Including factors in variable selection problems: a model selection perspective.

Gonzalo García-Donato[1] and Rui Paulo[2]

[1] Universidad de Castilla-La Mancha (Spain), [2] Universidade de Lisboa (Portugal)

July 2019 - O'Bayes Conference

Variable selection from a model selection perspective

The effect of parameterization

Priors over $\mathcal{M}$

A frequentist comparison with other methods

A final remark

Variable selection from a model selection perspective

The effect of parameterization

Priors over $\mathcal{M}$

A frequentist comparison with other methods

A final remark

Variable selection

- Which (numerical) variables of a given set

$$\{x_1, x_2, \ldots, x_k\}$$

explain a response variable y.
Focus here is on linear models and Gaussian response y.

- The model that contains all variables (full model) is

$$M: \; y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2), \; i = 1, \ldots, n.$$

or

$$M: \; \boldsymbol{y} = \mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Estimation vs. model selection

There are two main classes of solutions to this problem:

- Estimation-based: fit M, obtain $\pi(\boldsymbol{\beta} \mid \boldsymbol{y})$ and decide based on this posterior which variables are non-important: the ones that have a 'small enough' β_j according to some criterion
- Model selection-based: consider all possible 2^k models indexed by which variables are important

$$M_{\boldsymbol{\gamma}} \equiv \boldsymbol{\gamma}, \quad \boldsymbol{\gamma} \in \{0,1\}^k,$$

and based on the discrete posterior

$$f(\boldsymbol{\gamma} \mid \boldsymbol{y}) \propto B_{\boldsymbol{\gamma}}(\boldsymbol{y})\ f(\boldsymbol{\gamma})$$

decide which variables are relevant.

Sparsity vs. multiplicity

Estimation-based methods

- Sparsity must be induced, done by priors $\pi(\boldsymbol{\beta}, \beta_0, \sigma)$ that have $\boldsymbol{\beta} = 0$ a highly distinguished value.
- Examples of such priors: (continuous) spike and slab, double Laplace (LASSO) or the popular shrinkage priors: horseshoe, horseshoe++, etc.
- Only one model is fitted

Sparsity vs. multiplicity

Model selection-based methods

- Embedded in a model uncertainty context
- Results are automatically sparse because Bayes factors are Occam's razor.
- Multiplicity issues: many models of the same complexity are compared; by chance one of them may be deemed relevant when in fact it's not
- Computation can be more involved but the results are richer: inclusion probabilities of variables, model averaging, etc.

Model selection-based variable selection

- The paradigm in this talk is model-selection based variable selection from an objective point of view
- Recall that $f(\gamma \mid \boldsymbol{y}) \propto B_{\gamma}(\boldsymbol{y})\ f(\gamma)$ so that the two main ingredients are Bayes factors and prior model probabilities
- Let's examine each separately

$f(\gamma)$ and multiplicity

- Scott and Berger (2010) showed that the standard Constant (C) prior

$$f(\gamma) = 1/2^k,$$

  does not provide multiplicity control (spurious variables easily appear as signals).
- On the other hand, the hierarchical prior (from now on, SB)

$$f(\gamma \mid \tau) = \tau^{k_\gamma}(1-\tau)^{k-k_\gamma}, \quad \tau \sim \mathrm{U}(\tau \mid 0, 1)$$

  where $k_\gamma = \mathbf{1}^\top \gamma$, does provide the required multiplicity control.
- Marginally, the SB prior is

$$f(\gamma) \propto 1/\{\#\text{ of models of dimension } k_\gamma\}.$$

B_γ and rank deficiency

- The Bayes factor is: $B_{\boldsymbol{\gamma}} = m_{\boldsymbol{\gamma}}(\boldsymbol{y})/m_0(\boldsymbol{y})$ with

$$m_\gamma(\boldsymbol{y}) = \int f_{\boldsymbol{\gamma}}(\boldsymbol{y} \mid \beta_0, \boldsymbol{\beta}_\gamma, \sigma)\ \pi_\gamma(\beta_0, \boldsymbol{\beta}_\gamma, \sigma)\ d\beta_0\ d\boldsymbol{\beta}_\gamma\ d\sigma$$

- There are many (objective) proposals for π_γ (independent: spike-and-slab; g-prior; Jeffreys-Zellner-Siow; etc).
- In this talk, we use the robust prior of Bayarri et al. (2012), but the results apply to any of the so-called conventional priors (Berger and Pericchi, 2001)
- For the conventional priors:

$$B_\gamma = \mathcal{B}\left(\frac{SSE_\gamma}{SSE_0}, 1, k_\gamma + 1, n\right),$$

where SSE_γ is the sum of squared errors and $\mathcal{B}$ involves a univariate integral.

B_γ and rank deficiency

- Key assumptions in the development are $n > k$ and $\boldsymbol{X}$ full-column rank

Key result

Bayarri and García-Donato (2007) showed that, if $\boldsymbol{X}_{\boldsymbol{\gamma}}$ is rank deficient, the formula above still applies replacing $k_\gamma + 1$ by the rank of the design matrix of $M_{\boldsymbol{\gamma}}$.

An important remark

- The full model is

$$M_1 : \ \boldsymbol{y} = \mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

  and all other models being entertained are obtained from this model by either
  - Setting at zero components of $\boldsymbol{\beta}$
  - Removing columns of $\boldsymbol{X}$ and the corresponding entries of $\boldsymbol{\beta}$
- This is the way the class of all models being compared is generated.

This talk: including factors in variable selection problems

Main goal

Address the Bayesian variable selection problem from a model selection perspective but now selecting among k numerical variables and p factors:

$$\{x_1, \ldots, x_k, \Lambda_1, \ldots, \Lambda_p\},$$

where Λ_j is a factor (categorical variable) with $\ell_j \geq 2$ levels.

Using dummy variables, the full model can be written as

$$\boldsymbol{y} = \mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{Z} = \boldsymbol{Z}_1 \oplus \cdots \oplus \boldsymbol{Z}_p$, $\boldsymbol{\beta}_j = (\beta_{1j}, \ldots, \beta_{\ell_j, j})^\top$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p)^\top$ [the symbol $\oplus$ represents direct sum of matrices].

Challenges in the model selection approach

- It is tempting to feed a variable selection package, say, `BayesVarSel`, the matrix $[\boldsymbol{X}|\boldsymbol{Z}]$ and forget about the fact that there are factors in the regression
- The matrix will not be full-rank, so typically `R` will choose a full-rank parametrization
- The class of models will be generated by removing columns from the full-rank design matrix
- Does the parametrization matter? (Section 2)
- How should we now think about the prior on the model space? (Section 3)

Another important remark

- A numerical variable affecting the response means that the corresponding regression coefficient is non-zero
- What do we mean when we say that a factor affects the response?
- What is natural here is to say that a factor affects the response if *at least* one its levels has associated a non-zero regression coefficient

Variable selection from a model selection perspective

The effect of parameterization

Priors over $\mathcal{M}$

A frequentist comparison with other methods

A final remark

One factor, no covariates: a revealing example

Suppose we have only one factor Λ with $\ell = 3$ levels in the set of explanatory variables. The full model is

$$M : y_i = \beta_0 + \sum_{j=1}^{3} z_{ij}\beta_j + \varepsilon_i \tag{1}$$

with $z_{ij} = 1$ if individual i belongs to the j-th level of the factor and 0 otherwise.

The design matrix is rank deficient, a situation that is normally handled using a full-column rank parameterization. For instance `R` does one for you (even if you do not want!) So let's do it!

One factor no covariates: a revealing case

- Suppose we use the popular treatment parameterization of M and choose the first level as the baseline, hence $\beta_1 = 0$. The model space we get is:

One factor no covariates: a revealing case

- Suppose we use the popular treatment parameterization of M and choose the first level as the baseline, hence $\beta_1 = 0$. The model space we get is:

Baseline level = 1
$M_{b=1}(0,0) : \mu_{ij} = \beta_0$
$M_{b=1}(0,1) : \mu_{i1} = \beta_0,\ \mu_{i2} = \beta_0,\ \mu_{i3} = \beta_0 + \beta_3$
$M_{b=1}(1,0) : \mu_{i1} = \beta_0,\ \mu_{i2} = \beta_0 + \beta_2,\ \mu_{i3} = \beta_0$
$M_{b=1}(1,1) : \mu_{i1} = \beta_0,\ \mu_{i2} = \beta_0 + \beta_2,\ \mu_{i3} = \beta_0 + \beta_3$

One factor no covariates: a revealing case

- Suppose we use the popular treatment parameterization of M and choose the first level as the baseline, hence $\beta_1 = 0$. The model space we get is:
- Now repeat the exercise but choosing the second level as baseline (right column)

Baseline level = 1	Baseline level = 2
$M_{b=1}(0,0) : \mu_{ij} = \beta_0$	$M_{b=2}(0,0) : \mu_{ij} = \beta_0$
$M_{b=1}(0,1) : \mu_{i1} = \beta_0,\ \mu_{i2} = \beta_0,\ \mu_{i3} = \beta_0 + \beta_3$	$M_{b=2}(0,1) : \mu_{i1} = \beta_0,\ \mu_{i2} = \beta_0,\ \mu_{i3} = \beta_0 + \beta_3$
$M_{b=1}(1,0) : \mu_{i1} = \beta_0,\ \mu_{i2} = \beta_0 + \beta_2,\ \mu_{i3} = \beta_0$	$M_{b=2}(1,0) : \mu_{i1} = \beta_0 + \beta_1,\ \mu_{i2} = \beta_0,\ \mu_{i3} = \beta_0$
$M_{b=1}(1,1) : \mu_{i1} = \beta_0,\ \mu_{i2} = \beta_0 + \beta_2,\ \mu_{i3} = \beta_0 + \beta_3$	$M_{b=2}(1,1) : \mu_{i1} = \beta_0 + \beta_1,\ \mu_{i2} = \beta_0,\ \mu_{i3} = \beta_0 + \beta_3$

$M_{b=2}(1,0)$ is not in the first model space.

One factor no covariates: a revealing case

- Suppose we use the popular treatment parameterization of M and choose the first level as the baseline, hence $\beta_1 = 0$. The model space we get is:
- Now repeat the exercise but choosing the second level as baseline (right column)

Baseline level = 1	Baseline level = 2
$M_{b=1}(0,0) : \mu_{ij} = \beta_0$	$M_{b=2}(0,0) : \mu_{ij} = \beta_0$
$M_{b=1}(0,1) : \mu_{i1} = \beta_0,\ \mu_{i2} = \beta_0,\ \mu_{i3} = \beta_0 + \beta_3$	$M_{b=2}(0,1) : \mu_{i1} = \beta_0,\ \mu_{i2} = \beta_0,\ \mu_{i3} = \beta_0 + \beta_3$
$M_{b=1}(1,0) : \mu_{i1} = \beta_0,\ \mu_{i2} = \beta_0 + \beta_2,\ \mu_{i3} = \beta_0$	$M_{b=2}(1,0) : \mu_{i1} = \beta_0 + \beta_1,\ \mu_{i2} = \beta_0,\ \mu_{i3} = \beta_0$
$M_{b=1}(1,1) : \mu_{i1} = \beta_0,\ \mu_{i2} = \beta_0 + \beta_2,\ \mu_{i3} = \beta_0 + \beta_3$	$M_{b=2}(1,1) : \mu_{i1} = \beta_0 + \beta_1,\ \mu_{i2} = \beta_0,\ \mu_{i3} = \beta_0 + \beta_3$

$M_{b=2}(1,0)$ is not in the first model space.

If $M_{b=2}(1,0)$ is the true model, it cannot be chosen in the first parameterization (independently of the sample size!) simply because it is not in the model space.

The impact of parameterization in a real case

- A child obesity study in Spain; the main goal was to identify the relevance (if any) of the practice of sports (the factor) on the body mass index y of $n = 1002$ children. This factor has $\ell = 6$ levels coded ranging from no practice to very intense practice of sports

The impact of parameterization in a real case

- A child obesity study in Spain; the main goal was to identify the relevance (if any) of the practice of sports (the factor) on the body mass index y of $n = 1002$ children. This factor has $\ell = 6$ levels coded ranging from no practice to very intense practice of sports
- If the first level (no practice of sports) is used as baseline, the posterior probability of the factor is 0.26 while if the second level is used as baseline this probability increases up to 0.98 (a remarkable change).
- The model with the first level of the factor present and the others not (modeling the two situations of sendentary vs. not sedentary lifestyle) is quite a good model that simply disappears if this parameterization is used.

A possible solution with unexpected consequences

Full-column rank formulations does not allow us to span the whole class of models and the set of models that we miss is parameterization-dependent.

• Solution: not reparameterizing; using the original (rank deficient) matrix and the model space we obtain as we delete columns from it. We formally obtain 8 models.

Original (Rank deficient) parameterization	Rank
$M(0,0,0): \mu_{ij} = \beta_0$	1
$M(0,0,1): \mu_{i1} = \beta_0,\ \mu_{i2} = \beta_0,\ \mu_{i3} = \beta_0 + \beta_3$	2
$M(0,1,0): \mu_{i1} = \beta_0,\ \mu_{i2} = \beta_0 + \beta_2,\ \mu_{i3} = \beta_0$	2
$M(1,0,0): \mu_{i1} = \beta_0 + \beta_1,\ \mu_{i2} = \beta_0,\ \mu_{i3} = \beta_0$	2
$M(1,1,0): \mu_{i1} = \beta_0 + \beta_1,\ \mu_{i2} = \beta_0 + \beta_2,\ \mu_{i3} = \beta_0$	3
$M(1,0,1): \mu_{i1} = \beta_0 + \beta_1,\ \mu_{i2} = \beta_0,\ \mu_{i3} = \beta_0 + \beta_3$	3
$M(0,1,1): \mu_{i1} = \beta_0,\ \mu_{i2} = \beta_0 + \beta_2,\ \mu_{i3} = \beta_0 + \beta_3$	3
$M(1,1,1): \mu_{i1} = \beta_0 + \beta_1,\ \mu_{i2} = \beta_0 + \beta_2,\ \mu_{i3} = \beta_0 + \beta_3$	3

• No model is now missed,. . .

A possible solution with unexpected consequences

Full-column rank formulations does not allow us to span the whole class of models and the set of models that we miss is parameterization-dependent.

• Solution: not reparameterizing; using the original (rank deficient) matrix and the model space we obtain as we delete columns from it. We formally obtain 8 models.

Original (Rank deficient) parameterization	Rank
$M(0,0,0): \mu_{ij} = \beta_0$	1
$M(0,0,1): \mu_{i1} = \beta_0,\ \mu_{i2} = \beta_0,\ \mu_{i3} = \beta_0 + \beta_3$	2
$M(0,1,0): \mu_{i1} = \beta_0,\ \mu_{i2} = \beta_0 + \beta_2,\ \mu_{i3} = \beta_0$	2
$M(1,0,0): \mu_{i1} = \beta_0 + \beta_1,\ \mu_{i2} = \beta_0,\ \mu_{i3} = \beta_0$	2
$M(1,1,0): \mu_{i1} = \beta_0 + \beta_1,\ \mu_{i2} = \beta_0 + \beta_2,\ \mu_{i3} = \beta_0$	3
$M(1,0,1): \mu_{i1} = \beta_0 + \beta_1,\ \mu_{i2} = \beta_0,\ \mu_{i3} = \beta_0 + \beta_3$	3
$M(0,1,1): \mu_{i1} = \beta_0,\ \mu_{i2} = \beta_0 + \beta_2,\ \mu_{i3} = \beta_0 + \beta_3$	3
$M(1,1,1): \mu_{i1} = \beta_0 + \beta_1,\ \mu_{i2} = \beta_0 + \beta_2,\ \mu_{i3} = \beta_0 + \beta_3$	3

• No model is now missed,. . . but one of them is cloned three times!

A solution to the problem

Solution to the cloning issue

Kill the clones! (and collect in, say, $\mathcal{M}$ only non-repeated models). It does not matter which one you keep, as the Bayes factor is invariant to this selection.

- Result: The model space with only unique models, $\mathcal{M}$, has

$$\#\mathcal{M} = 2^k \times \prod_{j=1}^{p} [2^{\ell_j} - \ell_j].$$

• Choosing a full rank parameterization, the number of models that disappear is:

$$2^k \left(\prod_{j=1}^{p} (2^{\ell_j} - \ell_j) - \prod_{j=1}^{p} 2^{\ell_j - 1} \right).$$

- If $k = 0$ and $\ell = 3$ then $\#\mathcal{M} = 5$ and only one model would has disappeared in a full rank parameterization...
- If $k = 3$ and $p = 3$ with $\ell_1 = 3$, $\ell_2 = 4$ and $\ell_3 = 4$ we have $\#\mathcal{M} = 5760$, of which by choosing a full rank parameterization you make 3712 disappear (more than 60%!).
- If $\ell_j = 2$ for all j, no models are lost — no problem with binary factors

Variable selection from a model selection perspective

The effect of parameterization

Priors over $\mathcal{M}$

A frequentist comparison with other methods

A final remark

Notation for models in $\mathcal{M}$

Recall:

$$\{x_1, \ldots, x_k, \Lambda_1, \ldots, \Lambda_p\}.$$

Models in $\mathcal{M}$ can be identified by a parameter vector $M(\boldsymbol{\gamma}, \boldsymbol{\delta}) \equiv (\boldsymbol{\gamma}, \boldsymbol{\delta}) \in \{0,1\}^{k+\sum \ell_j}$, where

$$\boldsymbol{\gamma}^\top = (\gamma_1, \ldots, \gamma_k), \quad \boldsymbol{\delta}^\top = (\boldsymbol{\delta}_1^\top, \ldots, \boldsymbol{\delta}_p^\top), \quad \boldsymbol{\delta}_j^\top = (\delta_{j1}, \ldots, \delta_{j\ell_j}),$$

and $\gamma_i = 1$ if x_i is in the model (0 otherwise) and $\delta_{jh} = 1$ if level h of factor Λ_j is in the model (0 otherwise).

- This section is about the objective assignment of

$$f(\boldsymbol{\gamma}, \boldsymbol{\delta}), \quad (\boldsymbol{\gamma}, \boldsymbol{\delta}) \in \mathcal{M}.$$

Priors over $\mathcal{M}$: default extensions

The natural extensions of standard objective priors in variable selection are

- using a constant prior over the whole model space:

$$f(\boldsymbol{\gamma}, \boldsymbol{\delta}) = \left(\#\mathcal{M}\right)^{-1} \text{ if } (\boldsymbol{\gamma}, \boldsymbol{\delta}) \in \mathcal{M}, \qquad \text{(C)}$$

  or

- a prior that apportions the probability inversely proportional to the number of models of the same rank:

$$f(\boldsymbol{\gamma}, \boldsymbol{\delta}) \propto \Big(\#\text{models of same rank as } M(\boldsymbol{\gamma}, \boldsymbol{\delta})\Big)^{-1} \text{ if } (\boldsymbol{\gamma}, \boldsymbol{\delta}) \in \mathcal{M}. \qquad \text{(SB)}$$

Unsatisfactory properties

Both depend on the number of levels of the factors.
For instance, in a problem with just one factor with four levels, the inclusion prior probability of the factor is 0.92 in the case of (C) and 0.75 with (SB).

Priors over $\mathcal{M}$: two steps approach

To define more satisfactory priors, consider the following function of $\boldsymbol{\delta}$

$$\boldsymbol{\tau}^\top = (\tau_1, \ldots, \tau_p) = (h_1(\boldsymbol{\delta}), \ldots, h_p(\boldsymbol{\delta}))$$

where $\tau_j = h_j(\boldsymbol{\delta}) = 1$ if Λ_j is active (0 otherwise).

Priors over $\mathcal{M}$: two steps approach

To define more satisfactory priors, consider the following function of $\boldsymbol{\delta}$

$$\boldsymbol{\tau}^\top = (\tau_1, \ldots, \tau_p) = (h_1(\boldsymbol{\delta}), \ldots, h_p(\boldsymbol{\delta}))$$

where $\tau_j = h_j(\boldsymbol{\delta}) = 1$ if Λ_j is active (0 otherwise).

$f(\boldsymbol{\gamma}, \boldsymbol{\delta})$ can be decomposed as

$$f(\boldsymbol{\gamma}, \boldsymbol{\delta}) = f(\boldsymbol{\gamma}, \boldsymbol{\tau}) \times f(\boldsymbol{\delta} \mid \boldsymbol{\gamma}, \boldsymbol{\tau}),$$

where:

Priors over $\mathcal{M}$: two steps approach

To define more satisfactory priors, consider the following function of $\boldsymbol{\delta}$

$$\boldsymbol{\tau}^\top = (\tau_1, \ldots, \tau_p) = (h_1(\boldsymbol{\delta}), \ldots, h_p(\boldsymbol{\delta}))$$

where $\tau_j = h_j(\boldsymbol{\delta}) = 1$ if Λ_j is active (0 otherwise).

$f(\boldsymbol{\gamma}, \boldsymbol{\delta})$ can be decomposed as

$$f(\boldsymbol{\gamma}, \boldsymbol{\delta}) = f(\boldsymbol{\gamma}, \boldsymbol{\tau}) \times f(\boldsymbol{\delta} \mid \boldsymbol{\gamma}, \boldsymbol{\tau}),$$

where:

• The marginal for the parameters that indicate which covariates and factors are active:

$$f(\boldsymbol{\gamma}, \boldsymbol{\tau}).$$

Priors over $\mathcal{M}$: two steps approach

To define more satisfactory priors, consider the following function of $\boldsymbol{\delta}$

$$\boldsymbol{\tau}^\top = (\tau_1, \ldots, \tau_p) = (h_1(\boldsymbol{\delta}), \ldots, h_p(\boldsymbol{\delta}))$$

where $\tau_j = h_j(\boldsymbol{\delta}) = 1$ if Λ_j is active (0 otherwise).

$f(\boldsymbol{\gamma}, \boldsymbol{\delta})$ can be decomposed as

$$f(\boldsymbol{\gamma}, \boldsymbol{\delta}) = f(\boldsymbol{\gamma}, \boldsymbol{\tau}) \times f(\boldsymbol{\delta} \mid \boldsymbol{\gamma}, \boldsymbol{\tau}),$$

where:

• The marginal for the parameters that indicate which covariates and factors are active:

$$f(\boldsymbol{\gamma}, \boldsymbol{\tau}).$$

• The prior probabilities over the set of models that have a given set of active factors and variables, $\mathcal{M}(\boldsymbol{\gamma}, \boldsymbol{\tau})$:

$$f(\boldsymbol{\delta} \mid \boldsymbol{\gamma}, \boldsymbol{\tau})$$

We can think on four possible combinations:

$f(\gamma, \delta)$	$f(\gamma, \tau)$	$f(\delta \mid \gamma, \tau)$
CC	C	C
SBC	SB	C
CSB	C	SB
SBSB	SB	SB

By construction, within all these proposals it holds that $p(\tau_j = 1) = p(\gamma_i = 1) = 1/2$

Discriminating among the four possibilities

Through a series of simulation experiments we conclude that the winner is SBSB:

- Having SB in $f(\gamma, \tau)$ controls for multiplicity (if p increases the chance of declaring signals factors that are spurious is controled).
- The reason for using SB in $f(\delta \mid \gamma, \tau)$ is more subtle and does not have to do with multiplicity (since large ℓ does not affect the appearance of spurious signals).

Simulation experiments

- $n = 300$ observations and p factors all with the same number of levels $\ell_j = 4$.
- $\boldsymbol{\beta}_1^\top = (-1.08, -0.84, -0.74, 0.63)$,
  $\boldsymbol{\beta}_2^\top = (-0.51, 0.41, 0.18, 0.07)$.
- The rest of the $p - 2$ factors are spurious and hence $\boldsymbol{\beta}_j = \mathbf{0}$ for $3 \leq j \leq p$
- $p \in \{5, 15, 27\}$ repeating the experiment 10 times
- Computed inclusion probabilitites of the factors
- Active are in red; spurious are in green

Const-Const
SB-Const
SB-SB
Incl. Probab.
0.0
0.2
0.4
0.6
0.8
1.0
p=5
p=15
p=27

SBSB vs SBC

- $n = 300$ observations and $p = 4$ factors all with the same number of levels $\ell_j = \ell$.
- The first factor is active with only its first level having a coefficient different from zero:

$$\boldsymbol{\beta}_1^\top = (\beta_{11}, \mathbf{0}^\top),$$

- all the other vectors of regression coefficients are null
- $\ell \in \{5, 15\}$
- $\beta_{11} \in \{0.9, 1.0\}$

	(SBSB)		(SBC)	
$\beta_{11} = 1.0$	$\ell = 5$	$\ell = 15$	$\ell = 5$	$\ell = 15$
$p(\tau_1 = 1 \mid \boldsymbol{y})$	1.0000	1.0000	1.0000	1.0000
$p(\tau_2 = 1 \mid \boldsymbol{y})$	0.0442	0.1452	0.0305	0.0018
$p(\tau_3 = 1 \mid \boldsymbol{y})$	0.0830	0.1754	0.0664	0.0038
$p(\tau_4 = 1 \mid \boldsymbol{y})$	0.0355	0.1272	0.0233	0.0020
$\beta_{11} = 0.9$				
$p(\tau_1 = 1 \mid \boldsymbol{y})$	0.8546	0.9812	0.8703	0.6434
$p(\tau_2 = 1 \mid \boldsymbol{y})$	0.0446	0.1689	0.0322	0.0032
$p(\tau_3 = 1 \mid \boldsymbol{y})$	0.0851	0.2037	0.0719	0.0047
$p(\tau_4 = 1 \mid \boldsymbol{y})$	0.0361	0.1528	0.0238	0.0033

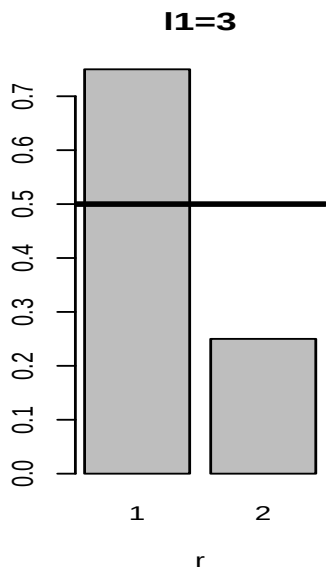
The inclusion probabilities of factors and covariates are a function of

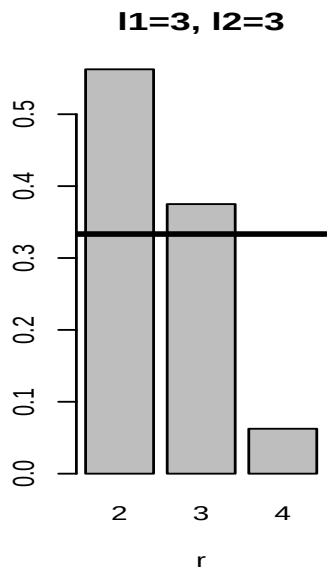
$$f(\boldsymbol{\gamma}, \boldsymbol{\tau} \mid \boldsymbol{y}) \propto \mathcal{B}_{(\gamma,\tau)}\, f(\boldsymbol{\gamma}, \boldsymbol{\tau}),$$

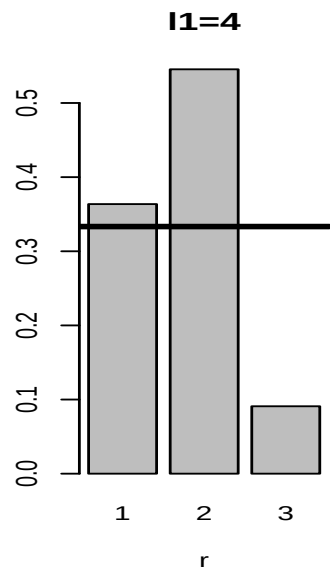
where

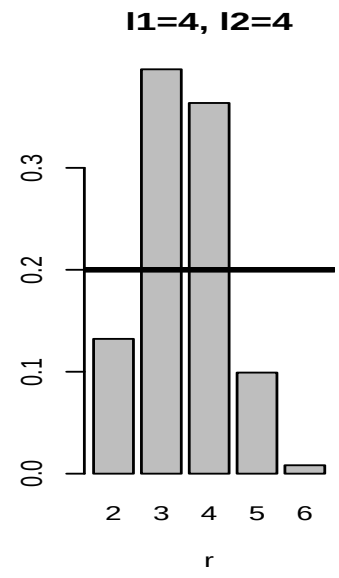
$$\mathcal{B}_{(\gamma,\tau)} = \sum_{r \in R(\gamma,\tau)} \overline{B_{\substack{(\gamma,\delta)\in\mathcal{M}(\gamma,\tau) \\ r(\gamma,\delta)=r}}} \times\; p(\kappa(\boldsymbol{\gamma}, \boldsymbol{\delta}) = r \mid (\boldsymbol{\gamma}, \boldsymbol{\delta}) \in \mathcal{M}(\boldsymbol{\gamma}, \boldsymbol{\tau})).$$

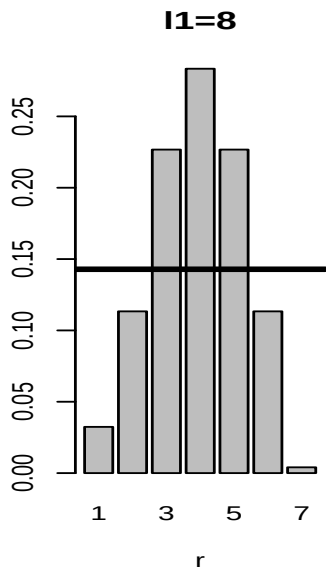
Which of $\mathcal{B}^{C}_{(\gamma,\tau)}$ or $\mathcal{B}^{SB}_{(\gamma,\tau)}$ behaves more like a sensible objective Bayes factor?

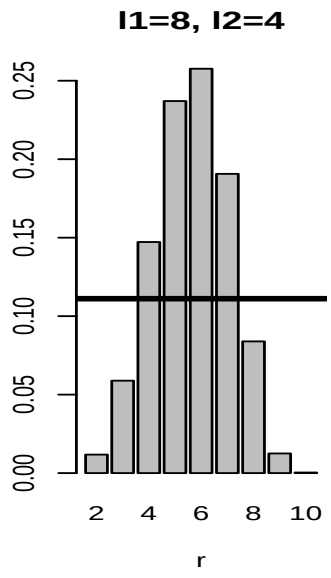

l1=3
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
1
2
r

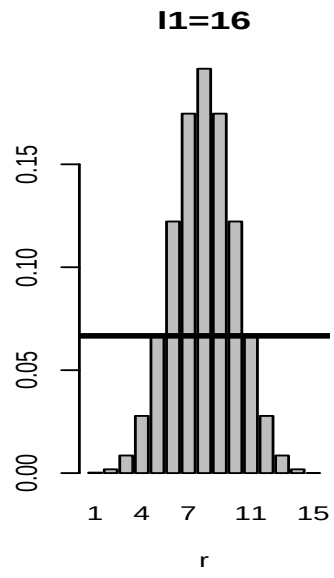

l1=3, l2=3
0.0
0.1
0.2
0.3
0.4
0.5
2
3
4
r

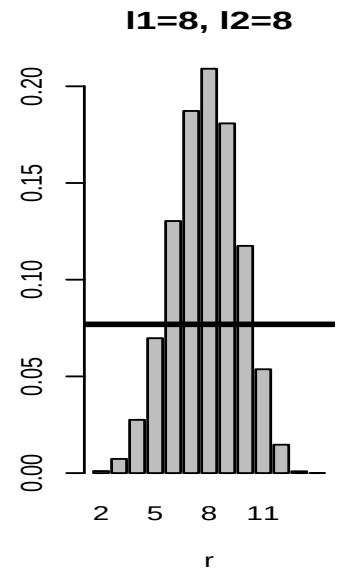

l1=4
0.0
0.1
0.2
0.3
0.4
0.5
1
2
3
r


l1=4, l2=4
0.0
0.1
0.2
0.3
2
3
4
5
6
r


l1=8
0.00
0.05
0.10
0.15
0.20
0.25
1
3
5
7
r


l1=8, l2=4
0.00
0.05
0.10
0.15
0.20
0.25
2
4
6
8
10
r


l1=16
0.00
0.05
0.10
0.15
1
4
7
11
15
r


l1=8, l2=8
0.00
0.05
0.10
0.15
0.20
2
5
8
11
r

About $f(\boldsymbol{\delta} \mid \boldsymbol{\gamma}, \boldsymbol{\tau})$

Roughly speaking:

- $\mathcal{B}^{C}_{\gamma,\tau}$ summarizes the evidence in favour of $\mathcal{M}(\boldsymbol{\gamma}, \boldsymbol{\tau})$ using models of medium size being, in general, quite conservative.
- $\mathcal{B}^{SB}_{\gamma,\tau}$ summarizes that evidence averaging the averages over all possible ranks. It is hence very robust and therefore it's our recommended choice.
- Not very important if ℓ is small

Variable selection from a model selection perspective

The effect of parameterization

Priors over $\mathcal{M}$

A frequentist comparison with other methods

A final remark

The other methods

We compared the frequentist performance of *our* methodology with several other methods

- BSGS: Bayesian Sparse Group Selection — Chen et al. (2016); `R` package `BSGS` by Lee and Chen (2015).
- eF: Effect Fusion — Pauger and Wagner (2017); `R` package `effectFusion` (Pauger et al. 2019).
- grLasso: Group Lasso (L2 penalty) — Yian and Lin (2006); `R` package `grpreg` (Breheny, 2019).
- grMCP: Group Lasso (minimax concave penalty) — Breheny and Huang (2009); `R` package `grpreg`.
- grSCAD: Group Lasso (smoothly clipped absolute deviation penalty) — Fan and Li (2001); `R` package `grpreg`.

Experiment

Similarly to the experiment in Pauger and Wagner (2017), we generated 100 datasets with a model with $n = 500$ observations and $p = 4$ factors:

- Λ_1 has $\ell_1 = 8$ levels with:

$$\boldsymbol{\beta}_1^\top = (0, 0, 1, 1, 1, 1, -2, -2),$$

- Λ_3 has $\ell_3 = 4$ levels

$$\boldsymbol{\beta}_3^\top = (0, 0, 2, 2),$$

- Λ_2 and Λ_4 are inert (with $\ell_2 = 8$ and $\ell_4 = 4$).

We counted the percentage of times the factors were truly declared as false (true negatives) or true (true positives).

	$\Lambda_1(\neq 0)$	$\Lambda_2(= 0)$	$\Lambda_3(\neq 0)$	$\Lambda_4(= 0)$
Bayesian (ours)	100	99	100	100
BSGS	100	98	100	74
eF	88	12	100	94
grLasso	100	13	100	21
grMCP	100	89	100	89
grSCAD	100	83	100	80

A final remark

- The class of models that we entertain to answer the question: does the factor affect the response? assume that the effect of each level, if non-zero, are different and unrelated
- It is therefore possible to imagine ways of a factor affecting the response in ways that are not captured by our class of models
- We can construct contrasts that cannot be tested in this framework
- A problem that is often considered is whether the effect of certain levels is exactly the same, which means we can fuse those levels
- Will this greatly affect our ability to detect an influential factor?

- For each value of $\beta_{11} \in \{0.20, 0.50, 0.75, 1.00, 1.25, 1.50\}$
- 10 simulated datasets from the model with no covariates ($k = 0$), only one factor $p = 1$ with effect:

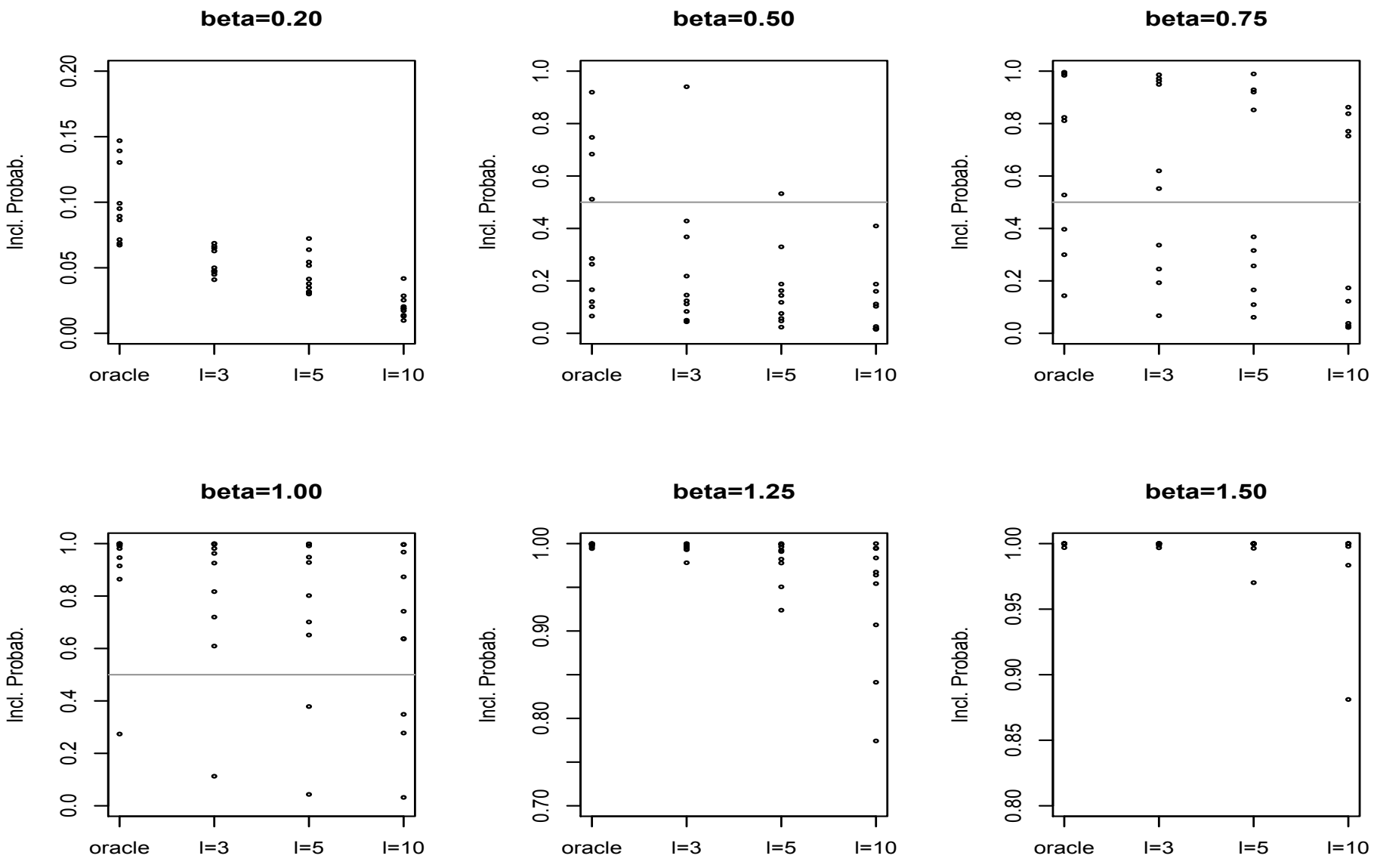
$$\boldsymbol{\beta}^\top = (\beta_{11}, \ldots, \beta_{11}, 0) \ .$$

[$\boldsymbol{\beta}$ has the first $\ell - 1$ components equal to β_{11} and the last is zero.]

- $n = 100$; 20 subjects assigned to the last level.
- The true model can be simply expressed as:

$$\begin{aligned} y_{i1} &= \beta_0 + \beta_{11} + \epsilon_{i1}, \quad \text{for} \quad i = 1, \ldots, 80, \\ y_{i2} &= \beta_0 + \epsilon_{i2}, \quad \text{for} \quad i = 81, \ldots, 100. \end{aligned}$$

We have considered three values of the number of levels, namely, $\ell \in \{3, 5, 10\}$.


beta=0.20
beta=0.50
beta=0.75
beta=1.00
beta=1.25
beta=1.50
Incl. Probab.
oracle
l=3
l=5
l=10

Summary

- When a linear regression model includes factors in the list of possible regressors, typically a full-rank parametrization is assumed
- The class of models generated by deleting columns of the full rank matrix will not be exhaustive; the results will be parametrization-dependent
- The solution is not to reparametrize and consider the class of unique models obtained with the rank deficient matrix
- The prior on the model space should be constructed hierarchically, and in each hierarchy the SB prior should be used
- This ensures multiplicity control on the number of predictors and a sensible summary of the evidence in favor of any combination of active variables and factors
- Fully automatic procedure

Thanks!