y}^t\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{C}(\boldsymbol{C}^t(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{C})^{-1}\boldsymbol{C}^t(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y}.$

Conventional prior

CPD's are: $\pi_1(\boldsymbol{\beta}_a, \boldsymbol{\beta}_b, \sigma) = \sigma^{-1} 1_k(\boldsymbol{\beta}_a = \boldsymbol{\beta}_b)$,

$\pi_2(\boldsymbol{\beta}_a, \boldsymbol{\beta}_b, \sigma) =$

$$\mathcal{K}\,\sigma^{-1}\left[1 + (\boldsymbol{\beta}_a - \boldsymbol{\beta}_b)^t \frac{((\boldsymbol{X}_a^t\boldsymbol{X}_a)^{-1} + (\boldsymbol{X}_b^t\boldsymbol{X}_b)^{-1}))^{-1}}{n\sigma^2}(\boldsymbol{\beta}_a - \boldsymbol{\beta}_b)\right]^{-(1+k)/2},$$

where the $\mathcal{K}$ is a Cauchy-type constant

These priors are nice and intuitive for MS:

- variances ('common' parameters) have same invariant non-informative prior under both models
- One of the regression coefficients, say $\boldsymbol{\beta}_a$, can be argued to be 'common' to both models $\rightsquigarrow$ CPD's are uniform under both models
- Conditional distribution of $\boldsymbol{\beta}_b$ given $(\boldsymbol{\beta}_a, \sigma)$ varies:
  - under the null model (no change point) $\rightsquigarrow$ degenerate on $\boldsymbol{\beta}_a = \boldsymbol{\beta}_b$
  - under the full model $\rightsquigarrow$ a Cauchy centered at $\boldsymbol{\beta}_a$ and with scale $n\sigma^2((\boldsymbol{X}_a^t\boldsymbol{X}_a)^{-1} + (\boldsymbol{X}_b^t\boldsymbol{X}_b)^{-1})$

- also, prior π_2 depends on the $\boldsymbol{\beta}$'s and the design matrices only through a quantity of the same functional form as the usual F-statistic:

$$F = (\hat{\boldsymbol{\beta}}_a - \hat{\boldsymbol{\beta}}_b)^t \ \frac{((\boldsymbol{X}_a^t \boldsymbol{X}_a)^{-1} + (\boldsymbol{X}_b^t \boldsymbol{X}_b)^{-1})^{-1}}{k\hat{\sigma_2^2}} \ (\hat{\boldsymbol{\beta}}_a - \hat{\boldsymbol{\beta}}_b),$$

VII - Divergence-based priors

General definition

For the problem

$$M_0 : f_0(\boldsymbol{y} \mid \boldsymbol{\alpha}), \quad M_1 : f_1(\boldsymbol{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}),$$

Bayarri and García-Donato (2008) proposed the Divergence-Based (DB) priors (a generalization of Jeffreys ideas):

$$\pi_1^D(\boldsymbol{\beta} \mid \boldsymbol{\alpha}) \propto g_q\Big(D(\boldsymbol{\beta}, \boldsymbol{\beta}_0, \boldsymbol{\alpha})\Big)\, \pi_1^N(\boldsymbol{\beta} \mid \boldsymbol{\alpha}), \quad \text{where}$$

- D is some 'distance' between f_1 and f_0,
- g_q is a real value decreasing function indexed by a parameter $q > 0$, and
- $\pi_1^N(\boldsymbol{\beta} \mid \boldsymbol{\alpha})$ is an objective estimation prior (possibly improper).

DB priors: recommended choices

This definition defines a vast family of prior distributions (depending on D, h_q and π_1^N)

Our specific recommendations:

- $D =$ symmetrized Kullback-Leibler divergence divided by n
- $g_q(x) = (1 + x)^{-q}$ (has polynomial tails),
- π_1^N the reference prior of Berger and Bernardo (1992),
- (partly our intuition)

$$q = \frac{1}{2} + \inf\{q > 0 \ : \ \pi_1^D() \text{ is proper}\}.$$

DB priors, the examples and the criteria

- For number of examples (included the previously shown), DB priors lead to proposals that meet the Desiderata
- no general results yet (work in progress), but partial results very promising