

Taylor & Francis
Taylor & Francis Group

Statistical Theory and Related Fields

ISSN: 2475-4269 (Print) 2475-4277 (Online) Journal homepage: www.tandfonline.com/journals/tstf20

On model prior probabilities in variable selection

James Berger, Gonzalo García-Donato & Luis Pericchi

To cite this article: James Berger, Gonzalo García-Donato & Luis Pericchi (22 Nov 2026): On model prior probabilities in variable selection, Statistical Theory and Related Fields, DOI: 10.1080/24754269.2026.2765046

To link to this article: https://doi.org/10.1080/24754269.2026.2765046

© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

Published online: 22 Nov 2026.

Submit your article to this journal

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=tstf20

STATISTICAL THEORY AND RELATED FIELDS
https://doi.org/10.1080/24754269.2026.2765046

Taylor & Francis
by informa

OPEN ACCESS Check for updates

On model prior probabilities in variable selection

James Berger [a,b], Gonzalo García-Donato[c] and Luis Pericchi[d]

[a]Department of Statistical Science, Duke University, Durham, NC, USA ; [b]Department of Statistics, Texas A&M University, College Station, TX, USA; [c]Department of Economics and Finance, University of Castilla-La Mancha, Ciudad Real, Spain; [d]Department of Mathematics, University of Puerto Rico, Rio Piedras, San Juan, PR, USA

ABSTRACT

For many years it was routine to use equal model prior probabilities in Bayesian model uncertainty analysis. At least twenty years ago, it became clear that this was problematic, as it tended to favour models that were far too large within the increasingly vast model spaces being explored in genomics and other fields. A popular replacement was to adopt a suggestion of Harold Jeffreys for the variable selection problem in which a total of k possible variables are being considered for inclusion in the model: give the collection of all models containing d variables ($d = 0, \ldots, k$) prior probability $1/(k+1)$ and then divide this prior probability equally among the models in the collection. Many other choices of model prior probabilities that impose severe parsimony have also been introduced. We begin by reviewing the problems with using equal model prior probabilities and then discuss some serious problems with the Jeffreys choice. Finally, we introduce and study a number of objective alternative choices of model prior probabilities, from both numerical and theoretical perspectives. The final choice of model prior probabilities is found to depend on the degree of sparsity that one requires.

ARTICLE HISTORY
Received 3 August 2026
Revised 13 October 2026
Accepted 12 November 2026

KEYWORDS
Bayesian variable selection; prior probability; objective choice; Jeffreys prior; parsimony; sparsity

1. Introduction and motivation

Consider the variable selection problem, in which variables $(x_1, x_2, \ldots, x_k)$ can or cannot be included in a model. Let $\gamma_i = 1$ if variable x_i is to be included in the model and $\gamma_i = 0$ otherwise. Then $\boldsymbol{\gamma} = (\gamma_1, \ldots \gamma_k)$ is the vector of zeroes and ones that defines which variables are in the model $M_{\boldsymbol{\gamma}}$, with $M_{\mathbf{0}}$ being the null model and $M_{\mathbf{1}}$ being the full model. Note that there are 2^k possible models and that the dimension of $M_{\boldsymbol{\gamma}}$ is $|\boldsymbol{\gamma}| = \sum \gamma_i$. We only consider prior specifications that depend on the dimensionality of $\boldsymbol{\gamma}$, that is, $P(M_{\boldsymbol{\gamma}})$ is a function of $|\boldsymbol{\gamma}|$, or equivalently,

$$P(M_{\boldsymbol{\gamma}}) = \frac{1}{\binom{k}{|\boldsymbol{\gamma}|}} \mathfrak{M}_k(|\boldsymbol{\gamma}|), \tag{1}$$

where $\mathfrak{M}_k(d)$ is the probability assigned to the set of models such that $|\boldsymbol{\gamma}| = d$. Clearly, $\mathfrak{M}_k(d)$ defines a probability mass function over $d \in \{0, \ldots, k\}$.

CONTACT James Berger berger@duke.edu Department of Statistical Science, Duke University, Box 90251 Durham, NC 27708-0251, USA

© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

Historically, the first two considered objective choices of model prior probabilities $P(M_{\boldsymbol{\gamma}})$ were *Uniform* probabilities

$$P(M_{\boldsymbol{\gamma}}) = 2^{-k} \tag{2}$$

and *Jeffreys* probabilities:

$$P(M_{\boldsymbol{\gamma}}) = \int_0^1 p^{|\boldsymbol{\gamma}|}(1-p)^{k-|\boldsymbol{\gamma}|}\,\pi(p)\,\mathrm{d}p, \tag{3}$$

with $\pi(p) = \mathrm{Un}(p \mid 0, 1)$, leading to

$$P(M_{\boldsymbol{\gamma}}) = \mathrm{Beta}(1+|\boldsymbol{\gamma}|, 1+k-|\boldsymbol{\gamma}|) = \frac{|\boldsymbol{\gamma}|!\,(k-|\boldsymbol{\gamma}|)!}{(k+1)!}. \tag{4}$$

One way in which the Jeffreys probabilities can be explained is to think of p as the prior inclusion probability of a variable (see Section 6 for the formal definition), assumed to be common across variables; to assume that the inclusions of different variables are independent; and to assume that p is assigned the natural objective uniform prior density. These probabilities were recommended by Jeffreys (1961), although the argument therein was to give each potential model size an equal probability of $1/(k+1)$, and then divide this probability up equally among all models of a given size; it is easy to show that this gives the same answer as (4).

The Uniform probabilities also arise from saying that the prior inclusion probability for each variable is $p = 0.5$ and that the components of γ are independent. Thus, there is no penalty for including variables, which can be problematical, particularly if k is large. For instance, suppose one is testing 1,000,000 genes to detect gene expression related to some disease. The Uniform probability assignment implies that there is an extremely large prior probability that the number of influential genes is around 500,000 (e.g., the models, for which the number of positive genes is in the interval [499,000, 501,000], have a total probability of about 0.95); this is typically very far from actual prior opinions. This issue with Uniform probabilities was also described as a lack of *multiplicity control* in Scott and Berger (2005) and Scott and Berger (2010). It is also shown therein that the Jeffreys probabilities do induce multiplicity control.

A study of gene association involving $k = 4088$ genes (García-Donato & Steel, 2021) illustrated the dramatic difference in the use of the two priors. With the Jeffreys choice, only one gene is clearly positive (posterior inclusion probability of 0.97), while the great majority are clearly ruled out (97% have an inclusion probability below 0.005). With the Uniform prior, the posterior distribution tends to highly mimic the prior, suggesting that about half of the genes may be important (with most of the posterior inclusion probabilities hovering around 0.5). For other discussions of the choice of prior probabilities of models, see Clyde et al. (1996); George and McCulloch (1993, 1997); Ley and Steel (2009); Raftery et al. (1997); Scott and Berger (2005, 2010); Smith and Kohn (1996).

One concern with the Jeffreys probabilities is that they give large models as much weight as small models. For instance, the null model and the full model both have equal prior probability of $1/(k+1)$. Typically, some type of parsimony is desired in which smaller models are given more weight than large models. Indeed, the use of Jeffreys probabilities has been found to produce disturbing results in application. For instance, in the infant obesity study given in Zurriaga et al. (2011) that involved $k = 47$ possible variables, use of the Jeffreys probabilities resulted in the very surprising result that the highest posterior probability model was

the full model. This is certainly not compatible with prior scientific beliefs in the obesity study and presumably arose because the full model had the largest prior probability of 1/48. (The details of this application are given in the Appendix, so as not to disrupt the flow of the paper.)

Thus, we are faced with problems with both the Uniform probabilities and the Jeffreys probabilities. In Section 2, we present some objective choices of model prior probabilities that seek to overcome these problems. The suggested possibilities will be evaluated in the remainder of the paper, first numerically and then theoretically.

The use of the word 'objective' in this introduction deserves a comment. The Uniform and Jeffreys model prior probabilities can certainly be called objective, but we have seen that they have serious problems, centred around a lack of parsimony. At the other extreme, the literature is replete with efforts to induce severe parsimony, but we would not label those efforts as being objective. (Some of the efforts most closely related to the proposals studied in this paper are discussed in Section 2.6.) What we are seeking is choices of model prior probabilities that incorporate necessary parsimony but only minimally so, in order to reasonably view them as being objective. Satisfying a minimal requirement for objectivity, note that none of the choices of model prior probabilities that we consider depend on tuning parameters. As part of this focus on objectivity, note that we are assuming exchangeability in assessments of variable prior probabilities. In many scenarios, this assumption is untenable; a few of the many papers that address the need for non-exchangeable assessments are Chipman (1996), Casella et al. (2014), Berger et al. (2014), García-Donato and Paulo (2022), and Zhu et al. (2019).

All model priors discussed in this paper can be readily implemented within the R package BayesVarSel (García-Donato & Forte, 2018) by specifying user-defined model probabilities. To facilitate their use, we provide an open-source implementation of the priors considered here at the GitHub repository https://github.com/gongardo/modelpriorprobs.

2. More parsimonious prior probabilities

Apart from Uniform and Jeffreys' we consider six additional choices of model prior probabilities, the first three from the literature.

2.1. Harmonic prior probabilities

Berger and Pericchi (1996) mentioned the possibility of using the prior

$$P(M_{\boldsymbol{\gamma}}) = \frac{(1+|\boldsymbol{\gamma}|)^{-1}}{\binom{k}{|\boldsymbol{\gamma}|} H_{k+1}}, \tag{5}$$

where $H_k = \sum_{j=1}^{k} j^{-1}$ represents the kth harmonic number.

This prior is naturally called the *Harmonic prior* and assigns the collection of models of dimension d probability proportional to $(1+d)^{-1}$, which is decreasing almost as slowly as possible in the dimension and is thus a natural objective choice.

2.2. Casella-Moreno-Girón (CMG) prior probabilities

Borrowing ideas from intrinsic priors, Casella et al. (2014) proposed a mixture of a Poisson distribution

$$P(M_{\boldsymbol{\gamma}}) \propto \frac{1}{\binom{k}{|\boldsymbol{\gamma}|}} \int_0^\infty \text{Po}(|\boldsymbol{\gamma}| \mid \lambda)\, \pi(\lambda)\, \mathrm{d}\lambda, \tag{6}$$

with λ being assigned the density

$$\pi^I(\lambda) = \frac{\mathrm{e}^{-(\lambda+1)}}{\sqrt{\pi\lambda}} \cosh(2\sqrt{\lambda}), \quad \lambda \geq 0.$$

2.3. Half-k Jeffreys' prior probabilities

One simple correction to the Jeffreys model prior probabilities would be to use the Jeffreys probabilities for models up to size $k/2$ if k is even and $(k-1)/2$ if k is odd, and then give the remaining models equal prior probability. Note that the number of large models receiving this constant prior probability is

$$N_k = \begin{cases} \frac{1}{2}\left(2^k - \binom{k}{k/2}\right), & \text{if } k \text{ is even,} \\ 2^{k-1}, & \text{if } k \text{ is odd.} \end{cases}$$

The resulting model prior probabilities unfortunately have an unappealing discontinuity, so we present here a version which smooths the discontinuity.

When k is even,

$$P(M_{\boldsymbol{\gamma}}) = \begin{cases} C_1^e \dfrac{1}{(k+1)\binom{k}{|\boldsymbol{\gamma}|}}, & \text{if } |\boldsymbol{\gamma}| \leq k/2, \\ C_2^e \dfrac{k}{2(k+1)N_k}, & \text{if } |\boldsymbol{\gamma}| \geq k/2+1, \end{cases} \tag{7}$$

where C_1^e and C_2^e are such that $P(M_{\boldsymbol{\gamma}})$ is a probability mass function and

$$P(M_{\boldsymbol{\gamma}} : |\boldsymbol{\gamma}| = k/2) = P(M_{\boldsymbol{\gamma}'} : |\boldsymbol{\gamma}'| = k/2+1).$$

Solving the corresponding system of equations we arrive at

$$C_1^e = \frac{2(k+1)\binom{k}{k/2}}{2^k + (k+1)\binom{k}{k/2}}, \quad C_2^e = \frac{2(k+1)\left(2^k - \binom{k}{k/2}\right)}{k\left(2^k + (k+1)\binom{k}{k/2}\right)}.$$

Similarly, when k is odd,

$$P(M_{\boldsymbol{\gamma}}) = \begin{cases} C_1^o \dfrac{1}{(k+1)\binom{k}{|\boldsymbol{\gamma}|}}, & \text{if } |\boldsymbol{\gamma}| \leq (k-1)/2, \\ C_2^o \dfrac{1}{2N_k}, & \text{if } |\boldsymbol{\gamma}| \geq (k+1)/2, \end{cases} \tag{8}$$

and imposing

$$P(M_{\boldsymbol{\gamma}} : |\boldsymbol{\gamma}| = (k-1)/2) = P(M_{\boldsymbol{\gamma}'} : |\boldsymbol{\gamma}'| = (k+1)/2),$$

we arrive at

$$C_1^o = \frac{2(k+1)\binom{k}{(k-1)/2}}{2^k + (k+1)\binom{k}{(k-1)/2}}, \quad C_2^o = \frac{2^{k+1}}{2^k + (k+1)\binom{k}{(k-1)/2}}.$$

2.4. Half-p Jeffreys' prior probabilities

The Half-k prior seems sensible and objective but it is rather cumbersome. The analogue of this prior in terms of p – the prior inclusion probability – is to use a Uniform density on half of its parameter space, that is, on $[0, 1/2]$. We actually propose the slight modification

$$\pi(p) = \text{Un}\left(p \mid 0, \frac{1}{2}\frac{k+1}{k}\right),$$

since, for $k = 1$, this reduces to the original proposal, and in particular gives 0.5 probability to the null (and to the alternative). We call this prior Half-p.

The prior probability of a model is

$$P(M_{\boldsymbol{\gamma}}) = \frac{2k}{k+1}\int_0^{\frac{1}{2}\frac{k+1}{k}} p^{|\boldsymbol{\gamma}|}(1-p)^{k-|\boldsymbol{\gamma}|}\,\mathrm{d}p = \frac{2k}{k+1}\,\text{Beta}\left(\frac{1}{2}\frac{k+1}{k}, |\boldsymbol{\gamma}|+1, k-|\boldsymbol{\gamma}|+1\right), \tag{9}$$

which can be easily computed by means of the incomplete Beta function.

The Half-p is very similar to using $p \sim \text{Un}(0, 1/2)$, a prior recently proposed by Villa and Lee (2020) following a different reasoning (briefly summarized in Section 2.6). A small difference is that the Half-p assigns slightly more probability to complex models when k is large, being for instance e times larger for the full model as k tends to infinity.

2.5. Beta and hierarchical beta prior probabilities

An obvious possibility for the inclusion probability p is to use a Beta$(1, \alpha)$, with $\alpha > 1$, and $\alpha = 2$ seemingly giving a simple parsimonious choice. The Beta$(1, 2)$ choice results in prior model probabilities

$$P(M_{\boldsymbol{\gamma}}) = 2\int_0^1 p^{|\boldsymbol{\gamma}|}(1-p)^{1+k-|\boldsymbol{\gamma}|}\,\mathrm{d}p = 2\,\text{Beta}(|\boldsymbol{\gamma}|+1, k-|\boldsymbol{\gamma}|+2). \tag{10}$$

We also look at a hierarchical prior, assigning α the $\pi(\alpha) = \frac{1}{2}\alpha^{-3/2}$ prior density on $(1, \infty)$ (about as flat as possible while maintaining propriety and simplicity). The resulting prior for

Table 1. Names and defining equations of the eight priors considered.

Name	$P(M_\gamma)$	Name	$P(M_\gamma)$
Uniform	(2)	Jeffreys	(4)
Harmonic	(5)	CMG	(6)
Half-k	(7),(8)	Half-p	(9)
Beta(1,2)	(10)	Hierarchical Beta	(11)

p is

$$\pi^{\mathrm{HB}}(p) = \int_1^\infty \mathrm{Beta}(p \mid 1, \alpha)\frac{1}{2}\alpha^{-3/2}\,\mathrm{d}\alpha = \frac{\sqrt{\pi}\,\Phi\left(-\sqrt{-2\log(1-p)}\right)}{(1-p)\sqrt{(-\log(1-p))}}$$

and will be called the hierarchical Beta prior. The resulting prior probability of a model is

$$P(M_\gamma) = \int_0^1 p^{|\gamma|}(1-p)^{k-|\gamma|}\pi^{\mathrm{HB}}(p)\,\mathrm{d}p. \tag{11}$$

This requires numerical integration, but it only has to be done k times.

2.6. Overly parsimonious choices from the perspective of objectivity

Several authors have focussed on alternative distributions for p in (3); Wilson et al. (2010) proposed $\pi(p) = \mathrm{Beta}(p \mid 1, k)$ and Castillo et al. (2015) suggested $\pi(p) = \mathrm{Beta}(p \mid a, k^u)$, for certain $a > 0$ and $u > 1$. Since the means of these priors are decreasing at least as fast as $O(1/k)$, they force concentration on very small models, which we would not label as objective.

Other authors have worked directly with the probability assigned to all models with the same dimension, $\mathfrak{M}_k(d)$. Following this path, Womack et al. (2025) recently introduced the 'Matryoshka doll prior' that depends on a parameter η. This prior does not have a closed-form, but the authors prove that $\mathfrak{M}_k(d)$ is given by the Poisson law, $\mathrm{Po}(d \mid \log(1+\eta^{-1}))$ as its limiting distribution as k increases. Their default proposals are $\eta = 1$ and $\eta = 1/(\mathrm{e}-1)$. Thus, $\mathfrak{M}_k(d)$ behaves like $1/d!$, which is too sharply decreasing in model size to be labelled objective.

Finally, Villa and Lee (2020) proposed $P(M_\gamma) \propto \mathrm{e}^{-c|\gamma|}$ with $c = 1.2785$, but acknowledged that it does not provide needed parsimony. To fix this issue they suggested using $c = \log(k)$ or using a hyperprior on c. Following this last path, and after an interesting discussion, the authors arrive at a distribution on c which is equivalent to using $\pi(p) = \mathrm{Un}(p \mid 0, 1/2)$ that we already mentioned in a previous section.

3. Basic comparisons

With the model prior probabilities described in the previous section plus the Uniform and Jeffreys', we have a total of 8 different model prior probabilities. These are listed in Table 1 and their corresponding functions $\mathfrak{M}_k(d)$ are collected in Table 2, along with their expectations that will be needed later. We begin their comparison by looking at some essential properties.

Table 2. Expressions for $\mathfrak{M}_k(d)$ and the expectation of $d \sim \mathfrak{M}_k(d)$ for the various priors.

Prior	$\mathfrak{M}_k(d)$	$E^{\mathfrak{M}_k}(d)$
Uniform	$\binom{k}{d}/2^k$	$\frac{k}{2}$
Jeffreys	$1/(k+1)$	$\frac{k}{2}$
Beta(1, 2)	$\frac{2}{k+2}(1-\frac{d}{k+1})$	$\frac{k}{3}$
Half-p	$\binom{k}{d}\frac{2k}{k+1}\text{Beta}(\frac{1}{2}\frac{k+1}{k}, d+1, k-d+1)$	$k\frac{k+1}{4k} \approx \frac{k}{4}$
Hierarchical	$\binom{k}{d}\int_0^1 p^d(1-p)^{k-d}\pi^{\text{HB}}(p)\,\text{d}p$	$k\int_0^1 p\,\pi^{\text{HB}}(p)\,\text{d}p$
CMG	$\frac{\int_0^\infty \text{Po}(d\|\lambda)\pi^I(\lambda)\,\text{d}\lambda}{\sum_{d=0}^k \int_0^\infty \text{Po}(d\|\lambda)\pi^I(\lambda)\,\text{d}\lambda}$	$\frac{\int_0^\infty \text{Po}(d\|\lambda)\pi^I(\lambda)\,\text{d}\lambda}{\sum_{d=0}^k \int_0^\infty \text{Po}(d\|\lambda)\pi^I(\lambda)\,\text{d}\lambda} \approx \int_0^\infty \lambda\,\pi^I(\lambda)\,\text{d}\lambda$
Harmonic	$\frac{(1+d)^{-1}}{H_{k+1}}$	$\frac{k+1}{H_{k+1}} - 1 \approx \frac{k}{\log(k)}$
Half-k (k even)	$\begin{cases} C_1^e \frac{1}{(k+1)}, & \text{if } d \le k/2, \\ C_2^e \binom{k}{d}\frac{k}{2(k+1)N_k}, & \text{if } d \ge k/2+1. \end{cases}$	$k\frac{2^{k-1}+\frac{k+1}{4}\binom{k}{k/2}}{2^k+(k+1)\binom{k}{k/2}} \approx \frac{k}{4}$
Half-k (k odd)	$\begin{cases} C_1^o \frac{1}{(k+1)}, & \text{if } d \le (k-1)/2, \\ C_2^o \binom{k}{d}\frac{1}{2N_k}, & \text{if } d \ge (k+1)/2. \end{cases}$	$k\frac{2^{k-1}+\frac{k^2-1}{4k}\binom{k}{(k-1)/2}}{2^k+(k+1)\binom{k}{(k-1)/2}} \approx \frac{k}{4}$

Notes: $\approx$ indicates the first order term in the large k approximation to the expectation. The computation of $E^{\mathfrak{M}_k}(d)$ for the mixture over p priors follows from the straightforward identity $E^{\mathfrak{M}_k}(d) = kE^{\pi}(p)$. The derivation of the large k approximations is straightforward, often using Stirling's approximation. Numerical integration yields $\int_0^\infty \lambda\,\pi^I(\lambda)\,\text{d}\lambda \cong 1.5$ and $\int_0^1 p\,\pi^{\text{HB}}(p)\,\text{d}p \cong 0.22$.

3.1. When k = 1

If $k = 1$ then it seems natural to require $P(M_0) = P(M_1) = 1/2$, the default choice for hypothesis testing. The Uniform, Jeffreys, Half-p and Half-k prior probabilities satisfy this. The rest do assign significant mass to the null model: CMG, Harmonic and Beta(1,2) assign 2/3 to M_0 and the Hierarchical 0.78.

3.2. Simplicity

All the model prior probabilities are in closed form, except the Hierarchical and CMG priors, which require evaluation of k one-dimensional integrals. The Half-k prior is rather cumbersome and is thus somewhat less appealing.

3.3. Graphs of the priors

Figure 1 contains two graphs of the prior model probabilities for a total number of covariates $k = 47$ (the number of explanatory variables in the infant obesity dataset of Section A.2). The left graph gives the log of the model prior probability for a model of the indicated size d while the right graph gives $\mathfrak{M}_k(d)$.

In this graph, we observe that the CMG prior kills off all models of dimension bigger than 14; the Half-k and Half-p priors kill off models of size bigger than 34, while the hierarchical, Beta(1,2) and Harmonic choices do give non-negligible mass to all model sizes. Interestingly, all the prior probability choices (except the Uniform) clearly give more mass to smaller models than does the Jeffreys' prior. We will use these plots to illustrate the general arguments in the next sections.

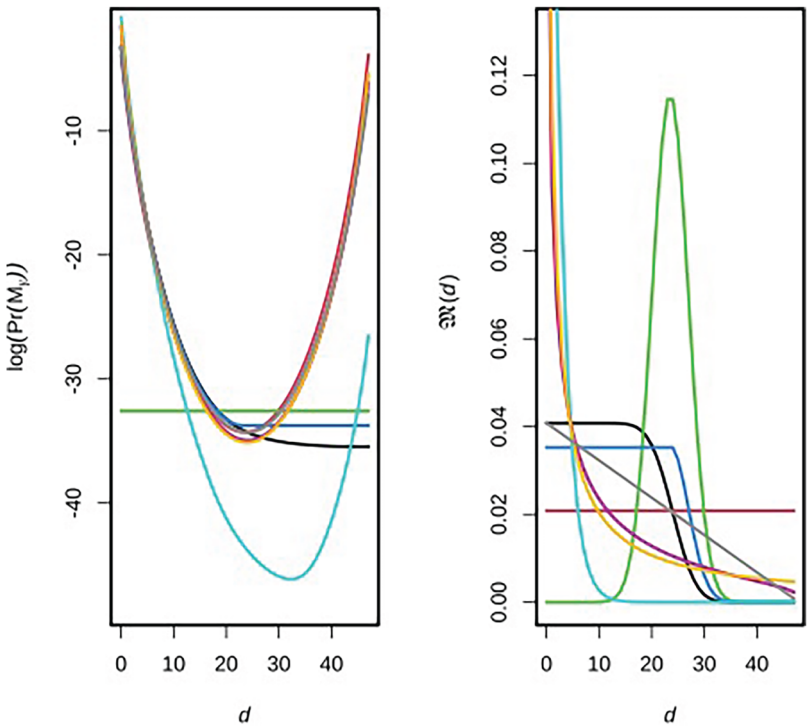


Figure 1. For $k = 47$, left: $\log(P(M_{\boldsymbol{\gamma}}))$ with $|\boldsymbol{\gamma}| = d$; right: $\mathfrak{M}_k(d)$. Priors: Half-p (black); Jeffreys (red); Uniform (green); Half-k (blue); CMG (cyan); Hierarchical (magenta); Beta(1,2) (grey); Harmonic (yellow).

4. Two notions of parsimony

4.1. Weak parsimony

It is natural to require that

$$\mathfrak{M}_k(d+1) \leq \mathfrak{M}_k(d) \quad \text{for all } d, \tag{12}$$

so that the sum of model prior probabilities of a given size is non-increasing in size. We call this *weak parsimony (WP)*.

The right graph in Figure 1 illustrates this, with all priors considered (except the Uniform) satisfying WP. Inspection of Table 2 confirms that the Jeffreys, Beta(1,2), Harmonic and Half-k priors always satisfy WP. For priors that are obtained as mixtures over p or over the Poisson mass function, the next two results show that non-increasing mixture functions produce priors that are WP. This confirms that the CMG, Half-p and Hierarchical prior probabilities are WP.

Lemma 4.1: *Weak parsimony (WP) holds for priors of the form* (3) *with* $\pi(p)$ *non-increasing on* $[0, 1]$.

Proof: In this case (12) reduces to showing that

$$\binom{k}{d+1} \int_0^1 p^{d+1}(1-p)^{k-d-1}\pi(p)\,\mathrm{d}p \leq \binom{k}{d} \int_0^1 p^{d}(1-p)^{k-d}\pi(p)\,\mathrm{d}p,$$

for all $d \in \{0, \ldots, k-1\}$. After straightforward algebra we obtain that this is equivalent to

$$\int_0^1 p^d (1-p)^{k-d-1}\, p\, \pi(p)\, \mathrm{d}p \leq \frac{d+1}{k+1} \int_0^1 p^d (1-p)^{k-d-1}\, \pi(p)\, \mathrm{d}p. \tag{13}$$

To show that this inequality holds under the conditions of the result, we use Chebyshev's integral inequality, which states that, for f, g integrable functions with one non-increasing and the other increasing, and for a probability density function μ,

$$\int_0^1 f(p)\, g(p)\, \mu(p)\, \mathrm{d}p \leq \int_0^1 f(p)\, \mu(p)\, \mathrm{d}p \int_0^1 g(p)\, \mu(p)\, \mathrm{d}p. \tag{14}$$

Choosing $f(p) = p$ (increasing), $g(p) = \pi(p)$ (non-increasing by hypothesis) and $\mu(p) = \text{Beta}(p \mid d+1, k-d)$ results in the inequality (13). ■

Lemma 4.2: *Weak parsimony (WP) holds for priors of the form* (6)*, with* $\pi(\lambda)$ *non-increasing on* $[0, \infty)$.

Proof: In this case (12) reduces to showing that

$$\int_0^\infty \frac{\mathrm{e}^{-\lambda} \lambda^{d+1}}{(d+1)!} \pi(\lambda)\, \mathrm{d}\lambda \leq \int_0^\infty \frac{\mathrm{e}^{-\lambda} \lambda^{d}}{d!} \pi(\lambda)\, \mathrm{d}\lambda,$$

for all $d \in \{0, \ldots, k-1\}$, or equivalently

$$\int_0^\infty \mathrm{e}^{-\lambda} \lambda^d\, \lambda\, \pi(\lambda)\, \mathrm{d}\lambda \leq (d+1) \int_0^\infty \mathrm{e}^{-\lambda} \lambda^d\, \pi(\lambda)\, \mathrm{d}\lambda.$$

Using Chebyshev's integral inequality (14) with $f(\lambda) = \lambda$ (increasing), $g(\lambda) = \pi(\lambda)$ (non-increasing by hypothesis), and $\mu(\lambda) = \text{Gamma}(\lambda \mid d+1, 1)$ establishes the result. ■

To conclude that the CMG prior satisfies weak parsimony we show that

$$\pi(\lambda) = \frac{(\mathrm{e}^{2\sqrt{\lambda}} + \mathrm{e}^{-2\sqrt{\lambda}})\mathrm{e}^{-\lambda}}{2\mathrm{e}\sqrt{\pi}\sqrt{\lambda}}$$

is decreasing. The derivative of $\log \pi(\lambda)$ is

$$\frac{(\mathrm{e}^{2\sqrt{\lambda}} - \mathrm{e}^{-2\sqrt{\lambda}})}{\sqrt{\lambda}\,(\mathrm{e}^{2\sqrt{\lambda}} + \mathrm{e}^{-2\sqrt{\lambda}})} - 1 - \frac{1}{2\lambda} < \frac{1}{\sqrt{\lambda}} - 1 - \frac{1}{2\lambda} = -\frac{1}{\lambda}(\lambda - \sqrt{\lambda} + \frac{1}{2}).$$

$(\lambda - \sqrt{\lambda} + \frac{1}{2})$ is a quadratic in $\sqrt{\lambda}$ that is easily shown to have no real roots, and hence is always positive. The conclusion follows.

4.2. Strong parsimony

We define *strong parsimony (SP)* as the requirement that individual model prior probabilities should be non-increasing as model size grows, that is

$$P(M_{\boldsymbol{\gamma}:|\boldsymbol{\gamma}|=d+1}) \leq P(M_{\boldsymbol{\gamma}:|\boldsymbol{\gamma}|=d}) \quad \text{for all } d. \tag{15}$$

The left plot in Figure 1 illustrates this for the case $k = 47$. The Uniform prior obviously satisfies SP (since it is constant) and, among the rest, only the Half-k and Half-p prior probabilities

are candidates to satisfy SP. Looking at the expressions (7) and (8) we straightforwardly conclude that Half-k satisfies SP. The result for the Half-p is more involved. Obviously, mixtures of the type (3) with non-increasing $\pi(p)$ do not necessarily guarantee SP, since the Jeffreys' prior violates strong parsimony. The following result establishes a characterization of SP for this class of priors.

Lemma 4.3: *Strong parsimony (SP) holds for priors of the form* (3) *if and only if*

$$\int_0^1 p^{k-1}(1-2p)\pi(p)\,\mathrm{d}p \geq 0. \tag{16}$$

Proof: Notice that

$$P(M_{\boldsymbol{\gamma}:|\boldsymbol{\gamma}|=d}) = \int_0^1 \left(\frac{p}{1-p}\right)^d (1-p)^k \pi(p)\,\mathrm{d}p.$$

Applying the Cauchy-Schwarz inequality yields

$$\begin{aligned}
P(M_{\boldsymbol{\gamma}:|\boldsymbol{\gamma}|=d})^2 &= \left(\int_0^1 \left(\frac{p}{1-p}\right)^d (1-p)^k \pi(p)\,\mathrm{d}p\right)^2 \\
&= \left(\int_0^1 \left(\frac{p}{1-p}\right)^{(d+1)/2} (1-p)^{k/2} \left(\frac{p}{1-p}\right)^{(d-1)/2} (1-p)^{k/2} \pi(p)\,\mathrm{d}p\right)^2 \\
&\leq \int_0^1 \left(\frac{p}{1-p}\right)^{d+1} (1-p)^k \pi(p)\,\mathrm{d}p \int_0^1 \left(\frac{p}{1-p}\right)^{d-1} (1-p)^k \pi(p)\,\mathrm{d}p \\
&= P(M_{\boldsymbol{\gamma}:|\boldsymbol{\gamma}|=d+1}) P(M_{\boldsymbol{\gamma}:|\boldsymbol{\gamma}|=d-1}).
\end{aligned}$$

Equivalently, consecutive ratios are non-decreasing

$$\frac{P(M_{\boldsymbol{\gamma}:|\boldsymbol{\gamma}|=1})}{P(M_{\boldsymbol{\gamma}:|\boldsymbol{\gamma}|=0})} \leq \frac{P(M_{\boldsymbol{\gamma}:|\boldsymbol{\gamma}|=2})}{P(M_{\boldsymbol{\gamma}:|\boldsymbol{\gamma}|=1})} \leq \cdots \leq \frac{P(M_{\boldsymbol{\gamma}:|\boldsymbol{\gamma}|=k})}{P(M_{\boldsymbol{\gamma}:|\boldsymbol{\gamma}|=k-1})}.$$

The above inequalities allow us to assert that SP holds if and only if the largest consecutive ratio is at most one, namely

$$\frac{P(M_{\boldsymbol{\gamma}:|\boldsymbol{\gamma}|=k})}{P(M_{\boldsymbol{\gamma}:|\boldsymbol{\gamma}|=k-1})} \leq 1.$$

Algebra shows that this is equivalent to (16). ■

A consequence of the above lemma is that the Half-p satisfies SP since

$$\int_0^1 p^{k-1}(1-2p)\,\mathrm{Un}\left(p \mid 0, \frac{1}{2}\frac{k+1}{k}\right) = 0.$$

For $p \sim \mathrm{Beta}(a,b)$, (16) can be shown to hold only if $b - a \geq k-1$. However, these priors do not satisfy WP and, for increasing a and b, lead to degenerate priors (e.g., the prior with

$a = k$, $b = 2k-1$ converges to a point mass at $p = 1/2$ leading to the Uniform assignment of probabilities).

An interesting observation arises when we compare both types of parsimony. It can be easily shown that, if SP holds, then (12) is satisfied for $d \in \{0, \ldots, k/2 - 1\}$ if k is even and for $d \in \{0, \ldots, (k-1)/2\}$ if k is odd. The following result focuses on the remaining model sizes, showing that imposing non-increasing individual model prior probabilities forces the collective probabilities to decrease at a fast rate.

Lemma 4.4: *Suppose k is even and* (15) *holds for all $d \in \{k/2, \ldots, k-1\}$. Then*

$$\mathfrak{M}_k(d+1) \leq \mathfrak{M}_k\left(\frac{k}{2}\right) \frac{\binom{k/2}{d+1-k/2}}{\binom{d+1}{d+1-k/2}}, \tag{17}$$

for all $d \in \{k/2, \ldots, k-1\}$.

If k is odd and (15) *holds for all $d \in \{(k+1)/2, \ldots, k-1\}$, then*

$$\mathfrak{M}_k(d+1) \leq \mathfrak{M}_k\left(\frac{k+1}{2}\right) \frac{\binom{(k-1)/2}{d-(k-1)/2}}{\binom{d+1}{d-(k-1)/2}},$$

for all $d \in \{(k-1)/2, \ldots, k-1\}$.

Proof: The proof follows by induction. Suppose k is even. Under the conditions of the lemma, first notice that (17) is trivially true for $d = k/2$. Supposing that it is satisfied for some d in the interval $[k/2+1, k-2]$, we now show that it is also satisfied for $d+1$:

$$\mathfrak{M}_k(d+2) = P(M_{\boldsymbol{\gamma}:|\boldsymbol{\gamma}|=d+2})\binom{k}{d+2} \leq P(M_{\boldsymbol{\gamma}:|\boldsymbol{\gamma}|=d+1})\binom{k}{d+2} = \frac{\mathfrak{M}_k(d+1)}{\binom{k}{d+1}}\binom{k}{d+2}$$

$$\leq \mathfrak{M}_k\left(\frac{k}{2}\right) \frac{\binom{k/2}{d+1-k/2}}{\binom{d+1}{d+1-k/2}} \frac{\binom{k}{d+2}}{\binom{k}{d+1}} = \mathfrak{M}_k\left(\frac{k}{2}\right) \frac{\binom{k/2}{d+2-k/2}}{\binom{d+2}{d+2-k/2}}.$$

The proof for k odd is similar and is omitted. ■

Based on the result above, a necessary condition for SP is $\mathfrak{M}_k(k) \leq \mathfrak{M}_k(k/2)/\binom{k}{k/2}$ and, since $\mathfrak{M}_k(k/2) \leq 1$, we have $\mathfrak{M}_k(k) \leq 1/\binom{k}{k/2} = O(2^{-k}\sqrt{k})$. Since $\mathfrak{M}_k(k) = P(M_k)$, this means that the probability of the full model can be at most $O(2^{-k}\sqrt{k})$, a very small number. Returning to the illustrative case with $k = 47$, the fact that the Half-k and Half-p priors essentially kill off all models of size greater than 34 is in fact a consequence of SP.

5. Large *k* behaviour and the control for multiplicity

5.1. Large k behaviour

In this section we focus on the behaviour of model priors in particular model sizes: $d = 0$, 1 (small models), $d = k/2$ (medium sized model) and $d = k$ (full model), to understand extreme behaviours. Table 3 gives values of $\mathfrak{M}_k(d)$ at these values of d for the various prior probabilities. Of particular interest is the situation where k grows. We have included large k

Table 3. Values of $\mathfrak{M}_k(d)$ for the various priors and specified values of d.

d	Uniform	Jeffreys	Beta(1,2)	Half-p	Hierarchical	CMG	Harmonic	Half-k
0	$= 2^{-k}$	$= \frac{1}{k+1}$	$= \frac{2}{k+2}$	$\approx \frac{2}{k}$	$\approx \frac{\pi}{2\sqrt{k}}$	≈ 0.43	$\approx \frac{1}{\log(k+1)}$	$\approx \frac{2}{k}$
1	$= k2^{-k}$	$= \frac{1}{k+1}$	$= \frac{2k}{(k+1)(k+2)}$	$\approx \frac{2}{k}$	$\approx \frac{\pi}{4\sqrt{k}}$	≈ 0.21	$\approx \frac{1}{2\log(k+1)}$	$\approx \frac{2}{k}$
$k/2$	$\approx \sqrt{\frac{2}{k\pi}}$	$= \frac{1}{k+1}$	$= \frac{1}{k+1}$	$\approx \frac{1}{k}$	$\approx \frac{0.51}{k}$	$< O(\frac{ke^{\sqrt{k}}}{2^{k/2}})$	$\approx \frac{2}{k\log k}$	$\approx \frac{2}{k}$
k	$= 2^{-k}$	$= \frac{1}{k+1}$	$= \frac{2}{(k+1)(k+2)}$	$\approx \frac{e}{k2^k}$	$\approx \frac{1}{2k\log k}$	$< O(\frac{ke^{\sqrt{2k}}}{2^k})$	$\approx \frac{1}{k\log k}$	$\approx \sqrt{\frac{2\pi}{k}}\frac{1}{2^k}$

Note: $\approx$ indicates that the ratio of the table entry and $\mathfrak{M}_k(d)$ goes to 1 as k grows.

approximations to $\mathfrak{M}_k(d)$. The validity of these approximations is proved in the following result.

Lemma 5.1: *The values of $\mathfrak{M}_k(d)$ in Table 3 are, as indicated, exact, given as an upper bound, or approximated ($\approx$) in the sense that the ratio of the table entry and $\mathfrak{M}_k(d)$ goes to 1 as k increases.*

***Proof*:** For Jeffreys' prior, $\mathfrak{M}_k(d) = 1/(k+1)$ for all d. For the Uniform prior, $\mathfrak{M}_k(d) = \binom{k}{d}/2^k$ and Stirling's approximation yields the result for $d = k/2$.

For the Half-p,

$$\mathfrak{M}_k(0) = \frac{2k}{k+1}\,\text{Beta}\left(\frac{1}{2}\frac{k+1}{k}, 1, k+1\right) \approx \frac{2}{k},$$

where the approximation of the integral follows from explicitly evaluating the integral corresponding to the incomplete Beta function and then using standard limiting arguments. Similarly

$$\begin{aligned}\mathfrak{M}_k(1) &= k\,\frac{2k}{k+1}\,\text{Beta}\left(\frac{1}{2}\frac{k+1}{k}, 2, k\right)\\ &= \frac{2k^2}{k+1}\left(\frac{1}{k(k+1)} - \frac{1}{k}\left(\frac{k-1}{2k}\right)^k + \frac{1}{k+1}\left(\frac{k-1}{2k}\right)^{k+1}\right)\\ &\approx \frac{2k^2}{k+1}\frac{1}{k(k+1)} \approx \frac{2}{k}.\end{aligned}$$

For $d = k/2$,

$$\begin{aligned}\mathfrak{M}_k(k/2) &= \binom{k}{k/2}\frac{2k}{k+1}\int_0^{\frac{k+1}{2k}} \big(p(1-p)\big)^{k/2}\,\mathrm{d}p\\ &= \binom{k}{k/2}\frac{2k}{k+1}\left(\int_0^{1/2} \big(p(1-p)\big)^{k/2}\,\mathrm{d}p + \int_{1/2}^{\frac{k+1}{2k}} \big(p(1-p)\big)^{k/2}\,\mathrm{d}p\right).\end{aligned}$$

The integrand is symmetric with respect to $p = 1/2$ and decreasing with $p \in [1/2, 1]$ and hence

$$\int_0^{1/2} \big(p(1-p)\big)^{k/2}\,\mathrm{d}p = \frac{1}{2}\,\text{Beta}\,(k/2+1, k/2+1) \approx \sqrt{\frac{2\pi}{k+2}}2^{-k-2}$$

(applying Stirling's approximation), and

$$\int_{1/2}^{\frac{k+1}{2k}} \left(p(1-p)\right)^{k/2} \mathrm{d}p \leq \left(\frac{1}{4}\right)^{k/2} \frac{1}{2}\left(\frac{k+1}{k} - 1\right).$$

It follows that

$$\begin{aligned}
\mathfrak{M}_k(k/2) &= \binom{k}{k/2} \frac{2k}{k+1}\left(\frac{1}{2}\,\mathrm{Beta}\,(k/2+1, k/2+1) + O\left(\frac{2^{-k}}{k}\right)\right) \\
&\approx \binom{k}{k/2} \frac{2k}{k+1}\left(\sqrt{\frac{2\pi}{k+2}}2^{-k-2} + O\left(\frac{2^{-k}}{k}\right)\right) \\
&\approx \binom{k}{k/2} \frac{2k}{k+1}\left(\sqrt{\frac{2\pi}{k+2}}2^{-k-2}\right) \approx \frac{1}{k},
\end{aligned}$$

since $\binom{k}{k/2} \approx \sqrt{\frac{2}{\pi k}}2^k$ by Stirling's approximation. Finally, the approximation for $\mathfrak{M}_k(k)$ follows by exact integration and standard asymptotic results.

To obtain the results concerning the Hierarchical Beta, we start by deriving a useful result for the mixture over p priors and if k and d are large,

$$\begin{aligned}
\mathfrak{M}_k(d) &= \binom{k}{d} \int_0^1 p^d (1-p)^{k-d} \pi(p)\,\mathrm{d}p \\
&= \frac{1}{k+1}\int_0^1 \mathrm{Beta}(p \mid d+1, k-d+1)\pi(p)\,\mathrm{d}p \approx \frac{\pi(d/k)}{k}
\end{aligned} \tag{18}$$

if $\pi(\nu)$ is continuous and nonzero. This follows because the mean of the Beta distribution is approximately $\mu = \frac{d}{k}$ and the standard deviation is smaller than $\frac{\mu}{\sqrt{d}}$, so the Beta distribution is concentrated around the mean. Hence $\mathfrak{M}_k(d) = \binom{k}{d}\int_0^1 p^d(1-p)^{k-d}\pi^{\mathrm{HB}}(p)\,\mathrm{d}p \approx \frac{\pi^{\mathrm{HB}}(\frac{d}{k})}{k}$ for large d and d/k bounded away from zero (necessary because π^{HB} is infinite at zero). Note that, as $\frac{d}{k} \to 1$, $\pi^{\mathrm{HB}}(\frac{d}{k}) \approx \frac{1}{-2\log(1-\frac{d}{k})}$. Similarly, $\mathfrak{M}_k(k/2) \approx \frac{\pi^{\mathrm{HB}}(0.5)}{k} \approx \frac{0.51}{k}$.

The results for the Harmonic follow from $\mathfrak{M}_k(d) = \frac{(1+d)^{-1}}{H_{k+1}} \approx \frac{(1+d)^{-1}}{\log(k+1)}$ if k is large.

The approximations for the Half-k are obtained by noticing that $C_1^e \approx 2$ and $C_2^e \approx \frac{\sqrt{2\pi}}{k}$, and using similar standard large k approximations to the remaining expressions.

The analysis for the CMG prior, when $d = k$, begins with the bound

$$\int_0^\infty \mathrm{Po}(k \mid \lambda)\pi^I(\lambda)\,\mathrm{d}\lambda < \frac{1}{k!}\int_0^\infty \lambda^{k-0.5}\mathrm{e}^{-2(\lambda-\sqrt{\lambda})}\,\mathrm{d}\lambda,$$

which follows from $\cosh(2\sqrt{\lambda}) < \mathrm{e}^{2\sqrt{\lambda}}$. Making the change of variables $\lambda = x^2$ results in the expression

$$\frac{1}{k!}\int_0^\infty x^{2k}\mathrm{e}^{-2(x^2-x)}\,\mathrm{d}x < \frac{1}{k!}\int_{-\infty}^\infty x^{2k}\mathrm{e}^{-2(x-0.5)^2}\,\mathrm{d}x = \frac{1}{k!}\sum_{j=0}^k \binom{2k}{2j}\frac{(2j)!}{2^{2k+j}j!},$$

the last step following from the expression for the $2k$-th moment of a normal distribution with mean 1/2 and variance 1/4. Using Stirling's approximation and dropping a term $\sqrt{k-j}$ in the

denominators (the results being larger) yields a summation whose terms can be maximized over j, the maximum term being $O(e^{\sqrt{2k}}/2^k)$. Since there are only k terms in the summation and the denominator for the expression for the CMG in Table 3 can be shown to converge to 1 as k increases, the result follows. The same argument works for $d = k/2$. ■

5.2. Evaluation of models based on large k behaviour

5.2.1. Full and null model

Looking at Table 3, it can be argued that the Uniform, Half-p, and Half-k priors give too little probability to the full model. In fact, these are more conservative (in support of the full model) than the bound $O(2^{-k}\sqrt{k})$ that we showed is a necessary condition for SP (Half-p being the most conservative, then Half-k and then the Uniform). The CMG prior can also be argued to give too little probability to the full model, even though the table entry does not violate the SP condition; this is almost certainly due to the entry being too large an upper bound, since the numerical results all suggested that the CMG prior penalizes the full model far more than the others. The Hierarchical and Harmonic priors give essentially the same weight to the full model and are slightly less conservative than the Beta(1,2) prior.

For the null model and the models with just one covariate, the CMG prior gives by far the most weight to small models. The Harmonic prior and then the Hierarchical Beta prior are next in terms of favouring small models. Jeffreys', Beta(1,2), Half-p and Half-k behave similarly in support for small models and the Uniform provides the least weight to small models.

5.2.2. Midsize models

The most populated model size is $d = k/2$ and, hence, $\mathfrak{M}_k(k/2)$ is intuitively an indicator of multiplicity adjustment; if not controlled by multiplicity adjustment, the midsize models will dominate in terms of prior probabilities (and often posterior probabilities). It was shown in Scott and Berger (2010) that the Uniform (with $\mathfrak{M}_k(k/2) \approx (2/k\pi)^{1/2}$) does not correct for multiplicity, while the Jeffreys' prior, with $\mathfrak{M}_k(k/2) = 1/(k+1)$ does. Taking the Jeffreys' prior as a standard for multiplicity control, we can evaluate the degree of multiplicity control provided by the rest of the proposals: priors with $\mathfrak{M}_k(k/2) \approx \frac{C}{k}$ provide a Jeffreys' type adjustment for multiplicity. This is the case for all the mixture over p priors and the Half-k. The Harmonic prior has an additional $1/\log k$ penalty while the CMG probabilities are essentially zero, killing all models of midsize.

5.3. Combining parsimony and control for multiplicity

The previous notion of Jeffreys' adjustment for multiplicity combined with strong parsimony (SP) provides an interesting characterization of model prior probabilities. According to Lemma 4.4, for priors with $\mathfrak{M}_k(k/2) \approx C\frac{1}{k}$ a necessary condition for SP becomes $\mathfrak{M}_k(k) = O(2^{-k}/\sqrt{k})$. Thus a prior satisfying $\mathfrak{M}_k(k) \approx C2^{-k}/\sqrt{k}$ would be the least parsimonious among the class of priors that satisfy SP and that provide a Jeffreys' type adjustment for multiplicity. Among the priors studied here, only the Half-k has this property.

Table 4. Prior variable inclusion probabilities for the various priors.

k	Uniform	Jeffreys	Beta(1,2)	Half-p	Hierarchical	CMG	Harmonic	Half-k
1	0.50	0.50	0.33	0.50	0.22	0.33	0.33	0.50
3	0.50	0.50	0.33	0.33	0.22	0.57	0.31	0.40
5	0.50	0.50	0.33	0.30	0.22	0.43	0.29	0.37
7	0.50	0.50	0.33	0.29	0.22	0.34	0.28	0.35
9	0.50	0.50	0.33	0.28	0.22	0.28	0.27	0.34
20	0.50	0.50	0.33	0.26	0.22	0.13	0.24	0.31
200	0.50	0.50	0.33	0.25	0.22	0.01	0.17	0.27
∞	0.50	0.50	0.33	0.25	0.22	0	0	0.25

6. Prior variable inclusion probabilities

The prior inclusion probability for a variable $x_j, j \in \{1, \ldots, k\}$ is

$$P(x_j) := P(\cup\{M_{\boldsymbol{\gamma}} : \gamma_j = 1\}),$$

i.e., the overall prior probability that the given variable occurs in models. The following result relates these probabilities to the expectation of $d \sim \mathfrak{M}_k(d)$.

Lemma 6.1: *For priors satisfying* (1), $P(x_j) = E^{\mathfrak{M}_k}(d)/k$.

Proof: Condition on the set of models of size d. Without loss of generality choose a model by sequential selection of variables. Because of exchangeability, the probability that variable x_j is chosen first is $1/k$. Indeed, the probability that x_j is chosen in any position is $1/k$. Since these are mutually exclusive events, the overall conditional inclusion probability of x_j is d/k. Probabilistically averaging over d gives the result. ■

Table 2 presents $E^{\mathfrak{M}_k}(d)$ for the priors, together with their large k approximations. For the mixture over p priors and since $E^{\mathfrak{M}_k}(d) = k\,E^{\pi}(p)$, it is immediate that $P(x_j) = E^{\pi}(p)$. Table 4 gives the prior inclusion probabilities for various values of k.

Initial thoughts were that objectivity would imply that the prior variable inclusion probability should be 0.5. We label this property Hard objectivity (HO). Of the priors studied, this is only satisfied by the Uniform and Jeffreys' priors, with the other priors having $P(x_j)$ strictly smaller than 0.5. Thus, having a prior inclusion probability of 0.5 seems to be incompatible with weak parsimony (WP). In Lemma 4.1, we showed that $\pi(p)$ being non-increasing guarantees WP but these priors have an expectation bounded above by 0.5 (with $p \sim \text{Un}(0, 1)$ being the limiting case).

Slightly less demanding is having a strictly positive limiting value for $P(x_j)$, as k grows. We name this characteristic Soft objectivity (SO). This property is satisfied by all priors studied except the CMG and Harmonic priors.

Our view is that WP is considerably more important than HO, so we would still rule out the Uniform and Jeffreys' prior probabilities. The situation with SO is not so clear; arguments can be made both ways. If one is including only variables that are viewed as being of legitimate scientific interest, then adding more legitimate variables should not cause the prior variable inclusion probabilities to decrease. But, if one is including lots of 'junk' variables, then it would be natural to have the prior inclusion probabilities decrease.

Table 5. Desirable properties of prior probability assignments for the various priors studied, with ✓ indicating that the prior has the property and ✗ indicating that it does not.

	Uniform	Jeffreys	Beta(1,2)	Hierarchical	CMG	Harmonic	Half-p	Half-k
Philosophical considerations								
1. Hypothesis testing	✓	✓	✗	✗	✗	✗	✓	✓
2. Simplicity	✓	✓	✓	✗	✗	✓	✓	✗
Parsimony								
3. Weak (WP)	✗	✓	✓	✓	✓	✓	✓	✓
4. Strong (SP)	✓	✗	✗	✗	✗	✗	✓	✓
Multiplicity								
5. Control	✗	✓	✓	✓	✓	✓	✓	✓
6. Jeffreys' Control	✗	✓	✓	✓	✗	✗	✓	✓
Objectivity in inclusion probabilities								
7. Hard (HO)	✓	✓	✗	✗	✗	✗	✗	✗
8. Soft (SO)	✓	✓	✓	✓	✗	✗	✓	✓

7. Conclusions

Our goal, because of the severe problems with the Uniform and Jeffreys' prior probabilities, was to find an objective prior distribution over model space that could be generally recommended. Unfortunately, we were not able to arrive at a definitive conclusion.

In Table 5 we have collected the main criteria for prior probabilities that have been studied in this paper. One could peruse this list and decide what is optimal in a given situation, but that is not realistic. So we provide our recommendations.

We definitely do not recommend the Uniform prior because it violates WP and multiplicity control. We do not recommend the Jeffreys' prior because it badly violates SP, with the full model having the largest prior probability (which is shown in the Appendix to possibly and erroneously lead to the full model having the highest posterior probability). We definitely do not recommend the CMG prior because it is much too parsimonious to be labelled objective.

For the remaining priors, the biggest difference is SP, with the Half-p and Half-k priors satisfying SP and the rest not. This would seem to suggest choosing between the two, with the Half-p prior perhaps winning because of simplicity. But recall that these two priors might be viewed as too parsimonious; they kill off larger models, with the full model having only probability $O(2^{-k}/k)$ and $O(2^{-k}/\sqrt{k})$, respectively.

If one does not wish to so severely penalize larger models, the remaining choices are the Beta(1,2), Hierarchical and Harmonic priors, all of which have very similar properties, except that the Harmonic does not satisfy SO. As we argued in the previous section, there are situations where not wanting SO is natural, and then the Harmonic prior would be the choice. If one does want SO, the choice would be between the Beta(1,2) and Hierarchical priors. The Beta(1,2) prior obviously has simplicity in its favour, but has the somewhat negative feature of more highly favouring mid-size models in Table 3.

In conclusion, if one desires SP and is okay with larger models being rapidly killed off, choose the Half-p prior. If one does not want to rapidly kill off large models but wants the variable inclusion probabilities to decline as k grows, choose the Harmonic prior. If one does not want to rapidly kill off large models and doesn't want the inclusion probabilities to go to 0 as k grows, use the Beta(1,2) or Hierarchical priors (with the Beta(1,2) being simpler).

As a final comment, note that there is a benefit in not having a firm conclusion as to the choice of model prior probabilities; one can do a sensitivity or robustness study over the possible choices, gaining considerable confidence if the answer does not vary much. Indeed, in the example studied in the Appendix, it is observed that there is considerable robustness

to the choice of the prior model probabilities if the *median probability model* (defined in the Appendix) is utilized as the final model.

Acknowledgments

We are grateful to Elias Moreno for helpful discussions on the material in the paper and to reviewers for excellent suggestions to improve the manuscript. Note that no generative AI tool was used in the creation of this manuscript.

Author contributions

CRediT: **James Berger:** Methodology, Writing – original draft; **Gonzalo García-Donato:** Methodology, Writing – original draft; **Luis Pericchi:** Methodology

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Agencia Estatal de Investigacion (AEI, Spain) [grant numbers PID2022-138201NB-I00 and PID2025-170720NB-I00].

ORCID

James Berger http://orcid.org/0000-0002-6413-2876

References

Berger, J., García-Donato, G., Moreno, E., & Pericchi, L. (2024). Objective Bayesian testing and model uncertainty. In *Handbook of Bayesian, Fiducial, and Frequentist Inference*. Chapman and Hall/CRC.

Berger, J., & Pericchi, L. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, *91*(433), 109–122. https://doi.org/10.1080/01621459.1996.10476668

Berger, J. O., Wang, X., & Shen, L. (2014). A Bayesian approach to subgroup identification. *Journal of Biopharmaceutical Statistics*, *24*(1), 110–129. https://doi.org/10.1080/10543406.2013.856026

Casella, G., Moreno, E., & Girón, F. J. (2014). Cluster analysis, model selection and prior distributions on models. *Bayesian Analysis*, *9*(3), 613–658. https://doi.org/10.1214/14-BA869

Castillo, I., Schmidt-Hieber, J., & van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics*, *43*(5), 1986–2018. https://doi.org/10.1214/15-AOS1334

Chipman, H. (1996). Bayesian variable selection with related predictors. *The Canadian Journal of Statistics*, *24*(1), 17–36. https://doi.org/10.2307/3315687

Clyde, M., DeSimone, H., & Parmigiani, G. (1996). Prediction via orthogonalized model mixing. *Journal of the American Statistical Association*, *91*(435), 1197–1208. https://doi.org/10.1080/01621459.1996.10476989

Furnival, G. M., & Wilson R. W. (1974). Regressions by leaps and bounds. *Technometrics*, *16*(4), 499–511. https://doi.org/10.1080/00401706.1974.10489231

Garca-Donato, G., & Martínez-Beneito, M. A. (2013). On sampling strategies in Bayesian variable selection problems with large model spaces. *Journal of the American Statistical Association*, *108*(501), 340–352. https://doi.org/10.1080/01621459.2012.742443

García-Donato, G., & Forte, A. (2018). Bayesian testing, variable selection and model averaging in linear models using R with BayesVarSel. *The R Journal*, *10*(1), 155–174. https://doi.org/10.32614/RJ-2018-021

García-Donato, G., & Paulo, R. (2022). Variable selection in the presence of factors: A model selection perspective. *Journal of the American Statistical Association*, *117*(540), 1847–1857. https://doi.org/10.1080/01621459.2021.1889565

García-Donato, G., & Steel, M. (2021). Bayes factors based on g-priors for variable selection. In M. G. Tadesse & M. Vanucci (Eds.), *Handbook of Bayesian Variable Selection* (Chap. 15). Chapman and Hall/CRC.

George, E. I., & McCulloch, R. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, *7*(2), 339–374.

George, E. I., & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, *88*(423), 881–889. https://doi.org/10.1080/01621459.1993.10476353

Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press.

Ley, E., & Steel, M. F. J. (2009). On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics*, *24*(4), 651–674. https://doi.org/10.1002/jae.v24:4

Lumley, T. (2024). *Leaps: Regression Subset Selection* (R Package Version 3.2) [Computer software].

Raftery, A. E., Madigan, D., & Hoeting, J. A. (1997). Bayesian model averaging for regression models. *Journal of the American Statistical Association*, *92*(437), 179–191. https://doi.org/10.1080/01621459.1997.10473615

Scott, J., & Berger, J. (2005). An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference*, *136*(7), 2144–2162. https://doi.org/10.1016/j.jspi.2005.08.031

Scott, J., & Berger, J. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, *38*(5), 2587–2619. https://doi.org/10.1214/10-AOS792

Smith, M., & Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, *75*(2), 317–344. https://doi.org/10.1016/0304-4076(95)01763-1

Villa, C., & Lee, J. E. (2020). A loss-based prior for variable selection in linear regression methods. *Bayesian Analysis*, *15*(2), 533–558. https://doi.org/10.1214/19-BA1162

Wilson, M. A., Iversen, E. S., Clyde, M. A., Schmidler, S. C., & Schildkraut, J. M. (2010). Bayesian model search and multilevel inference for SNP association studies. *Annals of Applied Statistics*, *4*(3), 1342–1364. https://doi.org/10.1214/09-AOAS322

Womack, A., Taylor-Rodriguez, D., & Fuentes, C. (2025). The Matryoshka Doll prior – principled multiplicity correction in Bayesian model comparison (Technical report). arXiv.

Zhu, L., Huo, Z., Ma, T., Oesterreich, S., & Tseng, G. C. (2019). Bayesian indicator variable selection to incorporate hierarchical overlapping group structure in multi-omics applications. *The Annals of Applied Statistics*, *13*(4), 2611–2636. https://doi.org/10.1214/19-AOAS1271

Zurriaga, O., Perez-Panades, J., Izquierdo, J., Gil, M., Anes, Y., Quiñones, C., Margolles, M., Lopez-Maside, A., Vega-Alonso, A. T., & Espí, M. T. M., & Recent OBICE Research Group. (2011). Factors associated with childhood obesity in Spain. The OBICE study: A case–control study based on sentinel networks. *Public Health Nutrition*, *14*(6), 1105–1113. https://doi.org/10.1017/S1368980010003770

Appendix. Formal variable selection and a discriminating example

A.1 Variable selection in the general linear model

Let X_0 and X be matrices of dimension $n \times k_0$ and $n \times k$, respectively. The columns in X are the observed values of each of the entertained variables while those in X_0 contain covariates common to all models (e.g., an intercept). Assume $(X_0\ X)$ has rank $k_0 + k$ and let y be the n-dimensional vector of observations.

For $\gamma \neq 0$, $M_{\boldsymbol{\gamma}}$ is

$$y = X_0\beta_0 + X_{\boldsymbol{\gamma}}\beta_{\boldsymbol{\gamma}} + \sigma\epsilon, \quad \epsilon \sim N_n(0, I_n),$$

where X_γ is the sub-matrix of X obtained by selecting the columns having a one in $\boldsymbol{\gamma}$. In this paper we use the intrinsic priors derived in Berger et al. (2024). These lead to a Bayes factor of $M_{\boldsymbol{\gamma}}$ to M_0:

$$B_{\boldsymbol{\gamma}} = \int_0^\infty \left(1 + g\frac{\mathrm{SSE}_{\boldsymbol{\gamma}}}{\mathrm{SSE}_0}\right)^{-(n-k_0)/2} (1+g)^{(n-|\boldsymbol{\gamma}|-k_0)/2}\, p^I(g)\, \mathrm{d}g, \tag{A1}$$

Table A1. For $d = 0$ (the null model), $d = 15$ (the model size with the largest Bayes factor), and $d = 46$, the Bayes factor, prior odds and posterior odds of the indicated model to the full model under the Jeffreys' prior probabilities in the obesity example.

d	$B_{d,47}$	$\text{Prior}_{d,47}$	$\text{Post}_{d,47}$
0	8.18×10^{-24}	1	8.18×10^{-24}
15	3.72×10^{10}	1.33×10^{-12}	4.95×10^{-2}
46	2.93	2.13×10^{-2}	6.24×10^{-2}

where $p^I(g)$ is the mixing (density) function

$$p^I(g) = \left(\frac{n-k_0}{|\boldsymbol{\gamma}|+1}\right)^{1/2} \frac{1}{g\pi} \left(g - \frac{n-k_0}{|\boldsymbol{\gamma}|+1}\right)^{-1/2}, \quad g > \frac{n-k_0}{|\boldsymbol{\gamma}|+1},$$

and $\text{SSE}_{\boldsymbol{\gamma}}$ is the sum of squared errors for $M_{\boldsymbol{\gamma}}$. Posterior probabilities of models are $P(M_{\boldsymbol{\gamma}} \mid y) \propto B_{\boldsymbol{\gamma}} P(M_{\boldsymbol{\gamma}})$.

In our numerical examples computations are assisted by the `BayesVarSel` package of R (García-Donato & Forte, 2018) which incorporates the above Bayes factor with the option `prior.betas` = '`IHG`'. When enumeration is feasible (say $k \leq 25$) the package provides exact values of the posterior inclusion probabilities; the highest posterior probability model (HPM) and the median probability model (MPM), given, respectively, by

$$\text{HPM} = \arg\max_{\boldsymbol{\gamma}} P(M_{\boldsymbol{\gamma}} \mid y), \quad \text{MPM} = M_{(1_{\{P(x_1|y)>0.5\}},\ldots,1_{\{P(x_k|y)>0.5\}})},$$

where 1_A is the indicator function and

$$P(x_j \mid y) = \sum_{\boldsymbol{\gamma}:\gamma_j=1} P(M_{\boldsymbol{\gamma}} \mid y), \quad (1 \leq j \leq k), \tag{A2}$$

are the posterior inclusion probabilities. When enumeration is not possible, posterior inclusion probabilities (and the MPM) are estimated using the Gibbs sampling algorithm studied in García-Donato and Martinez-Beneito (2013). This method samples models from their posterior distribution and so more probable models are expected to be sampled. Nevertheless, the algorithm is not specifically designed to identify the HPM. To identify the HPM we use the branch-and-bound method proposed by Furnival and Wilson (1974) and incorporated in the R package `leaps` (Lumley, 2024) with the function `regsubsets`. This algorithm identifies, for each possible dimension $1 \leq d \leq k$, the model with the lowest SSE, say $M_{\boldsymbol{\gamma}^\star(d)}$. Clearly, if $B_{\boldsymbol{\gamma}}$ decreases as a function of $\text{SSE}_{\boldsymbol{\gamma}}$ (for fixed n, k_0, SSE_0 and $|\boldsymbol{\gamma}|$), then $M_{\boldsymbol{\gamma}^\star(d)}$ is also the model with the largest posterior probability of dimension d since $P(M_{\boldsymbol{\gamma}})$ is constant for all models sharing the same dimension. From here, the HPM can be easily obtained as

$$\text{HPM} = \arg\max_{1\leq d\leq k} B_{\boldsymbol{\gamma}^\star(d)} h(d).$$

In the obesity example with $k = 47$, the HPM is identified in less than 5 minutes with a standard personal computer. In the following lemma we show the required monotonicity of $B_{\boldsymbol{\gamma}}$.

Lemma A.1: *Using* (A1)*, consider* $B_{\boldsymbol{\gamma}}(\text{SSE})$ *as a function of SSE for fixed* n, k_0, SSE_0 *and* $|\boldsymbol{\gamma}|$*. If* $\text{SSE}_1 \leq \text{SSE}_2$ *then* $B_{\boldsymbol{\gamma}}(\text{SSE}_1) \geq B_{\boldsymbol{\gamma}}(\text{SSE}_2)$.

Proof: It suffices to show that the inequality holds for the factor $(1 + g\frac{\text{SSE}}{\text{SSE}_0})^{-(n-k_0)/2}$ in (A1), which follows from

$$\frac{\partial}{\partial\,\text{SSE}} \left(1 + g\frac{\text{SSE}}{\text{SSE}_0}\right)^{-(n-k_0)/2} < 0.$$

■

Table A2. For the obesity dataset with $k = 47$ variables and $n = 591$, posterior inclusion probabilities obtained with different prior distributions $P(M_\gamma)$.

Variable	Jeffreys	Uniform	Half-k	Half-p	Hierarchical	Harmonic	Beta(1,2)	CMG
ActivDepor	0.67H/M	0.70H/M	0.63H/M	0.61H/M	0.63H/M	0.63H/M	0.64H/M	0.40-/-
Chuches	1.00H/M	1.00H/M	1.00H/M	1.00H/M	1.00H/M	1.00H/M	1.00H/M	1.00H/M
CincoComidas	0.54H/M	0.56-/M	0.51-/M	0.48-/-	0.51-/M	0.51H/M	0.51-/M	0.29-/-
Daceite	0.22H/-	0.23-/-	0.18-/-	0.15-/-	0.19-/-	0.20H/-	0.19-/-	0.05-/-
Dcereal	0.34H/-	0.34-/-	0.28-/-	0.25-/-	0.30-/-	0.29H/-	0.30-/-	0.09-/-
Desayuno	0.98H/M	0.98H/M	0.98H/M	0.97H/M	0.98H/M	0.98H/M	0.98H/M	0.97H/M
Dgalleta	0.30H/-	0.31-/-	0.23-/-	0.20-/-	0.25-/-	0.25H/-	0.26-/-	0.07-/-
Dislipemias	0.56H/M	0.59-/M	0.51-/M	0.45-/-	0.51-/M	0.51H/M	0.52-/M	0.22-/-
Dleche	0.28H/-	0.29-/-	0.23-/-	0.21-/-	0.25-/-	0.25H/-	0.25-/-	0.10-/-
Dotros	0.42H/-	0.46-/-	0.37-/-	0.33-/-	0.38-/-	0.38H/-	0.39-/-	0.12-/-
Dpan	0.31H/-	0.31-/-	0.25-/-	0.22-/-	0.26-/-	0.26H/-	0.26-/-	0.08-/-
Dzumoenv	0.82H/M	0.85H/M	0.80H/M	0.78H/M	0.81H/M	0.81H/M	0.81H/M	0.60H/M
Dzumonat	0.21H/-	0.21-/-	0.16-/-	0.14-/-	0.17-/-	0.18H/-	0.18-/-	0.05-/-
Faperitivos	0.37H/-	0.38-/-	0.30-/-	0.27-/-	0.32-/-	0.32H/-	0.32-/-	0.10-/-
Farroz	0.29H/-	0.30-/-	0.24-/-	0.21-/-	0.25-/-	0.25H/-	0.26-/-	0.09-/-
Fcarnes	0.38H/-	0.36-/-	0.31-/-	0.28-/-	0.32-/-	0.33H/-	0.33-/-	0.13-/-
Fchucherias	0.33H/-	0.35-/-	0.27-/-	0.23-/-	0.28-/-	0.29H/-	0.30-/-	0.08-/-
Fdulces	0.99H/M	0.99H/M	0.99H/M	0.99H/M	0.99H/M	0.99H/M	0.99H/M	0.98H/M
Ffiambres	0.24H/-	0.24-/-	0.18-/-	0.16-/-	0.20-/-	0.20H/-	0.21-/-	0.06-/-
Ffritos	0.28H/-	0.30-/-	0.23-/-	0.20-/-	0.25-/-	0.25H/-	0.25-/-	0.07-/-
Ffruta	0.21H/-	0.22-/-	0.17-/-	0.14-/-	0.18-/-	0.18H/-	0.19-/-	0.05-/-
Fhuevos	0.64H/M	0.67-/M	0.61-/M	0.57-/M	0.61H/M	0.61H/M	0.61H/M	0.37-/-
Flacteos	0.38H/-	0.40-/-	0.33-/-	0.29-/-	0.34-/-	0.33H/-	0.35-/-	0.13-/-
Flegumbres	0.31H/-	0.32-/-	0.26-/-	0.22-/-	0.28-/-	0.27H/-	0.28-/-	0.10-/-
Fpan	0.57H/M	0.62-/M	0.53-/M	0.50-/M	0.54-/M	0.54H/M	0.56-/M	0.28-/-
Fpescado	0.22H/-	0.21-/-	0.17-/-	0.14-/-	0.19-/-	0.18H/-	0.19-/-	0.05-/-
Fprecocina	0.23H/-	0.25-/-	0.18-/-	0.16-/-	0.20-/-	0.20H/-	0.21-/-	0.06-/-
Frefrescos	0.44H/-	0.47-/-	0.37-/-	0.34-/-	0.39-/-	0.38H/-	0.39-/-	0.14-/-
Fruta	0.21H/-	0.21-/-	0.16-/-	0.13-/-	0.18-/-	0.17H/-	0.17-/-	0.05-/-
Fverduras	0.42H/-	0.45-/-	0.34-/-	0.29-/-	0.36-/-	0.35H/-	0.38-/-	0.08-/-
HorasPCsem1	0.28H/-	0.28-/-	0.22-/-	0.19-/-	0.24-/-	0.24H/-	0.24-/-	0.07-/-
HorasPCsem2	0.20H/-	0.21-/-	0.16-/-	0.14-/-	0.18-/-	0.17H/-	0.18-/-	0.05-/-
HorasTVsem2	0.33H/-	0.33-/-	0.29-/-	0.26-/-	0.30-/-	0.29H/-	0.30-/-	0.15-/-
HoraSuenyo	0.32H/-	0.34-/-	0.27-/-	0.23-/-	0.29-/-	0.28H/-	0.29-/-	0.08-/-
HTA	0.30H/-	0.32-/-	0.25-/-	0.21-/-	0.26-/-	0.26H/-	0.27-/-	0.08-/-
LactMater	0.22H/-	0.22-/-	0.17-/-	0.15-/-	0.19-/-	0.19H/-	0.20-/-	0.05-/-
MadreObesa	0.99H/M	0.99H/M	0.99H/M	0.99H/M	0.99H/M	0.99H/M	0.99H/M	0.99H/M
NumHnosOb	0.89H/M	0.91H/M	0.88H/M	0.87H/M	0.88H/M	0.88H/M	0.89H/M	0.73H/M
OtrosPatol	0.20H/-	0.21-/-	0.16-/-	0.13-/-	0.17-/-	0.17H/-	0.18-/-	0.05-/-
PadreObeso	1.00H/M	1.00H/M	1.00H/M	1.00H/M	1.00H/M	1.00H/M	1.00H/M	1.00H/M
PesoNac	0.84H/M	0.88H/M	0.81H/M	0.79H/M	0.81H/M	0.80H/M	0.82H/M	0.46-/-
ProbOsteo	0.85H/M	0.89H/M	0.83H/M	0.80H/M	0.83H/M	0.82H/M	0.83H/M	0.59-/M
ProbPsico	0.31H/-	0.32-/-	0.25-/-	0.20-/-	0.28-/-	0.26H/-	0.28-/-	0.07-/-
ProbResp	0.29H/-	0.30-/-	0.23-/-	0.21-/-	0.26-/-	0.26H/-	0.26-/-	0.08-/-
TallaNac	0.92H/M	0.95H/M	0.91H/M	0.89H/M	0.90H/M	0.90H/M	0.91H/M	0.61-/M
TVDiario	0.87H/M	0.88H/M	0.85H/M	0.84H/M	0.86H/M	0.86H/M	0.86H/M	0.77H/M
Verduras	0.24H/-	0.24-/-	0.18-/-	0.15-/-	0.20-/-	0.21H/-	0.21-/-	0.05-/-
$\lvert HPM \rvert / \lvert MPM \rvert$	47/16	15/16	12/16	12/14	12/16	47/16	12/16	8/10

Notes: It is also indicated whether the variable belongs to the HPM (designated by H) and/or to the MPM (designated by M). The last row is the dimension of the HPM and the MPM.

A.2 A discriminating example

We consider the $k = 47$ variable obesity example for all prior model choices. Bayesian model uncertainty was implemented with these eight prior model probability choices and the intrinsic prior for the model parameters. Reported in Table A2 are the posterior inclusion probabilities, defined in (A2), for each of the possible variables.

Again, the most striking feature of this table is that the HPM (denoted H in the table) is the full model. This is not scientifically viable and indicates that the Jeffreys' prior, giving maximum probability

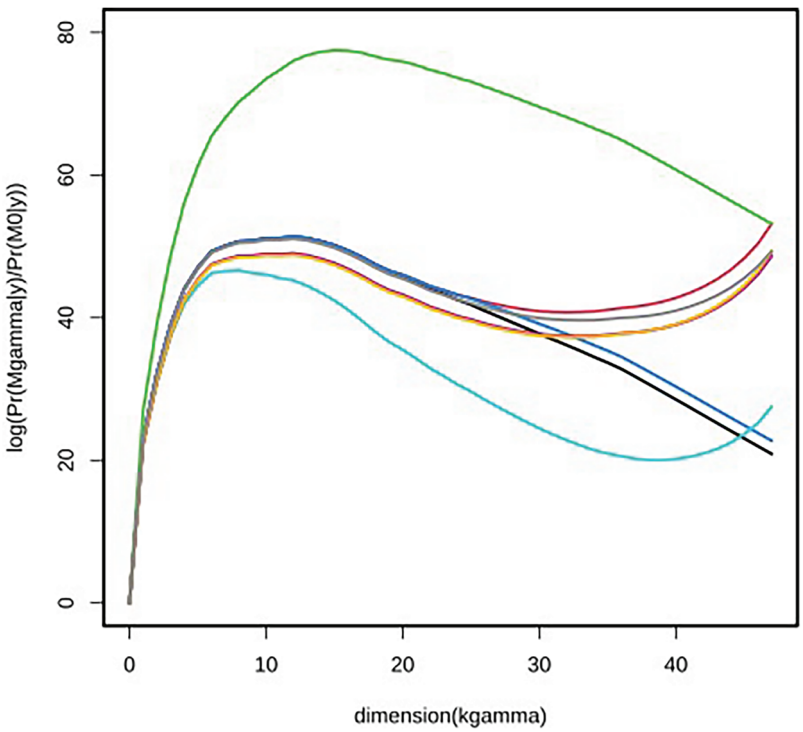


Figure A2. For the obesity dataset, $k = 47$, log ratio of the posterior probability of the most probable model of each dimension to that of the null model for different priors. Half-*p* (black), Jeffreys (red), Uniform (green), Half-*k* (blue), CMG (cyan), Hierarchical (magenta), Beta(1,2) (grey), Harmonic (yellow).

to the full model, is problematic. Surprisingly this is also the case for the Harmonic prior probabilities, suggesting that it is also not parsimonious enough. The HPM also varies considerably over the choice of prior probabilities. The Uniform prior resulted in the HPM containing 15 variables, while the other choices of prior probabilities resulted in smaller models with 8 to 12 variables. This all suggests that the HPM is non-robust with respect to the choice of prior probabilities.

In contrast, the posterior inclusion probabilities do not vary much across the choice of prior probabilities, except for the Uniform prior, where the inclusion probabilities are systematically higher, and the CMG prior with substantially lower inclusion probabilities. Finally, and quite curiously, the MPM (denoted M in the table) is quite stable over all prior probability choices, never differing by more than two variables, except for the CMG prior. This lends support to the large literature (not reported here) that suggests that one should focus attention on posterior inclusion probabilities and the resulting MPM, rather than focussing on the HPM.

Figure A2 presents, for this example, the log of the ratio of the posterior probability of the most probable model of each dimension to the null model, for the different prior probabilities. All the priors exhibit a bimodal behaviour, except for the Uniform, Half-*k* and Half-*p* priors. A curiosity is that the Hierarchical and Harmonic prior probabilities seem to exactly track each other, yet the Harmonic ended up with the HPM being the full model and the Hierarchical did not.

It is also useful to look at $B_{d,47}$, the largest Bayes factor among the models of size d compared to the full model of size 47, and $\text{Prior}_{d,47}$, the prior odds of models of size d to the full model; note that the product of these is $\text{Post}_{d,47}$, the posterior odds of the best model of size d to the full model. Table A1 presents these values for three choices of d. The null model has the same prior probability as the full model but a minuscule Bayes factor. The model with the largest Bayes factor is of size 15, but this large Bayes factor is overwhelmed by the model's small prior odds. Interestingly, the model of size 15 has roughly the same posterior odds as the best model of size 46.