Objective priors and search strategies in large variable selection problems

Gonzalo Garcia-Donato

Universidad de Castilla-La Mancha (Spain)

Zurich - February 2011
PART I: 
The problem

PART II: 
Objective priors

PART III: 
Inferences in large problems
PART I:
The problem of variable selection and initial considerations
The variable selection problem
Part I: The problem

The variable selection problem

\[ M_1: Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_p X_p + \varepsilon \]
The variable selection problem

\[ M_1: Y = \alpha + \beta_3 X_3 + \beta_5 X_5 + \beta_p X_p + \varepsilon \]

\[ M_2: Y = \alpha + \beta_4 X_4 + \beta_2 X_2 + \varepsilon \]
The variable selection problem

\[ M_1: Y = \alpha + \beta_3 X_3 + \beta_5 X_5 + \varepsilon \]

\[ M_2: Y = \alpha + \beta_4 X_4 + \beta_2 X_2 + \varepsilon \]

\[ M_3: Y = \alpha + \beta_3 X_3 + \beta_6 X_6 + \varepsilon \]
Part I: The problem

The variable selection problem

\[
\begin{align*}
M_1: & \quad Y = \alpha + \beta_3 X_3 + \beta_5 X_5 + \beta_p X_p + \varepsilon \\
M_2: & \quad Y = \alpha + \beta_4 X_4 + \beta_2 X_2 + \varepsilon \\
M_3: & \quad Y = \alpha + \beta_1 X_1 + \beta_6 X_6 + \varepsilon \\
M_4: & \quad Y = \alpha + \beta_1 X_1 + \varepsilon
\end{align*}
\]
Part I: The problem

The variable selection problem

There is uncertainty about which model provides a better explanation for a dependent variable $Y$. 

$$
\begin{align*}
M_1: & \quad Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_p X_p + \epsilon \\
M_2: & \quad Y = \alpha + \beta X_1 + \beta_2 X_2 + \epsilon \\
M_3: & \quad Y = \alpha + \beta_3 X_3 + \beta_5 X_5 + \epsilon \\
M_4: & \quad Y = \alpha + \beta_1 X_1 + \epsilon \\
\{X_1\} & \\
\{X_3, X_5\} & \\
\{X_1, X_2\} & \\
\{X_3, X_5, X_6\} & \\
\end{align*}
$$
There is uncertainty about which model provides a better explanation for a dependent variable $Y$. This uncertainty is explicitly considered in the analysis.
The variable selection problem

There is uncertainty about which model provides a better explanation for a dependent variable $Y$. This uncertainty is explicitly considered in the analysis.
Part I: The problem

The variable selection problem

We have $2^p$ candidate models:

$$M_i(y \mid \alpha, \beta_i, \sigma) : Y = 1\alpha + X_i\beta_i + \epsilon, \quad \epsilon \sim N_n(0, \sigma^2 I_n),$$

where

- $X_i$ are $n \times k_i$ design matrices (corresponding to $k_i$ of the $p$ initially considered), $i = 0, 1, 2, \ldots, 2^p - 1$,
- (by default) the simplest model $M_0$ (null model) only contains the intercept.
- $\mathcal{M}$ represents the model space.
The variable selection problem

We have $2^p$ candidate models:

$$M_i(y \mid \alpha, \beta_i, \sigma) : Y = 1\alpha + X_i\beta_i + \epsilon, \ \epsilon \sim N_n(0, \sigma^2 I_n),$$

where

- $X_i$ are $n \times k_i$ design matrices (corresponding to $k_i$ of the $p$ initially considered), $i = 0, 1, 2, \ldots, 2^p - 1$,
- (by default) the simplest model $M_0$ (null model) only contains the intercept.
- $\mathcal{M}$ represents the model space.
- $(\alpha, \sigma)$ are common (but not same) parameters,
The variable selection problem

We have $2^p$ candidate models:

$$M_i(y \mid \alpha, \beta_i, \sigma) : Y = 1\alpha + X_i\beta_i + \epsilon, \; \epsilon \sim N_n(0, \sigma^2 I_n),$$

where

- $X_i$ are $n \times k_i$ design matrices (corresponding to $k_i$ of the $p$ initially considered), $i = 0, 1, 2, \ldots, 2^p - 1$,
- (by default) the simplest model $M_0$ (null model) only contains the intercept.
- $\mathcal{M}$ represents the model space.
- $(\alpha, \sigma)$ are common (but not same) parameters,
- $\beta_i$ is a new $k_i$-dimensional parameter.
Posterior probabilities and Bayes factors

\[
Pr(M_j \mid y) = \frac{Pr(M_j)}{\sum_{i=0}^{2^p} B_{ij} Pr(M_i)}, \quad B_{ij} = \frac{m_i(y)}{m_j(y)},
\]

where \( m_i(y) \) are the integrated likelihoods (or prior predictive marginals):
Posterior probabilities and Bayes factors

\[ Pr(M_j \mid y) = \frac{Pr(M_j)}{\sum_{i=0}^{2^p} B_{ij} Pr(M_i)}, \quad B_{ij} = \frac{m_i(y)}{m_j(y)}, \]

where \( m_i(y) \) are the integrated likelihoods (or prior predictive marginals):

\[ m_i(y) = \int M_i(y \mid \alpha, \beta_i, \sigma) \pi_i(\alpha, \beta_i, \sigma) d\alpha d\beta_i d\sigma. \]
Posterior probabilities and Bayes factors

\[ Pr(M_j | y) = \frac{Pr(M_j)}{\sum_{i=0}^{2p} B_{ij} Pr(M_i)}, \quad B_{ij} = \frac{m_i(y)}{m_j(y)} , \]

where \( m_i(y) \) are the integrated likelihoods (or prior predictive marginals):

\[ m_i(y) = \int M_i(y | \alpha, \beta_i, \sigma) \pi_i(\alpha, \beta_i, \sigma) d\alpha d\beta_i d\sigma. \]

Above,

<table>
<thead>
<tr>
<th>( Pr(M_i) )</th>
<th>Difficulty/Concern</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conceptual</td>
<td></td>
</tr>
</tbody>
</table>
### Posterior probabilities and Bayes factors

\[
Pr(M_j \mid y) = \frac{Pr(M_j)}{\sum_{i=0}^{2^p} B_{ij} Pr(M_i)}, \quad B_{ij} = \frac{m_i(y)}{m_j(y)},
\]

where \( m_i(y) \) are the integrated likelihoods (or prior predictive marginals):

\[
m_i(y) = \int M_i(y \mid \alpha, \beta_i, \sigma) \pi_i(\alpha, \beta_i, \sigma) \, d\alpha \, d\beta_i \, d\sigma.
\]

Above,

<table>
<thead>
<tr>
<th></th>
<th>Difficulty/Concern</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Pr(M_i) )</td>
<td>Conceptual</td>
</tr>
<tr>
<td>( \pi_i(\alpha, \beta_i, \sigma) )</td>
<td>Part II: model selection priors are intriguing</td>
</tr>
</tbody>
</table>
Posterior probabilities and Bayes factors

$$Pr(M_j | y) = \frac{Pr(M_j)}{\sum_{i=0}^{2^p} B_{ij} Pr(M_i)}, \quad B_{ij} = \frac{m_i(y)}{m_j(y)},$$

where $m_i(y)$ are the integrated likelihoods (or prior predictive marginals):

$$m_i(y) = \int M_i(y | \alpha, \beta_i, \sigma) \pi_i(\alpha, \beta_i, \sigma) d\alpha d\beta_i d\sigma.$$
PART II:
Objective priors in variable selection problems

Joint work with: S. Bayarri, A. Forte (Universidad de Valencia) and J. Berger (Duke University).
In model selection

There is no consensus about which approach should be used to elicit default prior distributions.
In model selection

There is no consensus about which approach should be used to elicit default prior distributions. Several interesting proposals

- the conventional priors (Jeffreys 1961; Zellner and Siow 1980)
- the Intrinsic Priors (Berger and Pericchi 1996);
- the Expected posterior Priors (Pérez 1998) and the
- Divergence based Priors (Bayarri and García-Donato 2008).
In model selection

There is no consensus about which approach should be used to elicit default prior distributions. Several interesting proposals

- the conventional priors (Jeffreys 1961; Zellner and Siow 1980)
- the Intrinsic Priors (Berger and Pericchi 1996);
- the Expected posterior Priors (Pérez 1998) and the
- Divergence based Priors (Bayarri and García-Donato 2008).

There are no clear guidelines for evaluating such objective priors.
Two main challenges

• In this work we propose a new prior $\pi_i(\alpha, \beta_i, \sigma)$, which we will argue is a solid alternative to existing proposals. This is important for the problem at hand.

• This proposal is based on formal arguments, some of them new. These could be relevant for other model selection problems.
Integral representation

Without loss of generality we can write

\[ \pi_i(\alpha, \beta_i, \sigma) = \pi_i(\alpha, \sigma) \pi_i(\beta_i \mid \alpha, \sigma). \]
Integral representation

Without loss of generality we can write

\[ \pi_i(\alpha, \beta_i, \sigma) = \pi_i(\alpha, \sigma) \pi_i(\beta_i | \alpha, \sigma). \]

Warning!

\[ \pi_i(\beta_i | \alpha, \sigma) \] has to be a proper density. Further, arbitrarily vague proper priors cannot be used: both give rise to arbitrary answers.
Integral representation

Without loss of generality we can write

\[ \pi_i(\alpha, \beta_i, \sigma) = \pi_i(\alpha, \sigma) \pi_i(\beta_i | \alpha, \sigma). \]

**Warning!**

\( \pi_i(\beta_i | \alpha, \sigma) \) has to be a proper density. Further, arbitrarily vague proper priors cannot be used: both give rise to arbitrary answers.

Using this conditional, the original problem with competing models:

\[ M_i(y | \alpha, \beta_i, \sigma) \]

transforms in a problem with models:

\[ \int M_i(y | \alpha, \beta_i, \sigma) \pi_i(\beta_i | \alpha, \sigma) \, d\beta_i. \]
Integral representation

Without loss of generality we can write

$$\pi_i(\alpha, \beta_i, \sigma) = \pi_i(\alpha, \sigma) \pi_i(\beta_i \mid \alpha, \sigma).$$

**Warning!**

$\pi_i(\beta_i \mid \alpha, \sigma)$ has to be a proper density. Further, arbitrarily vague proper priors cannot be used: both give rise to arbitrary answers.

Using this conditional, the original problem with competing models:

$$M_i(y \mid \alpha, \beta_i, \sigma)$$

transforms in a problem with models:

$$M'_i(y \mid \alpha, \sigma) = \int M_i(y \mid \alpha, \beta_i, \sigma) \pi_i(\beta_i \mid \alpha, \sigma) \, d\beta_i.$$
Integral representation

Without loss of generality we can write

\[ \pi_i(\alpha, \beta_i, \sigma) = \pi_i(\alpha, \sigma) \pi_i(\beta_i | \alpha, \sigma). \]

**Warning!**

\( \pi_i(\beta_i | \alpha, \sigma) \) has to be a proper density. Further, arbitrarily vague proper priors cannot be used: both give rise to arbitrary answers.

Using this conditional, the original problem with competing models:

\[ M_i(y | \alpha, \beta_i, \sigma) \]

transforms in a problem with models:

\[ M_i^!(y | \alpha, \sigma) = \int M_i(y | \alpha, \beta_i, \sigma) \pi_i(\beta_i | \alpha, \sigma) d\beta_i. \]

Call \( \mathcal{M}^! \) the set of transformed models.
Argument I: conserve the invariance structure

All models in $\mathcal{M}$ are invariant under the group of transformations:

$$G = \{ y \rightarrow y^* = cy + 1b, \ b \in \mathcal{R}, \ c > 0 \}.$$ 

It seems natural to ask for the same type of invariance to the set of integrated models.

Result

All models in $\mathcal{M}^I$ are invariant under $G$ iff

$$\pi_i(\beta_i \mid \alpha, \sigma) = \frac{1}{\sigma^{k_i}} f_i\left(\frac{\beta_i}{\sigma}\right),$$

where $f_i$ is a proper density over $\mathcal{R}^{k_i}$.

Note: this provides a justification for the use of $\sigma$ to scale $\beta_i$. 
We have

$$\pi_i(\alpha, \beta_i, \sigma) = \pi_i(\alpha, \sigma) \frac{1}{\sigma^{k_i}} f_i\left(\frac{\beta_i}{\sigma}\right).$$
We have

\[ \pi_i(\alpha, \beta_i, \sigma) = \pi_i(\alpha, \sigma) \frac{1}{\sigma^{k_i}} f_i\left(\frac{\beta_i}{\sigma}\right). \]

What about \( \pi_i(\alpha, \sigma) \)?
Argument II: Common parameters

\[ \pi_i(\alpha, \beta_i, \sigma) = \pi_i(\alpha, \sigma) \frac{1}{\sigma^k_i} f_i(\frac{\beta_i}{\sigma}). \]

What about \( \pi_i(\alpha, \sigma) \)?
With the prior above, integrated models \( M^I \) share the same group of invariance (again \( G \)).
Part II: Objective priors

Argument II: Common parameters

\[ \pi_i(\alpha, \beta_i, \sigma) = \pi_i(\alpha, \sigma) \frac{1}{\sigma^{k_i}} f_i \left( \frac{\beta_i}{\sigma} \right). \]

What about \( \pi_i(\alpha, \sigma) \)?
With the prior above, integrated models \( M^I \) share the same group of invariance (again \( G \)).

Result (Berger, Pericchi and Varshavsky (1998))

*When models are invariant under the the same group of transformations, there are theoretical reasons that justify the use of the right Haar density for common parameters.*
Argument II: Common parameters

\[ \pi_i(\alpha, \beta_i, \sigma) = \pi_i(\alpha, \sigma) \frac{1}{\sigma^{k_i}} f_i\left(\frac{\beta_i}{\sigma}\right). \]

What about \( \pi_i(\alpha, \sigma) \)?
With the prior above, integrated models \( M^I \) share the same group of invariance (again \( G \)).

Result (Berger, Pericchi and Varshavsky (1998))

*When models are invariant under the the same group of transformations, there are theoretical reasons that justify the use of the right Haar density for common parameters.*

In our problem, this would lead us to use \( \pi_i(\alpha, \sigma) = \sigma^{-1} \).
We have

\[ \pi_i(\alpha, \beta_i, \sigma) = \sigma^{-1} \frac{1}{\sigma^k_i} f_i(\frac{\beta_i}{\sigma}). \]
Argument III: Scale matrix for $f_i$

\[
\pi_i(\alpha, \beta_i, \sigma) = \sigma^{-1} \frac{1}{\sigma^{k_i}} f_i\left(\frac{\beta_i}{\sigma}\right).
\]

The function $f_i$ is the (conditional) density on $\mathcal{R}^{k_i}$ of $\beta_i/\sigma$. We borrow the covariance of the mle of $\beta_i$ to provide $f_i$ with a dependency structure:

\[\text{Cov}\left(\frac{\beta_i}{\sigma}\right) \propto n \text{Cov}\left(\frac{\hat{\beta}_i}{\sigma}\right)\]
Argument III: Scale matrix for $f_i$

\[\pi_i(\alpha, \beta_i, \sigma) = \sigma^{-1} \frac{1}{\sigma^{k_i}} f_i\left(\frac{\beta_i}{\sigma}\right).\]

The function $f_i$ is the (conditional) density on $\mathcal{R}^{k_i}$ of $\beta_i/\sigma$. We borrow the covariance of the mle of $\beta_i$ to provide $f_i$ with a dependency structure:

\[\text{Cov}\left(\frac{\beta_i}{\sigma}\right) \propto n \text{Cov}(\frac{\hat{\beta}_i}{\sigma}) = n \left(X_i^t(I - \frac{1}{n}11^t)X_i\right)^{-1}\]

- $n$ ‘scales’ to a unitary size.
Argument III: Scale matrix for $f_i$

$$
\pi_i(\alpha, \beta_i, \sigma) = \sigma^{-1} \frac{1}{\sigma^{k_i}} f_i(\frac{\beta_i}{\sigma}).
$$

The function $f_i$ is the (conditional) density on $\mathcal{R}^{k_i}$ of $\beta_i/\sigma$. We borrow the covariance of the mle of $\beta_i$ to provide $f_i$ with a dependency structure:

$$
\text{Cov}(\frac{\beta_i}{\sigma}) \propto n \text{Cov}(\frac{\hat{\beta}_i}{\sigma}) = n (X_i^t(I - \frac{1}{n}11^t)X_i)^{-1}
$$

- $n$ ‘scales’ to a unitary size.
- Justification 1: This proposal remains unchanged under reparameterizations of the common parameters (since the mle is itself invariant under such transformations).
Argument III: Scale matrix for $f_i$

\[ \pi_i(\alpha, \beta_i, \sigma) = \sigma^{-1} \frac{1}{\sigma^{k_i}} f_i\left(\frac{\beta_i}{\sigma}\right). \]

The function $f_i$ is the (conditional) density on $\mathcal{R}^{k_i}$ of $\beta_i/\sigma$. We borrow the covariance of the mle of $\beta_i$ to provide $f_i$ with a dependency structure:

\[ \text{Cov}\left(\frac{\beta_i}{\sigma}\right) \propto n \text{Cov}\left(\frac{\hat{\beta}_i}{\sigma}\right) = n \left(X_i^t(I - \frac{1}{n}11^t)X_i\right)^{-1} \]

- $n$ ‘scales’ to a unitary size.
- Justification 1: This proposal remains unchanged under reparameterizations of the common parameters (since the mle is itself invariant under such transformations).
- Justification 2: Weak predictive matching (more on this later).
Argument IV: A functional form for $f_i$

We looked for a density

With heavy tails,
Argument IV: A functional form for $f_i$

We looked for a density

With heavy tails, analytically tractable,
Argument IV: A functional form for $f_i$

We looked for a density

With heavy tails, analytically tractable, flexible.
Argument IV: A functional form for $f_i$

We looked for a density

With heavy tails, analytically tractable, flexible.
Argument IV: A functional form for $f_i$

We looked for a density

With heavy tails, analytically tractable, flexible.

Strawderman (1971), Berger (1980). Standard form:

$$f(z) = \int_0^1 N_k(z \mid 0, \lambda^{-1} \rho \frac{1+n}{n} - \frac{1}{n}) Beta(\lambda \mid 1, 1/2) \, d\lambda \quad (\rho \geq 1/(1+n)).$$
How $f_i$ looks like

\[ f(\rho=0.1, n=100) \]

- Normal(0,1)
- Cauchy(0,1)
We have

\[ \pi_i(\alpha, \beta_i, \sigma) = \sigma^{-1} \int_0^1 N_{k_i} \left( \beta_i \mid 0, \text{Cov}(\hat{\beta}_i) \left( \frac{\rho_i(1+n)}{\lambda} - 1 \right) \right) B(\lambda \mid 1, 1/2) \, d\lambda. \]
We have

\[
\pi_i(\alpha, \beta_i, \sigma) = \sigma^{-1} \int_0^1 N_{k_i}(\beta_i \mid 0, \text{Cov}(\hat{\beta}_i)(\frac{\rho_i(1 + n)}{\lambda} - 1)) B(\lambda \mid 1, 1/2) \, d\lambda.
\]

It has appealing properties:

- several ways of consistency,
We have

\[ \pi_i(\alpha, \beta_i, \sigma) = \sigma^{-1} \int_0^1 N_{k_i} \left( \beta_i \mid 0, \text{Cov}(\hat{\beta}_i) \left( \frac{\rho_i(1 + n)}{\lambda} - 1 \right) \right) B(\lambda \mid 1, 1/2) \, d\lambda. \]

It has appealing properties:

- several ways of consistency,
- **NEW!** Weak predictive matching

For data of minimal size, the Bayes factor for two models \( M_i \) and \( M_j \) of the same complexity (ie \( k_i = k_j \)) should be as close as possible to one.
We have

\[
\pi_i(\alpha, \beta_i, \sigma) = \sigma^{-1} \int_0^1 N_{k_i}(\beta_i \mid 0, \text{Cov}(\hat{\beta}_i)(\frac{\rho_i(1 + n)}{\lambda} - 1)) B(\lambda \mid 1, 1/2) d\lambda.
\]

It has appealing properties:

- several ways of consistency,

**NEW!** Weak predictive matching

For data of minimal size, the Bayes factor for two models \(M_i\) and \(M_j\) of the same complexity (ie \(k_i = k_j\)) should be as close as possible to one.

And still \(\rho_i\) needs to be assigned
Argument V: Fine tuning for $\rho_i$

Null predictive matching (Spiegelhalter and Smith 1982; Ghosh and Samanta 2002)

For data of minimal size which is strongly compatible with $M_0$, the Bayes factor of $M_i$ to $M_0$ should be as close as possible to one.
Argument V: Fine tuning for $\rho_i$

Null predictive matching (Spiegelhalter and Smith 1982; Ghosh and Samanta 2002)

For data of minimal size which is strongly compatible with $M_0$, the Bayes factor of $M_i$ to $M_0$ should be as close as possible to one.

The rationale behind NullPM is that we ask for matching when:

- data is compatible with $M_0$, this model $M_0$ should be preferred
- data is of minimal size then $M_i$ provides almost perfect fit.
Argument V: Fine tuning for $\rho_i$

Null predictive matching (Spiegelhalter and Smith 1982; Ghosh and Samanta 2002)

For data of minimal size which is strongly compatible with $M_0$, the Bayes factor of $M_i$ to $M_0$ should be as close as possible to one.

The rationale behind NullPM is that we ask for matching when:
- data is compatible with $M_0$, this model $M_0$ should be preferred
- data is of minimal size then $M_i$ provides almost perfect fit.

Result

For our proposal to be Null predictive matching:

$$\rho_i = 1/(2 + k_i).$$
In summary

Our proposed prior is:

$$\pi_i(\alpha, \beta_i, \sigma) =$$

$$= \sigma^{-1} \int_0^1 N_{k_i}(\beta_i | 0, \text{Cov}(\hat{\beta}_i)\left(\frac{(1 + n)}{\lambda(2 + k_i)} - 1\right)) B(\lambda | 1, 1/2) d\lambda.$$

which leads to a Bayes factor of $M_i$ to $M_0$ of:
In summary

Our proposed prior is:

\[
\pi_i(\alpha, \beta_i, \sigma) = \sigma^{-1} \int_0^1 N_{k_i} \left( \beta_i \left| 0, \text{Cov}(\hat{\beta}_i) \left( \frac{(1 + n)}{\lambda(2 + k_i)} - 1 \right) \right. \right) B(\lambda \mid 1, 1/2) \, d\lambda.
\]

which leads to a Bayes factor of \( M_i \) to \( M_0 \) of:

\[
B_{i0} = \frac{1}{k_i + 1} \left( \frac{n + 1}{k_i + 2} \right)^{-k_i/2} Q_{i0}^{-(n-1)/2} \times 2F_1 \left( \frac{k_i + 1}{2}, \frac{n - 1}{2}, \frac{k_i + 3}{2}, \frac{(1 - Q_{0i})(k_i + 3)}{(1 + n)} \right),
\]

where \( Q_{i0} \) is the ratio of sum of squared errors and \( 2F_1 \) is the hypergeometric function.
Results in selected data sets

Results in selected data sets

Posterior probabilities and Bayes factors

\[
Pr(M_j \mid y) = \frac{Pr(M_j)}{\sum_{i=0}^{2^p} B_{ij} Pr(M_i)}, \quad B_{ij} = \frac{m_i(y)}{m_j(y)},
\]

where \( m_i(y) \) are the integrated likelihoods (or prior predictive marginals):

\[
m_i(y) = \int \mathcal{M}_i(y \mid \alpha, \beta_i, \sigma) \pi_i(\alpha, \beta_i, \sigma) \, d\alpha \, d\beta_i \, d\sigma.
\]

Above,

<table>
<thead>
<tr>
<th></th>
<th>Difficulty/Concern</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Pr(M_i) )</td>
<td>Conceptual</td>
</tr>
<tr>
<td>( \pi_i(\alpha, \beta_i, \sigma) )</td>
<td>Part II: model selection priors are intriguing</td>
</tr>
<tr>
<td>( \sum_{i=0}^{2^p} )</td>
<td>Part III: too many models</td>
</tr>
</tbody>
</table>
PART III:
Inferences in large variable selection problems
Joint work with M.A. Martinez-Beneito (CSISP, Valencia).
Inferences

• Once the prior has been assigned, drawing any type of inferences is just a question of summarizing the posterior distribution over the model space $Pr(M_i \mid y)$. 
Inferences

- Once the prior has been assigned, drawing any type of inferences is just a question of summarizing the posterior distribution over the model space $Pr(M_i | y)$.
- For a number of priors in the literature (including of course ours), inferences can be exactly obtained with few lines of R code (is a simple loop!).
Inferences

- Once the prior has been assigned, drawing any type of inferences is just a question of summarizing the posterior distribution over the model space \( Pr(M_i \mid y) \).
- For a number of priors in the literature (including of course ours), inferences can be exactly obtained with few lines of R code (is a simple loop!).

About the \( \sum_i \):

Other challenges appear when the number of covariates considered is moderate to large (say \( p \geq 30 \)): exhaustive enumeration of all models would be impossible!
Part III: Inferences in large problems

Large model spaces

Consequence:
When $p$ is large, inferences are to be based on a very small set

$$S = \{M(1), M(2), \ldots, M(N)\} \subset \mathcal{M},$$

of models visited.
Normally $N << 2^p$. 
The nature of inferences: our main interest

\[ S = \{ M_1, M_2, \ldots, M_N \} \subset \mathcal{M}, \]
The nature of inferences: our main interest

\[ S = \{ M_1, M_2, \ldots, M_N \} \subset \mathcal{M}, \]

Two different philosophies to produce inferences:

**Approach I: we label Empirical**
- \( S \) is a sample of models from the posterior distribution over the model space (via MCMC).
- Inferences are based on the frequency of visits.
The nature of inferences: our main interest

\[ S = \{ M_1, M_2, \ldots, M_N \} \subset \mathcal{M}, \]

Two different philosophies to produce inferences:

**Approach I: we label Empirical**
- \( S \) is a sample of models from the posterior distribution over the model space (via MCMC).
- Inferences are based on the frequency of visits

**Approach II: we label re-normalized**
- \( S \) is a sample of good models (with high Bayes factors).
- Inferences are based on re-normalizing the analytic expression of Bayes factors.
Estimation of posterior probabilities

For $M_j \in S$:

**Empirical approach**

$$
Pr(M_j | y) = \frac{\# \text{ Frequency of } M_j \text{ in } S}{\# \text{ of models in } S}
$$
Estimation of posterior probabilities

For $M_j \in S$:

**Empirical approach**

$$Pr(M_j \mid y) = \frac{\# \text{ Frequency of } M_j \text{ in } S}{\# \text{ of models in } S}$$

**Re-normalized approach**

$$Pr(M_j \mid y) = \frac{Pr(M_j)}{\sum_{i=1}^{N} B(i)(j) \cdot Pr(M_i)}, \quad B(i)(j) = \frac{m(i)(y)}{m(j)(y)}.$$
Part III: Inferences in large problems

Literature

- For the empirical, different proposals are just a question of the different MCMC strategies proposed: George and McCulloch (1993), George and McCulloch (1997); Kuo and Mallick (1998); Dellaportas et al (2000); Nott and Kohn (2005); Ntzoufras (2002, 2009); Casella and Moreno (2006).
For the empirical, different proposals are just a question of the different MCMC strategies proposed: George and McCulloch (1993), George and McCulloch (1997); Kuo and Mallick (1998); Dellaportas et al (2000); Nott and Kohn (2005); Ntzoufras (2002, 2009); Casella and Moreno (2006).

For the re-normalized approach: the Bayesian Adaptive Sampling (BAS) algorithm (Clyde et al. 2010) and the Stochastic Search of Berger and Molina (2005), Carvalho and Scott (2009) and Scott and Carvalho (2009).
A large problem with exact answer

Our experiment: to *compare* empirical and re-normalized approaches in a real dataset with a size on the boundaries of feasibility.

More details:

- Ozone 35 with $p = 35$ and $n = 178$ observations. $\mathcal{M}$ has $2^{35} \approx 3 \times 10^{10}$ models.
A large problem with exact answer

Our experiment: to compare empirical and re-normalized approaches in a real dataset with a size on the boundaries of feasibility. More details:

- Ozone 35 with $p = 35$ and $n = 178$ observations. $\mathcal{M}$ has $2^{35} \approx 3 \times 10^{10}$ models.
- $\pi_i(\alpha, \beta_i, \sigma)$: Zellner-g prior (it produces the simplest marginals). The conclusions are, to a great extent, independent of the priors.
A large problem with exact answer

Our experiment: to compare empirical and re-normalized approaches in a real dataset with a size on the boundaries of feasibility.

More details:

- Ozone 35 with $p = 35$ and $n = 178$ observations. $\mathcal{M}$ has $2^{35} \approx 3 \times 10^{10}$ models.

- $\pi_i(\alpha, \beta_i, \sigma)$: Zellner-$g$ prior (it produces the simplest marginals). The conclusions are, to a great extent, independent of the priors.

- The program was written in C and parallelized in a cloud with 150 cores. It took 20 hours.
A large problem with exact answer

In order to illustrate the type of results we found, here we present the results for the inclusion probabilities of the $x_P$ the pressure gradient, and $x_T x_H$ the interaction between Temperature and Humidity

$$q_{x_P} = Pr(x_P \in M^T \mid y) = \sum_{M_i: x_P \in M_i} Pr(M_i \mid y) = 0.29.$$
In order to illustrate the type of results we found, here we present the results for the inclusion probabilities of the $x_P$ the pressure gradient, and $x_T x_H$ the interaction between Temperature and Humidity.

$$q_{x_P} = Pr(x_P \in M_T^T \mid y) = \sum_{M_i : x_P \in M_i} Pr(M_i \mid y) = 0.29.$$ $q_{x_T x_H} = Pr(x_T x_H \in M_T^T \mid y) = \sum_{M_i : x_T x_H \in M_i} Pr(M_i \mid y) = 0.64.$

Similar conclusions for the other covariates.
The comparison

We consider

- For the empirical approach: the Gibbs sampling algorithm in George and McCulloch (1997),

We run the three methods 5 times each with 10,000 iterations:
ozone35: $q_{xp}$
ozone35: $q_{XP}$

- Gibbs sampling + empirical approach
ozone35: $q_{XP}$

- Bayesian Adaptive Sampling + re-normalized approach
Part III: Inferences in large problems

\( q_{XP} \)

- Stochastic Search+renormalized approach
ozone35: $q_{XP}$
We run the three methods 5 times each with 10,000 iterations:
We run the three methods 5 times each with 10,000 iterations:

- Gibbs sampling + empirical approach
We run the three methods 5 times each with 10,000 iterations:

- Bayesian Adaptive Sampling + re-normalized approach
ozone35: $q_{X_T X_H}$

We run the three methods 5 times each with 10,000 iterations:

![Graph showing runs and methods]

- Stochastic Search + re-normalized approach
ozone$_{35}$: $q_{X_T X_H}$

We run the three methods 5 times each with 10,000 iterations:

![Graph showing the results of the three methods run 5 times each with 10,000 iterations. The x-axis represents the run number and the y-axis represents the probability of $XX|y$. The graph includes points for Gibbs+Empirical, BAS+Renormalized, and SSBM+Renormalized methods.]
In summary

- The empirical approach produces more reliable results.
In summary

- The empirical approach produces more reliable results.
- The observed differences across runs in the re-normalized methods analyzed suggest biased responses.
In summary

- The empirical approach produces more reliable results.
- The observed differences across runs in the re-normalized methods analyzed suggest biased responses.

A simple likely explanation:

- Empirical approach: empirical estimators are proportional to size sampling (PPS) estimators. Hence, they are unbiased and with a small variance (with the usual hypothesis of the MCMC).
In summary

- The empirical approach produces more reliable results.
- The observed differences across runs in the re-normalized methods analyzed suggest biased responses.

A simple likely explanation:

- Empirical approach: empirical estimators are proportional to size sampling (PPS) estimators. Hence, they are unbiased and with a small variance (with the usual hypothesis of the MCMC).
- Re-normalized approach: look only for good models produce biased inferences towards these models.
What if I am only interested in finding high probable posterior models?

- The underlying methods for discovering high probable models in BAS and SSBM are quite effective (the R library for BAS is extremely fast).
What if I am only interested in finding high probable posterior models?

- The underlying methods for discovering high probable models in BAS and SSBM are quite effective (the R library for BAS is extremely fast).
- The ability in detecting high probable models when sampling from the posterior distribution is also very high.
What if I am only interested in finding high probable posterior models?

- The underlying methods for discovering high probable models in BAS and SSBM are quite effective (the R library for BAS is extremely fast).
- The ability in detecting high probable models when sampling from the posterior distribution is also very high.

All details in
Garcia-Donato and Martinez-Beneito (2011); *Inferences in Bayesian variable selection problems with large model spaces*, arXiv:1101.4368v1

Thanks!