Objective bayes criteria for variable selection

Article · January 2011

Source: OAI

CITATIONS
10

READS
281

1 author:

Anabel Forte
University of Valencia
88 PUBLICATIONS **985** CITATIONS

SEE PROFILE

UNIVERSITAT DE VALÈNCIA

Objective Bayes Criteria for Variable Selection

by

Anabel Forte Deltell

A thesis submitted in partial fulfillment for the
degree of Doctor en Ciencias Matemáticas

Supervised by:
Mª Jesús Bayarri García and
Gonzalo García-Donato Layrón

in the
Facultad de Ciencias Matemáticas
Departamento de Estadística e Investigación Operativa

2011

Declaración de Autoría

Mª Jesús Bayarri García, Doctora en Ciencias Matemáticas por la Universitat de València y Catedrática del departamento de Estadística e Investigación operativa de la Universitat de València.

Gonzalo García-Donato Layrón, Doctor en Ciencias Matemáticas por la Universitat de València y profesor titular del departamento de Análisis Económico y Finanzas de la Universidad de Castilla La Mancha.

CERTIFICAMOS: que la presente memoria con el título:

"Objective Bayes Criteria for Variable Selection"

ha sido realizada por Anabel Forte Deltell, bajo nuestra dirección y consitituye la tesis para optar al grado de Doctor en Ciencias Matemáticas.

Y para que conste firmamos este certificado en Valencia a 22 de Diciembre de 2010

Firmado:

Mª Jesús Bayarri García

Gonzalo García-Donato Layrón

"Which way I ought to go from here?"
"That depends a good deal on where you want to get to," said the cat.
"I don't much care where..." said Alice
"Then it doesn't matter which way you go," said the cat.

Lewis Carrol, *Alice in Wonderland*. Cited by Jeffreys (1961)

UNIVERSITAT DE VALÈNCIA

Abstract

Facultad de Ciencias Matemáticas
Departamento de Estadística e Investigación Operativa

Doctor en Ciencias Matemáticas

by Anabel Forte Deltell

Variable selection typically involves choosing among a large number of models, so that fast computation of Bayes factors is highly desirable. This desideratum has made common practice the use of g-priors and Laplace expansions, specially in large dimensions. It is well known, however, that priors with heavier tails often result in better performance for model selection. In this thesis, we use the Conventional approach of Jeffreys (1961) and generalize some ideas in Strawderman (1971, 1973) and Berger (1976, 1980, 1985) to propose a prior distribution for variable selection. We show that this choice is, to the best of our knowledge, the first proposal for variable selection which is fully justified from a theoretical point of view. This justification is heavily based on the invariance ideas in Berger et al. (1998). Moreover, it has Student-like tails and many optimal properties for model selection. It also generalizes previous proposals in the literature. In addition, for specific choices of the hyper-parameters, it produces closed-form marginal likelihoods (and hence, Bayes factors). We demonstrate its behavior in a couple of small problems and in a couple of large, but enumerable, ones.

Agradecimientos

Emprender la aventura de hacer una tesis no fue una decisión fácil, como tampoco lo ha sido su desarrollo, pero llegados a este punto y echando la vista atrás, lo único que me queda es una experiencia maravillosa.

Además de la investigación en si misma, lo mejor de esta tesis ha sido el camino recorrido y en particular los compañeros de viaje, personas encantadoras que siempre han tenido una palabra de apoyo y de ánimo para mi. Desde aquellos que cuando terminaba el instituto me animaban encarecidamente a que fuese a la universidad y han estado y estarán ahí siempre: Sandra (tu ya lo sabes), Mamen (ánimo que ahora te toca a ti), mil gracias; pasando por todos los que estuvieron a mi lado durante la carrera, compañeros inseparables, mis chicos: Rober, Mario (grandes compañeros, mejores amigos), Raul, Dioni, Diego, Victor, Pablo, Pascu, Jose Ramón, Juanje, Rafa... El tándem Rosaura y Jannet (que junto con Adrián también me aguantaron en casa), Elena (con una sonrisa siempre lista y una amabilidad inmensa), Rebeca, Isa, Marta, Ana, Lledó... Mis compañeras de piso, Irene, M. Jose, Amparo, Macarena... y muchos otros que me acompañaron durante esos primeros años en Valencia y que, a fuerza de interacción, influyeron en mi carácter y me ayudaron a llegar hasta aquí.

Indudablemente tengo que dar las gracias a los que confiaron en mi para lanzarme a esta aventura, Antonio, Carmen y David, nunca os estaré lo suficientemente agradecida. A los que me han aguantado (y mucho) durante los años de doctorado y de tesis, con los que he discutido de todo y he aprendido mucho, Rafa, Alejandro, Toni, Hector, Rubén, Maria, y sobre todo a Facu, gracias por tus consejos, tus opiniones y todo lo que me has aportado en estos años. A todo el personal del departamento de Estadística e investigación operativa, Juana, Tere y Teten, gracias por hacernos siempre el trabajo más fácil, y al resto de compañeros, gracias por todos esos cafés de las diez. También a aquellos que me

han acompañado en mis andanzas por Durham y Cambridge, Migue, Virgilio, Murali y Myriam, gracias por hacer que mis días fuera de casa fueran más amenos y agradables. Sin olvidar a todos aquellos que me han acompañado en algún momento de este viaje y que han estado ahí para escuchar y dar una palabra de apoyo o un consejo; gracias en especial a Xavi y a Paloma por su predisposición a ayudar siempre.

I want also to thank Jim Berger for his valuable comments. Also, I would like to thank Jim and his wife Ann for taking care of me during my time in Durham.

Gracias por supuesto a mis directores Gonzalo García-Donato y Mª Jesús Bayarri. Gon y Sus... no tengo palabras. Esta experiencia nunca habría sido lo mismo sin vosotros, gracias por disfrutarlo como si fuese vuestra propia tesis, por presumir de mi, por tirarme a la piscina sin flotador cuando hacia falta y por enseñarme cosas (de la investigación y de la vida) que no podría haber aprendido por mi misma ni en mil años que viviese. De todo corazón, Gracias.

Y el agradecimiento más importante, para aquellos que han estado a mi lado desde el principio de mis días y para los que, no estando desde el principio, estarán hasta el final... Gracias a mi familia. A mis padres y mi hermano, por su apoyo incondicional y desmedido, su confianza en mi, por esperarme siempre con los brazos abiertos. A mis tíos, abuelos y primos gracias por hacerme crecer en lo personal y empujarme a ser lo que hoy soy (que al final lo de estudiar en Valencia va a servir para algo...je je je). A la familia de Fran, que es también la mía, por tener siempre una palabra de aliento para mi. Y en especial gracias a Fran. Gracias a ti "peque", porque esta tesis es tan tuya como mía, por sufrir cada reunión, cada corrección, cada llanto, y cada alegría como si fueran tuyas, por dejarme volar y quedarte a cargo de todo sin poner nunca ni una sola pega. Ni en mil años juntos podré devolverte lo que has dado y das por mi cada día. Te quiero.

Contents

List of Figures

List of Tables

To Fran

Chapter 1

The Bayesian approach to the model selection problem

1.1 Introduction

Developing suitable theories to explain phenomena of interest is a main scientific goal. Often different theories are proposed to explain the same phenomenon, and the important issue of choosing among them arises. A key question is which of the entertained theories is the most likely explanation of reality. This is the core background scenario in which we frame our work.

Choosing among competing theories is usually carried out in the light of data which we denote simply by $\boldsymbol{y}$. We assume that each theory can be well represented by a statistical model or probability distribution explaining the joint random behavior of $\boldsymbol{y}$. These models usually depend on unknown parameters $\boldsymbol{\theta}$, and we adopt the usual convention of representing model M by its probability density function $f(\boldsymbol{y} \mid \boldsymbol{\theta})$ (a density over $\boldsymbol{y}$ given the value of the unknown $\boldsymbol{\theta}$).

Suppose that m different models are entertained and let

$$\mathcal{M} = \{M_1, \ldots, M_m\}$$

be the set of all of them; $\mathcal{M}$ is usually referred to as *model space*. Hence, the general scientific problem, succinctly introduced previously, simply becomes that of selecting one of the models in $\mathcal{M}$. In the sequel we refer to this problem as *model selection*.

It is important to remark that each model M_i specifies a joint probability distribution for the whole set of observations $\boldsymbol{y}$. Therefore, each model is a different, and complete explanation of reality, implicitly considering issues such as dependency among observed values in $\boldsymbol{y}$, finite versus infinite population, etc.

An important particular case of model selection is hypothesis testing. In this thesis, we focus on a specific hypothesis testing problem (*the variable selection problem*) described in detail in Chapter 2. In the rest of this chapter we outline our preferred Bayesian approach to model selection.

1.2 The Bayesian approach

Quoting Kass and Raftery (1995), "the Bayesian approach to hypothesis testing was developed by Jeffreys (1939) as a major part of his program for scientific inference". Jeffreys' solution is based on posterior probabilities or equivalently, on Bayes factors (see Kass and Raftery, 1995, and references therein).

Choosing a model among those in $\mathcal{M}$ based on posterior probabilities seems intuitively sensible. This approach to model selection also arises formally in decision theory frameworks for specific choices of the loss function. As in Jeffreys' solution, in this thesis model selection is based on posterior probabilities expressed in terms of Bayes factors which are

defined and studied in Section 1.2.1 and will be the main focus of our work.

1.2.1 Bayes factors and posterior probabilities

Bayes factors (Jeffreys, 1961) were introduced to quantify the evidence in the data in favor of a model M_i and against another model M_j, but of course they can be (and are) used to compare a set of $m > 2$ models.

Definition 1.1. Let M_i and M_j be two competing models. The Bayes factor in favor of M_i and against M_j given data $\boldsymbol{y}$ is defined as:

$$B_{ij} = \frac{P(M_i \mid \boldsymbol{y})/P(M_j \mid \boldsymbol{y})}{P(M_i)/P(M_j)}, \tag{1.1}$$

where, $P(M_l)$ represents the prior probability of model M_l and $P(M_l \mid \boldsymbol{y})$ is its corresponding posterior probability, for $l = i, j$. The Bayes factor B_{ij} is, hence, the ratio between posterior and prior odds in favor of M_i and against M_j. Therefore, a Bayes factor $B_{ij} = 10$ means that prior odds in favor of M_i have been multiplied by 10 after observing the data.

If one of the models in $\mathcal{M}$ is the true model, by Bayes theorem, the posterior probabilities are

$$P(M_i \mid \boldsymbol{y}) = \frac{m_i(\boldsymbol{y})P(M_i)}{\sum_{l=1}^{m} m_l(\boldsymbol{y})P(M_l)}.$$

where $m_l(\boldsymbol{y})$ is the prior predictive distribution (or marginal likelihood) under model M_l,

$$m_l(\boldsymbol{y}) = \int f_l(\boldsymbol{y} \mid \boldsymbol{\theta}_l)\, \pi_l(\boldsymbol{\theta}_l)\, d\boldsymbol{\theta}_l, \quad l = i, j \tag{1.2}$$

and $\pi_l(\boldsymbol{\theta}_l)$ is the prior density of $\boldsymbol{\theta}_l$ under M_l. It thus follows that the posterior odds in favor of M_i and against M_j are

$$\frac{P(M_i \mid \boldsymbol{y})}{P(M_j \mid \boldsymbol{y})} = \frac{m_i(\boldsymbol{y})}{m_j(\boldsymbol{y})} \cdot \frac{P(M_i)}{P(M_j)}, \tag{1.3}$$

so that the Bayes factors can also be expressed as

$$B_{ij} = \frac{m_i(\boldsymbol{y})}{m_j(\boldsymbol{y})} \tag{1.4}$$

In practice, (1.4) is usually taken as the definition of the Bayes factor, not requiring use of the prior odds.

To interpret Bayes factors between two models, Jeffreys (1961) suggested taking half-units on the $\log_{10}$ scale. In Table 1.1 we present a summary of Jeffreys' interpretations:

$log_{10}(B_{ij})$	B_{ij}	Evidence against M_j
0 to 1/2	1 to 3.2	Not worth more than a bare mention
1/2 to 1	3.2 to 10	Substantial
1 to 2	10 to 100	Strong
> 2	> 100	Decisive

TABLE 1.1: *Bayes factors interpretation.*

There are several possibilities for choosing the prior probabilities of models $P(M_l)$. One obvious possibility is to choose them to reflect genuine subjective prior belief. When this is not possible or not desired, a popular choice is to give each model in $\mathcal{M} = \{M_1, \ldots, M_m\}$ equal prior probability, that is $P(M_l) = 1/m$. This is not always an optimal choice. Our preferred choice for $P(M_i)$, in the particular scenario of variable selection, is given in Section 2.3.4.

Posterior probabilities can easily be expressed in terms of Bayes factors. In fact, summing over i in (1.3) gives

$$P(M_j \mid \boldsymbol{y}) = \left(\sum_{l=1}^{m} B_{lj} \cdot \frac{P(M_l)}{P(M_j)} \right)^{-1}.$$

Since Bayes factors are transitive in the sense that $B_{ij} = B_{il}B_{lj}$ for any model M_l, the usual approach to compare $m > 2$ models consists in comparing each model with a fixed one M_d which we call *base model*. Hence, since $B_{ij} = B_{ji}^{-1}$

$$P(M_j \mid \boldsymbol{y}) = \frac{B_{jd}\, P(M_j)}{\sum_{l=1}^{m} B_{ld}\, P(M_l)} = \left[1 + \sum_{l \neq j} \frac{P(M_l)}{P(M_j)} \cdot \frac{B_{ld}}{B_{jd}}\right]^{-1}. \qquad (1.5)$$

Bayes factors also arise formally in decision theoretical formulations. Indeed, decision theory provides the most complete framework for model selection. Here the action (decision) space is $\mathcal{M}$. To keep with the usual notation a specific decision is denoted by "a". The loss for deciding model a when the true model is M is $L(M, a)$. Often the loss also depends on the true parameter $\boldsymbol{\theta}_M$ under the true model M in which case

$$L(M, a) = \mathrm{E}^{\theta_M|y}\left[L(M, \boldsymbol{\theta}_M, a)\right] = \int_{\Theta_M} L(M, \boldsymbol{\theta}_M, a)\, \pi_M(\boldsymbol{\theta}_M \mid \boldsymbol{y})\, d\boldsymbol{\theta}_M,$$

where $\pi_M(\boldsymbol{\theta}_M \mid \boldsymbol{y})$ is the posterior distribution under model M corresponding to the prior density $\pi_M(\boldsymbol{\theta}_M)$, that is

$$\pi_M(\boldsymbol{\theta}_M \mid \boldsymbol{y}) = \frac{f_M(\boldsymbol{y} \mid \boldsymbol{\theta}_M)\, \pi_M(\boldsymbol{\theta}_M)}{m_M(\boldsymbol{y})}.$$

The optimal decision, or "Bayes action" (see Berger, 1985, Section 4.4) is the "a" minimizing the posterior expected loss

$$\mathrm{E}[L(M, a)] = \sum_{i=1}^{m} L(M_i, a) \cdot P(M_i \mid \boldsymbol{y}).$$

A common loss for hypothesis testing problems is the "0 - k_i" loss function in which k_i is the loss for incorrectly choosing a_i, and correct decisions have 0 loss; that is, $L(M, a_i) = 0$ if $a_i = M$ and $L(M, a_i) = k_i$ if $a_i \neq M$.

In particular, consider a hypothesis testing scenario where

$$y \mid \boldsymbol{\theta} \sim f(y \mid \boldsymbol{\theta}),$$

with $\boldsymbol{\theta} \in \mathbb{R}^k$, and it is desired to test the m hypotheses $H_i : \boldsymbol{\theta} \in \Theta_i$ for $\{\Theta_i\}_{i=1}^m$ a partition of Θ. Here the actions (or decisions) are a_i, $i = 1, \ldots, m$, where a_i is choosing H_i so the 0 - k_i loss can be expressed as

$$\begin{aligned} L(\boldsymbol{\theta}, a_i) &= 0 && \text{if } \boldsymbol{\theta} \in \Theta_i, \\ L(\boldsymbol{\theta}, a_i) &= k_i && \text{if } \boldsymbol{\theta} \notin \Theta_i. \end{aligned}$$

Often the k_i's are taken to be equal for $i = 1, \ldots, m$, indicating that the loss for a wrong decision is the same for all decisions.

The posterior expected loss for taking action a_i is

$$\begin{aligned} \mathrm{E}^{\theta|y}[L(\boldsymbol{\theta}, a_i)] &= \int_{\Theta} L(\boldsymbol{\theta}, a_i)\, \pi(\boldsymbol{\theta} \mid \boldsymbol{y})\, d\boldsymbol{\theta} = k_i \int_{\Theta_i^c} \pi(\boldsymbol{\theta} \mid \boldsymbol{y})\, d\boldsymbol{\theta} \\ &= k_i \left[1 - P(\Theta_i \mid \boldsymbol{y})\right] = k_i \left[1 - P(H_i \mid \boldsymbol{y})\right], \end{aligned}$$

that is, the expected loss for choosing H_i is proportional to the posterior probability of $\overline{H}_i$ (the combined hypothesis "H_i is not true"). Hence, it can be easily seen that the Bayes action is a_i if for all $j \neq i$

$$\frac{\mathrm{E}^{\theta|y}[L(\boldsymbol{\theta}, a_i)]}{\mathrm{E}^{\theta|y}[L(\boldsymbol{\theta}, a_j)]} = \frac{k_i}{k_j} \cdot \frac{1 - P(H_i \mid \boldsymbol{y})}{1 - P(H_j \mid \boldsymbol{y})} = \frac{k_i}{k_j} \cdot \frac{\sum_{l \neq i} B_{ld} P(H_l)}{\sum_{l \neq j} B_{ld} P(H_l)} < 1,$$

where B_{ld} is the Bayes factor of model l to model d. So the Bayes action is defined in terms of Bayes factors or posterior probabilities. In particular if all the k_i's are equal, then the optimal action is to choose the hypothesis with the maximum posterior probability.

Notice that in this decision theoretical framework we implicitly assume that one of the models is true. This requirement is often used as an argument against Bayes factors because this is not always the case. However

García-Donato (2003) shows that, even when the true model is not in $\mathcal{M}$, the evidence in favor of a model given some data is always proportional to its posterior probability (for which Bayes factors are essential ingredients). Also Dmochowski (1996) shows that, in this same scenario, the Bayes factors select the closest model to the true one in Kullback-Leibler sense (see Kullback, 1999). Wasserman (2000) also advocates use of posterior probabilities for comparing the relative evidence of models, even when they can not be considered "true". Quoting Wasserman (2000):

> *Newtonian physics and general relativity are both wrong. Yet it make sense to compare the relative evidence in favor of one or the other. Our conclusion would be: "under the tentative working hypothesis that one of these two theories is correct, we find that the evidence strongly favors general relativity." It is understood that the working hypothesis that "one of the models is correct" is wrong. But it is a useful, tentative hypothesis and proceeding under that hypothesis, it makes sense to evaluate the relative posterior probabilities of those hypotheses.*

1.3 Objective Bayesian model selection

To compute Bayes factors, prior distributions for parameters under each model, $\pi_i(\boldsymbol{\theta}_i)$ $i = 1, \ldots, m$, are required. These priors play an important role in model selection. There are two main approaches to the assignment of prior distributions: the subjective or informative prior elicitation, in which $\pi_i(\boldsymbol{\theta}_i)$ quantifies the prior believes about $\boldsymbol{\theta}_i$; and the objective approach in which no subjective information is explicitly introduced, apart from that required to define the models.

There has been a long debate in the Bayesian community as to the roles of subjective and objective Bayesian analysis (see Berger, 2006; Goldstein, 2006, and the discussion there). An objective Bayesian would argue that

appearance of objectivity is often needed and it is usually very difficult to get the required subjective information from experts. In model selection, with a large number of models, this argument becomes even stronger, since it is unfeasible to assess subjective prior distributions for every parameter under each model. In particular, in a variable selection problem with p covariates one has to choose among 2^p models, and 2^p subjective assessments for 2^p vectors of parameters have to be done. This becomes a basically impossible task for even moderate values of p. For this reason this thesis focus on objective Bayes methods for assessing prior distributions in the variable selection problem. Moreover, as commented in Berger and Pericchi (2001), it is quite more useful to utilize the limited time of subject experts for model formulation than for the subjective elicitation of priors. An extensive discussion about objectivity and objective Bayes methods can be found in Berger and Pericchi (2001) and Berger (2006).

1.3.1 Objective priors for model selection

The elicitation of prior distributions is always a delicate issue. In the context of estimation there seems to be general agreement on the priors that should be used, whether a subjective or an objective approach is adopted (see Berger et al., 2009; Garthwaite et al., 2005; O'Hagan, 1988; Press, 2003). In contrast, choice of suitable priors for model selection, is not obvious and should be carefully addressed. In particular:

- *Bayes factors can be very sensitive to the choice of prior distributions.* Moreover, and in contrast to the situation in estimation problems, the influence of the prior on the Bayes factors remains even asymptotically (as the number of observations grows, see Kass and Greenhouse, 1989; Kass, 1993; Kass and Raftery, 1995, and references therein).

- *Use of improper, non-informative priors often yields indeterminate Bayes Factors.* As an illustration, let H_1 and H_2 be two competing hypotheses. If we take improper non-informative priors under each hypothesis, $\pi_1(\boldsymbol{\theta}_1)$ and $\pi_2(\boldsymbol{\theta}_2)$ it might seem that we could use (1.1) to (formally) compute the Bayes factor B_{21}. However, since these priors are improper, we could just as well use $c_1\,\pi_1(\boldsymbol{\theta}_1)$ y $c_2\,\pi_2(\boldsymbol{\theta}_2)$ obtaining $(c_2/c_1)B_{21}$. Since the choice of c_1/c_2 is arbitrary, the Bayes factor is indeterminate. Nevertheless, there are situations (as invariant problems, see Berger et al., 1998) in which the use of objective, typically improper priors, is justified in the sense that the resulting Bayes factors are well defined (more details are given in Section 3.4.1; for a full exposition justification and further details see Berger et al., 1998)

- *Use of "vague proper prior" does not solve the difficulties arising with improper priors.* Indeed, as shown in Berger and Pericchi (2001), using a vague proper prior (a proper prior but with an arbitrarily large scale) is never better than using an improper prior. Liang et al. (2008) also show the danger in choosing an arbitrarily large variance (see the description of Bartlett's paradox in Liang et al., 2008).

Jeffreys (1961) dealt with the indeterminacy of Bayes factors arising from use of non-informative priors by using (under some conditions) default proper priors (but never arbitrarily vague priors) for parameters that occur in one model but not in the others and non-informative priors only for common parameters (parameters appearing in all models). Many authors followed this recommendation, as for instance Zellner and Siow (1980, 1984) (see Berger and Pericchi, 1996, for more references).

In the sequel, and following Berger and Pericchi (2001), we call *Conventional approach* to any method for choosing the prior distribution based on Jeffreys' pioneering ideas.

A more elegant way of dealing with indeterminacy is by exploiting invariance properties. Berger et al. (1998) describe conditions under which the Bayes factor obtained from a specific improper prior (right Haar measure) is well defined. These conditions are related to invariance in the sense described in Section 3.4.1.

In this work we use ideas from both the Conventional and the invariance approach, to propose a prior distribution for variable selection (see Chapter 3).

Still another way of dealing with the indeterminacy of Bayes factors arising from objective priors is to use *default Bayes factors* like *fractional Bayes factors*, developed by O'Hagan (1995) or the *intrinsic Bayes factors*, defined in Berger and Pericchi (1996). (See Berger and Pericchi, 2001, for an extensive review of these methods). Although these are not actual Bayes factors, (i.e. they are not computed directly from a explicit prior distribution), they can actually be shown to asymptotically correspond to Bayes factors arising from proper priors called *intrinsic priors*. Hence, these methods can also be seen as a way of eliciting suitable objective priors for model selection.

1.4 Final remarks

As we commented at the beginning of Section 1.2, our preferred solution to model selection is based on posterior probabilities and Bayes factors. However, there exists other approaches to model selection which do not assign any prior probabilities to models and consequently are not based on posterior probabilities. Such approaches are taken for instance in Ibrahim and Laud (1994); Bernardo and Smith (1994); Gelfand and Ghosh (1998); Goutis and Robert (1998); Ibrahim et al. (2001) and references therein. These methods are usually very difficult to calibrate (in the sense that the results are difficult to interpret). Moreover, they are

clearly inappropriate when the alternative models are believable scientific theories which is common in many scientific fields. For example an astronomer may want to investigate whether a recently discovered star system has zero, one, ore more planets.

In certain scenarios, like prediction, one is not required to chose a single model from $\mathcal{M}$, but rather to do the statistical analysis incorporating the uncertainty about models. This is the *model averaging* framework in which posterior probabilities of models (and hence Bayes factors) also play a crucial role (see Hoeting et al., 1999; Kass and Raftery, 1995; Leamer, 1978). Often, however, one is required to choose one single model from $\mathcal{M}$, as in the astronomer example mentioned above, or in many social sciences scenarios. In any case, the tools developed in this work are also relevant for model averaging.

The problem of *model checking* (or *model validation*) is also related to model selection. This approach tries to quantify the compatibility of a specific model with the observed data, without consideration of any alternative model or theory. Model selection instead compares different theories. Model checking will not be considered here, for further information see Bayarri and Berger (1998, 2000); O'Hagan (2003); Bayarri and Castellanos (2007); Bayarri and Morales (2003) and references therein.

Chapter 2

Variable selection

2.1 Introduction

Variable selection is an important problem often encountered in applied statistics aimed at explaining a response variable Y using a set of explanatory variables. In the problem we consider it is known that Y is affected by a given set of k_0 known variables and the goal is to find out which additional variables from a set $\{X_1, \ldots, X_p\}$ of potential ones are also relevant to explain Y.

The variable selection problem can be seen as a particular model selection problem where each entertained model M_i corresponds to a particular subset of covariates. The simplest model, hereafter denoted by M_0, has k_0 covariates, while the most complex one contains $k_0 + p$ covariates. In this problem the model space $\mathcal{M}$ contains a total of 2^p models.

In this thesis we address variable selection in the framework of linear regression. However, variable selection also appears in many other scenarios such as generalized linear models and non-parametric function estimation (see George, 2000, and references therein).

In the rest of this chapter we state the problem of variable selection and review important concepts which will be needed later on. In particular, we introduce some of the ideas motivating our proposed solution in Chapter 3 and review some tools needed for its development.

2.2 The problem

The problem of variable selection can be stated as follows: let $\boldsymbol{y} = (y_1, \ldots, y_n)^t$ be a sample of size n from the distribution of Y. Consider the set $\mathcal{M} = \{M_0, \ldots, M_{2^p-1}\}$ of 2^p possible models, where M_0 denotes the *simplest* model explaining Y:

$$M_0 : \; f_0(\boldsymbol{y} \mid \boldsymbol{\beta}_0, \sigma) = \mathcal{N}_n(\boldsymbol{y} \mid \boldsymbol{X}_0\boldsymbol{\beta}_0, \sigma^2\boldsymbol{I}_n)\,. \tag{2.1}$$

Here $\boldsymbol{X}_0$ is a $n \times k_0$ matrix of variables assumed to enter for sure the model explaining Y, and $\boldsymbol{\beta}_0$ is a k_0-vector of regression coefficients for the variables in $\boldsymbol{X}_0$. The matrix $\boldsymbol{X}_0$ is usually taken to contain at least the intercept (recall that in Bayesian model selection each model has to be a plausible explanation of reality, and we usually need at least the intercept for M_0 to be so). In the frequent particular case of it containing only the intercept, then $\boldsymbol{X}_0 = \boldsymbol{1}_n$ the n-vector of ones and β_0 is a scalar value representing the overall mean of Y under M_0. We refer to M_0 as the "null" model. For the $2^p - 1$ extra models we consider p extra variables, apart from the variables in $\boldsymbol{X}_0$, which can be involved in the studied process. For simplicity and slightly abusing notation, we follow the common convention and express each of these $2^p - 1$ models as

$$M_i : \; f_i(\boldsymbol{y} \mid \boldsymbol{\beta}_i, \boldsymbol{\beta}_0, \sigma) = \mathcal{N}_n(\boldsymbol{y} \mid \boldsymbol{X}_0\boldsymbol{\beta}_0 + \boldsymbol{X}_i\boldsymbol{\beta}_i, \sigma^2\boldsymbol{I}_n), \tag{2.2}$$

where $\boldsymbol{X}_i$ is a $n \times k_i$ matrix containing the observed values of a subset of k_i covariates out of the p entertained, and the k_i-vector $\boldsymbol{\beta}_i$ is the corresponding vector of regression coefficients. Parameters $(\boldsymbol{\beta}_0, \sigma)$ are usually known as common parameters or, in Jeffreys' terminology, the "old" parameters, while the $\boldsymbol{\beta}_i$'s are known as the extra or "new" parameters. Note that M_0 is nested in all the entertained models M_i, $i = 1, \ldots, 2^p - 1$.

Let us consider a simple example:

Example 2.1. Let Y represent the price of a house. A bank wants to study what is affecting house pricing and considers five possible explanatory variables: the location, the size, the views, the proximity to a shopping mall and the age. The bank has data about the price and the five variables of interest for n houses. The simplest explanation for Y is a normal distribution with unknown constant mean and variance. Hence, the null model here just contains the intercept (i.e. $\boldsymbol{X}_0 = \mathbf{1}_n$) and β_0 represents an overall mean for the house pricing. The rest of the $2^5 - 1 = 31$ models are defined as containing all the different combinations of the five variables (always including the intercept, β_0).

One of the difficulties of variable selection is the large number of models in $\mathcal{M}$ when p is even of moderate size. Due to the high dimensionality of the model space:

1. It becomes virtually impossible, as commented in Section 1.3, to subjectively elicit prior distributions to the 2^p vectors of parameters. To cope with this difficulty we adopt, in this thesis, an objective point of view heavily inspired in the ideas of Jeffreys (1961) (see Section 2.3.3) and shown to have desirable theoretical properties.

2. Sometimes the model space can not (for all practical purposes) even be enumerated and hence, posterior probabilities for all models can not be computed. A possible solution is then to explore these huge model spaces in search for "good" models accounting for a large

proportion of posterior probability. Some approaches for searching huge model spaces can be found in George and McCulloch (1993, 1997); Carlin and Chib (1995); Miller (2001); Robert and Casella (2004); Berger and Molina (2005) and references therein. However, searching in the model space is often a daunting task, and simple expressions for the computation of Bayes factors are preferred since they make it possible to devote computational resources to further exploration of the model space instead of to numerical computations. In this spirit, the proposal developed in this thesis produces closed-form expressions for marginal likelihoods and Bayes factors (see Chapters 3 and 4) being then specially suitable for solving large variable selection problems.

3. With such a large number of models, multiplicity issues arise. Indeed as observed by Scott and Berger (2010), in many fields of science as for example genetics, the number of entertained variables is enormous and scientists do not really trust any of the models but just wish to point out some interesting relations (e.g. which genes produce a certain cancer). Detecting signals (variables actually related to Y) in the presence of so much noise not only becomes a very difficult task, but also multiplicity correction is needed. In this thesis we consider the ideas in Scott and Berger (2010) who control for multiplicity with a suitable choice of the priors over the model space (see Section 2.3.4).

The next section introduces the methodology that will be used in Chapter 3 for developing our proposed prior.

2.3 Objective Bayes variable selection

As shown in Section 1.2.1, posterior probabilities for model selection can be expressed as

$$P(M_i \mid \boldsymbol{y}) = \frac{m_i(\boldsymbol{y})P(M_i)}{\sum_{l=0}^{2^p-1} m_l(\boldsymbol{y})P(M_l)} = \frac{B_{id}P(M_i)}{\sum_{l=0}^{2^p-1} B_{ld}P(M_l)} \tag{2.3}$$

where $P(M_i)$ for $i = 0, \ldots, 2^p - 1$ are the prior probabilities over the model space, and $m_i(\boldsymbol{y})$ is the prior predictive distribution under model M_i computed at the observed $\boldsymbol{y}$. For the variable selection problem defined in (2.1) and (2.2)

$$m_i(\boldsymbol{y}) = \int f_i(\boldsymbol{y} \mid \boldsymbol{\beta}_i, \boldsymbol{\beta}_0, \sigma)\, \pi_i(\boldsymbol{\beta}_i, \boldsymbol{\beta}_0, \sigma)\, d(\boldsymbol{\beta}_i, \boldsymbol{\beta}_0, \sigma), \tag{2.4}$$

for $i = 0 \ldots 2^p - 1$, where $\pi_i(\boldsymbol{\beta}_i, \boldsymbol{\beta}_0, \sigma)$ is the prior distribution for parameters under each model M_i. Chapter 3 is devoted to introduce and study our proposal for assigning this distribution, which is partly inspired in the Conventional approach of Jeffreys (1961) (which was briefly described in Section 1.3.1 and will be studied further in Section 2.3.3).

For the elicitation of $\pi_i(\boldsymbol{\beta}_i, \boldsymbol{\beta}_0, \sigma)$ as well as for the study of Bayes factors we will need some theoretical concepts of linear models that we review next in Section 2.3.2.

The elicitation of prior probabilities over the model space $P(M_i)$ is not the central part of this work. Hence we content ourselves with briefly reviewing in Section 2.3.4, several proposals, to account for multiplicity including the one by Scott and Berger (2010).

2.3.1 Base model

In (2.3) we only consider Bayes factors comparing each model M_i with a fixed model M_d, referred to as the base model (see Section 1.2.1). In

principle, the choice of this base model should be irrelevant (as it is the case in a subjective Bayes approach) but for the objective Bayes approach it is not so. In particular, for the Conventional approach of Jeffreys (1961) adopted in this thesis, the choice of $\pi_i(\boldsymbol{\beta}_i, \boldsymbol{\beta}_0, \sigma)$ depends on the model we are comparing M_i with, in the sense that it depends on which ones are the common and extra parameters and these are different for different models.

Our choice for the base model M_d is the null model M_0. We consider this to be the most sensible choice since then every Bayes factor is computed between the nested models M_i and M_0 with the common parameters clearly being those of M_0, $(\boldsymbol{\beta}_0, \sigma)$ while $\boldsymbol{\beta}_i$ is the new parameter. Hence, this choice produces unique model-specific priors $\pi_i(\boldsymbol{\beta}_i, \boldsymbol{\beta}_0, \sigma)$ under every model M_i. In this spirit, and without loss of generality we express the prior $\pi_i(\boldsymbol{\beta}_i, \boldsymbol{\beta}_0, \sigma)$ as:

$$\pi_i(\boldsymbol{\beta}_i, \boldsymbol{\beta}_0, \sigma) = \pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma)\pi_i(\boldsymbol{\beta}_0, \sigma).$$

Note that for any other choice of M_d the common and uncommon parameters will change depending on the specific comparison M_i vs M_d and so will do the prior distribution under M_d, $\pi_d(\boldsymbol{\beta}_d, \boldsymbol{\beta}_0, \sigma)$ which does not seem desireable. Other choices for the base model and some discussion about this issue can be found in Pérez (1998); Casella and Moreno (2006); Liang et al. (2008) and references therein.

2.3.2 Some theoretical aspects of linear models

We briefly review next some theoretical concepts of linear regression which will be needed both in the choice of $\pi_i(\boldsymbol{\beta}_i, \boldsymbol{\beta}_0, \sigma)$ and in the derivation of the properties of the resulting B_{i0}. We refer the interested reader to, for example, Guttman (1982) or Rao (1965) for notation and a deep treatment of linear models.

Estimators

Consider the model M_i in (2.2). Let $\boldsymbol{\beta}$ denote its entire $(k_0 + k_i)$-vector of regressors, that is, $\boldsymbol{\beta}^t = [\boldsymbol{\beta}_0^t \mid \boldsymbol{\beta}_i^t]$ with $\boldsymbol{X} = [\boldsymbol{X}_0 \mid \boldsymbol{X}_i]$ being its full $n \times (k_0 + k_i)$ design matrix. The maximum likelihood estimator for $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y},$$

and its sampling distribution is normal, centered at $\boldsymbol{\beta}$ and with covariance matrix

$$\mathrm{Cov}[\hat{\boldsymbol{\beta}}] = \sigma^2(\boldsymbol{X}^t\boldsymbol{X})^{-1}. \tag{2.5}$$

Note that $(\boldsymbol{X}^t\boldsymbol{X})^{-1}$ exists if and only if $\boldsymbol{X}$ is a full rank matrix so that n needs to be at least $k_0 + k_i$.

In objective Bayes model selection, appropriate choice of the scale of the prior is crucial. For reasons that will become clear in the next chapter, we base our choice for the scale of $\pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma)$ on the marginal covariance matrix of $\hat{\boldsymbol{\beta}}_i$. This matrix is, of course, the corresponding block of $\mathrm{Cov}[\hat{\boldsymbol{\beta}}]$ in (2.5) and can be easily derived with simple algebraic manipulations. This covariance matrix also arises when considering an orthogonal parameterization of M_i (i.e. a parameterization for which the Fisher information matrix is block diagonal). Jeffreys, and after him, Zellner and Siow popularized such reparameterization, which has been broadly followed in virtually all Conventional approaches to model selection. Indeed in the Conventional approach, as well as in many other objective Bayes approaches to model selection, the orthogonal parameterization is usually stated to be "assumed without loss of generality". Until now this has been, in fact, a main requirement for assessing a common prior for common parameters $\pi(\boldsymbol{\beta}_0, \sigma)$ but wether or not it is also required for assessing a Conventional Prior for $\pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma)$ seems to not have been discussed. We hope to provide some insights in this regard. In Section 3.4 we show that interestingly, the orthogonal parameterization

is not required and that the choice of $\pi(\boldsymbol{\beta}_0, \sigma)$ can be fully justified from invariance arguments.

The orthogonal reparameterization of model M_i for Conventional Bayes analysis is:

$$\boldsymbol{y} = \boldsymbol{X}_0\boldsymbol{\gamma} + \boldsymbol{V}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}, \;\; \boldsymbol{\varepsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n), \tag{2.6}$$

where $\boldsymbol{V}_i = (\boldsymbol{I}_n - \boldsymbol{P}_0)\boldsymbol{X}_i$ with $\boldsymbol{P}_0 = \boldsymbol{X}_0(\boldsymbol{X}_0^t\boldsymbol{X}_0)^{-1}\boldsymbol{X}_0^t$, and $\boldsymbol{\gamma} = \boldsymbol{\beta}_0 + (\boldsymbol{X}_0^t\boldsymbol{X}_0)^{-1}\boldsymbol{X}_0^t\boldsymbol{X}_i\,\boldsymbol{\beta}_i$.

Notice that, in the new parameterization $(\boldsymbol{\gamma}, \boldsymbol{\beta}_i, \sigma)$ the parameter $\boldsymbol{\beta}_i$ defining the different models remains unchanged, the new design matrix is $\boldsymbol{X}^* = [\boldsymbol{X}_0 \mid \boldsymbol{V}_i]$, and $\boldsymbol{X}_0^t\boldsymbol{V}_i = \boldsymbol{0}$ so that $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}_i$ are orthogonal in Fisher sense. In fact, the Fisher matrix in this parameterization is proportional to:

$$\boldsymbol{X}^{*t}\boldsymbol{X}^* = \left[\begin{array}{c|c} \boldsymbol{X}_0^t\boldsymbol{X}_0 & \boldsymbol{0} \\ \hline \boldsymbol{0} & \boldsymbol{V}_i^t\boldsymbol{V}_i \end{array}\right],$$

whose inverse is:

$$(\boldsymbol{X}^{*t}\boldsymbol{X}^*)^{-1} = \left[\begin{array}{c|c} (\boldsymbol{X}_0^t\boldsymbol{X}_0)^{-1} & \boldsymbol{0} \\ \hline \boldsymbol{0} & (\boldsymbol{V}_i^t\boldsymbol{V}_i)^{-1} \end{array}\right].$$

Hence, in this reparameterization

$$\hat{\boldsymbol{\beta}}_i = (\boldsymbol{V}_i^t\boldsymbol{V}_i)^{-1}\boldsymbol{V}_i^t\,\boldsymbol{y},$$

and

$$\mathrm{Cov}[\hat{\boldsymbol{\beta}}_i] = \sigma^2(\boldsymbol{V}_i^t\boldsymbol{V}_i)^{-1}, \tag{2.7}$$

which is equal to the corresponding block of (2.5). In fact, for any reparameterization such that $\boldsymbol{\beta}_i$ remains unchanged, the corresponding maximum likelihood estimator, $\hat{\boldsymbol{\beta}}_i$ doesn't change and neither do the covariance matrix of $\hat{\boldsymbol{\beta}}_i$. Nevertheless, it is important to remark that $\sigma^2(\boldsymbol{X}_i^t\boldsymbol{X}_i)^{-1}$, with $\boldsymbol{X}_i$ being the original design matrix, is not the covariance matrix of $\hat{\boldsymbol{\beta}}_i$ unless $\boldsymbol{X}_i^t\boldsymbol{X}_0 = \boldsymbol{0}$.

Hence, for the purposes of choosing the scale for the model-specific conditional prior for $\boldsymbol{\beta}_i$, *the orthogonalization is not needed* as long as this scale is expressed in terms of $\text{Cov}[\hat{\boldsymbol{\beta}}_i]$ which, as remarked above, is invariant to transformations that leave $\boldsymbol{\beta}_i$ unchanged (as the orthogonal reparameterization) and not in terms of the original design matrices, which are not. This important point seems to have gone unnoticed in the relevant literature. For simplicity in notation we write this covariance matrix in terms of $\boldsymbol{V}_i$.

Test statistics

The model selection problem of choosing between M_0 in (2.1) and M_i in (2.2) can be alternatively expressed as the hypothesis testing of

$$H_0 : \boldsymbol{\beta}_i = \mathbf{0} \quad \text{vs} \quad H_i : \boldsymbol{\beta}_i \neq \mathbf{0}.$$

The usual test statistic for solving this testing is:

$$F = \frac{SSE_0 - SSE_i}{SSE_i} \cdot \frac{n - k_i - k_0}{k_i}, \tag{2.8}$$

where $SSE_0 = \boldsymbol{y}^t(\boldsymbol{I}_n - \boldsymbol{P}_0)\,\boldsymbol{y}$ is the residual sum of squares under model M_0 and $\boldsymbol{P}_0 = \boldsymbol{X}_0(\boldsymbol{X}_0^t\boldsymbol{X}_0)^{-1}\boldsymbol{X}_0^t$. Similarly $SSE_i = \boldsymbol{y}^t(\boldsymbol{I}_n - \boldsymbol{P}_i)\,\boldsymbol{y}$ is the residual sum of squares under M_i, where $\boldsymbol{P}_i = \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t$ for $\boldsymbol{X} = [\boldsymbol{X}_0 \mid \boldsymbol{X}_i]$ (or equivalently $\boldsymbol{X} = [\boldsymbol{X}_0 \mid \boldsymbol{V}_i]$). Under H_0, F follows a Snedecor's F-distribution with $(k_i,\ n - k_0 - k_i)$ degrees of freedom (more details in Guttman, 1982).

The F-statistic can also be written in terms of the ratio of the residual sum of squares under each model, $Q_{i0} = SSE_i/SSE_0$ as:

$$F = \frac{n - k_i - k_0}{k_i}(Q_{i0}^{-1} - 1).$$

We will repeatedly use Q_{i0} in future chapters.

Since SSE_i is always smaller than SSE_0 (more complex models always produce smaller residuals than simpler ones), Q_{i0} takes values in $[0, 1]$. Values of $Q_{i0} \approx 0$ indicate that $SSE_i << SSE_0$ and hence, intuitively, that the data supports M_i. On the other hand, values of $Q_{i0} \approx 1$ indicate $SSE_i \approx SSE_0$ and hence data gives as much support to the simpler model M_0 as possible.

The ratio of residual sum of squares Q_{i0} is extensively studied in Moreno et al. (2009) and Casella et al. (2009). In particular, in each of the 2^p comparisons M_i vs M_0 it is shown that the corresponding distribution of Q_{i0} given the "true" model

$$M_T : \mathcal{N}_n(\boldsymbol{X}_T\boldsymbol{\beta}_T,\, \sigma^2\boldsymbol{I}_n); \;\; M_T \in \{M_0, \ldots, M_{2^p-1}\},$$

is

$$Q_{i0} \mid M_T \sim \mathcal{B}e(\frac{n-k_i-k_0}{2},\; \frac{k_i}{2},\; n\delta_1,\; n\delta_2),$$

a doubly non-central beta distribution with non centrality parameters:

$$\begin{aligned} \delta_1 &= \boldsymbol{\beta}_T^t \, \frac{\boldsymbol{X}_T^t(\boldsymbol{I}_n - \boldsymbol{P}_i)\boldsymbol{X}_T}{n\,\sigma^2} \, \boldsymbol{\beta}_T \\ \delta_2 &= \boldsymbol{\beta}_T^t \, \frac{\boldsymbol{X}_T^t(\boldsymbol{P}_i - \boldsymbol{P}_0)\boldsymbol{X}_T}{n\,\sigma^2} \, \boldsymbol{\beta}_T. \end{aligned}$$

(See Appendix A for a description of the doubly non-central beta distribution.)

When the true model is one of the models in the specific comparison M_i vs M_0 (for which Q_{i0} is computed) these non-centrality parameters take the following values:

- If $M_T = M_0$, then $\delta_1 = \delta_2 = 0$ and the sampling distribution of Q_{i0} is

$$Q_{i0} \mid M_0 \sim \mathcal{B}e(\frac{n-k_i-k_0}{2},\; \frac{k_i}{2})$$

with mean and mode given respectively by

$$\mathrm{E}[Q_{i0} \mid M_0] = \frac{n - k_i - k_0}{n - k_0}; \tag{2.9}$$
$$\mathrm{Mode}[Q_{i0} \mid M_0] = \frac{n - k_i - k_0 - 2}{n - k_0 - 4}, \quad \text{for } k_i \geq 3 \text{ and } n \geq k_i + k_0 + 3.$$

- If $M_T = M_i$, we have $\delta_1 = 0$ and $\delta_2 = \delta_{i0}$, where for $\boldsymbol{\beta}^t = [\boldsymbol{\beta}_0^t \mid \boldsymbol{\beta}_i^t]$ and $\boldsymbol{X} = [\boldsymbol{X}_0 \mid \boldsymbol{X}_i]$

$$\delta_{i0} = \boldsymbol{\beta}^t \, \frac{\boldsymbol{X}^t(\boldsymbol{I}_n - \boldsymbol{P}_0)\boldsymbol{X}}{n\,\sigma^2} \, \boldsymbol{\beta}.$$

  Moreno et al. (2009) consider δ_{i0} as a measure of the distance between M_i and M_0 for a given sample of size n.

The next lemma (which follows from Lemma 1 in Moreno et al., 2009) shows the asymptotic behavior of the distribution of Q_{i0} under the true model (when n grows).

Lemma 2.1. *The sampling distribution of Q_{i0} degenerates to a point mass at*

$$q_M = \frac{1 + \delta_1}{1 + \delta_1 + \delta_2}$$

as $n \to \infty$.

Where $q_M = 1$ for $M_T = M_0$ and $q_M = 1/(1 + \delta)$ for $M_T = M_i$, with $\delta = \lim_{n\to\infty} \delta_{i0}$.

Proof. see Appendix D.1 □

When M_i is the true model, Casella et al. (2009) and Moreno et al. (2009) interpret $\delta = \lim_{n\to\infty} \delta_{i0}$ as the limiting "distance" between M_i and M_0. In fact, assuming that

$$S = \lim_{n\to\infty} \frac{\boldsymbol{X}^t(\boldsymbol{I}_n - \boldsymbol{P}_0)\boldsymbol{X}}{n}$$

is a constant semi-definite positive matrix (which as shown in Casella et al. (2009) is not a too demanding condition) δ depends just on the value of $\boldsymbol{\beta}$. In particular, for a fixed value of $\boldsymbol{\beta}_0$, it grows with $||\boldsymbol{\beta}_i||$ (indeed, when $||\boldsymbol{\beta}_i||^2 \to \infty$ with $||\boldsymbol{\beta}_i||^2 = \boldsymbol{\beta}_i^t \boldsymbol{X}_i^t \boldsymbol{X}_i \boldsymbol{\beta}_i$, then $\delta \to \infty$). This gives support to the intuitive idea of considering δ as a limiting distance between M_0 and M_i.

From Lemma 2.1 it is easy to see that if $\delta \to \infty$ then $q_M \to 0$, so that the sampling distribution of Q_{i0} degenerates to a point mass at 0.

Figure 2.1 shows the density of Q_{i0} when the true model is M_i for varying n (left) and δ_2 (right) (recall that in this case $\delta_1 = 0$) when $k_i = 3$ and $k_0 = 1$. It can be seen in the left plot that the distribution concentrates sharply around $1/(1 + \delta_2)$ as n grows (approximately around 0.01 for $\delta_2 = 100$). In the right plot we can see that this point mass moves to 0 as δ_2 grows.

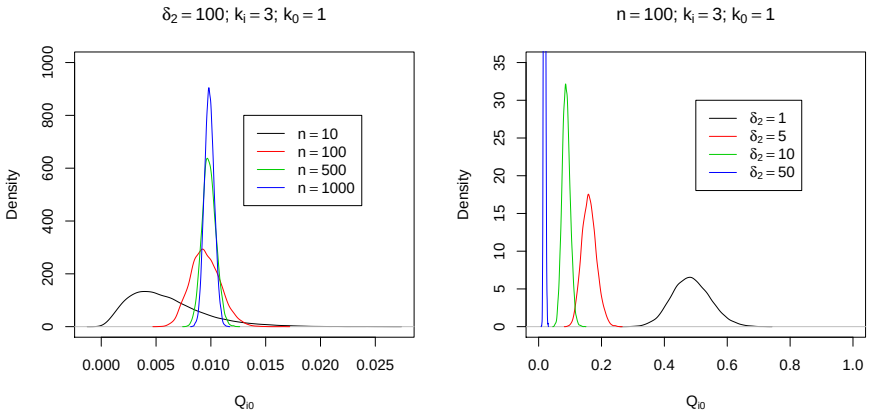


FIGURE 2.1: *Comparisons of non-central beta densities with non centrality parameter $n\delta_2$ for varying n and $\delta_2 = 100$ (left) and for varying δ_2 for $n = 100$ (right)*

Note that the fact that Q_{i0} tends to concentrate around 0 under M_i and around 1 under M_0 agrees with the intuition that values of $Q_{i0} \approx 0$ support M_i and values of $Q_{i0} \approx 1$ support M_0.

2.3.3 Conventional priors for variable selection in the literature

In this section we review some of the main ideas in the Conventional approach of Jeffreys (1961) for the elicitation of $\pi_i(\boldsymbol{\beta}_i, \boldsymbol{\beta}_0, \sigma)$. As previously mentioned, we express this prior as

$$\pi_i(\boldsymbol{\beta}_i, \boldsymbol{\beta}_0, \sigma) = \pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma)\pi_i(\boldsymbol{\beta}_0, \sigma).$$

In Chapter 1 we refer to Conventional priors as those model specific distributions based on Jeffreys' ideas, and particularly, on his arguments for testing a normal mean ($y_i \sim \mathcal{N}(\lambda, \sigma^2)$, $H_0 : \lambda = 0$ vs $H_1 : \lambda \neq 0$). These ideas were extended to variable selection by Zellner and Siow (1980).

Specifically, Jeffreys' idea was to assign, conditional on the "old" parameters, a proper prior distribution for the "new" parameters, $\pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma)$ and, a non-informative prior (possibly improper) for the "old" parameters $\pi_i(\boldsymbol{\beta}_0, \sigma)$. His arguments for doing so were heavily based on orthogonality. In particular, Jeffreys (and many authors after him) argue that using an improper prior for common parameters is intuitively justified only if the old and new parameters are orthogonal in Fisher sense (see Hsiao, 1997; Kass and Vaidyanathan, 1992). This extended practice has become the agreed upon "default" objective choice for common parameters, but there seems not to be any theoretical arguments behind it.

As prior distributions for common parameters, Jeffreys (1961) and Zellner and Siow (1980) use the reference or independent Jeffreys' prior, $\pi_i(\boldsymbol{\beta}_0, \sigma) = 1/\sigma$, under each model M_i, $i = 0, \ldots, 2^p - 1$. In the next chapter we also recommend the use of this prior as the objective prior for common parameters under every model. Our choice, however, does not require orthogonality and is justified on theoretical basis (see Section 3.4.1).

Jeffreys also states some specific requirements for the Conventional choice of $\pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma)$:

- *It should be proper.* To avoid the indeterminacy of Bayes factors.
- *It should be symmetric around the null hypothesis.* Quoting Jeffreys (1961):

  > *... we must say that the mere fact that it has been suggested that λ is zero corresponds to some presumption that it is fairly small.*

  Hence Jeffreys centered his prior at $\lambda = 0$. Also Zellner and Siow, considering nested models in the variable selection problem, stated their null hypothesis as $H_0 : \boldsymbol{\beta}_i = \mathbf{0}$ and so centered their choice at zero.
- *It should be scaled, in some sense, by the scale of the entertained models.* Jeffreys (1961) remarks that

  > *for consideration of similarity it [the prior for λ under H_1] must depend on σ since there is nothing in the problem except σ to give scale for λ.*

  In particular, for the variable selection problem, Zellner and Siow (1980) propose to scale by "... a matrix suggested by the form of the Fisher information matrix". Their actual choice for the scale matrix was n times the matrix $\sigma^2(\boldsymbol{V}_i^t \boldsymbol{V}_i)^{-1}$ where $\boldsymbol{V}_i$ is the design matrix of model M_i in the orthogonal parameterization, see (2.6). This scale matrix is, in fact, the inverse of the Fisher information matrix when the model is in its orthogonal parameterization.
- *It should have no finite moments.* In the context of hypothesis testing regarding a univariate normal mean, λ, and unknown variance σ^2, Jeffreys realized of some undesirable features when the normal

distribution $N(\lambda \mid \mu, \sigma^2)$ with known μ is used as a prior for testing hypothesis. These are in part a consequence of the shape of the normal tails. He proposed instead taking a prior distribution with no moments and hence heavier tails than those of the normal. Having no finite moments is quite a strong requirement. In fact, a Student's t-distribution with two degrees of freedom is considered a heavy tailed distribution but it does have a mean.

Jeffreys, Zellner and Siow, agreed that the "simplest" distribution satisfying these requirements is a Cauchy distribution. Specifically, for the variable selection problem Zellner and Siow (1980)'s proposal takes the form:

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

This distribution can also be written as

$$\pi_i^{ZS}(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma) = \int_0^\infty \mathcal{N}_{k_i}(\boldsymbol{\beta}_i \mid 0, g\,\sigma^2(\boldsymbol{V}_i^t \boldsymbol{V}_i)^{-1})\, h_n^{ZS}(g)\, dg \qquad (2.10)$$

where $h_n^{ZS}(g)$ is

$$h_n^{ZS}(g) = IGa(g \mid \frac{1}{2}, \frac{n}{2}), \quad g \geq 0.$$

Distributions following the structure in (2.10) are known as *scale mixture of normals* and the corresponding $h_n(g)$ is referred to as *mixing function*. This can also be interpreted as building the prior in a hierarchical way (first using a normal prior for $\boldsymbol{\beta}_i$ given g and then using a proper prior $h_n(g)$ over g). Berger (1985) points out that eliciting priors in this hierarchical way usually induce heavier tails.

Other authors also use this hierarchical structure for their proposal for prior distributions. For instance, Liang et al. (2008) use, in (2.10), the mixing function

$$h_n^L(g) = \frac{1}{2n}(1 + \frac{g}{n})^{-\frac{3}{2}}, \quad g \geq 0.$$

Previously, Zellner (1986) takes a degenerated point mass for g (i.e. choosing a specific value for g) in order to keep the normal form and thus obtaining closed-forms expression for the Bayes factors. These are popularly referred to as *g-priors*. Obviously the choice of g in g-priors becomes crucial (Liang et al., 2008) and it can not be taken to be arbitrarily large (see the problems with vague priors in Section 1.3.1). There have been several proposals for suitable choice of g, some of which are discussed in Liang et al. (2008). The closed-form expressions for the resulting Bayes factors have made the g-priors very popular despite its not entirely satisfactory behavior for some values of g.

Our particular proposal for $\pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma)$ introduced in Chapter 3, follows the spirit of Conventional Priors described here and, as will be shown in Section 3.3.1, it also can be expressed as a scale mixture of normals.

2.3.4 Prior distributions over the model space. Multiplicity issues

Prior probabilities over the model space $P(M_i)$ are important ingredients for computing posterior probabilities. When no information is available, it might seem reasonable to take them equal for each model. In particular, for variable selection this is $P(M_i) = 1/2^p$ for $i = 0, \ldots, 2^p - 1$. However, a closer inspection makes it obvious that with this uniform prior, the most probable models are those for which the number of extra covariates, k_i is around $p/2$ (since there are many more models of this complexity). This effect becomes more pronounced as p grows. This is clearly inadequate if the possible number of explanatory variables have been chosen intentionally very large (trying to "discover" influential variables in a, somewhat blind, way) whereas it is expected that only some few variables affect the response. Interestingly these ideas are connected to multiplicity as it is shown in Scott and Berger (2010). Specifically they show how $P(M_i)$ can be chosen to account for multiplicity control.

Multiplicity control is particularly needed in scenarios, like variable selection, with huge model spaces. Indeed, Scott and Berger (2010) argue that it is important to account for the increasing number of models and hence for the difficulty of detecting influential covariates when p and, presumably the background noise, grow.

This multiplicity penalty must not be confused with Occam's razor penalty. Occam's razor is a penalty against complexity of models and is inherent to Bayesian analysis due to the behavior of prior predictive distributions. But the Bayes factor between any two models remains fixed no matter how many models we are comparing. Hence, a constant and equal $P(M_i)$ across models, producing Bayes factors as posterior odds (see Section 1.2.1), accounts for Occam's razor but it does not account for multiplicity.

As Scott and Berger (2010) point out, a standard practice in variable selection is to give probability q to each variable being in the model, and consider their inclusion in a model as exchangeable Bernoulli trials. That is

$$P(M_i \mid q) = q^{k_i}(1-q)^{p-k_i}. \tag{2.11}$$

A fixed value of q (independent of p) does not control multiplicity. For instance, selecting $q = 1/2$ gives the same results as giving an equal prior probability to each model. Scott and Berger (2010) show that treating q as an unknown parameter and allowing learning from data results in an automatic penalty for multiplicity. Choosing a uniform prior for q in (2.11), and integrating it out results (Scott and Berger, 2010) in

$$P(M_l) = \frac{1}{p+1}\binom{p}{k_l}^{-1}. \tag{2.12}$$

The corresponding posterior probabilities are:

$$P(M_i \mid \boldsymbol{y}) = \frac{B_{i0}\binom{p}{k_i}^{-1}}{\sum_{l=0}^{2^p-1} B_{l0}\binom{p}{k_l}^{-1}}.$$

Note that the prior (2.12) is equivalent to assessing an uniform prior to each dimension k, that is $P(k) = 1/(p+1)$ for $k = 0, \ldots, p$, and then dividing this probability equally among the $\binom{p}{k}$ models of dimension $k_l = k$.

It is interesting to remark that Scott and Berger proposal results in marginal prior inclusion probability of $1/2$ for each variable, the same as the one for the constant prior $P(M_i)$ but the behavior is very different due to the way of apportioning the probability among models.

Chapter 3

Prior specification through Conventional arguments and invariance

3.1 Introduction

In this chapter we present and formally justify a novel proposal for prior distributions in the variable selection scenario introduced in Chapter 2. Recall that this problem consists of 2^p comparisons, M_i as given in (2.2) vs M_0 in (2.1). In each of these comparisons we need to specify a prior distribution $\pi_i(\boldsymbol{\beta}_i, \boldsymbol{\beta}_0, \sigma)$, under model M_i. As commented in Section 2.3.1, following Jeffreys' Conventional approach it is convenient to express these priors as:

$$\pi_i(\boldsymbol{\beta}_i, \boldsymbol{\beta}_0, \sigma) = \pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma)\, \pi_i(\boldsymbol{\beta}_0, \sigma) \tag{3.1}$$

For the specification of $\pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma)$ we extend proposals by Strawderman (1971, 1973) and Berger (1976, 1980, 1985). These proposals were

developed in a context of robust and minimax estimation. In this chapter we generalize these priors and adapt them to the scenario of variable selection (see Sections 3.2 and 3.3). Interestingly, these prior distributions achieve good properties, also in this scenario for which they were not originally developed. Their suitability for model selection remained unnoticed until Berger et al. (2010a) rescued them for developing BIC-like expressions. Many of these desirable properties are due to the robust, thick tails of these priors (recall that having thick tails is related to the fourth requirement of Jeffreys' conventional approach, see Section 2.3.3). Priors with thick tails are typically difficult to work with; in particular they usually do not produce closed-form expressions for the marginal likelihood; interestingly, as we show in Section 3.5.1, our proposal does. This attractive property also makes it particularly suitable for model selection, specially for very large model spaces, which is often the case in variable selection problems.

After choosing $\pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma)$, we show (see Section 3.4) that the usual (but somewhat ad hoc) choice of $\pi_i(\boldsymbol{\beta}_0, \sigma) = 1/\sigma$ can be formally derived and its use in our approach to variable selection is justified with invariance arguments.

Finally in Section 3.5 we show some good properties of our ultimate proposal for the joint prior which, along with the previous arguments, justifies all the components in the prior elicitation (3.1).

3.2 An extension of Berger's Robust priors

3.2.1 Original idea

We begin by succinctly reviewing the arguments in Strawderman (1971, 1973) and Berger (1976, 1980, 1985). Their work was developed for the

estimation of a k-variate normal mean, $\boldsymbol{\theta}$,

$$\boldsymbol{X} \sim \mathcal{N}_k(\boldsymbol{\theta}, \boldsymbol{\Sigma}) \tag{3.2}$$

where $\boldsymbol{\Sigma}$ is a known covariance matrix.

Strawderman original prior distribution was proposed to derive admissible minimax estimators for $\boldsymbol{\theta}$ when $\boldsymbol{\Sigma} = \boldsymbol{I}_k$, Berger (1976, 1980, 1985) generalizes this idea to any known $\boldsymbol{\Sigma}$. Berger realizes that the resulting prior distribution is extremely well suited for Bayesian robust analyses, producing estimators for $\boldsymbol{\theta}$ which are still minimax and also robust.

To introduce this robust prior we follow the notation and developments in Berger (1980, 1985), where prior beliefs are incorporated through a guess $\boldsymbol{\mu}$ for $\boldsymbol{\theta}$, and a matrix $\boldsymbol{A}$ reflecting the accuracy of this guess. This prior is defined in a hierarchical way. Specifically:

$$\begin{aligned} &\pi(\boldsymbol{\theta} \mid \lambda) = \mathcal{N}_k(\boldsymbol{\theta} \mid \boldsymbol{\mu},\, \lambda^{-1}\rho\,(\boldsymbol{\Sigma} + \boldsymbol{A}) - \boldsymbol{\Sigma}), \\ &\pi(\lambda) = a\lambda^{a-1};\;\; \lambda \in [0, 1]. \end{aligned} \tag{3.3}$$

The rather odd looking form of the covariance matrix of the normal distribution considerably simplifies calculations. Indeed, despite the fact that this prior has not itself a closed-form, it results in simple expressions for the corresponding estimators and also for the corresponding prior predictive distribution (see Berger, 1980, 1985).

Notice that this prior is defined up to two parameters, a and ρ, which values need to be chosen based on desirable properties of the resulting inferences. In this sense, it is important to remark that this prior distribution is a proper density only for values of $a > 0$. We briefly summarize here several choices of a and ρ which were considered for estimation purposes:

- Parameter a. Berger (1980) chooses $a = -1$ (resulting in an improper distribution), because the corresponding prior produces a robust minimax estimator with good frequentist (coverage) properties. In fact, the resulting estimator will be admissible for any $a \geq -1$ but good coverage properties require $a \leq -1$, hence, a natural choice for a is $a = -1$. However, in his 1985 book, for a purely robust Bayesian analysis, Berger's choice is $a = 1/2$. This value results in a proper robust prior which also produces robust minimax estimators.

- Parameter ρ. Berger (1980) argues that a good choice is

$$\rho = \frac{2a + k}{2a + 2 + k}.$$

  This choice was based on the similarity of the resulting estimator to the best linear estimator in scenarios where the latter proved to be reasonable. For the two choices of a above this gives $\rho = (k-2)/k$ in Berger (1980) and $\rho = (k+1)/(k+3)$ in Berger (1985). Anyway, Berger (1985, p. 240, Theorem 6) and Berger (1980, Theorem 2.2.1) show that the resulting Bayes estimator will be minimax for $k \geqslant 5$ no matter which ρ is used.

Although these choices of a and ρ have very good properties for estimation we can not adopt them blindly without investigating optimal properties for model selection. In particular, we can not choose $a = -1$ since the resulting prior is improper (see Section 1.2.1).

3.2.2 Generalized formulation

We generalize (3.3) by introducing a new (adjustable) parameter b as follows

$$\begin{aligned} &\pi(\boldsymbol{\theta} \mid \lambda) = \mathcal{N}_k(\boldsymbol{\theta} \mid \boldsymbol{\mu},\, \lambda^{-1}\rho\,(b\,\boldsymbol{\Sigma} + \boldsymbol{A}) - b\,\boldsymbol{\Sigma}); \\ &\pi(\lambda) = a\lambda^{a-1};\;\; \lambda \in [0,1]. \end{aligned} \tag{3.4}$$

Berger's prior (3.3) corresponds to the particular choice $b = 1$. By introducing this new parameter, the Robust prior in (3.4) can be seen to generalize other proposals in literature (further details in Sections 3.3.1 and 3.3.2).

In the next section we adapt the generalized version of Berger's Robust prior in (3.4) to be used as model specific priors in variable selection.

3.3 Adapting Berger's Robust priors for variable selection

In adapting (3.4) for variable selection first note that variable selection fits the estimation of a multivariate normal mean framework considered by Strawderman and Berger. Indeed, for each model M_i and given $(\boldsymbol{\beta}_0,\, \sigma)$, it suffices to consider the sampling distribution of the maximum likelihood estimator $\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ of $\boldsymbol{\beta}_i$ (see Section 2.3.2, Estimators):

$$\hat{\boldsymbol{\beta}}_i \sim \mathcal{N}_{k_i}(\boldsymbol{\beta}_i, \sigma^2\,(\boldsymbol{V}_i^t \boldsymbol{V}_i)^{-1}),$$

where $\boldsymbol{V}_i = (\boldsymbol{I}_n - \boldsymbol{P}_0)\boldsymbol{X}_i$ is the design matrix in the orthogonal parameterization of the model. Here the multivariate normal mean of interest is $\boldsymbol{\beta}_i$ (of dimension k_i).

Hence, following the ideas of Strawderman and Berger we use the generalized version of the Robust prior in (3.4) as our proposal for $\pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma)$:

$$\pi_i^R(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma) = \int_0^1 \mathcal{N}_{k_i}(\boldsymbol{\beta}_i \mid \boldsymbol{\mu}, \lambda^{-1}\rho_i\,(b\,\boldsymbol{\Sigma} + \boldsymbol{A}) - b\,\boldsymbol{\Sigma})\,\pi(\lambda)\,d\lambda, \quad (3.5)$$

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

Note that (3.5) does not depend on $\boldsymbol{\beta}_0$ so we are implicitly assuming that $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_0$ are independent given σ.

Also note that in (3.5) the parameter ρ has a subindex i, reflecting the fact that we will be considering differing ρ_i for each model M_i. This is in agreement with the choices in Berger (1980, 1985) which depend on the dimension of $\boldsymbol{\beta}$ (see Section 3.2.1), and hence, different for each model in the particular scenario of variable selection. But ρ_i is not the only parameter in (3.5). In fact we distinguish two sets of parameters:

1. Subjective parameters ($\boldsymbol{\mu}$, $\boldsymbol{A}$). These were introduced in the original proposal to incorporate prior beliefs. In this thesis, we assign them from an objective Bayes point of view (see below).

2. Adjustable parameters (a, b, ρ_i). These are chosen so as to endow the prior with properties which are desirable for variable selection. The ultimate choice for (a, b, ρ_i) will rely on further theoretical results and is delayed till Chapter 5.

In objective Bayesian analyses $\boldsymbol{\mu}$ and $\boldsymbol{A}$ are not chosen based on any subjective prior information. To guide our choice, we follow instead the Conventional approach desiderata, that is:

- *the resulting prior should be centered at and symmetric around the null hypothesis.* In our approach the null model in every comparison is M_0, which is equivalent to $H_0 : \boldsymbol{\beta}_i = \mathbf{0}$. Hence we take the prior guess to be $\boldsymbol{\mu} = \mathbf{0}$;

- *the resulting prior should be scaled (in some sense) by the variance of the entertained models.* We take $\boldsymbol{A} = n\,\boldsymbol{\Sigma}$ where $\boldsymbol{\Sigma} = \sigma^2(\boldsymbol{V}_i^t\boldsymbol{V}_i)^{-1}$ is the covariance matrix of the maximum likelihood estimator of $\boldsymbol{\beta}_i$. Note that this choice is the popular scale matrix in Zellner and Siow (1980), but our formulation in terms of the covariance matrix of $\hat{\boldsymbol{\beta}}_i$ does not require the preliminary orthogonalization (see Section 2.3.2, Estimators). We have not seen this alternative formulation so far and we think that it is more elegant than the usual one.

In the proposal for $\boldsymbol{A}$, the covariance matrix of the maximum likelihood estimator, $\boldsymbol{\Sigma} = \mathrm{Cov}[\hat{\boldsymbol{\beta}}_i]$ is corrected by the sample size n, in the hope of having the information to roughly be of unitary size. This simple choice, which is appropriate for i.i.d. observations, might not work well in complex situations. Indeed, we think that a better choice would be to correct by an appropriately chosen *effective sample size* (see Berger et al., 2010a,b, and references therein). This attractive idea is still under investigation and there is not general agreement in literature about which is the best choice for effective sample size. For this reason, in this work, we consider the default choice of taking the sample size n as effective sample size. Note that all the properties and developments use n and may not be applicable to correction by other effective sample sizes. Anyway, in the examples (see Chapter 6) we investigate the impact of correcting by other definitions of effective sample size using the ideas in Berger et al. (2010b).

With the choices for $\boldsymbol{\mu}$ and $\boldsymbol{A}$ detailed above our proposal is as follows:

Definition 3.1 (Conventional Robust prior)**.** The conditional prior distribution under M_i in (3.1) is taken to be

$$\pi_i^R(\boldsymbol{\beta_i} \mid \boldsymbol{\beta}_0, \sigma) = \int_0^1 \mathcal{N}_{k_i}(\boldsymbol{\beta_i} \mid \mathbf{0}, \left(\lambda^{-1}\rho_i\,(b+n) - b\right)\,\boldsymbol{\Sigma})\,\pi(\lambda)\,d\lambda, \quad (3.6)$$

where $\boldsymbol{\Sigma} = \text{Cov}[\hat{\boldsymbol{\beta}}_i]$. The prior distribution for λ is taken as in (3.4). Recall that $\boldsymbol{\Sigma} = \sigma^2(\boldsymbol{V}_i^t \boldsymbol{V}_i)^{-1}$.

The restrictions required for the propriety of the resulting prior result in the following parametric space for $(a,\, b,\, \rho_i)$:

$$\mathcal{A} = \{(a, b, \rho_i) : \;\; a > 0, \;\; 1 \leq b \leq n, \;\; \rho_i \geq b/(b+n)\}. \tag{3.7}$$

We remark that as long as the assessment of the scale of $\pi_i^R(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0,\, \sigma)$ is done in terms of $\text{Cov}[\hat{\boldsymbol{\beta}}_i]$ this prior would remain the same for any reparameterization of the type

$$(\boldsymbol{\beta}_0,\, \boldsymbol{\beta}_i,\, \sigma) \to (\boldsymbol{\gamma},\, \boldsymbol{\beta}_i,\, \sigma)$$

that leaves $\boldsymbol{\beta}_i$ unchanged since then the $\text{Cov}[\hat{\boldsymbol{\beta}}_i]$ would not change. Notice that the orthogonal parameterization is of this type. Note also that reparameterizations involving $\boldsymbol{\beta}_i$ would alter the definition and interpretation of the models and does not seem to make sense in variable selection.

Summarizing, for the purpose of uniquely defining the scale of $\pi_i^R(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0,\, \sigma)$, the preliminary orthogonalization is not needed. We will later see that it is not needed either for the purposes of choosing a common objective prior $\pi_i(\boldsymbol{\beta}_0,\, \sigma)$ for all models, (the usual main reason for orthogonalizing in Objective Bayes variable selection).

In the rest of this section and, in general in the rest of this thesis we exhaustively study all the good properties that this prior achieves for variable selection as well as its relationship with other Conventional Priors in the literature.

3.3.1 Connections with other Conventional priors

In a similar way as many other Conventional priors (see Section 2.3.3), the Conventional Robust prior in Definition 3.1 can also be expressed as

the *scale mixture of normals* distribution given in next proposition:

Proposition 3.1. *The Conventional Robust prior in* (3.6) *can be expressed as the following scale mixture of normals*

$$\pi_i^R(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma) = \int_0^\infty \mathcal{N}_{k_i}(\boldsymbol{\beta}_i \mid \mathbf{0}, g\,\sigma^2\,(\boldsymbol{V}_i^t\boldsymbol{V}_i)^{-1})\,h_n^R(g)\,dg \tag{3.8}$$

with

$$h_n^R(g) = \begin{cases} a(\rho_i(b+n))^a(g+b)^{-(a+1)} & g > \rho_i(b+n) - b \\ 0 & \textit{otherwise} \end{cases} \tag{3.9}$$

Proof. It suffices to make the change of variables:

$$g = \lambda^{-1}\rho_i(b+n) - b$$

□

Corollary 3.1. *Liang et al. Hyper-g/n prior (see Liang et al., 2008) is a particular case of the Conventional Robust prior in* (3.6) *for* $b = n$, $a = 1/2$ *and* $\rho_i = 1/2$. *In this case* $\rho_i\,(b+n) - b = 0$ *and the support of* g *in the mixing density is* $[0, \infty)$.

Proof. Just recall (from Section 2.3.3) that

$$h_n^L(g) = \frac{1}{2n}\,(1 + \frac{g}{n})^{-\frac{3}{2}} = \frac{1}{2}\,n^{\frac{1}{2}}\,(g+n)^{-(\frac{1}{2}+1)}.$$

□

An important remark about the shape of the mixing function should be made at this point. Note that for every scale mixture of normals, that is, priors of the form (3.8) for a general mixing density $h_n(g)$, the normal distribution part degenerates to a point mass at $\boldsymbol{\beta}_i = \mathbf{0}$ for $g = 0$. This would in principle lead to Bayes factors close to 1 for mixing densities accumulating a lot of mass in the neighborhood of $g = 0$ when

the likelihood is concentrated around $\boldsymbol{\beta}_i = \mathbf{0}$ (or equivalently, for data supporting M_0). This is because, in such a scenario, the corresponding marginal likelihood $m_i(\boldsymbol{y})$ is large and so is B_{i0} (but still $B_{i0} < 1$ because the data is compatible with M_0). Zellner-Siow proposal avoids (at least partially) this effect because its mixing density is an inverse gamma, which goes to zero when g goes to zero. Similarly, the Conventional Robust prior given by (3.6) and (3.8) avoids concentrating prior mass in $g = 0$ by effectively truncating the mixing density away from $g = 0$. Unlike these two Conventional priors, for the prior in Liang et al. (2008) the mixing density has a positive mass arbitrarily close to $g = 0$. As a consequence, at least for the univariate case $k_i = 1$ Liang et al. conditional prior for β_i is not differentiable at 0. Therefore, Liang et al. prior would seem to produce Bayes factors larger than the Zellner-Siow or robust ones under data compatible with M_0 (larger meaning closer to 1). We will revisit this effect and its consequences when studying our final choice and the resulting Bayes factor in Chapter 5.

In Figure 3.1 we show mixing densities for some values of the hyper-parameters. Dashed lines represent the lower bound of the support of the mixing density for the Conventional Robust prior.

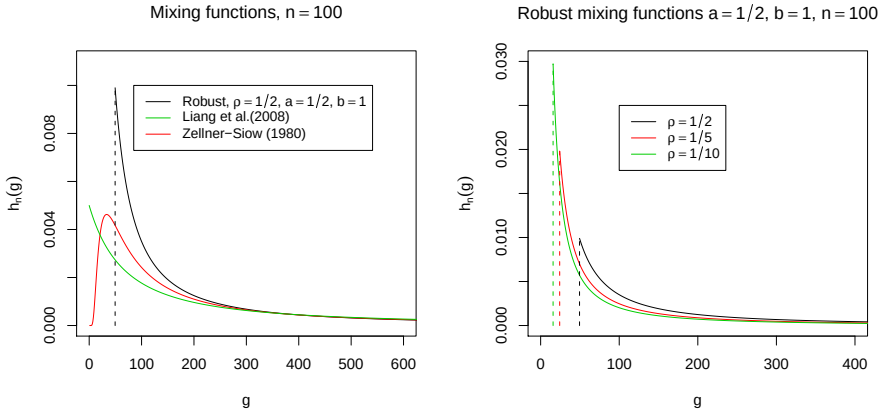


FIGURE 3.1: *Comparisons of different mixing functions.*

3.3.2 Behavior on the tails

One of Jeffreys' requirements for a Conventional prior distribution was that it should have no finite moments, a property that is strongly related to the heaviness of the prior tails. Jeffreys also noticed that the behavior on the tails is related to the notion of information consistency, which we will explore later in Section 4.3.2. Then, having thick tails becomes a very important feature for a model selection prior. Jeffreys (1961) and Zellner and Siow (1980) achieved this property by taking a Cauchy prior.

Berger (1980) mentions that his proposed Robust prior behaves in the tails as a multivariate Student's t-distribution. Indeed, one of his main motivations for using this prior was to keep the robustness properties of Student's tails while considerably simplifying the computations. This remains true for our particular adaptation of his Robust prior to variable selection (3.6), as we show in the following result:

Proposition 3.2. *Let* $\|\boldsymbol{\beta}_i\|^2 = \boldsymbol{\beta}_i^t(\boldsymbol{V}_i^t\boldsymbol{V}_i)\boldsymbol{\beta}_i$*, then*

$$\lim_{\|\boldsymbol{\beta}_i\|^2\to\infty} \frac{\pi_i^R(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0,\, \sigma)}{\mathcal{S}t_{k_i}(\boldsymbol{\beta}_i \mid \mathbf{0},\, \boldsymbol{C}_i^*,\, 2a)} \;=\; 1,$$

where:

$$\boldsymbol{C}_i^* = \frac{c\rho_i\, \boldsymbol{B}_i^*(b,\sigma)}{a},$$

$c = (a\,\Gamma(a))^{1/a}$*, and* $\boldsymbol{B}_i^*(b,\sigma) = \sigma^2(b+n)(\boldsymbol{V}_i^t\boldsymbol{V}_i)^{-1}$

Proof. See Appendix E.1 □

Corollary 3.2. π^R *has no moments for* $a \leq \frac{1}{2}$.

Proof. It follows trivially from the the tail's behavior in Proposition 3.2 □

Notice that Corollary 3.2 indicates that for achieving the fourth requirement of Jeffreys we need to take $a \in (0, 1/2]$. Interestingly for $a = 1/2$,

$b = 1$, $\rho_i = 2/\pi$ and large n, $\pi_i^R(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma)$ and Zellner-Siow's Cauchy prior have equal tails.

Summarizing, our proposed conditional prior in Definition 3.1 follow Jeffreys' desiderata, that is:

1. It is proper (for $n \geq k_0 + k_i$, which is the sample size required for $\text{Cov}[\hat{\boldsymbol{\beta}}_i]$ to be defined, see Section 2.3.2, Estimators).
2. It is centered on the simpler model M_0.
3. It is scaled by $\sigma^2(\boldsymbol{V}_i^t \boldsymbol{V}_i)^{-1}$.
4. It does not have any finite moment (for $a \leq 1/2$).

According to the pioneering work of Jeffreys (1961) these are basic properties for objective priors in model selection to have. These desiderata have been adopted and advocated by many other authors after him (see for example Zellner and Siow, 1980; Berger and Pericchi, 2001; Casella and Moreno, 2006; Bayarri and García-Donato, 2007; Liang et al., 2008, and references therein).

We illustrate the behavior of the prior distributions described above in the univariate case $k_i = 1$. Consider a simple example with $k_0 = 1$, $\sigma = 1$, $n = 100$ and $\boldsymbol{X}_i$ being a simulated vector from a $\mathcal{N}(5, 9)$. In Figure 3.2 we represent the Conventional Robust prior with: i) Berger (1985)'s choices of the parameters ($a = 1/2$, $b = 1$, $\rho_i = 1/2$); ii) Liang et al. (2008) choices ($a = 1/2$, $b = n$, $\rho_i = 1/2$); and the Cauchy prior distribution of Zellner and Siow (1980) similar in the tails to π^R with ($a = 1/2$, $b = 1$, $\rho_i = 2/\pi$). Notice that, as commented in Section 3.3.1 Liang et al. prior has a peak at $\beta_i = 0$ that makes it not differentiable at this point. It is also remarkable that Zellner Siow prior and Liang et al. one are quite close in the tails (see right picture). We will revisit this effect on Chapter 6.

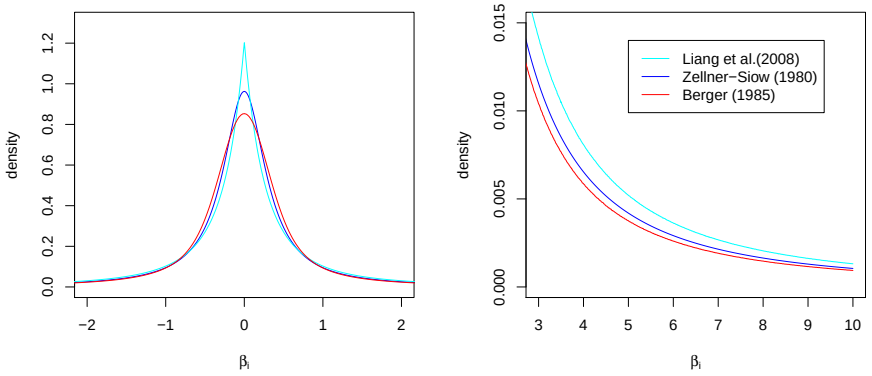


FIGURE 3.2: *Comparisons of different prior distributions in the univariate case. The left picture presents the center of the distributions and the right one the tails (notice that the scales are different).*

3.4 An invariant prior for "common" parameters

Once $\pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma)$ has been chosen as in Definition 3.1, we now turn to the choice of the marginal prior for the common parameters $\pi_i(\boldsymbol{\beta}_0, \sigma)$ and thus complete the specification of our prior distribution under each model M_i

$$\pi_i^R(\boldsymbol{\beta}_i, \boldsymbol{\beta}_0, \sigma) = \pi_i^R(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma)\, \pi_i^R(\boldsymbol{\beta}_0, \sigma).$$

Here we depart from the usual justification which (dating back to Jeffreys, 1961) was based on:

1. Orthogonalizing $\boldsymbol{\beta}_i$ and $(\boldsymbol{\beta}_0, \sigma)$ (in Fisher sense) as in (2.6).

2. Intuitively arguing that in this case, $(\boldsymbol{\beta}_0, \sigma)$ could be taken to have similar meaning in all models.

3. Arguing that under these conditions a common prior distribution $\pi(\boldsymbol{\beta}_0, \sigma)$ could be taken under each of the models.

4. Going one step further and arguing that in the case of taking improper priors for the common $\pi(\boldsymbol{\beta}_0, \sigma)$, the same arbitrary constant can be used in all models.

This intuitively reasonable (but ad-hoc) procedure results in very sensible priors for model selection and is adopted by virtually every Conventional approach to variable selection. However it has not been formally justified.

We have found that invariance arguments, developed in next section, provide a powerful and solid basis to chose the corresponding right Haar measure as prior distribution for the common parameters under each model. In particular, for our framework the right Haar measure turns out to be, $\pi(\boldsymbol{\beta}_0, \sigma) = 1/\sigma$. Hence we take $\pi_i^R(\boldsymbol{\beta}_0, \sigma) = \pi(\boldsymbol{\beta}_0, \sigma) = 1/\sigma$ for $i = 0, \ldots, 2^p - 1$. This prior distribution is also the usual choice in literature but we i) formally justify its choice by invariance arguments and ii) do so without requiring preliminary orthogonalization.

3.4.1 Invariance

Invariance and its implications are thoughtfully explored by many authors as for example by Berger (1985), where it is also shown that invariance is strongly related to objective Bayesian analysis. In this work we use invariant-based arguments to guide our choice of an appropriate prior distribution. We briefly review some needed concepts of invariance and refer to Berger (1985) and Eaton (1989) for notation and further details.

We begin with the definition of an invariant family of densities:

Definition 3.2. The family of densities for $\boldsymbol{y} \in \mathbb{R}^n$, $\mathfrak{F} := \{f(\boldsymbol{y} \mid \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ is said to be *invariant under the group of transformations* $\mathfrak{G} := \{g : \mathbb{R}^n \to \mathbb{R}^n\}$ if for every $g \in \mathfrak{G}$ and $\boldsymbol{\theta} \in \Theta$, there exists a unique $\boldsymbol{\theta}^* \in \Theta$ such that $\boldsymbol{X} = g(\boldsymbol{Y})$ has density $f(\boldsymbol{x} \mid \boldsymbol{\theta}^*)$. In such a situation, $\boldsymbol{\theta}^*$ will be denoted $\bar{g}(\boldsymbol{\theta})$.

Example 3.1 (Location-Scale invariance)**.** Let Y be normally distributed with $f_Y(y \mid \mu, \sigma) = \mathcal{N}(y \mid \mu, \sigma^2)$ and consider the group of location scale transformations, $\mathfrak{G} = \{g_{c,b} : g_{c,b}(y) = cy + b;\ c > 0\}$. Then the density for $X = g_{b,c}(Y)$ is

$$f_X(x \mid \mu, \sigma) = c^{-1} f_Y(\frac{x-b}{c} \mid \mu, \sigma) = \mathcal{N}(x \mid \mu^*, \sigma^*)$$

with $\mu^* = c\mu + b$ and $\sigma^* = c\sigma$. That is, the normal density is invariant under location-scale transformations.

We are interested in the impact of invariance in the elicitation of improper non-informative priors for model selection. In this sense Berger et al. (1998) show that, even though Bayes factors can not usually be defined with improper objective priors (see Section 1.2.1), when the problem is invariant under a group of transformations, the use of the right Haar density (see Berger, 1985), although improper, results in well defined Bayes factors (in the sense studied in Berger et al., 1998).

We next show that the (conditional) marginal likelihood, $m_i^R(\boldsymbol{y} \mid \boldsymbol{\beta}_0, \sigma)$ is invariant, thus providing a powerful justification of the use of the right Haar prior on these problems.

Proposition 3.3. *The likelihood $m_i(\boldsymbol{y} \mid \boldsymbol{\beta}_0, \sigma)$ for $(\boldsymbol{\beta}_0, \sigma)$ under model M_i for $i = 1, \dots 2^p - 1$ derived by integrating out $\boldsymbol{\beta}_i$ with any prior of the form:*

$$\pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma) = \sigma^{-k_i} f_i(\frac{\boldsymbol{\beta}_i}{\sigma}), \tag{3.10}$$

for any known density on $\mathbb{R}^{k_i}$, f_i, is invariant under the group of transformations

$$\mathfrak{G} = \{g_{c,b} : g_{c,b}(\boldsymbol{y}) = c\,\boldsymbol{y} + \boldsymbol{X}_0\boldsymbol{b};\ \boldsymbol{b} \in \mathbb{R}^{k_0};\ c > 0\}. \tag{3.11}$$

Proof. See Appendix E.2. □

Corollary 3.3. *The marginal likelihood of* $(\boldsymbol{\beta}_0, \sigma)$ *for the Conventional Robust prior in* (3.6) *is invariant under the group of transformations in* (3.11).

Proof. Corollary 3.3 follows trivially since $\pi_i^R(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma)$ is of the form (3.10). □

Note that $m_0(\boldsymbol{y} \mid \boldsymbol{\beta}_0, \sigma) = \mathcal{N}_n(\boldsymbol{y} \mid \boldsymbol{X}_0\boldsymbol{\beta}_0, \sigma^2\boldsymbol{I}_n)$ is also invariant under the group in (3.11).

Berger et al. (1998) justify the use of the corresponding right Haar density when all the models are invariant with respect to the same group of transformations. They also show that the corresponding right Haar density under this group of transformation is $\pi(\boldsymbol{\beta}_0, \sigma) = 1/\sigma$. This provides a formal justification for the, very popular use, of the reference prior or independent Jeffreys' prior

$$\pi_i^R(\boldsymbol{\beta}_0, \sigma) = \pi(\boldsymbol{\beta}_0, \sigma) = \frac{1}{\sigma}$$

for variable selection and under all models $i = 0, \ldots, 2^p - 1$.

Another important remark has to be done here. Note that in order to have the above invariance structure, σ needs to be a scale parameter in $\pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma)$. This very extended practice is virtually always adopted in Conventional priors since the pioneering work of Jeffreys, but again to the best of our knowledge it has never been formally justified before.

3.5 Appealing properties arising from the proposed prior

The model specific (joint) prior that we propose for the parameters in M_i is

$$\pi_i^R(\boldsymbol{\beta}_i, \boldsymbol{\beta}_0, \sigma) = \frac{1}{\sigma}\, \pi_i^R(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma), \tag{3.12}$$

with $\pi_i^R(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma)$ given in Definition 3.1.

This is an improper joint Conventional Robust prior with many desirable properties for variable selection. Note that properties of $\pi_i^R(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma)$ were crucial in deriving $\pi_i(\boldsymbol{\beta}_0, \sigma)$. Some of the highlights of the formulation of our proposal are:

- It justifies the commonly adopted inclusion of σ in the scale matrix of $\pi_i^R(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma)$. Indeed the fact that σ is a scale parameter in $\pi_i^R(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma)$ was needed to achieve the invariance result in Corollary 3.3.
- The orthogonal parameterization is not longer a required prerequisite
  - neither for the specification of the scale of the conditional prior for $\boldsymbol{\beta}_i$,
  - nor for the specification of $\pi_i(\boldsymbol{\beta}_0, \sigma) = 1/\sigma$, which now is fully justified in terms of invariance.
- The use of $\boldsymbol{V}_i^t\boldsymbol{V}_i$ is justified by the choice of the scale matrix for $\pi_i^R(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma)$ in terms of $\mathrm{Cov}[\hat{\boldsymbol{\beta}}_i]$, which, exclusively for simplicity of notation is represented as $\sigma^2(\boldsymbol{V}_i^t\boldsymbol{V}_i)^{-1}$.

The joint prior (3.12) has additional attractive properties as we next show.

3.5.1 Closed-form expressions for prior predictive distributions

From a computational point of view, an interesting characteristic of our proposed prior distribution is that it results in closed-form expressions for the prior predictive distribution under each model and hence for the corresponding Bayes factors. Indeed the prior predictive distribution (marginal likelihood or evidence) can be expressed in terms of the hypergeometric function of two variables (also known as Appell hypergeometric function, see Appell, 1925):

Proposition 3.4. *For any* $(a,\, b,\, \rho_i) \in \mathcal{A}$, *(where* $\mathcal{A}$ *is the parametric espace in (3.7)) and* $n \geq k_i + k_0$, *the prior predictive distribution for* $\boldsymbol{y}$ *under* M_i *using the Conventional Robust prior is:*

$$m_i^R(\boldsymbol{y}) = m_0^R(\boldsymbol{y})\, Q_{i0}^{-\frac{n-k_0}{2}}\, \frac{2a}{k_i + 2a}\, (\rho_i\,(n+b))^{-\frac{k_i}{2}}\; AP_{i0}, \tag{3.13}$$

where

$$m_0^R(\boldsymbol{y}) = \frac{1}{2}\, \pi^{-\frac{n-k_0}{2}}\, |\boldsymbol{X}_0^t \boldsymbol{X}_0|^{-\frac{1}{2}}\, \Gamma\left[\frac{n-k_0}{2}\right]\; SSE_0^{-\frac{n-k_0}{2}}, \tag{3.14}$$

and AP_{i0} *is a hypergeometric function of two variables or Appell hypergeometric function:*

$$AP_{i0} = F_1\Big[a + \frac{k_i}{2};\, \frac{k_0 + k_i - n}{2},\; \frac{n-k_0}{2};$$
$$a + 1 + \frac{k_i}{2};\, \frac{(b-1)}{\rho_i\,(b+n)};\, \frac{b - Q_{i0}^{-1}}{\rho_i\,(b+n)}\Big].$$

Recall that Q_{i0} *is the ratio of residual sum of squares under each model* SSE_i/SSE_0 *and was defined in Section 2.3.2.*

Note that for computing $m_i(\boldsymbol{y})$ *a sample of size* $n \geq k_i + k_0$ *is needed. Indeed, if we want to compute* $m_i(\boldsymbol{y})$ *for every model* M_i *with* $i = 0, \ldots, 2^p - 1$ *we need a sample of size* $n \geq p + k_0$.

Proof. See Appendix E.3. □

The hypergeometric function is considered a closed-form expression in the sense that it is originally defined as an infinite sum (the interested reader can find a careful definition in Appendix C, for further information see Appell, 1925). The expression of $m_i^R(\boldsymbol{y})$ is substantially simplified with some specific values of the parameters as we will see in Chapter 4. Practical benefits of this simpler formulation in variable selection will be studied in detail in Chapter 6 with some real and simulated examples.

3.5.2 Predictive Matching

The motivating argument for what is, generally, known as *predictive matching* can informally be stated as follows:

> *It seems intuitively reasonable that when the information in the sample is barely enough for estimating the model specific parameters there is not enough information for distinguishing among models. In such a situation model comparison can not really be conclusive.*

Predictive matching has been studied in literature from many different points of view. Berger and Pericchi (2001) describe predictive matching as follows: "in comparing two models M_i and M_j, π_i and π_j should be chosen so that for a sample of minimal sample size $m_i(\boldsymbol{y})$ and $m_j(\boldsymbol{y})$ are as close as possible". Spiegelhalter and Smith (1982) and Ghosh and Samanta (2002) consider that predictive matching holds when the Bayes factor B_{i0} is exactly 1 for any sample of "minimal sample size" obtained under M_0. This is a particular case of Berger and Pericchi (2001)'s predictive matching. Similar ideas are used in, for example, Kadane et al. (1980); Suzuki (1983); Ibrahim and Laud (1994) and Laud and Ibrahim (1995).

A common concept in all approaches to predictive matching is the concept of "minimal sample size". In Berger and Pericchi (2001) the minimal sample size is defined as the size of a *proper minimal training sample* defined as follows:

Definition 3.3. A training sample $\boldsymbol{y}^*$ is called proper if, given a non-informative (usually improper) prior for the parameters $\pi_i^N(\boldsymbol{\theta}_i)$, the corresponding marginal likelihood $m^N(\boldsymbol{y}^*)$ is $0 < m_i^N(\boldsymbol{y}^*) < \infty$ for all M_i; and minimal if it is proper and not subset is proper.

The size of a minimal proper training sample can thus depend on the prior π^N used. In particular it will be different for the usual Jeffreys independent or reference prior $\pi(\boldsymbol{\beta}_i, \boldsymbol{\beta}_0, \sigma) = 1/\sigma$ and our Conventional Robust prior.

Specifically, in variable selection, Berger and Pericchi show that, using the usual objective joint prior $\pi_i(\boldsymbol{\beta}_i, \boldsymbol{\beta}_0, \sigma) = 1/\sigma$, for $i = 0, \ldots, 2^p - 1$, the size of a proper minimal training sample is $n^* = p + k_0 + 1$, but as we show below this is not the minimal sample size when using our proposed joint prior (recall that our joint proposal, despite being improper, is defined in two steps with the first part $\pi_i^R(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma)$ being proper for any sample of size $n \geq k_i + k_0$). It is anyway important to remark that Berger and Pericchi's minimal sample size $n^* = p + k_0 + 1$ is still the sample size needed from a purely frequentist point of view to estimate all the parameters in all the models.

We now revise the concept of minimal sample size as well as predictive matching to the problem of variable selection.

Predictive matching for variable selection

To define our idea of predictive matching we first need to determine which is our minimal sample size.

Result 3.1. *Given any model M_i the corresponding minimal sample size for the Conventional Robust prior is $n_i^* = k_i + k_0$.*

Proof. Let $\boldsymbol{y}^*$ be a sample of size $n^* = k_i + k_0$ then $m_i^R(\boldsymbol{y}^* \mid \boldsymbol{\beta}_0, \sigma)$ is proper since $\pi_i^R(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma)$ it is so. Integrating over $(\boldsymbol{\beta}_0, \sigma)$ using $\pi(\boldsymbol{\beta}_0, \sigma) = 1/\sigma$ we need a sample of size $n \geq k_0 + 1$ for the result to be bounded. As we already have a sample of size $n \geq k_0 + 1$ ($n^* = k_i + k_0 \geq k_0 + 1$)

$$m_i^R(\boldsymbol{y}^*) = \int m_i^R(\boldsymbol{y}^* \mid \boldsymbol{\beta}_0, \sigma) \, \frac{1}{\sigma} \, d(\boldsymbol{\beta}_0, \sigma)$$

is finite. In fact the exact value of $m_i^R(\boldsymbol{y}^*)$ for a sample of size $n^* = k_i + k_0$ is given in Proposition 3.5. □

An important remark has to be done here. Note that in the definition of a minimal training sample of Berger and Pericchi (2001) an unique minimal sample size is defined for every model M_i. For our prior this would be $n^* = p + k_0$. For reasons that will become clear in the sequel we consider instead minimal sample sizes for each specific model M_i, so that $n_i^* = k_i + k_0$. We remark again that for estimating the parameters of model M_i from a purely frequentist point of view (or from a Bayes perspective with the usual improper estimation prior) the sample size needed is $n = k_i + k_0 + 1$, and not $n = k_i + k_0$.

Once that the minimal sample size in our scenario has been defined, we now introduce our idea of predictive matching which is closely related to the proposal in Berger and Pericchi (2001) mentioned above, but it incorporates some interesting twists. As a matter of fact, we are much less ambitious in our demands for predictive matching. Indeed whereas Berger and Pericchi would predictively match $m_i(\boldsymbol{y})$ and $m_j(\boldsymbol{y})$ no matter the complexity of M_i and M_j, we consider that such matching makes the most sense when it is entertained among models of the same complexity $k_i = k_j$, and we will only aim for that.

Definition 3.4 (Weak Predictive Matching.)**.** In a model selection problem entertaining m models $M_j, j = 1, \ldots, m$, we say that model specific priors $\pi_j, j = 1, \ldots, m$ result in weak predictive matching, if the evidence (prior predictive) $m_i(\boldsymbol{y}_i^*)$ is the same for all models M_i of the same complexity and for all data $\boldsymbol{y}_i^*$ of minimal sample size for that complexity (that is, the minimal sample size for which the posteriors for those models are proper). We assume that equally complex models have the same minimal sample size (an intuitively sound requirement for objective Bayes model selection).

In the context of variable selection for linear models, the complexity of model M_i is clearly characterized by the dimension k_i of the extra parameter. As previously established, with the Conventional Robust prior the minimal sample size for a model with k_i extra parameters is $k_i + k_0$, so that indeed all models of the same complexity have the same minimal sample size. This sample size is also increasing with model complexity, as intuitively expected.

The motivation behind weak predictive matching is to have a less restrictive criteria than the full predictive matching of Berger and Pericchi (2001) but still allowing assessment of wether the priors are well balanced across models (an important criteria in objective Bayes model selection). Weak predictive matching only indicates that the priors for models of the same complexity are well "calibrated" among themselves. It does not say anything one way or the other for comparisons of models of differing complexity. This might be "too weak" and some other requirement might have to be investigated in order to have all priors "well balanced" across all models.

The next Proposition and its Corollary show that the Conventional Robust prior results in weak predictive matching. In addition they also show that the base model M_0 is matched to every model M_i for the minimal sample size for that model.

Proposition 3.5. *Given a model M_i with k_i extra covariates, for any sample $\boldsymbol{y}^*$ of size $n^* = k_i + k_0$ we have*

$$m_i^R(\boldsymbol{y}^*) = m_0^R(\boldsymbol{y}^*) = \frac{1}{2}\,\pi^{-\frac{k_i}{2}}\,|\boldsymbol{X}_0^t\boldsymbol{X}_0|^{-\frac{1}{2}}\,\Gamma\left[\frac{k_i}{2}\right]\,SSE_0^{-\frac{k_i}{2}},$$

which only depends on SSE_0, $\boldsymbol{X}_0$ and k_i.

Proof. See Appendix E.4 □

Corollary 3.4 (Weak Predictive Matching)**.** *For any sample of size $n_i^* = k_0 + k_i$ there is predictive matching for all models M_i with k_i extra covariates in the following sense: for* any *sample $\boldsymbol{y}^*$ of size n_i^**

1. *The prior predictive evaluated at $\boldsymbol{y}^*$, $m_i^R(\boldsymbol{y}^*)$ is the same for all models with k_i extra covariates; hence, all models of dimension k_i are predictively matched for all such $\boldsymbol{y}^*$.*

2. *Each Bayes factor $B_{i0}^R = m_i^R(\boldsymbol{y}^*)/m_0^R(\boldsymbol{y}^*) = 1$, and hence all models M_i of dimension k_i and M_0 are predictively matched for all such $\boldsymbol{y}^*$.*

Proof. It's trivial from Proposition 3.5. □

The first part of the Corollary gives weak predictive matching, thus showing well calibrated priors within all models of the same dimension. The second part provides the additional property that all such models are also well "balanced" with the base model M_0 in the sense that, for their minimal training sample, there is predictive matching. Notice that each M_i matches M_0 for possibly different sample sizes n_i^*. Since all Bayes factors B_{ij} are defined through comparisons of each model with the base model M_0, Corollary 3.4 is in fact showing that the Conventional Robust prior is calibrated across all models; here this 'calibration' is in a much weaker sense than usually required in the literature.

A major advance is the recognition that "minimal training sample size" needs to reflect the necessary proper priors for "uncommon" parameters, and be based only on models of the same complexity; Corollary 3.4 is a very strong statement in this context.

Chapter 4

Conventional Robust Bayes Factors: Closed-form expression and consistency issues

4.1 Introduction

The model posterior probabilities can be expressed in terms of the 2^p Bayes factors B_{i0} in favor of model M_i and against model M_0 as:

$$P(M_i \mid \boldsymbol{y}) = \frac{B_{i0}\, P(M_i)}{\sum_{l=0}^{2^p-1} B_{l0}\, P(M_l)}.$$

for $i = 0, \ldots, 2^p - 1$.

In this chapter, we focus on the Bayes factors resulting from the Conventional Robust prior defined in Chapter 3, referred to as *Conventional Robust Bayes factors* and denoted by B_{i0}^R.

4.2 Definition, closed-form and posterior probabilities

As it was shown in Section 1.2.1 Bayes factors can also be expressed as the ratio of marginal likelihoods:

$$B_{i0} = \frac{m_i(\boldsymbol{y})}{m_0(\boldsymbol{y})}$$

where $m_i(\boldsymbol{y})$ for $i = 0, \ldots, 2^p - 1$ is

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

When using our prior distribution $\pi_i^R(\boldsymbol{\beta}_i, \boldsymbol{\beta}_0, \sigma)$ in (3.12) this marginal likelihood has a closed-form expression, as shown in Proposition 3.4. Hence, the resulting Conventional Robust Bayes factor also has a closed-form expression.

Proposition 4.1. *The Conventional Robust Bayes factor in favor of model M_i and against model M_0 is*

$$B_{i0}^R = Q_{i0}^{-\frac{n-k_0}{2}} \frac{2a}{k_i + 2a} \left(\rho_i \left(n + b\right)\right)^{-\frac{k_i}{2}} AP_{i0}, \tag{4.1}$$

where AP_{i0} is the hypergeometric function of two variables defined in Proposition 3.4, and $Q_{i0} = SSE_i/SSE_0$ is the ratio of residual sum of squares under each model (see Section 2.3.2).

Proof. This expression follows directly from Proposition 3.4. □

Expression (4.1) is further simplified for $b = 1$ as shown in Corollary 4.1.

Corollary 4.1. *For $b = 1$,*

$$B_{i0}^R = Q_{i0}^{-\frac{n-k_0}{2}} \frac{2a}{k_i + 2a} \left(\rho_i \left(n + 1\right)\right)^{-\frac{k_i}{2}} HG_{i0}, \tag{4.2}$$

where HG_{i0} is the hypergeometric function of one variable:

$$HG_{i0} = {}_2F_1\Big[a + \frac{k_i}{2}; \frac{n-k_0}{2}; a + 1 + \frac{k_i}{2}; \frac{1 - Q_{i0}^{-1}}{\rho_i\,(1+n)}\Big].$$

Proof. It follows directly from Proposition 4.1 □

Having closed-form expressions is a very appealing characteristic specifically for problems with large model spaces, as is usually the case for variable selection. Consider, for example, a problem where the number of covariates is so large that the model space can not be enumerated (for all practical purposes). In such a problem we make a search over the model space trying to find the most probable models. For each visited model, B_{i0} should be computed. It is obvious that the simpler the expression of B_{i0} the faster the computation, and hence, for a given time of computation, the larger the number of visited models. Also, if the model space can be enumerated, the time needed to compute the 2^p Bayes factors can be substantially reduced if B_{i0} has a simple expression.

The hypergeometric function of two variables and hypergeometric function of one variable are both considered closed-form expressions and can be computed with several statistical and mathematical software. But the latter is implemented in a wider range of computer software. Moreover, its computation is noticeably easier and faster, and so, it is considered a simpler expression (in terms of computation) and thus preferred for variable selection.

For the Bayes factors B^R_{i0} in (4.2) the posterior probabilities are:

$$P(M_i \mid \boldsymbol{y}) = \frac{Q_{i0}^{-\frac{n-k_0}{2}} \frac{2a}{k_i+2a} \left(\rho_i \left(n+b\right)\right)^{-\frac{k_i}{2}} \mathrm{HG}_{i0}\, P(M_i)}{\sum_{j=0}^{2^p-1} Q_{j0}^{-\frac{n-k_0}{2}} \frac{2a}{k_j+2a} \left(\rho_j \left(n+b\right)\right)^{-\frac{k_j}{2}} \mathrm{HG}_{j0}\, P(M_j)} =$$

$$= \left[\sum_{j=0}^{2^p-1} (n+b)^{\frac{k_i-k_j}{2}} \left(\frac{Q_{i0}}{Q_{j0}}\right)^{\frac{n-k_0}{2}} \frac{k_i+2a}{k_j+2a} \frac{\rho_i^{k_i/2}}{\rho_j^{k_j/2}} \frac{\mathrm{HG}_{j0}}{\mathrm{HG}_{i0}} \frac{P(M_j)}{P(M_i)}\right]^{-1}$$

for $i = 0, \ldots 2^p - 1$.

A curiosity is that, for models M_i differing with M_0 in one covariate (i.e. $k_i = 1$), for certain values of a and b the Conventional Robust Bayes factor have a extremely simple expression:

Corollary 4.2. *If $a = 1/2$, $b = 1$ and $k_i = 1$ (i.e. M_i has one more covariate than M_0), then:*

$$B^R_{i0} = \frac{\sqrt{\rho_i(1+n)}\,(Q_{i0})^{-\frac{n-k_0}{2}} \left[1 - \left(1 - \frac{1-Q_{i0}^{-1}}{\rho_i(1+n)}\right)^{-\frac{n-k_0-2}{2}}\right]}{(n-k_0-2)(Q_{i0}^{-1}-1)}, \qquad (4.3)$$

for $n > k_0 + 2$ and

$$B^R_{i0} = \frac{1}{2}\sqrt{\rho_i(k_0+3)}\,(1-Q_{i0})^{-1} \log\left[1 + \frac{Q_{i0}^{-1}-1}{\rho_i(k_0+3)}\right],$$

for $n = k_0 + 2$.

Proof. It follows easily from Proposition 4.1 □

In a framework like ours where the base model is fixed (recall that we always consider M_0 as base model, see Section 2.3.1) the expression in (4.3) is not really useful because just few models in the problem would typically differ from the base model in just one covariate. Nevertheless, this simple closed-form expression can be convenient when considering

some numerical methodologies as some stepwise methods in which Bayes factors are always computed between models differing in one covariate (see for example Berger and Molina, 2005).

4.3 Consistency of Conventional Robust Bayes factors

Studying the consistency of any statistical procedure is commonly associated with studying its behavior when the sample size tends to infinity. In model selection, this is known in the literature as *model selection consistency*. In this section we study model selection consistency for the Conventional Robust Bayes factors.

We also study other types of consistency, in particular we study the behavior of Conventional Robust Bayes factors for a fixed n when:

1. the data overwhelmingly support M_i. This type of consistency is referred to as *information consistency* (see Section 4.3.2).
2. the "information" in the data is very large in favor of M_0; we refer to this as *null information consistency* (see Section 4.3.3).

4.3.1 Model selection consistency

Informally stated, model selection consistency requires that: (quoting O'Hagan, 1994)

> *"[...] as the number of observations tends to infinity, the probability of selecting the correct model tends to one" .*

More precisely, if the true model is M_i, a model selection procedure (and in particular Bayes factors) is consistent if the posterior probability

$P(M_i \mid \boldsymbol{y})$ converges in probability to 1 as the sample size, n, grows (or equivalently if $\text{plim}_n\, B_{ji} = 0$ for all $M_j \neq M_i$).

Model selection consistency is an important property that has been amply studied in literature. A number of recent references include: Fernández et al. (2001); Berger et al. (2003); Liang et al. (2008); Casella et al. (2009); Guo and Speckman (2009); Moreno et al. (2009).

We next show that under very weak conditions, Conventional Robust Bayes factors are model selection consistent. This result is based on a previous one of Liang et al. (2008).

Proposition 4.2. *If* $\lim_{n\to\infty} \rho_i\ (b+n) = \infty$*, then the Conventional Robust Bayes factors are consistent.*

Proof. See Appendix F.4 □

This is a fundamental property of our methodology and the only requirement is that $\rho_i(b+n)$ tends to ∞ as n grows. When later on, in Chapter 5, we study the possible choices for ρ_i it will be important to keep this in mind and avoid values of ρ_i that make $\rho_i(b+n)$ go to a constant when $n \to \infty$ (as for example any expression of order n^{-1}).

4.3.2 Information consistency

When the observed data provides a lot of "information" in favor of model M_i, it is reasonable to expect that the corresponding Bayes factor B_{i0} would reflect this "information" by having a large value. Moreover, as the information favouring M_i grows without limits we would expect B_{i0} to tend to infinity. This desirable property is usually referred to as information consistency (see Bayarri and García-Donato, 2008). When a Bayes factor is not information consistent it is said to suffer from the *Information Paradox* (see Liang et al., 2008).

The information paradox phenomenon was first noted by Jeffreys (1961) and caused him to reject certain type of priors suffering from it. Conjugate prior, as the g-priors of Zellner (1986) suffer from information paradox (see Berger and Pericchi, 2001; Liang et al., 2008).

Of course, it is important to specify what "information in favor of model M_i" means in statistical terms. In particular, in terms of the statistic Q_{i0} because it is the only summary of the data used in the computation of B_{i0}. The statistic Q_{i0} was defined in Section 2.3.2 as the ratio of the residual sum of squares for M_i and M_0, $Q_{i0} = SSE_i/SSE_0$. There we also remarked that a value of $Q_{i0} \to 0$ indicates an increasing support for M_i.

It is thus easy to state information consistency in terms of Q_{i0}:

> The Bayes factor B_{i0}^R is said to be information consistent if it tends to infinity when Q_{i0} tends to 0. That is if
>
> $$\lim_{Q_{i0}\to 0} B_{i0}^R = \infty$$

The conditions under which Conventional Robust Bayes factors are information consistent are given in the following result.

Proposition 4.3. *The Bayes factor, B_{i0}^R is information consistent if and only if $n \geq k_i + k_0 + 2a$.*

Proof. See Appendix F.2 □

Similar results are derived by Liang et al. (2008) for Bayes factors arising from a "scale mixture of normals" prior (as in (3.8)) which, as shown in Proposition 3.3.1, includes our Conventional Robust prior distribution.

In Proposition 4.3 we see that the parameter a controls the sample size needed to achieve information consistency. In fact it follows easily that

in a problem of variable selection with p covariates, we need a sample of size $n \geq p + k_0 + 2a$ for the $2^p - 1$ Bayes factors B^R_{i0} to be information consistent. If, in particular, $a \in (0, 1/2]$ it suffices to have $n \geq p + k_0 + 1$.

Interestingly, recall that the parameter a also controls the thickness of $\pi^R_i(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma)$'s tails, which were shown to be like those of a Student's t-distribution with $2a$ degrees of freedom. This is indeed not a coincidence. Jeffreys realized that the tails of the conditional prior distribution for the "new" parameter is closely related to information consistency and that thick tails are required to obtain this desirable property. In fact this idea is reflected in his fourth desideratum "the conditional prior should have no finite moments" (see Section 2.3.3) and was his main reason for choosing a Cauchy distribution instead of the normal distribution (in spite of the latter resulting in simpler expressions).

4.3.3 Null information consistency

Along the same lines as in information consistency, we next show that B^R_{i0} is always smaller than 1 when the data strongly supports the simplest model, M_0. Recall that in Section 2.3.2 we interpreted this limiting support for M_0 with $Q_{i0} \to 1$.

Proposition 4.4. *For any n, B^R_{i0} is bounded above by a constant for $Q_{i0} \to 1$. This constant is smaller than 1, and depends only on k_i and a. Specifically,*

$$\lim_{Q_{i0} \to 1} B^R_{i0} \;\leq\; \left[1 + \frac{k_i}{2a}\right]^{-1} < 1.$$

Proof. See Appendix F.3 □

Since the upper bound is smaller than 1, B^R_{i0} always supports M_0 as $Q_{i0} \to 1$. Note that this bound is increasing with a. In particular if $a \in (0, 1/2]$ (the range of values that produces conditional priors with no moments, see Section 3.3.2) this bound is smaller than $1/2$ giving

even stronger support for M_0 (recall that the smaller B_{i0}^R the larger the support for M_0) and hence, being more consistent with the information provided by the data. It should also be noted that the larger the value of a the less parsimonious Bayes factors, in the sense that the larger a the more complex models are favored. Again the relationship between consistency and tail's shape arises since a controls the bound of the B_{i0}^R in this case, and hence the strength of the support for M_0 in the resulting Bayes factor.

This argument jointly with the specific expression of $\lim_{Q_{i0}\to 1} B_{i0}^R$ (given in Appendix F.3) is used in the next chapter to partly guide our recommended choices for a and ρ_i in the Conventional Robust priors.

Finally note that this bound is decreasing with k_i, thus showing clearly the penalty for complexity or Occam's razor effect.

Chapter 5

The hyper-parameters $(a,\, b,\, \rho_i)$

5.1 Introduction

So far we have proposed a prior distribution for addressing variable selection problems and studied its convenient properties in this framework. But recall that our proposal $\pi_i^R(\boldsymbol{\beta}_i,\, \boldsymbol{\beta}_0,\, \sigma)$ in (3.12) depends on three parameters $(a,\, b,\, \rho_i)$. It is quite remarkable that all the studied properties are achieved for all the values of $(a,\, b,\, \rho_i)$. To summarize, for any $(a,\, b,\, \rho_i)$ with: $a \in (0, \infty)$, $b \in [1, n]$, and $\rho_i \geq b/(b+n)$, the prior $\pi_i^R(\boldsymbol{\beta}_i,\, \boldsymbol{\beta}_0,\, \sigma)$ in (3.12):

- Follows most of Jeffreys' desiderata for model selection (see Section 2.3.3), and all of them if $a \in (0, 1/2]$.
- Has attractive and novel predictive matching properties (see Section 3.5.2).
- Results in well defined (see Section 3.4.1) closed-form (see Section 4.2) and consistent (see Section 4.3) Bayes factors.

In this chapter, we propose "optimal" objective choice of specific values for $(a,\, b,\, \rho_i)$. Indeed we show that our recommended choices result (in some sense to be defined later) in an improvement of the properties mentioned above as well as in achieving some new desirable ones.

5.2 The Parameter a and the behavior on the tails

As we have seen, the parameter a is closely related to tails behavior. In particular, it controls the degrees of freedom of the Student's distribution which define the tails of $\pi_i^R(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0,\, \sigma)$ in (3.6) (see Section 3.3.2) and hence the number of moments of this conditional prior. Recall that one of Jeffreys' requirements for a conditional prior for the new parameters was that it should have no finite moments. This gives us our first guideline for choosing suitable values for a:

1. The parameter a should be $a \leq 1/2$ so that $\pi_i^R(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0,\, \sigma)$ has no finite moments.

Moreover, information consistency studied in Section 4.3.2 is also related to this behavior in the tails and hence it is relevant in guiding the choice of a. Indeed in Proposition 4.3 it is shown that a controls the sample size needed for the $2^p - 1$ paired Bayes factors B_{i0} for $i = 1, \ldots, 2^p - 1$ to be information consistent. Specifically, $n \geq p + k_0 + 2a$. This gives us another reason to follow requirement 1 above and choose $a \leq 1/2$, because we then have information consistency even for samples of minimal size in the frequentist sense, that is $n = k_i + k_0 + 1$.

Also related with the value of a is null information consistency. In proposition 4.4 it is shown that for data the most compatible with M_0 the

corresponding Bayes factor (which we denote B^0_{i0}) is bounded above by

$$B^0_{i0} \leq \left[1 + \frac{k_i}{2a}\right]^{-1}.$$

This quantity is, for $a \in (0, \infty)$, bounded by 1 which as described in Section 4.3.3 is very reasonable since data is compatible with M_0. Moreover, when $a \leq 1/2$ as in the previous requirement, this bound becomes $B^0_{i0} < 1/2$ thus reflecting even clearer the support in the data for M_0. However, it is important to remark that a very small value of B^0_{i0} will result in very conservative procedures (which might favour simple models too much). This consideration along with the fact that the bound above is increasing with a gives us our second guideline for choosing a:

2. Smaller values of a result in smaller upper bounds for B^0_{i0} so too small values for a might result in excessively conservative Bayes factors.

Point 1 induces a choice of $a \leq 1/2$ while Point 2 indicates that we need to find a balance between being null information consistent (choosing a not too big) and not being too conservative (choosing a not too small). A compromise between these ideas is to choose $a = 1/2$ which is our final proposal for this parameter. This choice matches the proposal in the original paper by Berger (1980). Moreover, with this choice, π^R_i has Cauchy tails like the popular proposals of Jeffreys (1961), and Zellner and Siow (1980, 1984).

In Summary, taking $a = 1/2$ makes our proposal to achieve Jeffreys' desiderata, be information consistent (for $n \geq p + k_0 + 1$) and being null information consistent (with $B^0_{i0} < 1/2$) without being too conservative.

5.3 A computational convenient choice for b and a sensitivity study

One of the original motivations for this thesis was to develop a methodology for variable selection with good properties and simple expressions. As shown in Corollary 4.1 the simplest expressions are obtained with $b = 1$ which is our recommended choice for the value of parameter b. For this value of b the resulting Bayes factor has a closed-form expression in terms of the hypergeometric function of one variable.

Recall that for any other value of b the expression of the Bayes factors is in terms of the hypergeometric function of two variables. As commented in Section 4.2, there it is a noticeable difference between the hypergeometric function of one variable and the hypergeometric function of two variables in terms of computational availability and complexity. This makes the hypergeometric function of one variable way more suitable for variable selection, and hence, supports the choice of $b = 1$. Moreover, for this choice of b, Berger (1985) gives an alternative, very simple way of computing predictive distributions (and hence Bayes factors) and nothing similar seems to be available when $b \neq 1$.

However, this choice of b is not based on any strong theoretical argument and so, an obvious question arises: How large is the impact of b in the results? As an attempt to investigate this question, we present an empirical analysis to gain understanding on the sensitivity of Bayes factors with respect to b. In particular we study the ratio

$$R(b) = \frac{B^R_{i0}(b)}{B^R_{i0}(b=1)} \tag{5.1}$$

as b changes (recall that b can take values in $[1, n]$ as shown in (3.7)).

For this study we consider $n = 100$, $k_i = 3$ and $k_0 = 1$ taking $a = 1/2$ (our choice) and two different values for ρ_i:

- $\rho_i = 1/2$;
- $\rho_i = (k_i + 1)/(k_i + 3) = 2/3$ (the one proposed by Berger, 1985).

Given these parameters we consider three different scenarios for the values of Q_{i0}, considering little, medium, and strong support for M_i. In particular,

1. **Scenario 1:** The first scenario considers values of Q_{i0} giving little support for M_i. In particular, we take values of Q_{i0} close to 1 ($Q_{i0} \in [0.9, 1]$). This interval has, for the considered values of n, k_i and k_0, a 90% of the probability in the distribution of Q_{i0} under M_0. Indeed, in Section 2.3.2 we show that $Q_{i0} \mid M_0 \sim \mathcal{B}e((n - k_i - k_0)/2, k_i/2)$, which here is $Q_{i0} \mid M_0 \sim \mathcal{B}e(48, 1.5)$ for which $P(Q_{i0} \in [0.9, 1]) = 0.9$. Specifically, we use the values $Q_{i0} = 1, 0.95, 0.92, 0.9$.

2. **Scenario 2:** Our second scenario considers values of Q_{i0} which give medium support for M_i. Those are values of Q_{i0} not too close to 1 (where there is strong support for M_0) and not too close to 0 (which indicates strong support for M_i). In particular we use the values $Q_{i0} = 0.8, 0.6, 0.4, 0.2$.

3. **Scenario 3:** Our third and last scenario considers values of Q_{i0} strongly supporting M_i. As shown in Section 2.3.2 values of Q_{i0} giving strong support for M_i are values of Q_{i0} close to 0. Specifically in this scenario we take $Q_{i0} = 0.05, 0.01, 0.005, 0.001$.

A remark is in order here for the reader who might wonder why the sensitivity to b is not being studied for our recommended choice of ρ_i (to be discussed later). The reason is that our ultimate choice $\rho_i = 1/(k_i + k_0 + 1)$ will specifically be derived for $b = 1$. Moreover it can not be used with every value of b due to the restrictions of the parametric space in (3.7) ($\rho_i > b/(b + n)$), so it is not well suited for this exercise.

Results are presented in Figures 5.1, 5.2 and 5.3, which are commented below.

Scenario 1

Figure 5.1 shows that the impact of b depends considerably on the value of ρ_i. Indeed, for the choice of $\rho_i = 1/2$ (top) the sensitivity of $R(b)$ in (5.1) to b is larger than for the choice of $\rho_i = (k_i + 1)/(k_i + 3)$ (bottom). Specifically, for $\rho_i = 1/2$, $R(b)$ goes from 1 to 11.87 as b goes from 0 to 100, while the ratio for $\rho_i = (k_i + 1)/(k_i + 3)$ stays between 1 and 1.25.

Note that for all the values of Q_{i0} in this scenario, the ratio $R(b)$ grows with b taking its maximum value at $b = n$. Indeed, for $\rho_i = 1/2$, $B^R_{i0}(b = n)$ is more than 11 times larger than $B^R_{i0}(b = 1)$ when $Q_{i0} = 1$. In a sense, this was expected since B^R_{i0} for $b = n$, $a = 1/2$ and $\rho_i = 1/2$ corresponds to Liang et al. prior which, as exposed in Section 3.3.1, spikes sharply around $\boldsymbol{\beta}_i = 0$ (i.e. the null model) resulting in Bayes factors which, for data compatible with M_0, are larger (closer to 1) than any other Conventional Robust prior with the same a and ρ_i but smaller values of b.

Notice that, for $\rho_i = 1/2$, the various $R(b)$ increase slowly from $b = 1$ up to $b \approx 50$ when they start growing dramatically. This indicates that the sensitivity of this ratio to the value of b increases as b gets close to n and that small and moderate values of b seem to be quite more stable than the larger ones. This also supports, somewhat, the choice $b = 1$, for the parameter b.

In spite of the sensitivity of the ratio of Bayes factors to b in this scenario, it is worth noting that the values of all of these Bayes factors are always less than 1, as they should be for values of Q_{i0} the most compatible with M_0.

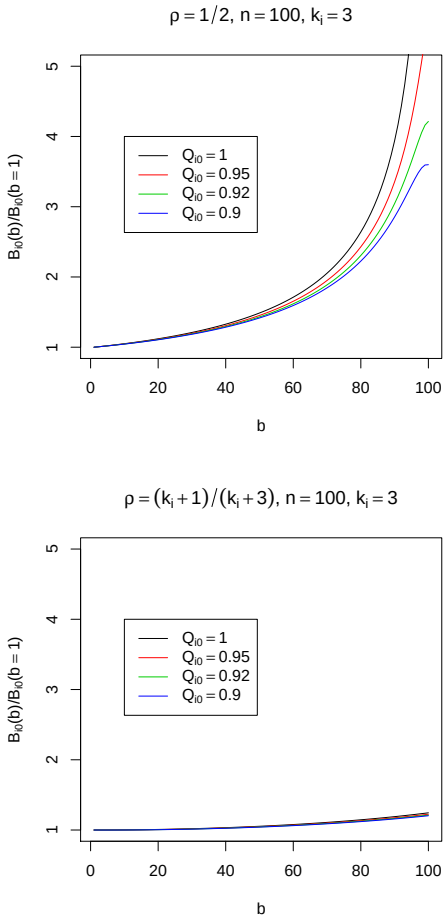


FIGURE 5.1: *Scenario 1: Ratio $R(b)$ of Conventional Robust Bayes factors as a function of b*

Scenario 2

Figure 5.2 shows that, similarly to the previous scenario, the sensitivity is larger for $\rho_i = 1/2$ (top) than for $\rho_i = (k_i + 1)/(k_i + 3)$ (bottom). However, in this scenario the sensitivity is considerably smaller than in scenario 1. In scenario 2 $R(b) \in (0.6, 2.15)$, so Conventional Robust Bayes factors are remarkably robust to the choice of b.

In any case, these values of the ratio do no really make a difference in practice since the resulting Bayes factors are all between 10^2 to 10^{30} and having $B^R_{i0}(b = 1)$ or $B^R_{i0}(b = n) \approx 2\, B^R_{i0}(b = 1)$ will result in the same very clear conclusion: data supports M_i.

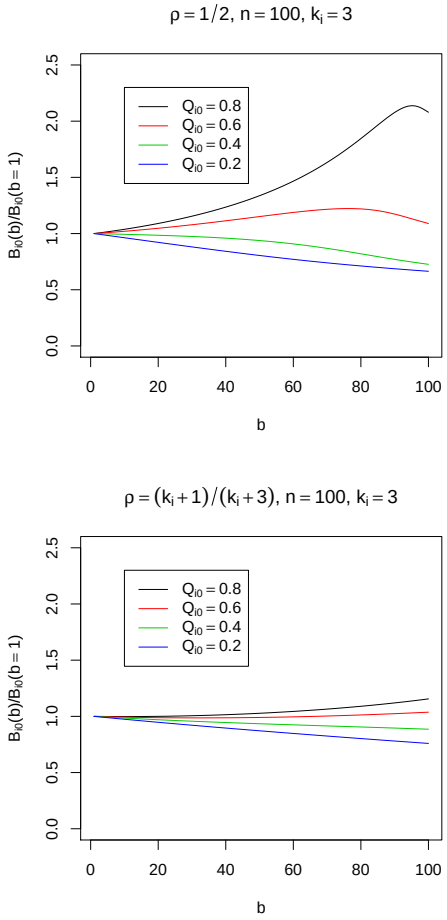


FIGURE 5.2: *Scenario 2: Ratio $R(b)$ of Conventional Robust Bayes factors as a function of b*

Scenario 3

Figure 5.3 show that the maximum absolute value of the ratio $R(b)$, that is $R(b = n)$, tends to 1.4 as $Q_{i0} \to 0$. Moreover, a ratio of 1.4 is of not practical significance at all, since in this scenario the order of Bayes factors is about 10^{40}, hence giving overwhelmingly support for M_i whatever the value of b, as expected.

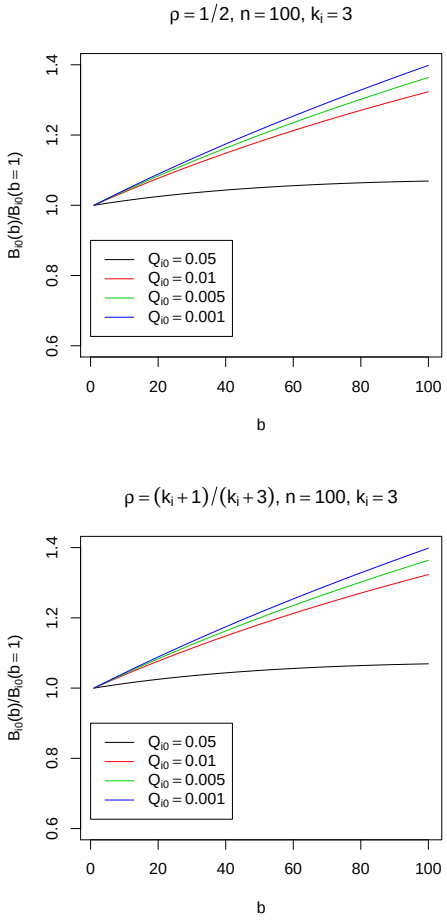


FIGURE 5.3: *Scenario 3: Ratio $R(b)$ of Conventional Robust Bayes factors as a function of b*

In summary,

1. Bayes factors are very sensitive to b only when i) the data is the most compatible with M_0; and ii) b is close to n.

2. Whatever the value of b, the decision taken based on the values of B_{i0} seems to be quite robust and intuitively correct.

Finally 1. and 2. together with the simple expression of the Bayes factor achieved suggest that our choice of $b = 1$ is a very sensible one.

5.4 The choice of ρ_i: revisiting the predictive matching criteria

In this section we explore the role of ρ_i and suggest a good choice for the value of ρ_i given our preferred choice for the other parameters: $a = 1/2$ and $b = 1$.

The assessment of ρ_i is quite a delicate issue specially because of its potentially large impact in the tail behavior of the prior distribution. This is because Conventional Robust priors behave in the tails as Student's t-distributions (see Section 3.3.2) with scale matrix:

$$\boldsymbol{\Sigma}_i^* = \rho_i \, a^{-1}[a\Gamma(a)]^{1/a} \, (b+n) \, \sigma^2 (\boldsymbol{V}_i^t \boldsymbol{V}_i)^{-1}. \tag{5.2}$$

Clearly ρ_i enters the scale matrix as a multiplicative constant and multiplicative terms in the scale of priors for model selection have the potential for a very large impact on Bayes factors. Since for $b = 1$, ρ_i can take values in $[1/(1+n), \infty)$, a choice of a specific value in this huge range does have a direct, and important impact on the scale matrix, and hence on the tails of $\pi_i^R(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma)$.

Our initial choice was to consider $\rho_i = (k_i + 1)/(k_i + 3)$ since this is the "optimal" value suggested in Berger (1985). However, this choice was developed for an estimation scenario so it may not be the optimal choice

for model selection. We therefore try here to use optimality criteria in model selection to guide our recommendation for ρ_i. Moreover, we study the implications of our choice later in this chapter as well as in the examples of Chapter 6. We also compare them with the Bayes factors from other approaches.

Recall that our prior distribution achieves Jeffreys' desiderata for model selection priors (see Section 2.3.3) as well as many other desirable properties reviewed at the beginning of this chapter, and it does so no matter the value of ρ_i. At the same time, we saw above that this parameter has the potential for a large impact in the results, so we need some criteria to choose a specific value for it. We base our choice for ρ_i on predictive matching ideas.

Weak predictive matching, studied in detail in Chapter 3, postulates that given a sample of minimal size $n = k_i + k_0$ the predictive distribution should be the same under any model containing k_i extra covariates. This is a nice property because, as commented in Section 3.5.2, with such a sample size, data can not really be expected to have enough information to discriminate between models of complexity k_i.

Weak predictive matching holds no matter the value of a, b and ρ_i so it is not of much help in choosing a suitable value for ρ_i. We hence look at some related criteria.

Since, $B^R_{i0} = 1$ for a sample of size $n = k_i + k_0$, a natural question arises: What should we expect for a sample of size $n = k_i + k_0 + 1$? Can one single observation provide enough information to *strongly* discriminate between M_i and M_0?

Our intuition is that it shouldn't. Therefore, for $a = 1/2$ and $b = 1$ we look for a value of ρ_i that still makes B^R_{i0} close to 1 for samples of size $n = k_i + k_0 + 1$. Intuitively, with this sample size the data itself can fit the model, but barely so, and hence there is no much information left for model comparison. Unfortunately, except for a reduced (but important)

set of problems, this criteria is not applicable since the Bayes factor associated with the minimal information depends on the specific data (see Berger and Pericchi, 2001, and references therein). We therefore take a more modest requirement aiming, instead, for Bayes factors as close to 1 as possible for data clearly compatible with the simpler model. This idea is similar to the predictive matching approach of Ghosh and Samanta (2002) and Spiegelhalter and Smith (1982). For the purposes of this chapter, we use it to help in choosing a good value for ρ_i. We refer to this type of predictive matching as *null predictive matching*. Notice that this criterion is not the same as the *weak predictive matching* discussed in Section 3.5.2.

On the other hand, it would not be reasonable to aim for perfect matching (i.e. $B^R_{i0} = 1$) for a sample of size $n = k_i + k_0 + 1$ since a sample of this size allows the frequentist estimation of parameters (but barely so, see Section 3.5.2) and so it gives some useful (even if small) information about the comparison when using our partially informative prior. Also, since the models under comparison are of different dimensions, a sample of this size could provide some information about the adequacy of M_0. Moreover, and since we are supposing data compatible with M_0, the Bayes factor should intuitively be $B_{i0} < 1$.

We combine both reasonings above and aim to be "close" to null predictive matching for $n = k_0 + k_i + 1$, but without the probably unreasonable requirement of "exact" null predictive matching.

Specifically, we choose ρ_i such that, meets previous requirements and makes the Bayes factor as close as possible to 1 when the sample size is $n = k_i + k_0 + 1$ and the test statistic $Q_{i0} \to 1$. That is, choose ρ_i to make $\underline{B}^R_{i0} = \lim_{Q_{i0}\to 1} B^R_{i0}$, as close as possible to 1. For $a = 1/2$, $b = 1$ and $n = k_i + k_0 + 1$ this means to make (see Proposition 4.4 and its proof in Appendix F.3)

$$\underline{B}^R_{i0} = \frac{1}{k_i + 1} \left[\rho_i (k_i + k_0 + 2)\right]^{-\frac{k_i}{2}}, \tag{5.3}$$

as close as possible to 1. It is easy to see that $\underline{B}^R_{i0}$ is a decreasing function of ρ_i, bounded above by 1/2. Therefore, ρ_i should be taken as small as possible. Since for ρ_i to be valid in complete generality, we need it to be $\rho_i \geq 1/(1+n)$ for any sample size n giving proper posteriors. That is $\rho_i \geq 1/(1+n_{\min})$ (so as to ensure propriety of the conditional prior for the "new" parameters). With the proposed partially proper prior the minimal sample size is $n_{\min} = k_i + k_0$ (see Section 3.5.2) which is precisely the sample size for which weak predictive matching was defined. The value of ρ_i minimizing (5.3) and giving proper conditional priors for all $n \geq k_i + k_0$ is then

$$\rho_i = 1/(k_i + k_0 + 1), \tag{5.4}$$

which is our recommended choice for ρ_i.

One could alternatively consider that the minimal sample size should be $n^*_{\min} = k_i + k_0 + 1$ since this gives proper posteriors with the usual objective estimation priors. Then the choice $\rho_i = 1/(k_i + k_0 + 2)$ makes the corresponding $\underline{B}^R_{i0}$ closer to 1 and hence more appropriate for the sole purpose of null predictive matching. Notice, however, that with this value of $n^*_{\min}$, weak predictive matching (Section 3.5.2) is lost. ($\rho_i = 1/(k_i + k_0 + 2)$ would not be a permissible value for ρ_i when $n = k_i + k_0$).

It is important to remark that the choice in (5.4) does make the Bayes factor model consistent. Recall that this happens whenever $\rho_i(b+n)$ tends to infinity with n (see Proposition 4.2) which holds for this choice of ρ_i.

When requiring "null predictive matching" we have taken "compatibility with the null" to mean $Q_{i0} \to 1$, but this is not the only possibility. Indeed, even though, $Q_{i0} \to 1$ intuitively provides the most evidence in favor of M_0, the distribution of $Q_{i0} \mid M_0$ for $n = k_i + k_0 + 1$ (a beta with parameters 1/2, $k_i/2$) accumulates very sharply around 0 making values of Q_{i0} close to 1 virtually impossible. This suggests that, in order to choose ρ_i we should also check the behavior of B^R_{i0} for data that is

compatible with the null model in the sense of being compatible with the distribution of Q_{i0} under M_0. We next explore other forms of compatibility of the data with M_0.

In particular, the next proposition shows that B_{i0}^R (for $a = 1/2$, $b = 1$ and $n = k_i + k_0 + 1$) is also decreasing with ρ_i for a wide range of Q_{i0} values compatible with M_0, including the mean under M_0, $\mathrm{E}[Q_{i0} \mid M_0] = 1/(k_i+1)$. Hence, our recommended choice is still $\rho_i = 1/(k_i+k_0+1)$ for a much wider range of data that could be considered "compatible" with M_0. Moreover, the Bayes factor can also be shown to be smaller than 1.

Proposition 5.1. *Let $\boldsymbol{y}$ be a sample of size $n = k_i + k_0 + 1$ with $Q_{i0} \geq (k_i + 1)^{-1}$. For $a = 1/2$ and $b = 1$, B_{i0}^R is decreasing with ρ_i.*

Moreover, for the distribution of Q_{i0} under M_0 the region $Q_{i0} \geq (k_i+1)^{-1}$ accumulates at least at 30% of probability and contains the mean (that is, $(k_i + 1)^{-1}$ is not larger than the 70% percentile).

Proof. See Appendix G.1 □

Proposition 5.2. *Let $\boldsymbol{y}$ be a sample of size $n = k_i + k_0 + 1$ with $Q_{i0} \geq (k_i + 1)^{-1}$. For $a = 1/2$, $b = 1$ and $\rho_i = 1/(k_i + k_0 + 1)$ the maximum value of B_{i0}^R is always less than 1.*

Proof. See Appendix G.2 □

We comment in passing that for samples of size $n = k_i + k_0 + 1$ and $Q_{i0} \geq (k_i + 1)^{-1}$, it can be shown that when $\rho_i = 1/(k_i + k_0 + 2)$, B_{i0}^R is decreasing as a function of k_i (see Proposition G.1 in Appendix G) which seems intuitively natural. We have not been able to proof that this property also holds for $\rho_i = 1/(k_i + k_0 + 1)$, although extensive numerical exploration suggest that it should also be true for this choice.

Another interesting remark concerning choice for ρ_i is in order. Recall that, when $a = 1/2$ and $b = 1$ the scale matrix in (5.2) is

$$\boldsymbol{\Sigma}_i^* = \rho_i \, \frac{2}{\pi} \, (1+n) \, \sigma^2 (\boldsymbol{V}_i^t \boldsymbol{V}_i)^{-1}.$$

That is, asymptotically, choices of ρ_i close to one roughly correspond to unit information types for the prior scale, while the proposed choice seems to correspond to dividing unit information by the number of parameters. This intriguing relation with unit information ideas is certainly worth of further research.

5.5 Recommended Conventional Robust Bayes factor: a comparison with other proposals

Our ultimate Conventional Robust Bayes factor is the one computed with the Conventional Robust prior with

$$a = 1/2, \; b = 1, \; \rho_i = 1/(k_i + k_0 + 1),$$

and it is given by:

$$\begin{aligned} B_{i0}^R = & \frac{1}{k_i + 1} \left[\frac{n+1}{k_i + k_0 + 1} \right]^{-k_i/2} Q_{i0}^{-\frac{n-k_0}{2}} \\ & {}_2F_1\Big[\frac{k_i+1}{2}; \frac{n-k_0}{2}; \frac{k_i+3}{2}; \frac{(1 - Q_{i0}^{-1})(k_i + k_0 + 1)}{(1+n)} \Big], \end{aligned} \tag{5.5}$$

a very simple, computationally fast, closed-form expression.

In what follows we compare the Bayes factor in (5.5) (to be denoted by R1) with other alternatives. Specifically, we compare R1 with the Conventional Robust Bayes factors of Berger (1985) (denoted by R2) and Liang et al. (2008) (referred to as Li) and also with the Bayes factor resulting from the use of the Cauchy distribution of Jeffreys (1961) and

Zellner and Siow (1980) (denoted here by JZS). We present in Table 5.1 the notation and the colour code for those approaches.

Notation	Color	Description
R1	■	our recommended B^R_{i0}, that is, $a = 1/2$, $b = 1$ and $\rho_i = 1/(k_i + k_0 + 1)$
R2	■	Berger's estimation proposal, that is, B^R_{i0} with $a = 1/2$, $b = 1$ and $\rho_i = (1+k_i)/(3+k_i)$
Li	■	Liang et al. B_{i0}, which corresponds to the B^R_{i0} with $a = 1/2$, $b = n$ and $\rho_i = 1/2$
JZS	■	B_{i0} resulting from the Cauchy prior distribution of Jeffreys (1961) and Zellner and Siow (1980)

TABLE 5.1: *Notation for the entertained approaches*

In Figure 5.4 we display $\log_{10}$ of the Bayes factors B_{i0} for the entertained approaches, computed at $\mathrm{E}[Q_{i0} \mid M_0]$. Pictures on the left column show the behavior of the different Bayes factors as k_i grows with: i) a sample of size $n = 1000$ (top); and ii) a sample of minimal size $n = k_i + k_0 + 1$ varying for each k_i (bottom). Because our approach (R1) and Liang et al. (2008)'s approach (Li) can not really be distinguished in the scale of the pictures on the left, we present, on the right, a "blown up" of these two approaches. Figure 5.5 show the same pictures but for B_{i0} computed at $Q_{i0} = 1$.

It is interesting to remark that, when n is fixed all of the B_{i0}'s computed at $\mathrm{E}[Q_{i0} \mid M_0]$ decrease with k_i until k_i is close to the corresponding minimal sample size ($k_i \approx n - k_0 - 1$). At this point B_{i0} grows to 1. This behavior is clearly reflecting the idea of null predictive matching under minimal sample size $n = k_i + k_0 + 1$.

On the other hand, when $Q_{i0} = 1$ this is no longer true, except for our proposal (R1). Since we consider that this behavior is according with

intuition, we take this as further support for our recommended choice for ρ_i.

Notice that in sharp contrast, Li stays close to 1 all the way. We explain this effect in light of the results in Section 3.3.1, where it was shown that the prior distribution of Liang et al. sharply spike around $\boldsymbol{\beta}_i = 0$. This results in Bayes factors that, under M_0, tend to be closer to 1 than any other Conventional Robust Bayes factor, and also than the Jeffreys, Zellner and Siow's Bayes factor (JZS).

Note that for a sample of size $n = k_i + k_0 + 1$ (bottom pictures) our proposal (R1) is always the closest to 1. This is, in fact, what we were aiming to with the choice of ρ_i in Section 5.4. We can also observe in these pictures that, for $n = k_i + k_0 + 1$, all the entertained configurations are decreasing with k_i, even R1 in $\mathrm{E}[Q_{i0} \mid M_0]$ (Figure 5.4 bottom pictures) but it decreases so slowly that it can not be noticed in the pictures.

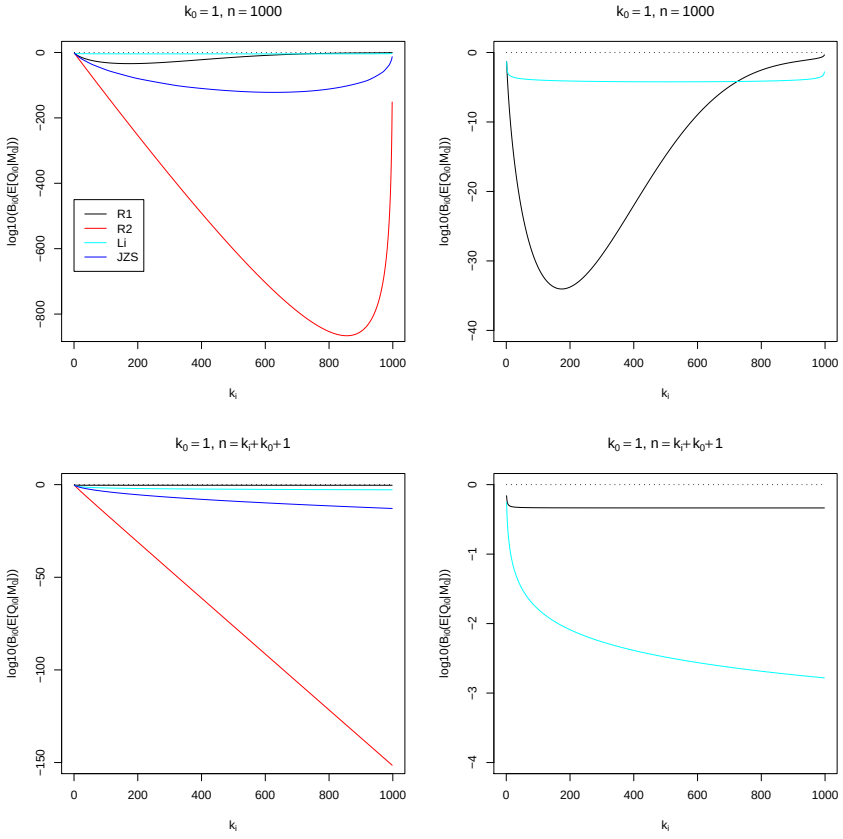


FIGURE 5.4: *Behavior of* $\log_{10}(B_{i0})$ *for the 4 entertained approaches (left) and a "blown up" for R1 and Li (right), with* k_i*, for fixed* $n = 1000$ *(top) and* $n = k_i + k_0 + 1$*(bottom) computed at* $E[Q_{i0} \mid M_0]$.

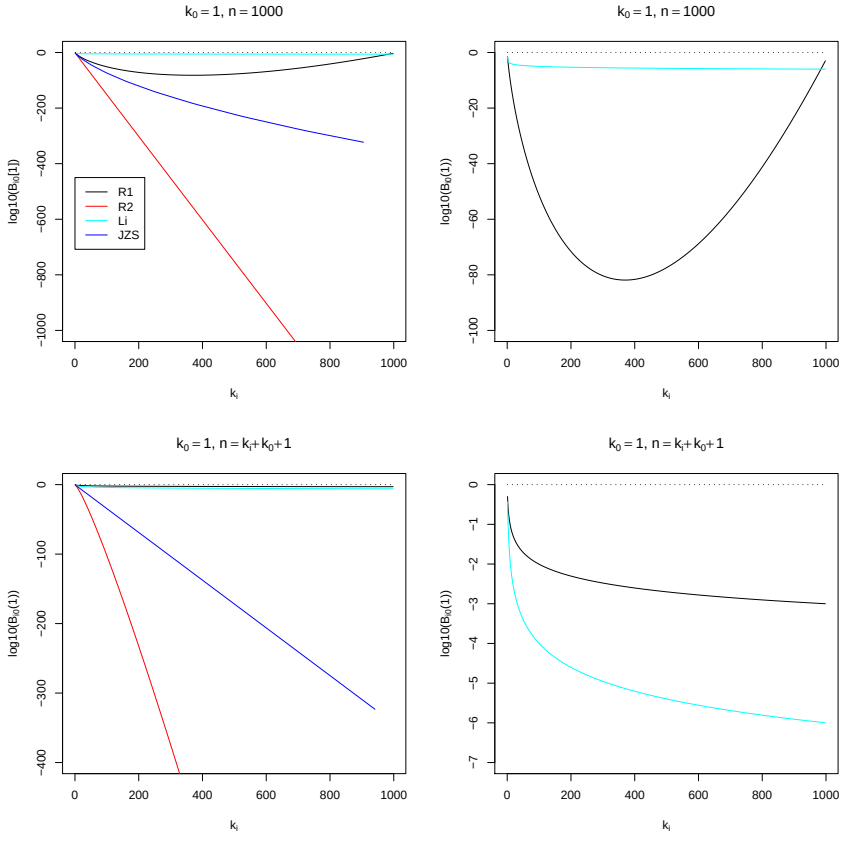


FIGURE 5.5: *Behavior of* $\log_{10}(B_{i0})$ *for the 4 entertained approaches (left) and a "blown up" for R1 and Li (right), with* k_i*, for fixed* $n = 1000$ *(top) and* $n = k_i + k_0 + 1$*(bottom) computed at* $Q_{i0} = 1$.

Chapter 6

Examples

6.1 Introduction

In this chapter our approach (described and studied in Chapters 3, 4 and 5) is illustrated in a number of data sets. We compare the results with those obtained from the use of standard Conventional (Bayes) approaches (see Section 2.3.3).

In particular we use four data sets, three of which are real data sets, widely studied in the literature, while the fourth one is a simulated (correlated) example. Two of the data sets entertain a small number of covariates ($p = 4$ and $p = 15$). The other two, consider a larger number of explanatory variables ($p = 27$ and $p = 30$).

In each one of these examples we have computed the corresponding posterior probabilities of the 2^p potential models $P(M_i \mid \boldsymbol{y})$. These posterior probabilities are summarized here through the highest posterior probability models, posterior probability of dimensions, and inclusion probabilities of covariates.

We interpret the posterior probabilities of dimensions as an interesting summary for measuring the preference of each approach for more or less

complex models. The posterior probability of each dimension k where $k \in \{0, \ldots, p\}$ is :

$$P(k \mid \boldsymbol{y}) = \sum_{\{M_l:\, k_l = k\}} P(M_l \mid \boldsymbol{y}).$$

The inclusion probability of each covariate is defined (see Barbieri and Berger, 2004) as the sum of the posterior probabilities of the models that include that covariate:

$$P(X_j \mid \boldsymbol{y}) = \sum_{\{M_l:\, X_j \in M_l\}} P(M_l \mid \boldsymbol{y}),$$

for $j = 1, \ldots, p$. These probabilities have interesting theoretical properties (see Barbieri and Berger, 2004). Also, they are useful for synthesizing the results, specially when the number of models is large and the posterior probabilities are so small that they become difficult to interpret. We also present the median probability model which contains all the covariates with inclusion probability higher than 0.5. This model (which does not necessarily equal the higher posterior probability model) has appealing predictive characteristics (see Barbieri and Berger, 2004, for details)

In the assignment of prior probabilities over the model space, $P(M_i)$, we assume two different approaches:

1. The default approach that considers every model equally probable a priori. That is: $P(M_i) = 1/2^p$ for $i = 0, \ldots, 2^p - 1$. This prior could also be alternatively interpreted as reporting Bayes factors (which are equal to posterior odds for this prior). In what follows we refer to this approach as PMD.

2. The multiplicity control approach presented by Scott and Berger (2010). That is: $P(M_i) = \binom{p}{k_i}^{-1}/(p+1)$ (see Section 2.3.4). In what follows we refer to this approach as PMSB.

Finally, for $\pi_i(\boldsymbol{\beta}_i, \boldsymbol{\beta}_0, \sigma)$ we use six different Conventional priors, including, of course, our final proposal in Chapter 5. These are explicitly introduced in the next section.

6.2 Entertained approaches

The six Conventional priors entertained for the computation of Bayes factors (and hence of posterior probabilities) are summarized in Table 6.1. In this table we also introduce the color code used in the figures presented throughout the chapter.

Notation	Color	Description
R1	■	Our proposal for the Conventional Robust prior, that is, $a = 1/2$, $b = 1$ and $\rho_i = 1/(k_i + k_0 + 1)$.
R2	■	Berger's prior, i.e. Conventional Robust prior with $a = 1/2$, $b = 1$ and $\rho_i = (1+k_i)/(3+k_i)$.
TESS1	■	Conventional Robust prior with $a = 1/2$, $b = 1$ and $\rho_i = 1/2$ using the effective sample size n_i^T instead of n.
TESS2	■	Same as TESS1 but taking $\rho_i = \max\{1/(k_i + k_0 + 1), 1/(1 + n_i^T)\}$.
Li	■	Liang et al. prior, i.e. Conventional Robust prior taking $a = 1/2$, $b = n$ and $\rho_i = 1/2$.
JZS	■	Cauchy prior distribution of Jeffreys (1961) and Zellner and Siow (1980).

TABLE 6.1: *Notation for the entertained approaches*

We next give a brief description of each prior.

6.2.1 Our recommended Conventional Robust prior (R1)

We denote by R1 the Conventional Robust prior in (3.12) with the recommended choices for the parameters given in Chapter 5:

$$a = 1/2,\ b = 1,\ \rho_i = 1/(k_i + k_0 + 1).$$

The final expression of the Bayes factor using this prior is:

$$B_{i0}^{R1} = \frac{1}{k_i+1}\left[\frac{n+1}{k_i+k_0+1}\right]^{-k_i/2} Q_{i0}^{-\frac{n-k_0}{2}} \\ {}_2F_1\Big[\frac{k_i+1}{2}; \frac{n-k_0}{2}; \frac{k_i+3}{2}; \frac{(1-Q_{i0}^{-1})\,(k_i+k_0+1)}{(1+n)}\Big].$$

6.2.2 Berger (1985)'s Prior (R2)

We call R2 the Robust prior of Berger (1985)'s, that is the Conventional Robust prior in (3.12) taking $a = 1/2$, $b = 1$, and $\rho_i = (k_i + 1)/(k_i + 3)$.

It is important to recall that Berger's choices of the parameters were chosen to be optimal for the robust estimation of a normal mean, not for model selection.

This choice of parameters also result in a Bayes factor depending on the simple hypergeometric function of one variable:

$$B_{i0}^{R2} = \frac{1}{k_i+1}\left[\frac{(k_i+1)\,(n+1)}{k_i+3}\right]^{-k_i/2} Q_{i0}^{-\frac{n-k_0}{2}} \\ {}_2F_1\Big[\frac{k_i+1}{2}; \frac{n-k_0}{2}; \frac{k_i+3}{2}; \frac{(1-Q_{i0}^{-1})\,(k_i+3)}{(k_i+1)\,(1+n)}\Big].$$

6.2.3 Correcting by the effective sample size (TESS1 and TESS2)

When choosing the appropriate scale for objective priors and, in particular, for our Conventional Robust priors, the sample size is needed. The sample size "corrects" (by multiplying) the information (variance) in the sample in an attempt of making it of unitary size. As discussed in Section 3.3, this correction may work well for i.i.d. observations but might not in more complex situations. An appropriate "correction" requires a suitable definition of an effective sample size, alas, this seems to still be under investigation.

So far we have assumed the default and simplest choice, which is to take the effective sample size equal to the sample size n. However, it is obviously interesting to gain some understanding in the impact of using other choices for the effective sample size. In particular, we assume here the novel definition of effective sample size of Berger et al. (2010b) referred to as *The Effective Sample Size* (TESS). The definition of TESS in Berger et al. (2010b) is adapted in Appendix B to be used in our specific framework. Interestingly, a different TESS, n_i^T is obtained for each model M_i depending on the specific design of the model (i.e. the included covariates).

The sample size n_i^T is a value between 1 and n reflecting the quantity of information provided by the data. Notice that a value of $n_i^T \approx n$ indicates that the observations are almost i.i.d. and equally informative, while a value of $n_i^T \approx 1$ might indicates large correlation in the data for a situation in which one observation provides most of the information see Berger et al. (2010b) for examples and discussions.

The definition of n_i^T and some more insight about this topic can be found in Appendix B.

A note of caution is needed here: all the theoretical developments in this thesis have been done using n as sample size and might not be applicable when using n_i^T instead. For instance, the permissible parametric space in (3.7) when using n_i^T becomes:

$$\mathcal{A}^T = \{(a, b, \rho_i) : \ a > 0, \ \ 1 \leq b \leq n_i^T, \ \ \rho_i \geq b/(b + n_i^T)\}.$$

Then, the choices for a and b in Chapter 5 ($a = 1/2$ and $b = 1$) are still valid. Unfortunately, given $b = 1$, ρ_i should be $\rho_i \geq 1/(1+n_i^T)$ and, since n_i^T can be smaller than $k_i + k_0 + 1$, the choice of $\rho_i = 1/(k_i + k_0 + 1)$ is not longer valid in general. Notice also, that the asymptotic behavior of n_i^T is not clear. In particular, it might happen that n_i^T does not grow to infinity with n. Therefore, the asymptotic properties in this thesis can not be assumed to hold when simply replacing n by n_i^T.

A detailed study of the optimal choice for ρ_i when using n_i^T is beyond the scope of this thesis, and hence, we don't have any theoretical basis for its choice. Anyway, in an attempt to study the impact of TESS in the results we present two possible choices of ρ_i (for a and b we consider the same choices as in R1 and R2, $a = 1/2$ and $b = 1$).

1. **TESS1**. A general choice valid for any $1 \leq n_i^T \leq n$, and any value of b (and particularly for our recommended value $b = 1$) is $\rho_i = 1/2$.

2. **TESS2**. Following the ideas in Chapter 5 (in the sense of maximizing the value of B_{i0} for a sample of minimal sample size $n = k_i + k_0 + 1$ and data compatible with M_0), another possible choice is to take ρ_i as small as possible. For $b = 1$ this leads to

$$\rho_i = \max\{1/(k_i + k_0 + 1), \, 1/(1 + n_i^T)\}.$$

For both, TESS1 and TESS2, we choose $b = 1$ so we still achieve the attractive property of having a closed-form Bayes factors which, depending

on the choice of ρ_i are:

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

for $j = 1, 2$, with $\rho_i^1 = 1/2$ and $\rho_i^2 = \max\{1/(k_i + k_0 + 1),\, 1/(1 + n_i^T)\}$

6.2.4 Liang et al. (2008)'s prior (Li)

Liang et al. (2008)'s prior distribution is a particular case of our Conventional Robust prior when taking $a = 1/2$, $b = n$, and $\rho_i = 1/2$. We denote this approach by Li.

In this case the expression of Bayes factors depend on the hypergeometric function of two variables here computed through numerical integration, using its original expression in Liang et al. (2008):

$$B_{i0}^{Li} = \int_0^\infty (1+g)^{\frac{n-k_0-k_i}{2}}\,(1+Q_{i0}\,g)^{-\frac{n-k_0}{2}}\,\frac{1}{2n}(1+\frac{g}{n})^{-\frac{3}{2}}\,dg.$$

6.2.5 Jeffreys and Zellner-Siow approach (JZS)

The Conventional approach of Jeffreys for testing a normal mean was extended to variable selection by Zellner and Siow (1980). Their proposal for $\pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0,\, \sigma)$ is a Cauchy distribution which has been shown to give very good results in this framework. We denote this approach by JZS.

The resulting Bayes factors do not have closed-form expression and must be computed by numerical integration. Its integral expression is:

$$B_{i0}^{JZS} = \int_0^\infty (1+g)^{\frac{n-k_0-k_i}{2}}\,(1+Q_{i0}\,g)^{-\frac{n-k_0}{2}}\;IGa(g \mid \frac{1}{2},\,\frac{n}{2})\,dg.$$

6.3 Computation

Our code to compute the 2^p posterior probabilities, the inclusion probabilities and the dimension probabilities, was programmed in ansi C.

To navigate the model space, the binary representation of the natural numbers proved to be a very efficient and powerful tool.

One of the slowest parts of the analysis was the computation of the residual sums of squares (SSE). For this calculation it was very useful to compute the SSE of a regression model (response $\boldsymbol{y}$ and $n \times k$ design matrix $\boldsymbol{X}$) as the sum of the squares of the last $n - k$ components of the vector $\boldsymbol{Q}'\boldsymbol{y}$, where $\boldsymbol{QR}$ is the qr-decomposition of $\boldsymbol{X}$. All algebraic expressions were evaluated using the *GNU Scientific library* (gsl) (see Galassi et al., 2009).

The gsl-library is also used for the computation of the Bayes factors. In particular, the approaches using $b = 1$, R1, R2, TESS1 and TESS2 are computed using the hypergeometric function included in gsl-library. Finally, for Li (Conventional Robust Bayes factor with $b = n$), and JZS we use numerical integration provided also by gsl-library.

The examples with $p = 4$ and $p = 15$ are computed in few seconds (even for the problem with $2^{15} = 32768$ models).

For the examples entertaining a larger number of covariates, in Sections 6.4.3 and 6.4.4, we preferred a parallel computation. The problem is trivially 'parallelized' by simply assigning a bunch of marginal likelihoods to be computed by each separate CPU. This is so obvious, that it is called "an embarrassing parallel problem" (Tierney et al., 2007). This solution reduces considerably the computational time. Still, the numerical integrations are very time consuming. Another issue with numerical integration in these examples is that it sometimes does not work properly and do not produce reasonable results. In particular, the routines for

computing Li and JZS that worked properly in the first two examples, worked very slowly or even would not work at all (for instance JZS do not produce numerical results in the Ozone example but *nan*'s) in the large examples. Hence, we decided just to compute R1, R2 and TESS1 for those problems.

6.4 Examples

We next present the results for the four entertained data sets. In all of this examples and since no other information is available, we suppose that the simpler model M_0 contains only the intercept (so $k_0 = 1$).

6.4.1 Hald data

The following data relates to an engineering application that was interested in the effect of the cement composition on heat evolved during hardening (for more details, see Woods et al., 1932). The response variable is the heat evolved per gram of cement (in calories). The entertained covariates ($p = 4$) are described in Table 6.2. This data has been analyzed by many authors (e.g. George and McCulloch, 1993; Hald, 1952; Laud and Ibrahim, 1995; Pérez, 1998, among others).

Covariate	Description
X_1	Amount of tricalcium aluminate
X_2	Amount of tricalcium silicate
X_3	Amount of tetracalcium alumino ferrite
X_4	Amount of dicalcium silicate

TABLE 6.2: *Hald data. Description of covariates*

As a summary of the results we give the following tables and figures:

- Tables 6.3(a) (PMD) and 6.3(b) (PMSB) display the five models with largest posterior probabilities according to R1.

- Tables 6.4(a) (PMD) and 6.4(b) (PMSB) provide the inclusion probabilities for each covariate.
- Figure 6.1 presents the dimension probabilities (top) and the cumulative dimension probabilities (bottom) for PMD (left) and PMSB (right)
- Figure 6.2 presents the distribution of the $\log_{10}$ of the posterior probabilities of the 2^p models through box-plots for PMD (left) and PMSB (right). It is important to keep in mind that they represent the distribution of the *model probabilities* and not the distribution of the models per se.

The effect of $\pi_i(\boldsymbol{\beta}_i, \boldsymbol{\beta}_0, \sigma)$

We do not find large differences among the approaches as posterior probabilities do not change much with $\pi_i(\boldsymbol{\beta}_i, \boldsymbol{\beta}_0, \sigma)$.

In Figure 6.2 we appreciate two slightly different ways of apportioning the probability. The approaches R2, Li and JZS seem to produce a flatter distribution of model probabilities, having slightly larger tails for small probabilities (that is, smaller values of the probability "appear" more often) than R1, TESS1 and TESS2 do. However, the shape of the tail for the large values of model probabilities is similar in all the six approaches.

The effect of $P(M_i)$

The choice of $P(M_i)$ seems to have a larger effect in the results than the choice of the prior in the parametric space. The most significative differences are observed in Figure 6.1. In this Figure the effect of the prior distribution ($p(k) = \sum_{k_i=k} P(M_i)$, represented by a dashed line) is clearly reflected in the posterior result. In this sense, the posterior

dimension probability using PMD is concentrated around 2 (i.e. $p/2$) while, it is a little bit flatter when using PMSB.

(a) *The 5 most probable (using R1) models with PMD*

Covariates	R1	R2	TESS1	TESS2	Li	JZS
$\{X_1\ X_2\}$	0.546	0.523	0.543	0.531	0.542	0.535
$\{X_1\ X_4\}$	0.174	0.167	0.157	0.170	0.167	0.169
$\{X_1\ X_2\ X_4\}$	0.090	0.100	0.098	0.096	0.095	0.096
$\{X_1\ X_2\ X_3\}$	0.089	0.099	0.097	0.095	0.094	0.095
$\{X_1\ X_3\ X_4\}$	0.072	0.079	0.073	0.076	0.075	0.076

(b) *The 5 most probable (using R1) models with PMSB*

Covariates	R1	R2	TESS1	TESS2	Li	JZS
$\{X_1,\ X_2\}$	0.464	0.437	0.456	0.446	0.458	0.451
$\{X_1,\ X_4\}$	0.148	0.139	0.131	0.143	0.141	0.142
$\{X_1,\ X_2,\ X_4\}$	0.115	0.125	0.123	0.121	0.120	0.121
$\{X_1,\ X_2,\ X_3\}$	0.114	0.123	0.122	0.120	0.119	0.119
$\{X_1,\ X_3,\ X_4\}$	0.091	0.099	0.092	0.096	0.095	0.096

TABLE 6.3: *Hald data. The 5 highest probability models according to R1 and their probabilities under the other approaches.*

(a) *Inclusion probabilities with PMD*

Covariate	R1	R2	TESS1	TESS2	Li	JZS
X_1	0.980	0.978	0.977	0.979	0.981	0.980
X_2	0.750	0.750	0.766	0.750	0.755	0.752
X_3	0.190	0.210	0.203	0.203	0.196	0.200
X_4	0.365	0.378	0.360	0.373	0.364	0.370

(b) *Inclusion probabilities with PMSB*

Covariate	R1	R2	TESS1	TESS2	Li	JZS
X_1	0.976	0.974	0.973	0.975	0.977	0.976
X_2	0.757	0.758	0.773	0.758	0.761	0.759
X_3	0.272	0.299	0.290	0.290	0.280	0.286
X_4	0.422	0.440	0.422	0.434	0.423	0.430

TABLE 6.4: *Hald data. Inclusion probabilities. The median probability model contains the covariates corresponding to the gray coloured rows.*

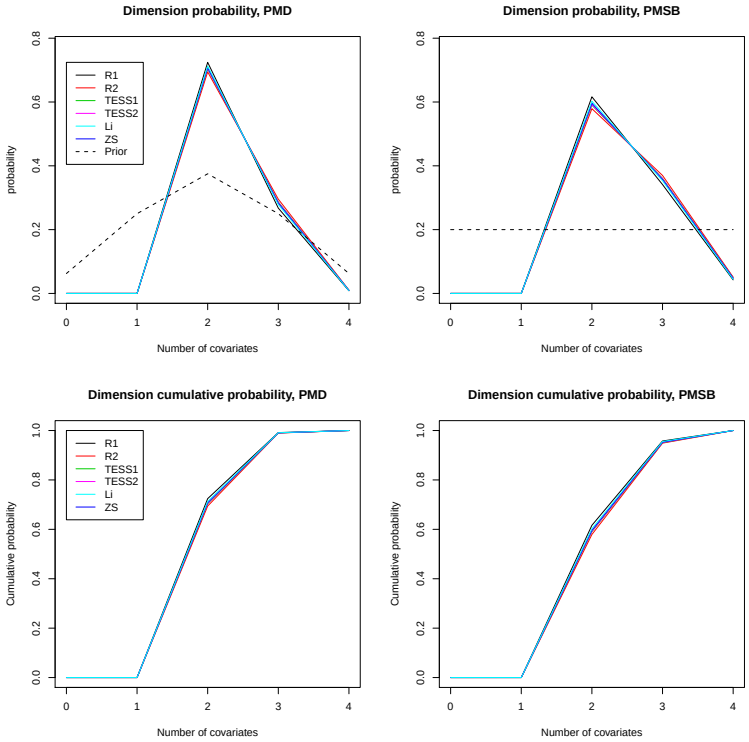


FIGURE 6.1: *Hald data. Posterior probabilities (top) and cumulative posterior probability (bottom) of the dimension of the true model for priors PMD (left) and PMSB (right). Induced prior over dimension is represented by a dashed line.*

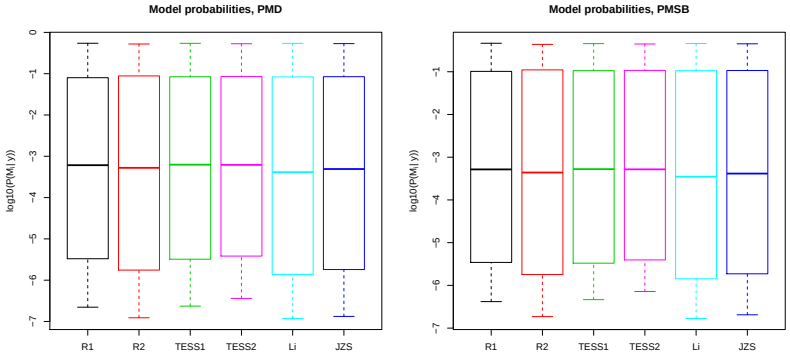


FIGURE 6.2: *Hald data. Distribution of model posterior probabilities for priors PMD (left) and PMSB (right).*

6.4.2 Crime data

Our second example relates USA rate crime data with several social explanatory variables. The experiment was designed and performed by Ehrlich (1973) and revised by Vandaele (1978). The data is distributed as part of the library `MASS` in R and has been analyzed from a Bayesian perspective by a number of authors, including Fernández et al. (2001); Hoeting et al. (1999); Liang et al. (2008); Raftery et al. (1997). The experiment consists on observations of rate crime during 1960 in $n = 47$ states in the US. Apart from the intercept, $p = 15$ covariates (briefly described in Table 6.5) are considered for explanation of the rate crime. Ehrlich (1973), based on theoretical arguments, concentrated on two regression models that included the covariates indicated with a "•" in Tables 6.7(a) and 6.7(b). As in the papers cited above, we transform the data into the logarithmic scale (except for the indicator variable `So`).

Covariate	Description
`M`	Percentage of males aged 14-24
`So`	Indicator variable for a southern state
`Ed`	Mean years of schooling
`Po1`	Police expenditure in 1960
`Po2`	Police expenditure in 1959
`LF`	Labour force participation rate
`MF`	Number of males per 1000 females
`Pop`	State population
`NW`	Number of nonwhites per 1000 people
`U1`	Unemployment rate of urban males 14-24
`U2`	Unemployment rate of urban males 35-39
`GDP`	Gross domestic product per head
`Ineq`	Income inequality
`Prob`	Probability of imprisonment
`Time`	Average time served in state prison

TABLE 6.5: *Crime data. Description of the covariates*

After computing the $2^{15} = 32\,768$ Bayes factors with each configuration we present the following tables and figures as a summary of the results:

- Tables 6.6 (PMD) and 6.7 (PMSB) display the 15 most probable models as ordered by R1. In each of these tables we include a column indicating the order that the model has when the other approach to prior probabilities over the model space is used instead (O. PMSB indicates the order in posterior probabilities of models when the multiplicity control prior is used, and O. PMD indicates the order for the default prior over the model space).

- Tables 6.8(a) (PMD) and 6.8(b) (PMSB) provide the inclusion probabilities for each covariate.

- Figure 6.3 presents the dimension probabilities (top) and the cumulative dimension probabilities (bottom) for PMD (left) and PMSB (right).

- Figure 6.4 presents the distribution of the $\log_{10}$ of the posterior probabilities of the 2^p models through box-plots for PMD (left) and PMSB (right).

- Figures 6.5(a) (PMD) and 6.5(b) (PMSB) present the number of models needed to achieve some fixed probability with each of the approaches. In particular, ordering the 2^p-vector of posterior probabilities within each approach we compute the minimum number of models needed for achieving a probability of (0.2, 0.5, 0.7, 0.9) respectively

The effect of $\pi_i(\boldsymbol{\beta}_i, \boldsymbol{\beta}_0, \sigma)$

In this example, prior choice seems to play an important role.

For description purposes we differentiate three types of behavior among the entertained approaches

- The first group is formed by R1, TESS1 and TESS2. These approaches seem to behave quite similarly in all the summaries presented here with two small differences. First we find that when using TESS as the sample size, the results seems to be slightly less conservative (giving larger probability to more complex models) as we can appreciate in Figure 6.3. Also, in Figure 6.4 we notice that the distribution of posterior probability of models is slightly favouring higher probabilities for TESS2 (note that the lower tail of the box-plot is shorter for TESS2 than for R1 and TESS1). The similitudes among these distributions are not extremely surprising because these approaches use almost the same prior distribution with the only differences being the definition of the sample size (here $n = 47$ while $n_i^T \approx 10$) and the choice of ρ_i and its seems that, in this specific example, those are not very influent.

- The second group comprises Li and JZS approaches. These two approaches present certain similitudes which can be appreciated mainly when looking at high model posterior probabilities. For instance, the highest probability models set probabilities which are very close to each other in both approaches (see Tables 6.6 and 6.7). Also, in Figure 6.4 we observe that the upper tails, representing high probabilities, are quite similar (note, however, that it is not so for lower tails). Looking back to Section 3.3.2, we find a likely explanation for this effect. Asymptotically (with large n) Li and JZS have similar tails. Recall that both distributions have Cauchy tails with asymptotically similar covariance matrices (JZS's covariance matrix is $n\,\sigma^2\,(\boldsymbol{V}_i^t\boldsymbol{V}_i)^{-1}$ and Li's one is $(\pi/2)\,n\,\sigma^2\,(\boldsymbol{V}_i^t\boldsymbol{V}_i)^{-1}$). These similitudes in the tails will be evident in problems, as this one, where there is clear support for complex models and little support for models very close to M_0 in complexity. Note that, indeed, the highest model posterior probabilities in this example correspond to complex models (containing around 8 covariates for PMD and

15 for PMSB, of the 15 entertained). We presume that this similarities will disappear in problems with data supporting the simplest models. (Recall that we expect Li and JZS priors to be very similar only on the tails; Li is quite more concentrated than JZS around 0.)

- The third group just comprises R2. Berger's Robust prior seems to behave in a completely different way. The most likely explanation is the fact that this prior was developed for estimation and not for model selection.

As we can observe in Figure 6.3, the first group is the less conservative among the three (not too far from Li and JZS), assigning larger probability to models with larger number of covariates. On the other hand, R2 is the most conservative and even more when using PMSB. This is clearly displayed by the red graph which is clearly located to the left of all the others.

The different behavior of R2 is reflected in a different ordering of the models, as we observe in Tables 6.6 and 6.7. We also find significative differences in the inclusion probabilities. Indeed, the median probability model for PMSB with R2 contains a considerable smaller number of covariates than the other approaches.

In Figure 6.4 we observe very different ways of apportioning probability. R2 presents longer tails than any of the other approaches, followed by JZS, TESS1 and R1. Note also that the central part of the distribution of model probabilities with R2 is placed towards the bottom of the rest of approaches. This effect is clarified in Figures 6.5(a) and 6.5(b). In these pictures we observe that R2 assigns very large probabilities to a small bunch of models (this explain that the upper tail in Figure 6.4 is larger for R2) and, hence, assigning smaller probabilities to the rest of models (what produces the displacement of the center towards the bottom and the long lower tail of the distribution of model probabilities). In fact,

when using PMD, for R2 to achieve a 90% of the probability, we only need around 1000 models while for JZS and Li we need around 1500 models and for R1, TESS1 and TESS2 we need around 2000 models.

Summarizing, whereas R1, TESS1, TESS2, Li and JZS give in general very similar results, we can find large differences for R2. This suggests the following considerations:

- Since R2 was developed in an estimation framework it may not be doing an optimal job in model selection, being maybe too conservative. On the other hand, the fact that it concentrates a lot of probability in a smaller number of models (which, except for little differences in the ordering, are the most probable models with all the approaches) may be useful for entertaining search methodologies over the model space, making it easier to quickly detect the most probable models.
- JZS has been proved to be a very good choice for model selection, hence having a methodology which give similar results to this one, as is the case of R1, which can be computed in closed-form, is a very appealing situation and encourages the use of our novel approach.

The effect of $P(M_i)$

The two choices of priors over the model space also have a large impact in the results. In Tables 6.8(a) and 6.8(b) we see that the order of the higher posterior probability models is quite different from PMD to PMSB. While PMSB seems to concentrate high posterior probabilities around the same models that PMD does: notice that the 15 most probable models with PMD are included in the 40 most probable models with PMSB. However, PMSB also gives high posterior probability to several models not "captured" by PMD. In particular, PMSB seems to prefer more complex models. Note that the most probable model with PMSB

contains all the 15 covariates while the most probable models with PMD (the 5th with PMSB) just contains 8 of them.

This preference for more complex models is also reflected in Figure 6.3 where we observe a displacement of the dimension probability distribution towards higher dimensions from PMD to PMSB. Notice that referring to PMSB as the "multiplicity control" approach may be confusing since, here, this approach is reflecting certain preference for complex models. In fact, PMSB will actually favor larger models in examples where larger models have significant weight. But, at the same time, this choice will penalize larger models when the posterior support is primarily on smaller models as we will see in the next example. This is what it is usually desired in classical "multiplicity control", hence we retain the name, but keeping in mind that this does not imply always penalizing for larger models, only when data does discourage them.

The inclusion probabilities also change from one approach to the other as we can see in Tables 6.8(a) and 6.8(b). Even the median probability models are different (except for R2 which seems to have a quite different behavior as commented above). The median probability model for PMB (and R2 in PMSB) consider 8 covariates and among them we find 4 of the 7 covariates indicated by Ehrlich, while the median probability model for PMSB includes 11 covariates and among them 6 of the 7 Ehrlich's covariates, again reflecting the preference of PMSB for more complex models.

These ideas indicates that the results are quite sensitive to the choice of priors over the model space, so this choice should be done carefully. Among other reasons, the fact that the approach in Scott and Berger (2010) PMSB accounts for multiplicity (see Section 2.3.4) makes this choice the most attractive to us.

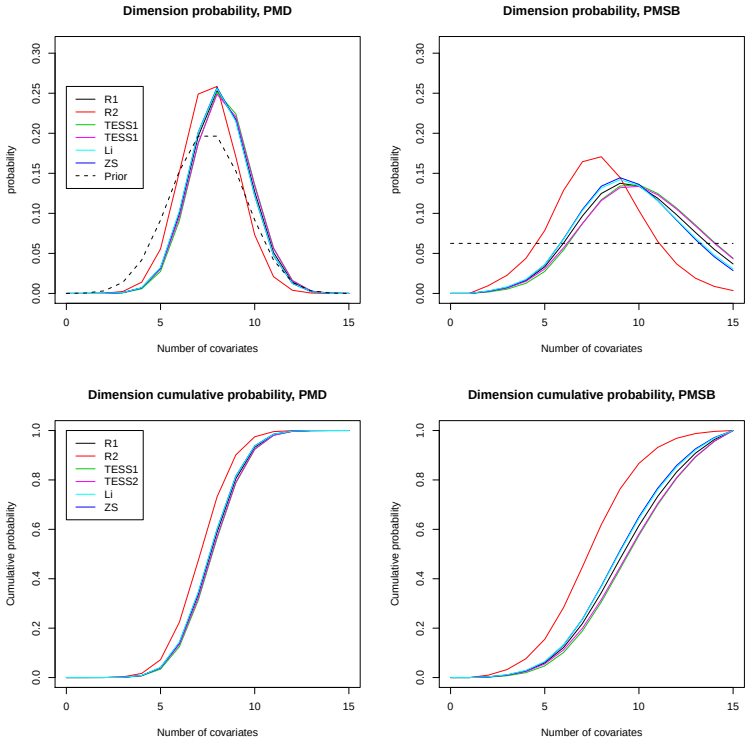


FIGURE 6.3: *Crime data. Posterior probabilities (top) and cumulative posterior probability (bottom) of the dimension of the true model for priors PMD (left) and PMSB (right). Induced prior over dimension is represented by a dashed line.*

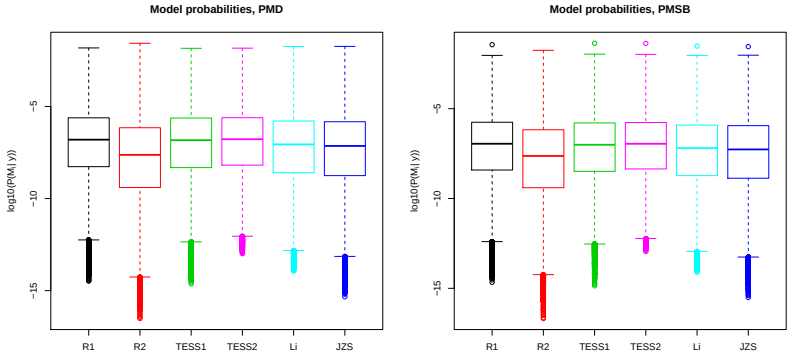


FIGURE 6.4: *Crime data. Distribution of model posterior probabilities for priors PMD (left) and PMSB (right).*

O. PMSB	Covariates	R1	R2	TESS1	TESS2	Li	ZS	k_i
5	{M, Ed, Po1, NW, U2, Ineq, Prob, Time}	0.015	0.024	0.014	0.015	0.018	0.018	8
6	{M, Ed, Po1, NW, U2, Ineq, Prob }	0.015	0.027	0.014	0.014	0.017	0.018	7
11	{M, Ed, Po2, NW, U2, Ineq, Prob}	0.010	0.018	0.009	0.009	0.012	0.012	7
14	{M, Ed, Po1, Pop, NW, U2, Ineq, Prob}	0.009	0.013	0.009	0.009	0.011	0.011	8
9	{M, Ed, Po1, U2, Ineq, Prob}	0.009	0.018	0.008	0.008	0.010	0.010	6
18	{M, Ed, Po1, NW, Ineq, Prob, Time}	0.008	0.013	0.007	0.007	0.009	0.009	7
10	{M, Ed, Po1, NW, U2, GDP, Ineq, Prob, Time}	0.008	0.010	0.008	0.008	0.009	0.009	9
25	{M, Ed, Po2, NW, U2, Ineq, Prob, Time}	0.007	0.010	0.007	0.007	0.008	0.008	8
36	{M, Ed, Po1, NW, U2, GDP, Ineq, Prob}	0.006	0.009	0.006	0.006	0.007	0.007	8
19	{M, Ed, Po2, U2, Ineq, Prob }	0.006	0.012	0.005	0.006	0.007	0.007	6
38	{M, Ed, Po1, Pop, NW, Ineq, Prob}	0.006	0.010	0.006	0.006	0.007	0.007	7
41	{M, Ed, Po2, NW, U2, GDP, Ineq, Prob}	0.006	0.008	0.005	0.006	0.006	0.007	8
23	{M, Ed, Po1, NW, U1, U2, Ineq, Prob, Time }	0.006	0.006	0.006	0.006	0.006	0.006	9
28	{Ed, Po1, Pop, NW, Ineq, Prob }	0.005	0.010	0.005	0.005	0.006	0.006	6
29	{M, Ed, Po1, NW, Ineq, Prob }	0.005	0.010	0.005	0.005	0.006	0.006	6

TABLE 6.6: *Crime data. The 15 highest probability models according to R1 and their probabilities under the other approaches for PMD.*

O. PMD	Covariates	R1	R2	TESS1	TESS2	Li	ZS	k_i
4447	{M, So, Ed, Po1, Po2, LF, M.F, Pop, NW, U1, U2, GDP, Ineq, Prob, Time }	0.037	0.004	0.044	0.043	0.031	0.028	15
2356	{M, So, Ed, Po1, LF, M.F, Pop, NW, U1, U2, GDP, Ineq, Prob, Time}	0.009	0.002	0.011	0.011	0.008	0.008	14
2452	{M, So, Ed, Po1, Po2, LF, M.F, Pop, NW, U2, GDP, Ineq, Prob, Time}	0.009	0.002	0.010	0.010	0.008	0.008	14
2490	{M, Ed, Po1, Po2, LF, M.F, Pop, NW, U1, U2, GDP, Ineq, Prob, Time}	0.008	0.002	0.009	0.010	0.008	0.007	14
1	{M, Ed, Po1, NW, U2, Ineq, Prob, Time}	0.007	0.016	0.007	0.007	0.009	0.009	8
2	{M, Ed, Po1, NW, U2, Ineq, Prob}	0.007	0.018	0.006	0.006	0.009	0.009	7
2842	{M, So, Ed, Po1, Po2, M.F, Pop, NW, U1, U2, GDP, Ineq, Prob, Time}	0.007	0.001	0.008	0.007	0.006	0.006	14
3063	{ M, So, Ed, Po2, LF, M.F, Pop, NW, U1, U2, GDP, Ineq, Prob, Time}	0.006	0.001	0.007	0.006	0.005	0.005	14
5	{M, Ed, Po1, U2, Ineq, Prob}	0.006	0.015	0.005	0.005	0.007	0.007	6
7	{M, Ed, Po1, NW, U2, GDP, Ineq, Prob, Time}	0.005	0.008	0.005	0.005	0.006	0.006	9
3	{M, Ed, Po2, NW, U2, Ineq, Prob}	0.005	0.012	0.004	0.004	0.006	0.006	7
1072	{M, So, Ed, Po1, LF, M.F, Pop, NW, U2, GDP, Ineq, Prob, Time}	0.005	0.002	0.006	0.005	0.005	0.005	13
1091	{M, Ed, Po1, LF, M.F, Pop, NW, U1, U2, GDP, Ineq, Prob, Time}	0.005	0.002	0.005	0.005	0.005	0.005	13
4	{M, Ed, Po1, Pop, NW, U2, Ineq, Prob}	0.005	0.009	0.004	0.004	0.005	0.006	8
3423	{M, So, Ed, Po1, Po2, LF, Pop, NW, U1, U2, GDP, Ineq, Prob, Time}	0.004	0.001	0.005	0.005	0.004	0.004	14

TABLE 6.7: *Crime data. The 15 highest probability models according to R1 and their probabilities under the other approaches for PMSB.*

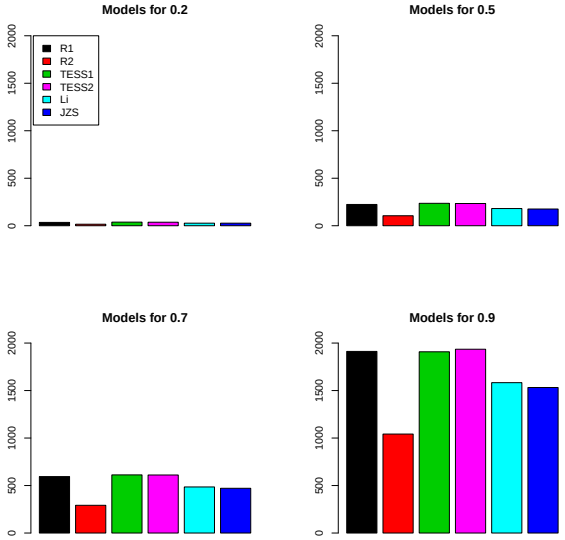
(a) *Inclusion probabilities with PMD*

Covariate	Eh1	Eh2	R1	R2	TESS2	TESS1	Li	ZS
M	•		0.836	0.836	0.839	0.837	0.848	0.850
So			0.288	0.217	0.311	0.299	0.272	0.270
Ed			0.965	0.974	0.969	0.966	0.972	0.973
Po1			0.661	0.663	0.663	0.662	0.664	0.664
Po2			0.462	0.416	0.465	0.464	0.449	0.448
LF	•		0.218	0.143	0.227	0.224	0.201	0.199
M.F			0.220	0.148	0.224	0.225	0.203	0.202
Pop			0.376	0.313	0.387	0.383	0.366	0.365
NW	•	•	0.675	0.654	0.679	0.677	0.686	0.688
U1			0.264	0.192	0.271	0.269	0.250	0.248
U2			0.596	0.576	0.607	0.602	0.607	0.609
GDP	•	•	0.366	0.290	0.371	0.371	0.355	0.355
Ineq	•	•	0.994	0.997	0.995	0.995	0.996	0.996
Prob	•	•	0.883	0.882	0.887	0.884	0.893	0.896
Time	•	•	0.369	0.309	0.376	0.376	0.365	0.366

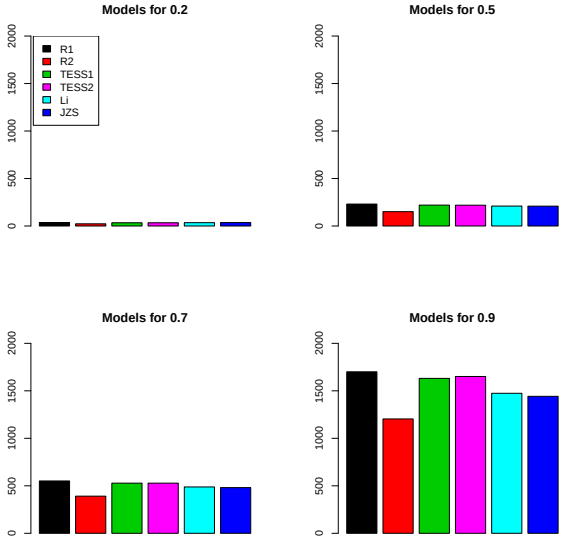
(b) *Inclusion probabilities with PMSB*

Covariate	Eh1	Eh2	R1	R2	TESS1	TESS2	Li	ZS
M	•		0.880	0.816	0.890	0.885	0.882	0.883
So			0.420	0.244	0.457	0.441	0.392	0.387
Ed			0.966	0.946	0.973	0.969	0.969	0.971
Po1			0.716	0.677	0.725	0.723	0.713	0.712
Po2			0.545	0.435	0.558	0.556	0.524	0.520
LF	•		0.384	0.191	0.411	0.405	0.354	0.348
M.F			0.404	0.208	0.427	0.425	0.375	0.370
Pop			0.527	0.353	0.554	0.546	0.506	0.503
NW	•	•	0.762	0.646	0.779	0.774	0.761	0.762
U1			0.415	0.233	0.439	0.434	0.389	0.384
U2			0.703	0.580	0.725	0.718	0.701	0.701
GDP	•	•	0.536	0.346	0.559	0.556	0.516	0.513
Ineq	•	•	0.995	0.995	0.996	0.995	0.996	0.996
Prob	•	•	0.904	0.840	0.915	0.909	0.906	0.908
Time	•	•	0.530	0.352	0.553	0.550	0.514	0.511

TABLE 6.8: *Crime data. Inclusion probabilities. The median probability model contains the covariates corresponding to the gray coloured rows.*



(a) *PMD*



(b) *PMSB*

FIGURE 6.5: *Crime data. Smallest number of models needed to achieve some pre-specified probability*

6.4.3 Ozone data

Our last real example uses the ground-level ozone data analyzed by Breiman and Friedman (1985). More recently it has also been studied by Miller (2001); Casella and Moreno (2006); Liang et al. (2008) and Scott and Berger (2010) among others.

The original dataset (which is distributed as part of the library `mlbench` in R) consists in a response variable, the daily measurements of the maximum ozone concentration near Los Angeles, and 12 meteorological variables described in Table 6.9. The total number of observations were 366, but after removing observations containing missing data we are left with a sample of size $n = 203$, (for the purpose of this exercise this seems appropriate).

Over the years, authors studying this data have considered different subsets of the 12 variables (removing some variables for different reasons), sometimes including the quadratic effect and the interactions of the entertained covariates. The intersection of the different considered sets of covariates consists in 7 variables. For our analysis we remove one of this 7 covariates (`wind`) based on its small posterior probability, as well as in the small probability of its quadratic effect and of its interactions with the rest of variables (as shown in Scott and Berger, 2010; García-Donato and Martínez-Beneyto, 2010). In Table 6.9 we present a summary of the original variables indicating which ones of them are considered here, as well as the ones considered in the papers cited above.

So, our data set entertains 6 covariates, their quadratic effects and interactions, with a total of $p = 27$ covariates. The motivation for considering only 6 meteorological variables (and p=27 covariates) is to obtain a model space which is on the one hand small enough so that it can be completely enumerated and all the models visited, but at the same time is large enough to show the potential of having tractable expressions to compare posterior probabilities.

Covariate	CM	L	SB	Here	Description
m	•		•		Month: 1=January, ..., 12=December
Dm	•		•		Day of month
Dw	•		•		Day of week: 1=Monday, ..., 7=Sunday
vh	•	•	•	•	500 millibar pressure height (m) measured at Vandenberg AFB
wind	•	•	•		Wind speed (mph) at Los Angeles International Airport (LAX)
hum	•	•	•	•	Humidity (%) at LAX
temp1	•	•	•	•	Temperature (degrees F) measured at Sandburg, CA
temp2					Temperature (degrees F) measured at El Monte, CA
ibh	•	•	•	•	Inversion base height (feet) at LAX
dpg	•	•	•	•	Pressure gradient (mm Hg) from LAX to Daggett, CA
ibt		•			Inversion base temperature (degrees F) at LAX
vis	•	•	•	•	Visibility (miles) measured at LAX

TABLE 6.9: *Ozone data. Description of covariates. "CM" represents the variables considered in Casella and Moreno (2006), "L" the ones in Liang et al. (2008), "SB" the covariates in Scott and Berger (2010) and "Here" the covariates in this work.*

We compute the posterior probabilities of the $2^{27} = 134217728$ models as well as the distribution of the dimension and the inclusion probabilities. This takes about 10 minutes with 135 cores or around 10 hours with a single core depending, of course, on the cores used. However, due to the huge dimension of the model space, we can not save the posterior probabilities for all the models. Instead, our program keeps probabilities of the 10000 most probable models (ordered by R1). In this specific problem this represents a $20 - 30\%$ of the total probability. Of course, the distribution of dimensions and inclusion probabilities are computed with all the models.

In a different run of the program we kept the probabilities of the 10000 highest probability models for the R2 approach.

We present the following summaries:

- Tables 6.10 (PMD) and 6.11 (PMSB) show the 25 most probable models ordered by R1.
- Tables 6.12 (PMD) and 6.13 (PMSB) provide the inclusion probabilities for each covariate.
- Figure 6.6 presents the dimension probabilities (top) and the cumulative dimension probabilities (bottom) for PMD (left) and PMSB (right).
- Figure 6.7(a) presents the distribution of the $\log_{10}$ of the model probabilities for the 10000 most probable models for R1 through box-plots for PMD (left) and PMSB (right). Figure 6.7(b) compare the distribution of model probabilities for the 10000 most probable models when those are chosen with R1 and when they are chosen with R2, both for priors PMD (left) and PMSB (right).

The effect of $\pi_i(\boldsymbol{\beta}_i, \boldsymbol{\beta}_0, \sigma)$

As in the Crime example, R1 and TESS1 show here a similar behavior while R2 presents a different one.

Again R2 appears to be more conservative with the red line in Figure 6.6 placed clearly to the left of the green and black ones.

But the most significant differences can be found in the way of apportioning probability in the distribution of the model posterior probabilities. R2 concentrates most of the probability mass in a small number of models. In particular, the 10000 most probable models with R1 and TESS1 accumulates around a 20% of the probability while the 10000 most probable models for R2 accumulates an 80%. Figures 6.7(a) and 6.7(b) show clearly this effect. In Figure 6.7(a) where the probability of the 10000 most probable models with R1 is represented, we can see that R2 is slightly displaced toward higher probabilities. This indicates that the

probabilities of the 10000 most probable models with R1 obtain slightly higher probabilities with R2. On the other hand a longer lower tail indicates that some of them obtain a smaller probability in R2 than in R1. This is because the most probable models with R1 are not necessarily the most probable models with R2. In fact, in Figure 6.7(b) we observe that the 10000 most probable models with R2 have a larger probability that the 10000 with R1. This is reflected in the distribution of model probabilities being displaced toward higher probabilities. This effect is less evident in PMSB, although again the upper tail is larger for R2.

Again, the different behavior of R2 maybe justified by the fact that this approach was developed in an estimation scenario.

We do not find important differences between our approach R1 and TESS1. The only small difference is that TESS1 seems to be a little bit less conservative than R1 as was also observed in the two previous examples.

The conclusions about the behavior of R2 are very similar to the ones obtained in the crime example.

This example provides a clear demonstration that the closed-form for the Bayes factors can be clearly crucial for "routine and easy" implementations. Indeed, JZS and Li could not even be computed with the same routines used successfully in the previous, simple, examples.

The effect of $P(M_i)$

The PMSB approach spreads more the probability among dimensions (this was previously observed also in the Hald and Crime examples). But contrary to the behavior showed in the two previous examples, here the PMSB approach moves the probability mass towards the simpler models instead of toward the most complex ones. However, note that in Hald and Crime data the distribution derived with PMD was slightly

displaced to the right while here it is displaced towards left, indicating a preference in the data for the simpler models. In fact in Hald and Crime data the models reported with PMD as the most probable ones contained a large number of covariates in contrast with the total number of entertained ones (2 of 4, and 8 of 15 respectively) while in this example the most probable model contains only 8 out of the 27 entertained variables. Hence, as commented in the previous example, PMSB seems to be amplifying the preference in the data for simpler models (as well as it did for complex models in Hald and Crime data examples).

This preference for simpler models of PMSB is also reflected in the inclusion probabilities (see Tables 6.12 and 6.13). Note that for PMSB the median probability model contains only three covariates, while for PMD it contains 6 of them (5 for R2, which again shows a different behavior as commented previously).

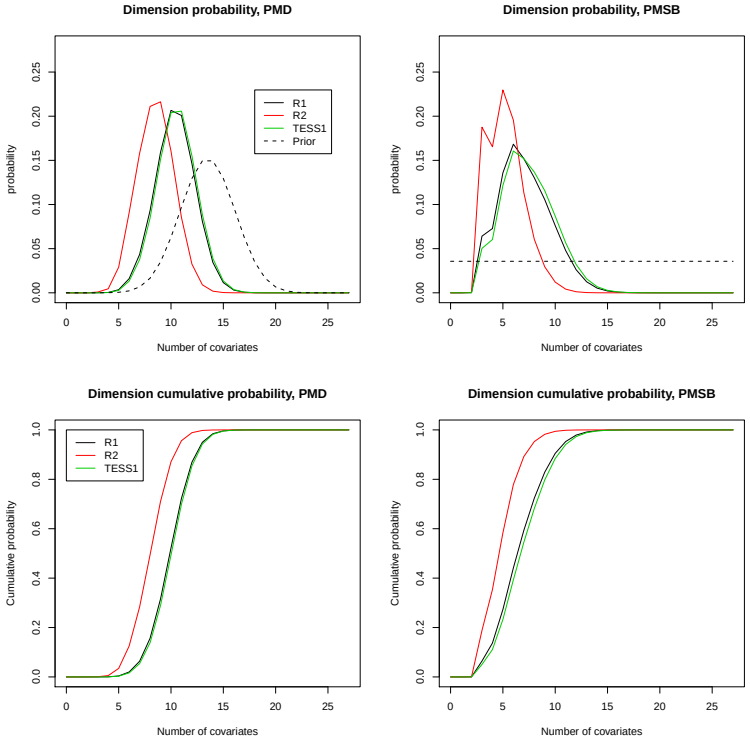


FIGURE 6.6: *Ozone Data. Posterior probabilities (top) and cumulative posterior probability (bottom) of the dimension of the true model for priors PMD (left) and PMSB (right). Induced prior over dimension is represented by a dashed line.*

Covariates	R1	R2	TESS1	dim
{hum^2, dpg.vh, vis.vh, temp1.hum, ibh.hum, dpg.temp1, vis.temp1, dpg.ibh}	0.0003	0.0010	0.0003	8
{vis, temp1^2, dpg^2, temp1.hum, ibh.hum, vis.temp1}	0.0003	0.0021	0.0002	6
{temp1^2, dpg^2, vis.vh, temp1.hum, ibh.hum, vis.temp1 }	0.0003	0.0020	0.0002	6
{dpg.vh, temp1.hum, ibh.temp1, dpg.temp1, dpg.ibh}	0.0003	0.0018	0.0002	6
{vis, hum^2, dpg.vh, temp1.hum, ibh.hum, dpg.temp1, vis.temp1, dpg.ibh}	0.0003	0.0008	0.0002	8
{hum^2, ibh^2, dpg.vh, vis.vh, temp1.hum, dpg.temp1, vis.temp1, dpg.ibh}	0.0003	0.0008	0.0003	8
{dpg, hum^2, vis.vh, temp1.hum, ibh.hum, dpg.temp1, vis.temp1, dpg.ibh}	0.0003	0.0007	0.0002	8
{temp1, hum^2, temp1.vh, dpg.vh, temp1.hum, ibh.hum, dpg.temp1, dpg.ibh}	0.0002	0.0007	0.0002	8
{temp1, hum^2, temp1.vh, dpg.vh, temp1.hum, ibh.temp1, dpg.temp1, dpg.ibh}	0.0002	0.0007	0.0002	8
{temp1, dpg, hum^2, temp1.vh, temp1.hum, ibh.hum, dpg.temp1 , dpg.ibh }	0.0002	0.0006	0.0002	8
{temp1, dpg, hum^2 , temp1.vh, temp1.hum, ibh.temp1, dpg.temp1, dpg.ibh}	0.0002	0.0006	0.0002	8
{dpg, hum^2 , temp1.hum, ibh.temp1, dpg.temp1, dpg.ibh }	0.0002	0.0014	0.0002	6
{ibh, hum^2, dpg.vh, vis.vh, temp1.hum, dpg.temp1, vis.temp1, pg.ibh}	0.0002	0.0006	0.0002	8
{hum^2, dpg.vh, vis.vh, temp1.hum, ibh.temp, dpg.temp1, vis.temp1, dpg.ibh}	0.0002	0.0006	0.0002	8
{vis, hum^2, ibh^2, dpg.vh, temp1.hum, dpg.temp1, vis.temp1, dpg.ibh}	0.0002	0.0006	0.0002	8
{hum^2, ibh.vh, dpg.vh, vis.vh, temp1.hum, dpg.temp1, vis.temp1, dpg.ibh}	0.0002	0.0006	0.0002	8
{dpg, vis, hum^2, temp1.hum, ibh.hum, dpg.temp1, vis.temp1, dpg.ibh }	0.0002	0.0006	0.0002	8
{dpg, hum^2, ibh^2, vis.vh, temp1.hum, dpg.temp1, vis.temp1, dpg.ibh}	0.0002	0.0005	0.0002	8
{vis, temp1^2, temp1.hum, ibh.hum, vis.temp1}	0.0002	0.0017	0.0001	5
{hum^2, dpg^2, vis.vh, temp1.hum, ibh.hum, dpg.hum, dpg.temp1, vis.temp1}	0.0002	0.0005	0.0001	8
{hum^2, dpg.vh, temp1.hum, ibh.temp1, dpg.temp1, vis.temp1, dpg.ibh, vis.ibh}	0.0002	0.0005	0.0002	8
{vis, temp1^2, dpg2, hum.vh, ibh.hum, vis.temp1 }	0.0002	0.0011	0.0001	6
{temp1^2, dpg^2, hum.vh, vis.vh, ibh.hum, vis.temp1}	0.0002	0.0011	0.0001	6
{vis, hum^2, dpg^2, temp1.hum, ibh.hum, dpg.hum, dpg.temp1, vis.temp1 }	0.0002	0.0005	0.0001	8
{vis, hum^2, dpg.vh, temp1.hum, ibh.temp1, dpg.temp1, vis.temp1, dpg.ibh}	0.0002	0.0005	0.0002	8

TABLE 6.10: *Ozone data. The 25 highest probability models according to R1 and their probabilities under the other approaches for PMD.*

Covariates	R1	R2	TESS1	dim
{hum^2, temp1.hum, ibh.hum }	0.0258	0.0766	0.0199	3
{temp1^2, temp1.hum, ibh.hum }	0.0147	0.0432	0.0107	3
{temp1, temp1^2, temp1.hum, ibh.hum}	0.0082	0.0195	0.0065	4
{vis, temp1^2, temp1.hum, ibh.hum, vis.temp1}	0.0073	0.0132	0.0058	5
{temp1^2, vis.vh, temp1.hum, ibh.hum, vis.temp1}	0.0066	0.0119	0.0053	5
{hum^2, dpg^2, temp1.hum, ibh.hum}	0.0063	0.0150	0.0047	4
{temp1^2,hum.vh, ibh.hum }	0.0055	0.0160	0.0045	3
{temp1^2, dpg^2, temp1.hum, ibh.hum}	0.0048	0.0112	0.0034	4
{vis, temp1^2, hum.vh, ibh.hum, vis.temp1}	0.0043	0.0076	0.0036	5
{temp1, temp1^2, dpg^2, temp1.hum, ibh.hum}	0.0041	0.0073	0.0032	5
{temp1^2, hum.vh, vis.vh, ibh.hum, vis.temp1}	0.0039	0.0070	0.0034	5
{hum, temp1^2, ibh.hum }	0.0035	0.0100	0.0028	3
{vis, temp1^2, dpg^2, temp1.hum, ibh.hum, vis.temp1}	0.0035	0.0045	0.0027	6
{temp1^2, dpg^2, vis.vh, temp1.hum, ibh.hum, vis.temp1}	0.0034	0.0044	0.0027	6
{vh, hum, temp1^2, hum.vh, ibh.hum}	0.0032	0.0057	0.0030	5
{hum^2, temp1.hum, ibh.temp1}	0.0030	0.0088	0.0025	3
{hum^2, dpg.vh, temp1.hum, ibh.temp1, dpg.temp1, dpg.ibh }	0.0030	0.0039	0.0028	6
{hum, vis, temp1^2, ibh.hum vis.temp1}	0.0028	0.0050	0.0024	5
{hum, temp1^2, vis.vh, ibh.hum, vis.temp1}	0.0028	0.0049	0.0024	5
{temp1^2, temp1.vh, temp1.hum, ibh.hum}	0.0027	0.0063	0.0022	4
{vh, hum, temp1^2, hum.vh, ibh.temp1}	0.0025	0.0045	0.0025	5
{dpg, hum^2, temp1.hum, ibh.temp1, dpg.temp1, dpg.ibh}	0.0024	0.0031	0.0023	6
{hum^2, temp1.hum, ibh.hum, vis.temp1}	0.0023	0.0054	0.0018	4
{hum, temp1.hum, ibh.hum}	0.0022	0.0063	0.0018	3
{hum, vh^2, temp1^2, hum.vh, ibh.hum}	0.0021	0.0037	0.0020	5

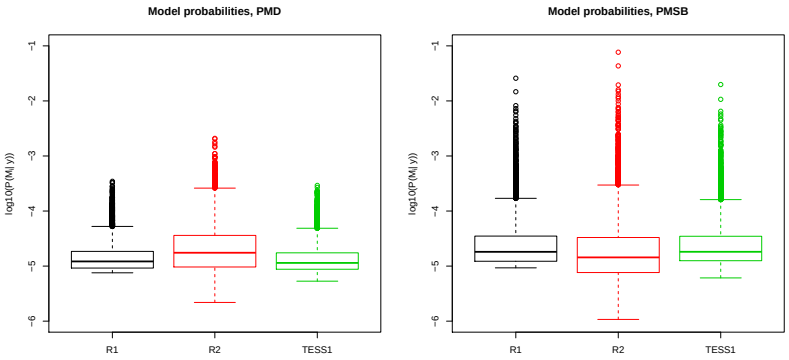
TABLE 6.11: *Ozone data. The 25 highest probability models according to R1 and their probabilities under the other approaches for PMSB.*

Covariate	R1	R2	TESS1
vh	0.306	0.230	0.308
hum	0.414	0.341	0.428
temp1	0.354	0.281	0.358
ibh	0.266	0.205	0.288
dpg	0.450	0.319	0.458
vis	0.339	0.284	0.340
vh^2	0.307	0.226	0.309
hum^2	0.615	0.533	0.621
$temp1^2$	0.455	0.467	0.454
ibh^2	0.207	0.139	0.224
dpg^2	0.617	0.523	0.614
vis^2	0.164	0.099	0.165
hum.vh	0.413	0.349	0.427
temp1.vh	0.353	0.280	0.358
ibh.vh	0.267	0.208	0.287
dpg.vh	0.460	0.336	0.468
vis.vh	0.346	0.293	0.348
temp1.hum	0.749	0.710	0.748
ibh.hum	0.576	0.612	0.564
dpg.hum	0.227	0.137	0.231
vis.hum	0.156	0.091	0.158
ibh.temp1	0.314	0.286	0.319
dpg.temp1	0.610	0.461	0.620
vis.temp1	0.589	0.539	0.584
dpg.ibh	0.491	0.365	0.501
vis.ibh	0.214	0.140	0.215
vis.dpg	0.136	0.079	0.139

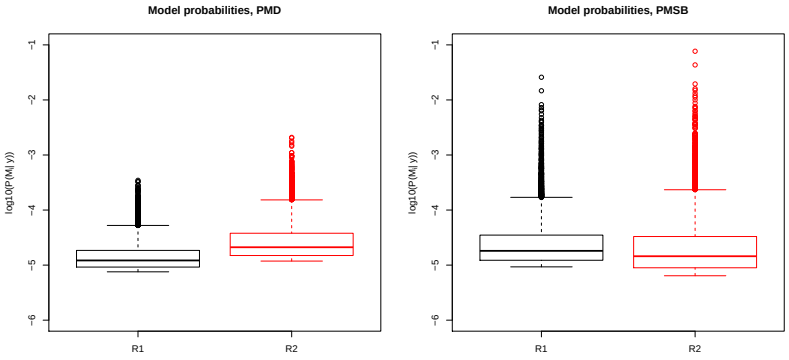
TABLE 6.12: *Ozone data. Inclusion probabilities with PMD. The median probability model contains the covariates corresponding to the gray coloured rows.*

Covariate	R1	R2	TESS1
vh	0.177	0.100	0.189
hum	0.300	0.225	0.323
temp1	0.229	0.156	0.239
ibh	0.164	0.099	0.192
dpg	0.206	0.083	0.226
vis	0.226	0.147	0.232
vh^2	0.171	0.092	0.182
hum^2	0.454	0.392	0.468
$temp1^2$	0.505	0.527	0.497
ibh^2	0.105	0.053	0.123
dpg^2	0.417	0.278	0.424
vis^2	0.075	0.033	0.079
hum.vh	0.314	0.244	0.335
temp1.vh	0.219	0.133	0.231
ibh.vh	0.167	0.102	0.194
dpg.vh	0.217	0.091	0.238
vis.vh	0.228	0.145	0.236
temp1.hum	0.704	0.723	0.697
ibh.hum	0.672	0.749	0.646
dpg.hum	0.097	0.038	0.104
vis.hum	0.072	0.033	0.076
ibh.temp1	0.263	0.215	0.280
dpg.temp1	0.298	0.135	0.327
vis.temp1	0.433	0.303	0.437
dpg.ibh	0.236	0.105	0.259
vis.ibh	0.105	0.050	0.110
vis.dpg	0.060	0.025	0.065

TABLE 6.13: *Ozone data. Inclusion probabilities with PMSB. The median probability model contains the covariates corresponding to the gray coloured rows.*



(a) *The 10000 most probable models ordered by R1*



(b) *The 10000 most probable models each approach with its own order*

FIGURE 6.7: *Ozone data. Distribution of model posterior probabilities for priors PMD (left) and PMSB (right).*

6.4.4 Simulated data with correlated covariates

With this simulated example we pretend to show the potential of having tractable expressions to compute posterior probabilities (that is, a Conventional Robust prior with $b = 1$) in a problem of considerable size. Again, as in the Ozone example, the routines used successfully for computation of Li and JZS in the two first examples do not behave properly in this example and so we just center our attention in R1, R2 and TESS1.

We choose a simulated data set with $n = 60$ observations and 30 explanatory variables, X_1 through X_{30}, apart from the intercept. This problem has more than one thousand millions of possible models. We simulated data as in Kuo and Mallick (1998); specifically, for $j = 1, \ldots, 30$, the *correlated* covariates are simulated as

$$\boldsymbol{X}_j = \boldsymbol{X}_j^* + \boldsymbol{Z},$$

where, independently

$$\boldsymbol{X}_1^*, \boldsymbol{X}_2^*, \ldots, \boldsymbol{X}_{30}^*, \boldsymbol{Z} \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{I}_n).$$

The vector of dependent observations is simulated as $\boldsymbol{Y} = \alpha \boldsymbol{1}_n + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\boldsymbol{0}, 4I_n)$ and

$$\boldsymbol{\beta}^t = (0, \overset{10)}{\ldots}, 0, 1, \overset{10)}{\ldots}, 1, 2, \overset{10)}{\ldots}, 2).$$

The correlations between the regressors in our simulated data varied in [0.32,0.73].

We keep the probabilities of 10000 models which concentrate a 99% of the total probability, this indicates a high degree of concentration in surprisingly few models, something that is not usual in real examples as we saw in the Ozone data.

The computations have been done approximately in 16 hours using 100 cores (note that this time obviously depends on the cores used).

We also present the posterior probabilities and dimension probability in the following summaries.

- Tables 6.14 (PMD) and 6.15 (PMSB) shows the 25 most probable models ordered by R1.
- Tables 6.16 (PMD) and 6.17 (PMSB) provide the inclusion probabilities for each covariate.
- Figure 6.8 presents the dimension probabilities (top) and the cumulative dimension probabilities (bottom) for PMD (left) and PMSB (right).
- Figure 6.9 presents the distribution of the $\log_{10}$ of the probabilities of the 10000 most probable models through box-plots for PMD (left) and PMSB (right).

The effect of $\pi_i(\boldsymbol{\beta}_i, \boldsymbol{\beta}_0, \sigma)$

This is a very special example in which the most probable models seem to be pretty clear and so, the results are very robust to the entertained approaches. In fact, as commented before, the 10000 most probable models with R1 accumulate more than a 99% of the probability with any of the entertained approaches. We do not appreciate any significative differences, in neither the posterior probabilities (see Tables 6.14 and 6.15, and Figure 6.9) neither in the inclusion probabilities (see Tables 6.16 and 6.17) nor dimension probabilities (see Figure 6.8).

The covariates $\boldsymbol{X}_{21}, \ldots, \boldsymbol{X}_{30}$ (whose regressor coefficient's value in the simulation is 2) have posterior inclusion probabilities of 1. On the other hand the covariates $\boldsymbol{X}_{11}, \ldots, \boldsymbol{X}_{20}$ (whose regressor coefficient's value is 1)

have lower probabilities, with just a half of them achieving a probability higher than 0.5. The rest of variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{10}$ (whose regressor coefficient's value is 0) have really small probabilities except for $\boldsymbol{X}_7$ which reaches a posterior inclusion probability over 0.5, thus being one of the variables in the median probability model.

The median probability model is finally formed by 16 variables from which 15 of them are actually in the model. This model is the second most probable model with every entertained approach.

The effect of $P(M_i)$

The differences among the approaches to assess the probabilities over the model space are also inexistent. Only in Figure 6.8 we appreciate that PMSB is spreading a little bit more the probability among dimensions, but it is hardly appreciable.

The main goal of this example is to show that our approach can deal with such a large problem directly without the necessity for search methods. And this is due to the closed-form expressions of Bayes factors.

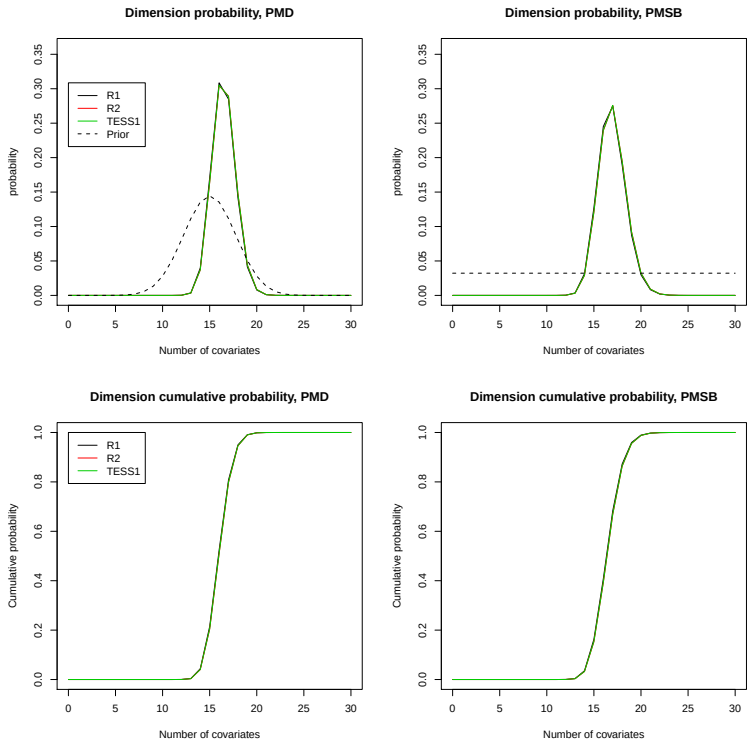


FIGURE 6.8: *Simulated data. Posterior probabilities (top) and cumulative posterior probability (bottom) of the dimension of the true model for priors PMD (left) and PMSB (right). Induced prior over dimension is represented by a dashed line.*

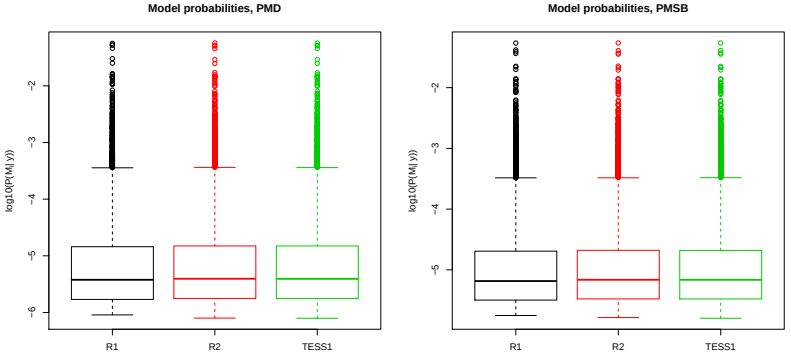


FIGURE 6.9: *Simulated data. Distribution of model posterior probabilities for priors PMD (left) and PMSB (right) for the 10000 most probable models.*

Covariates	R1	R2	TESS1	dim
$\{X_7, X_{11}, X_{13}, X_{16}, X_{17}, X_{18}, X_{19}, X_{21}, \ldots, X_{30}\}$	0.056	0.057	0.057	17
$\{X_7, X_{11}, X_{13}, X_{16}, X_{17}, X_{18}, X_{21}, \ldots, X_{30}\}$	0.052	0.052	0.052	16
$\{X_7, X_{11}, X_{13}, X_{16}, X_{17}, X_{21}, \ldots, X_{30}\}$	0.052	0.050	0.051	15
$\{X_7, X_{11}, X_{13}, X_{16}, X_{17}, X_{19}, X_{21}, \ldots, X_{30}\}$	0.046	0.045	0.045	16
$\{X_{11}, X_{13}, X_{16}, X_{17}, X_{18}, X_{21}, \ldots, X_{30}\}$	0.030	0.029	0.029	15
$\{X_{11}, X_{13}, X_{16}, X_{17}, X_{18}, X_{19}, X_{21}, \ldots, X_{30}\}$	0.025	0.025	0.025	16
$\{X_7, X_{11}, X_{13}, X_{15}, X_{16}, X_{17}, X_{18}, X_{19}, X_{21}, \ldots, X_{30}\}$	0.016	0.017	0.017	18
$\{X_{11}, X_{13}, X_{15}, X_{16}, X_{17}, X_{18}, X_{21}, \ldots, X_{30}\}$	0.016	0.016	0.016	16
$\{X_7, X_{11}, X_{13}, X_{16}, X_{17}, X_{20}, X_{21}, \ldots, X_{30}\}$	0.015	0.015	0.015	16
$\{X_7, X_{13}, X_{16}, X_{17}, X_{18}, X_{21}, \ldots, X_{30}\}$	0.015	0.014	0.014	15
$\{X_7, X_{11}, X_{13}, X_{16}, X_{17}, X_{18}, X_{20}, X_{21}, \ldots, X_{30}$	0.014	0.015	0.015	17
$\{X_7, X_{11}, X_{13}, X_{15}, X_{16}, X_{17}, X_{18}, X_{21}, \ldots, X_{30}\}$	0.014	0.015	0.015	17
$\{X_{11}, X_{13}, X_{15}, X_{16}, X_{17}, X_{18}, X_{19}, X_{21}, \ldots, X_{30}\}$	0.014	0.014	0.014	17
$\{X_7, X_{11}, X_{13}, X_{15}, X_{16}, X_{17}, X_{21}, \ldots, X_{30}\}$	0.013	0.013	0.013	16
$\{X_7, X_{11}, X_{13}, X_{15}, X_{16}, X_{17}, X_{19}, X_{21}, \ldots, X_{30}\}$	0.012	0.012	0.012	17
$\{X_{11}, X_{12}, X_{13}, X_{16}, X_{17}, X_{18}, X_{19}, X_{21}, \ldots, X_{30}\}$	0.011	0.011	0.011	17
$\{X_{11}, X_{13}, X_{16}, X_{17}, X_{18}, X_{20}, X_{21}, \ldots, X_{30}\}$	0.011	0.011	0.011	16
$\{X_{11}, X_{13}, X_{16}, X_{17}, X_{21}, \ldots, X_{30}\}$	0.011	0.010	0.010	14
$\{X_7, X_{13}, X_{16}, X_{17}, X_{21}, \ldots, X_{30}\}$	0.011	0.010	0.010	14
$\{X_{11}, X_{15}, X_{16}, X_{17}, X_{18}, X_{21}, \ldots, X_{30}\}$	0.008	0.008	0.008	15
$\{X_{11}, X_{13}, X_{16}, X_{17}, X_{19}, X_{21}, \ldots, X_{30}\}$	0.007	0.007	0.007	15
$\{X_{11}, X_{12}, X_{13}, X_{15}, X_{16}, X_{17}, X_{18}, X_{19}, X_{21}, \ldots, X_{30}\}$	0.007	0.007	0.007	18
$\{X_7, X_{11}, X_{12}, X_{13}, X_{16}, X_{17}, X_{18}, X_{19}, X_{21}, \ldots, X_{30}\}$	0.007	0.007	0.007	18
$\{X_2, X_7, X_{11}, X_{13}, X_{16}, X_{17}, X_{21}, \ldots, X_{30}\}$	0.006	0.006	0.006	16
$\{X_4, X_7, X_{11}, X_{13}, X_{16}, X_{17}, X_{18}, X_{19}, X_{21}, \ldots, X_{30}\}$	0.006	0.006	0.006	18

TABLE 6.14: *Simulated data. The 25 highest probability models according to R1 and their probabilities under the other approaches for PMD.*

Covariates	R1	R2	TESS1	dim
$\{X_7, X_{11}, X_{13}, X_{16}, X_{17}, X_{18}, X_{19}, X_{21}, \ldots, X_{30}\}$	0.054	0.054	0.054	17
$\{X_7, X_{11}, X_{13}, X_{16}, X_{17}, X_{18}, X_{21}, \ldots, X_{30}\}$	0.042	0.041	0.041	16
$\{X_7, X_{11}, X_{13}, X_{16}, X_{17}, X_{21}, \ldots, X_{30}\}$	0.039	0.037	0.037	15
$\{X_7, X_{11}, X_{13}, X_{16}, X_{17}, X_{19}, X_{21}, \ldots, X_{30}\}$	0.036	0.036	0.036	16
$\{X_{11}, X_{13}, X_{16}, X_{17}, X_{18}, X_{21}, \ldots, X_{30}\}$	0.022	0.021	0.021	15
$\{X_7, X_{11}, X_{13}, X_{15}, X_{16}, X_{17}, X_{18}, X_{19}, X_{21}, \ldots, X_{30}\}$	0.022	0.023	0.022	18
$\{X_{11}, X_{13}, X_{16}, X_{17}, X_{18}, X_{19}, X_{21}, \ldots, X_{30}\}$	0.020	0.019	0.019	16
$\{X_7, X_{11}, X_{13}, X_{16}, X_{17}, X_{18}, X_{21}, \ldots, X_{30}\}$	0.014	0.014	0.014	17
$\{X_7, X_{11}, X_{13}, X_{15}, X_{16}, X_{17}, X_{18}, X_{21}, \ldots, X_{30}\}$	0.014	0.014	0.014	17
$\{X_{11}, X_{13}, X_{15}, X_{16}, X_{17}, X_{18}, X_{19}, X_{21}, \ldots, X_{30}\}$	0.014	0.014	0.014	17
$\{X_{11}, X_{13}, X_{15}, X_{16}, X_{17}, X_{18}, X_{21}, \ldots, X_{30}\}$	0.013	0.012	0.012	16
$\{X_7, X_{11}, X_{13}, X_{16}, X_{17}, X_{20}, X_{21}, \ldots, X_{30}\}$	0.012	0.011	0.011	16
$\{X_7, X_{11}, X_{13}, X_{15}, X_{16}, X_{17}, X_{19}, X_{21}, \ldots, X_{30}\}$	0.012	0.012	0.012	17
$\{X_7, X_{13}, X_{16}, X_{17}, X_{18}, X_{21}, \ldots, X_{30}\}$	0.011	0.010	0.010	15
$\{X_{11}, X_{12}, X_{13}, X_{16}, X_{17}, X_{18}, X_{19}, X_{21}, \ldots, X_{30}\}$	0.011	0.011	0.011	17
$\{X_7, X_{11}, X_{13}, X_{15}, X_{16}, X_{17}, X_{21}, \ldots, X_{30}\}$	0.011	0.010	0.010	16
$\{X_{11}, X_{12}, X_{13}, X_{15}, X_{16}, X_{17}, X_{18}, X_{19}, X_{21}, \ldots, X_{30}\}$	0.009	0.010	0.010	18
$\{X_7, X_{11}, X_{12}, X_{13}, X_{16}, X_{17}, X_{18}, X_{19}, X_{21}, \ldots, X_{30}\}$	0.009	0.009	0.009	18
$\{X_{11}, X_{13}, X_{16}, X_{17}, X_{18}, X_{20}, X_{21}, \ldots, X_{30}\}$	0.009	0.009	0.009	16
$\{X_{11}, X_{13}, X_{16}, X_{17}, X_{21}, \ldots, X_{30}\}$	0.009	0.008	0.008	14
$\{X_7, X_{13}, X_{16}, X_{17}, X_{21}, \ldots, X_{30}\}$	0.008	0.008	0.008	14
$\{X_4, X_7, X_{11}, X_{13}, X_{16}, X_{17}, X_{18}, X_{19}, X_{21}, \ldots, X_{30}\}$	0.008	0.009	0.009	18
$\{X_{11}, X_{15}, X_{16}, X_{17}, X_{18}, X_{21}, \ldots, X_{30}\}$	0.006	0.006	0.006	15
$\{X_7, X_{11}, X_{13}, X_{16}, X_{17}, X_{18}, X_{19}, X_{20}, X_{21}, \ldots, X_{30}\}$	0.006	0.006	0.006	18
$\{X_{11}, X_{13}, X_{16}, X_{17}, X_{19}, X_{21}, \ldots, X_{30}\}$	0.006	0.005	0.005	15

TABLE 6.15: *Simulated data. The 25 highest probability models according to R1 and their probabilities under the other approaches for PMSB.*

Covariate	R1	R2	TESS1
X_1	0.054	0.056	0.055
X_2	0.067	0.068	0.068
X_3	0.036	0.036	0.036
X_4	0.070	0.072	0.072
X_5	0.067	0.068	0.068
X_6	0.036	0.037	0.037
X_7	0.651	0.654	0.652
X_8	0.045	0.046	0.046
X_9	0.039	0.040	0.040
X_{10}	0.042	0.043	0.043
X_{11}	0.896	0.899	0.899
X_{12}	0.125	0.129	0.129
X_{13}	0.923	0.926	0.926
X_{14}	0.038	0.039	0.039
X_{15}	0.258	0.263	0.262
X_{16}	1.000	1.000	1.000
X_{17}	0.988	0.989	0.989
X_{18}	0.584	0.589	0.588
X_{19}	0.403	0.411	0.410
X_{20}	0.145	0.148	0.147
$X_{21} \dots X_{30}$	1.000	1.000	1.000

TABLE 6.16: *Simulated data. Inclusion probabilities with PMD. The median probability model contains the covariates corresponding to the gray coloured rows.*

Covariate	R1	R2	TESS1
X_1	0.068	0.070	0.070
X_2	0.081	0.082	0.082
X_3	0.047	0.049	0.049
X_4	0.095	0.098	0.097
X_5	0.081	0.083	0.083
X_6	0.049	0.051	0.051
X_7	0.673	0.677	0.675
X_8	0.059	0.061	0.061
X_9	0.052	0.054	0.054
X_{10}	0.058	0.060	0.060
X_{11}	0.912	0.915	0.916
X_{12}	0.165	0.170	0.171
X_{13}	0.936	0.939	0.939
X_{14}	0.051	0.053	0.053
X_{15}	0.304	0.310	0.309
X_{16}	1.000	1.000	1.000
X_{17}	0.990	0.991	0.991
X_{18}	0.628	0.634	0.633
X_{19}	0.472	0.481	0.480
X_{20}	0.167	0.170	0.169
$X_{21} \dots X_{30}$	1.000	1.000	1.000

TABLE 6.17: *Simulated data. Inclusion probabilities with PMSB. The median probability model contains the covariates corresponding to the gray coloured rows.*

Chapter 7

Conclusions and future work

7.1 Thesis summary and conclusions

Many of today's scientific problems require identifying which variables from an entertained set are involved in a specific phenomenon. For instance, many public health studies require the identification of the causes of a certain disease.

This problem is referred to as variable selection and can be seen as a particular case of model selection. In this specific model selection problem each model contains a certain subset of the entertained covariates. This means a total of 2^p possible models for a problem with p potential covariates. The variable selection problem is difficult to address both from a theoretical and from a computational point of view.

In particular, in this work the problem of variable selection is addressed in the framework of linear regression, but it also appears in many other scenarios such as generalized linear models and non-parametric function estimation (see George, 2000, and references therein).

Our preferred Bayesian way for solving model selection, and, in particular variable selection, is to base the choice on the posterior probabilities of the competing models. These posterior probabilities can be expressed in terms of the prior probabilities of the models and the 2^p Bayes factors.

For the assignment of prior probabilities over the model space we entertain and compare some approaches, and state our preferred choice. However, this is not the main topic of this thesis.

Posterior probabilities require the computation of 2^p Bayes factors in favor of each model M_i and against a base model M_d for $i = 0, \dots 2^p - 1$. Our choice for M_d is the simplest model explaining the data, which as usual we denote M_0; M_0 is nested in every model M_i. The computation of those 2^p Bayes factors require the elicitation of priors for the corresponding parameters under each model. Subjective elicitation of priors assessed by experts knowledge in this scenario is practically impossible due to the very large number of models, and model-specific parameters. The idea is hence, to adopt an objective point of view (see Berger, 2006, and references therein) but the objective elicitation of priors in model selection has to be done carefully due to the high sensitivity of Bayes factors to the choice of objective priors. In fact, the usual non-informative (usually improper) priors, which work well in estimation problems do not always produce sensible results in model selection (see Berger and Pericchi, 2001, and references therein) often resulting in indeterminate Bayes factors.

The large number of models also posses a computational challenge since the numerical computation of the 2^p Bayes factors is required. When p is so large that the models space can not even be enumerated (for all practical purposes), many authors (see, for example, George and McCulloch, 1993; Carlin and Chib, 1995; George and McCulloch, 1997; Miller, 2001; Robert and Casella, 2004; Berger and Molina, 2005, and references therein) propose methods for searching over the model space trying to find models with high posterior probabilities. But usually Bayes factors

are hard to compute, so that even this solution can be computationally very demanding. This difficulty can be largely alleviated if simple expressions for Bayes factors are available.

The aim of this thesis is to propose a novel, suitable and rigorously justified prior distribution for the variable selection problem. In particular, we look for a prior distribution which achieves many desirable properties and provides simple expressions for the Bayes factors.

We follow the Conventional approach of Jeffreys (1961), who outlined a number of desiderata for a good objective prior distribution to have in the variable selection problem.

Following Jeffrey's Conventional scheme, the prior distribution under each model M_i is assessed in two steps. The first one consists in assigning a proper prior distribution for those parameters in M_i that are not in M_0 conditionally on those parameters in both models (in particular, as M_0 is nested in M_i this means conditionally on the parameters in M_0). The second step consists in assigning a non-informative prior for the parameters in M_0.

For assessing the conditional prior in the first step we found some interesting ideas in the work of Strawderman (1973, 1971) and Berger (1976, 1980, 1985). Their work, originally developed in a context of robust and minimax normal mean estimation, is extended and adapted here to solve the variable selection problem.

For the prior distribution of the parameters in M_0 (occurring in all models) we consider a prior which makes the problem invariant. In this case, it happens to coincide with the reference prior or independent Jeffreys' prior which is the usual choice in the literature. Hence, the usual choice gets fully justified.

The result is a joint prior distribution in the parametric space which, following Berger (1985) we call Conventional Robust prior. This prior

distribution is defined up to some parameters that can be tuned to achieve a number properties. Our specific proposal for these parameters is based in certain optimality properties of the resulting procedure.

The theoretical highlights of this distribution for variable selection are

- *The choice of the prior is justified from a theoretical point of view.* Jeffreys (1961)'s Conventional approach scheme for the elicitation of prior distributions was based on the orthogonal parameterization of the model. Our choice is instead completely justified by a sensible choice of the scale matrix and the use of invariance ideas in Berger et al. (1998). This fully theoretical justification makes the orthogonal parameterization no longer required.

- *It produces well defined Bayes factors with good consistency properties from many points of view.* The resulting Bayes factors are well defined in the sense that they are not indeterminate as is usually the case when using objective (improper) priors. This indeterminacy is avoided here through invariance arguments. On the other hand, the consistency properties of the resulting Bayes factor, closely related to the shape of the prior's tails, makes this choice a suitable prior for variable selection.

- *It agrees with the predictive matching idea.* In particular, our prior distribution accords with our preferred and weaker interpretation of predictive matching for this problem. Specifically, we require that, if the information in the sample is barely enough for estimating the specific parameters of *any* model entertaining k extra covariates (i.e. $n = k_0 + k$), then this information should not be enough to discriminate among those models.

In addition, our approach produces simple, tractable, closed-form expressions for Bayes factors considerably simplifying computation.

The outline of the thesis is as follows

Chapter 1 introduces the general problem of model selection and presents posterior probabilities based on Bayes factors as our preferred tool for solving it. Then, in Section 1.3, we introduce the objective Bayesian point of view for model selection, concluding with a brief review of some other approaches to model selection.

Chapter 2 is mainly devoted to the study of the variable selection problem and the objective Bayesian approaches to this problem, including the Conventional approach of Jeffreys (1961).

Chapters 3, 4 and 5 present our novel approach. First in Chapter 3 we review the work of Strawderman (1973, 1971) and Berger (1976, 1980, 1985) and, based on those, we define (up to three adjustable parameters) our proposal for the prior distribution in the parametric space. We theoretically justify this choice and study many of its good properties for variable selection. In Chapter 4 we define Conventional Robust Bayes factors using our proposed distribution. We show that they can be computed in closed-form and study the consistency properties that they achieve. Then, in Chapter 5 we complete the choice of the prior by assessing the adjustable parameters to endow the methodology with even better properties for variable selection.

Finally in Chapter 6 we apply this methodology to some real and simulated examples and compare the solution obtained with our proposed approach to the ones provided by other Conventional approaches in literature.

7.2 Suggestions for future work

In this thesis we address variable selection in a framework of linear regression. It might also be interesting to consider the suitability of this methodology for other scenarios as, for instance, for generalized linear models.

As we observe in the examples of Chapter 6 the use of TESS (Berger et al., 2010b) seems to produce slightly less conservative results than the use of the sample size n. This issue may be something worthy of further research. Moreover, understanding the implications of using TESS in the properties achieved by the resulting methodology is an important issue. For instance, to achieve model consistency when using n we need that $\rho_i(b+n)$ tends to infinity with n. But, when using TESS what we need is $\rho_i(b+n_i^T) \to \infty$ with the number of observations n, for $i = 1, \ldots, 2^p - 1$ (recall that n_i^T is the corresponding TESS under each model M_i as defined in Appendix B). Establishing the conditions under which this property holds is something that also deserves further work. At the same time, a good choice for ρ_i when using TESS needs to be addressed.

Appendix A

Usual Distributions

Bernoulli distribution

A random variable X has a Bernoulli distribution $X \sim \mathcal{B}(p)$ with parameter $p \in [0, 1]$ if its probability function is:

$$f(x \mid p) = p^x \, (1-p)^{1-x} \quad \text{for} \quad x \in \{0, 1\}.$$

The mean and variance are:

$$\begin{aligned} &\mathrm{E}[X] = p, \\ &\mathrm{Var}[X] = p \, (1-p). \end{aligned}$$

Binomial distribution

A random variable X has a binomial distribution $X \sim \mathcal{B}i(N, p)$ with parameters $p \in [0, 1]$ and N (a finite integer) $N \in \mathbb{N}$ if its probability

function is

$$f(x \mid N,\, p) = \binom{N}{x} p^x \,(1-p)^{N-x} \quad \text{for} \quad x \in \{0, 1, \ldots, N\}.$$

The mean and variance are:

$$\begin{aligned} \mathrm{E}[X] &= Np, \\ \mathrm{Var}[X] &= Np\,(1-p). \end{aligned}$$

Beta distribution

A random variable X has a beta distribution $X \sim \mathcal{B}e(\alpha,\, \beta)$ with parameters $\alpha > 0$ and $\beta > 0$ if its probability density function is:

$$f(x \mid \alpha,\, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\, x^{\alpha-1}\,(1-x)^{\beta-1} \quad \text{for} \quad x \in (0, 1).$$

The mean and variance are:

$$\begin{aligned} \mathrm{E}[X] &= \frac{\alpha}{\alpha+\beta}, \\ \mathrm{Var}[X] &= \frac{\alpha\,\beta}{(\alpha+\beta)^2\,(\alpha+\beta+1)}. \end{aligned}$$

Doubly non-central beta distribution

A random variable X has a doubly non-central beta distribution $X \sim \mathcal{B}e(\alpha,\, \beta;\, \lambda_1,\, \lambda_2)$ with parameters $\alpha > 0$, $\beta > 0$, $\lambda_1 > 0$ and $\lambda > 0$ if it can be expressed in terms of independent non-central chi-squared distributions as:

$$X = \frac{\chi^2(2\alpha;\, \lambda_1)}{\chi^2(2\alpha;\, \lambda_1) + \chi^2(2\beta;\, \lambda_2)};$$

where $\chi^2(k;\, \lambda)$ follow a non-central chi-square distribution with k degrees of freedom and non-centrality parameter λ. For further details about this distribution see Chattamvelli (1995).

Gamma distribution

A random variable X has a gamma distribution $X \sim Ga(\alpha,\, \beta)$ with parameters $\alpha > 0$ and $\beta > 0$ if its probability density function is:

$$f(x \mid \alpha,\, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\, x^{\alpha-1}\, e^{-\beta x} \quad \text{for} \quad x > 0.$$

The mean and variance are:

$$\begin{aligned} \mathrm{E}[X] &= \frac{\alpha}{\beta}, \\ \mathrm{Var}[X] &= \frac{\alpha}{\beta^2}. \end{aligned}$$

Inverse gamma distribution

A random variable X has a inverse gamma distribution $X \sim IGa(\alpha,\, \beta)$ with parameters $\alpha > 0$ and $\beta > 0$ if its probability density function is:

$$f(x \mid \alpha,\, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\, x^{-(\alpha+1)}\, e^{-\frac{\beta}{x}} \quad \text{for} \quad x > 0.$$

The mean and variance are:

$$\begin{aligned} \mathrm{E}[X] &= \frac{\beta}{\alpha - 1} \quad \text{if} \quad \alpha > 1, \\ \mathrm{Var}[X] &= \frac{\beta^2}{(\alpha-1)^2\,(\alpha-2)} \quad \text{if} \quad \alpha > 2. \end{aligned}$$

Multivariate normal distribution

A k-dimensional random vector $\boldsymbol{X}$ has a multivariate normal distribution $\boldsymbol{X} \sim \mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with location parameter $\boldsymbol{\mu} \in \mathbb{R}^k$ and scale matrix $\boldsymbol{\Sigma}$, for $\boldsymbol{x} \in \mathbb{R}^k$, if its probability density function is:

$$f(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}|^{-1/2}(2\pi)^{-k/2} exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2}\right).$$

The mean an variance are:

$$\begin{aligned} \mathrm{E}[\boldsymbol{X}] &= \boldsymbol{\mu}, \\ \mathrm{Var}[\boldsymbol{X}] &= \boldsymbol{\Sigma}. \end{aligned}$$

Multivariate Student's t-distribution

A k-dimensional random vector $\boldsymbol{X}$ has a Student's t-distribution $\boldsymbol{X} \sim \mathcal{S}t_k(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ with location parameter $\boldsymbol{\mu} \in \mathbb{R}^k$, scale matrix $\boldsymbol{\Sigma}$ and ν degrees of freedom, for $\boldsymbol{x} \in \mathbb{R}^k$, if its probability density function is:

$$f(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\Gamma(\frac{\nu+k}{2})}{\Gamma(\frac{\nu}{2})\,(\nu\,\pi)^{k/2}}\,|\boldsymbol{\Sigma}|^{-1/2}\left[1 + (\boldsymbol{x}-\boldsymbol{\mu})^t\frac{\boldsymbol{\Sigma}^{-1}}{n}\,(\boldsymbol{x}-\boldsymbol{\mu})\right]^{-\frac{\nu+k}{2}}.$$

The mean and variance are:

$$\begin{aligned} \mathrm{E}[\boldsymbol{X}] &= \boldsymbol{\mu} \;\text{ if }\; \nu \geq 2, \\ \mathrm{Var}[\boldsymbol{X}] &= \boldsymbol{\Sigma} \;\text{ if }\; \nu \geq 3. \end{aligned}$$

Multivariate Cauchy

A multivariate Student's t-distribution with 1 degree of freedom, $\boldsymbol{X} \sim \mathcal{S}t_k(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu = 1)$ is called a Cauchy distribution $\boldsymbol{X} \sim \mathcal{C}a_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Hence, the Cauchy distribution won't have mean nor variance.

Snedecor's F-distribution

A random variable X has a Snedecor's F-distribution $\boldsymbol{X} \sim \mathcal{F}(d_1,\, d_2)$ with parameters d_1 and d_2, positive integers, if for $\boldsymbol{x} \in [0, \infty)$ its probability density function is:

$$f(\boldsymbol{x} \mid d_1,\, d_2) = \left[x \operatorname{Beta}\left(\frac{d_1}{2},\, \frac{d_2}{2}\right)\right]^{-1} \sqrt{\frac{(d_1 x)^{d_1} (d_2)^{d_2}}{(d_1 x + d_2)^{d_1 + d_2}}}.$$

d_1 and d_2 are referred to as degrees of freedom and $\operatorname{Beta}(a,\, b)$ is the beta function:

$$\operatorname{Beta}(a,\, b) = \frac{\Gamma(a)\,\Gamma(b)}{\Gamma(a+b)}.$$

The mean and variance are:

$$\begin{aligned} \mathrm{E}[\boldsymbol{X}] &= \frac{d_2}{d_2 - 2} \quad \text{if} \quad d_2 > 2, \\ \operatorname{Var}[\boldsymbol{X}] &= \frac{2\,(d_2)^2\,(d_1 + d_2 - 2)}{d_1\,(d_2 - 2)^2\,(d_2 - 4)} \quad \text{if} \quad d_2 > 4. \end{aligned}$$

Appendix B

The efective sample size TESS in variable selection

Appropriate scale of objective priors for model selection and, in particular, for variable selection, are usually based on the "information provided by one observation". This desideratum usually implemented by dividing a measure of the information in the sample by n, the sample size. However, the information in the sample can basically be contained in a smaller number of observations *the effective sample size* (see Berger et al., 2010b, for examples and discussion).

The intuitive idea is that if all the observations are i.i.d. the "effective" number of observations containing the information in a sample is n. However, when for instance the observations are correlated, its seems intuitively reasonable that the effective sample size is less than n. Indeed, if all observations are perfectly correlated the effective sample size should be 1.

In an attempt of giving a definition for this measure of information, Berger et al. (2010b) define *the effective sample size* (TESS). TESS has

been analyzed in a broad variety of scenarios giving reasonable answers even where other proposals fail.

We adapt here their definition of TESS to the problem of variable selection.

Consider the model M, containing all the p potential covariates, with the orthogonal parameterization as in (2.6):

$$M : \boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{X}_0\boldsymbol{\gamma} + \boldsymbol{V}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}). \tag{B.1}$$

Where $\boldsymbol{V}$ is the $n \times p$ matrix of covariates, which has been orthogonalized to $\boldsymbol{X}_0$ ($\boldsymbol{X}_0^t\boldsymbol{V} = \boldsymbol{0}$), $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]^t$ is the p-vector of regression parameters.

The definition of TESS does not directly apply to the vector $\boldsymbol{\beta}$ but only to scalar parameters. In particular, for obtaining TESS for each component β_j Berger et al. (2010b) derive an "effective sample size matrix" which diagonal values correspond to the "effective sample sizes" of each β_j. Our strategy is to use this values to obtain an averaged TESS for the whole $\boldsymbol{\beta}$.

Following Berger et al. (2010b), TESS for each β_j is

$$n_j = \frac{\sum_{l=1}^{n} v_{jl}^2}{c_j^2}.$$

Where c_j's can be chosen in few different ways. In this work we adopt the choice of Berger et al. (2010b):

$$c_j = \max_l |v_{jl}|,$$

the maximum absolute value among all values of the jth variable. With this choice TESS goes from 1 to n which, as commented above, is intuitively appealing.

Other possible choices are:

1. Taking
$$c_j = \left(\frac{1}{n} \sum_{l=1}^{n} v_{jl}^2 \right)^{\frac{1}{2}}$$
which is inspired by the Fisher information matrix. With this value $n_j = n$. That is, the default choice for the sample size.

2. taking c_j as the mean for the corresponding covariate. But in this case n_j can be larger than n, which is not allways intuitive.

Hence, our ultimate choice of TESS for M is the mean of all the n_j corresponding to a β_j involved in the model.

$$n^T = \frac{1}{k} \sum_{\beta_j \in M} \frac{\sum_{l=1}^{n} v_{jl}^2}{c_j^2}. \tag{B.2}$$

Appendix C

Hypergeometric functions

Appell hypergeometric function

The *Appell hypergeometric function* (Appell, 1925) is an analytical function of a, b_1, b_2, c, z_1, z_2 defined in $\mathbb{C}^6$ usually denoted

$$F_1(a;\, b_1,\, b_2;\, c;\, z_1,\, z_2).$$

This function is also referred to as *hypergeometric function of two variables*. This function can be expressed as an infinite sum which converges only for $|z_k| < 1$ for $k = 1, 2$; However, convergence can be extended to any value of z_k by considering its integral representation which is valid whenever $\mathrm{Re}(a) > 0$ and $\mathrm{Re}(c - a) > 0$, namely:

$$F_1(a;\, b_1, b_2;\, c;\, z_1,\, z_2) = \frac{\Gamma(c)}{\Gamma(a)\,\Gamma(c-a)} \cdot \\ \cdot \int_0^1 t^{a-1}\,(1-t)^{c-a-1}\,(1-t\,z_1)^{-b_1}\,(1-t\,z_2)^{-b_2}\,dt.$$

For a deep study of this function see Wolfram (2010a) and references there.

Gauss hypergeometric function

When, in the hypergeometric function of two variables (or Appell hypergeometric function) $z_1 = 0$ or $z_2 = 0$, then F_1 becomes the *hypergeometric function of one variable* ${}_2F_1$ also referred to as *Gauss hypergeometric function*. As it is the case with the hypergeometric function of two variables, the Gauss hypergeometric function is considered a closed-form expression. Indeed for $|z| < 1$ this function is defined by an infinite convergent sum. Outside the unit circle, $|z| > 1$ it is defined as the analytic continuation with respect to z of this sum whose integral expression is:

$$ {}_2F_1(a;\, b;\, c;\, z) = \frac{\Gamma(c)}{\Gamma(a)\,\Gamma(c-a)} \int_0^1 t^{a-1}\,(1-t)^{c-a-1}\,(1-t\,z)^{-b}\,dt. $$

The hypergeometric function ${}_2F_1$ has been extensively studied (see e.g. Abramowitz and Stegun, 1964), and it is widely implemented in many computer software packages, so it has become quite a standard' function. Fast approximations are also available, see Wolfram (2010b) and references therein (see also the related H_m function in Berger, 1985).

Appendix D

Proofs of results in Chapter 2

D.1 Proof of Lema 2.1

Lemma. *The sampling distribution of Q_{i0} degenerates to a point mass at*

$$q_M = \frac{1+\delta_1}{1+\delta_1+\delta_2}$$

as $n \to \infty$.

Where $q_M = 1$ for $M_T = M_0$ and $q_M = 1/(1+\delta)$ for $M_T = M_i$, with $\delta = \lim_{n\to\infty} \delta_{i0}$.

Proof. This lemma is easy to proof from the following lemma taken from Casella et al. (2009)

Lemma. *Let $\{X_n, n \geqslant 1\}$ be a sequence of random variables with distribution*

$$X_n \sim \mathcal{B}e(\frac{n-p_i}{2},\ \frac{p_i-p_0}{2};\ n\delta_1,\ n\delta_2),$$

where p_0, p_i, δ_1 and δ_2 are positive constants. Then:

(i) the sequence X_n converges in probability to the constant

$$\frac{1+\delta_1}{1+\delta_1+\delta_2}$$

(ii) If $\delta_1 = \delta_2 = 0$, then X_n degenerates in probability to 1. However, the random variable $-n/2 \log X_n$ does not degenerate and has an asymptotic gamma distribution $Ga(p_i - p_0, 1)$

□

Appendix E

Proofs of results in Chapter 3

E.1 Proof of proposition 3.2

Proposition. *Let* $\|\boldsymbol{\beta}_i\|^2 = \boldsymbol{\beta}_i^t(\boldsymbol{V}_i^t\boldsymbol{V}_i)\boldsymbol{\beta}_i$, *then*

$$\lim_{\|\boldsymbol{\beta}_i\|^2\to\infty} \frac{\pi_i^R(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0,\, \sigma)}{\mathcal{S}t_{k_i}(\boldsymbol{\beta}_i \mid \mathbf{0},\, \boldsymbol{C}_i^*,\, 2a)} \;=\; 1,$$

where:

$$\boldsymbol{C}_i^* = \frac{c\rho_i\, \boldsymbol{B}_i^*(b,\sigma)}{a},$$

$c = (a\,\Gamma(a))^{1/a}$, *and* $\boldsymbol{B}_i^*(b,\sigma) = \sigma^2(b+n)(\boldsymbol{V}_i^t\boldsymbol{V}_i)^{-1}$

The proof requires the following lemma:

Lemma E.1. *Assume* $m > 1$ *and* $p > 0$, *then*

$$\lim_{z\to\infty}\; z^{a+k} \int_0^1 \lambda^{a-1}\Big(\frac{\lambda}{m-\lambda}\Big)^k e^{-\frac{\lambda}{m-\lambda}\cdot p\cdot z} d\lambda = m^a\,\Gamma(a+k)\,p^{-(a+k)}.$$

Proof. Clearly, for any $0 < \epsilon < 1$ and all $\lambda \in [\epsilon, 1]$ the limit

$$\lim_{z\to\infty}\; z^{a+k}\,\lambda^{a-1}\Big(\frac{\lambda}{m-\lambda}\Big)^k e^{-\frac{\lambda}{m-\lambda}\cdot p\cdot z} = 0$$

Then

$$\lim_{z\to\infty} z^{a+k} \int_0^1 \lambda^{a-1}\Big(\frac{\lambda}{m-\lambda}\Big)^k e^{-\frac{\lambda}{m-\lambda}\cdot p\cdot z}\, d\lambda =$$

$$= \lim_{z\to\infty} z^{a+k} \int_0^\epsilon \lambda^{a-1}\Big(\frac{\lambda}{m-\lambda}\Big)^k e^{-\frac{\lambda}{m-\lambda}\cdot p\cdot z}\, d\lambda +$$

$$+ \lim_{z\to\infty} z^{a+k} \int_\epsilon^1 \lambda^{a-1}\Big(\frac{\lambda}{m-\lambda}\Big)^k e^{-\frac{\lambda}{m-\lambda}\cdot p\cdot z}\, d\lambda, \tag{E.1}$$

where the limit and the integral in (E.1) can be interchanged because the integrand is continuous and the integral is over a compact set. Hence,

$$\lim_{z\to\infty} z^{a+k} \int_0^1 \lambda^{a-1}\Big(\frac{\lambda}{m-\lambda}\Big)^k e^{-\frac{\lambda}{m-\lambda}\cdot p\cdot z}\, d\lambda =$$

$$= \lim_{z\to\infty} z^{a+k} \int_0^\epsilon \lambda^{a-1}\Big(\frac{\lambda}{m-\lambda}\Big)^k e^{-\frac{\lambda}{m-\lambda}\cdot p\cdot z}\, d\lambda.$$

Next, make the change of variables $t = \lambda/(m-\lambda)$ to get

$$\int_0^\epsilon \lambda^{a-1}\Big(\frac{\lambda}{m-\lambda}\Big)^k e^{-\frac{\lambda}{m-\lambda}\cdot p\cdot z}\, d\lambda = m^a \int_0^{\frac{\epsilon}{m-\epsilon}} \frac{t^{k+a-1}}{(1+t)^{a+1}}\, e^{-t\cdot p\cdot z}\, dt.$$

Since for $t \in (0, \epsilon/(m-\epsilon))$

$$\frac{1}{(1+\epsilon/(m-\epsilon))^{a+1}} \le \frac{1}{(1+t)^{a+1}} \le 1,$$

the integral of interest can be bounded as follows

$$\frac{m^a\,(zp)^{-(a+k)}\Big(\Gamma(a+k) - \Gamma(a+k,\, \frac{\epsilon}{m-\epsilon}\, pz)\Big)}{(1+\epsilon/(m-\epsilon))^{a+1}} \le$$

$$\le m^a \int_0^{\frac{\epsilon}{m-\epsilon}} \frac{t^{k+a-1}}{(1+t)^{a+1}}\, e^{-t\cdot p\cdot z}\, dt \le \tag{E.2}$$

$$\le m^a\,(zp)^{-(a+k)}\Big(\Gamma(a+k) - \Gamma(a+k,\, \frac{\epsilon}{m-\epsilon}\, pz)\Big),$$

where $\Gamma(\nu_1,\, \nu_2)$ is the incomplete gamma function,

$$\Gamma(\nu_1,\, \nu_2) = \int_{\nu_2}^{\infty} t^{\nu_1 - 1} e^{-t}\, dt,$$

which goes to zero as ν_2 goes to infinity. Multiplying the three parts of the inequality in (E.2) by $z^{(a+k)}$ and taking limits as $z \to \infty$, gives

$$\frac{m^a\, p^{-(a+k)}\, \Gamma(a+k)}{(1+\epsilon/(m-\epsilon))^{a+1}} \;\leq\; \lim_{z\to\infty} m^a\, z^{a+k} \int_0^{\frac{\epsilon}{m-\epsilon}} \frac{t^{k+a-1}}{(1+t)^{a+1}}\, e^{-t\cdot p\cdot z}\, dt \leq$$

$$\leq\; m^a\, p^{-(a+k)}\, \Gamma(a+k).$$

Since this holds for every value $\epsilon > 0$, it follows that

$$\lim_{z\to\infty} m^a\, z^{a+k} \int_0^{\frac{\epsilon}{m-\epsilon}} \frac{t^{k+a-1}}{(1+t)^{a+1}}\, e^{-t\, p\, z}\, dt = m^a\, p^{-(a+k)}\, \Gamma(a+k),$$

or equivalently

$$\lim_{z\to\infty} z^{a+k} \int_0^1 \lambda^{a-1} \Big(\frac{\lambda}{m-\lambda}\Big)^k e^{-\frac{\lambda}{m-\lambda}\, p\, z}\, d\lambda = m^a\, \Gamma(a+k)\, p^{-(a+k)}.$$

□

Now we can prove proposition 3.2

Proof. In the sequel we remove the subindex i for simplicity in notation.

It can be easily shown that

$$\lim_{\|\beta\|^2\to\infty} \mathcal{S}t_k(\boldsymbol{\beta} \mid \mathbf{0},\, \boldsymbol{C}^*,\, 2a) =$$

$$= \lim_{\|\beta\|^2\to\infty} \frac{\Gamma(a+k/2)}{(2\pi)^{k/2}}\, a\, \big(\sigma^2 \rho\, (b+n)\big)^a\, |\boldsymbol{V}^t \boldsymbol{V}|^{1/2}\, 2^{a+k/2}\, \big(\|\boldsymbol{\beta}\|^2\big)^{-(a+k/2)}.$$

It then follows that

$$\lim_{\|\beta\|^2\to\infty} \frac{\pi^R(\boldsymbol{\beta} \mid \boldsymbol{\beta}_0,\, \sigma)}{\mathcal{S}t_k(\boldsymbol{\beta} \mid \mathbf{0},\, \boldsymbol{C}^*,\, 2a)} = \frac{(2\sigma^2)^{-(a+k/2)}\, b^{-k/2}}{\Gamma(a+k/2)\ (\rho(b+n))^a}\cdot$$

$$\cdot \lim_{\|\beta\|^2\to\infty} (\|\boldsymbol{\beta}\|^2)^{a+k/2} \int_0^1 \lambda^{a-1}\Big(\frac{\lambda}{m-\lambda}\Big)^{k/2} e^{-\frac{\lambda}{m-\lambda}\, p\, \|\beta\|^2}\, d\lambda,$$

where $m = (\rho\,(b+n))/b$ and $p = 1/(2\sigma^2 b)$. Since $\rho > b/(b+n)$, $m > 1$ we can apply Lemma E.1 and the result follows. □

E.2 Proof of proposition 3.3

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

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

for any known density on $\mathbb{R}^{k_i}$, f_i, is invariant under the group of transformations

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

Proof. We need to find $\boldsymbol{\beta}_0^*$ and σ^* such that $\boldsymbol{y}^* = c\,\boldsymbol{y} + \boldsymbol{X}_0\boldsymbol{b}$ with $c > 0$ and $\boldsymbol{b}^t \in \mathbb{R}^{k_0}$ has density $m_i(\boldsymbol{y}^* \mid \boldsymbol{\beta}_0^*,\, \sigma^*)$. We know that:

The distribution of $\boldsymbol{y}^*$ given $(\boldsymbol{\beta}_0,\, \boldsymbol{\beta}_i, \sigma)$ is

$$\mathcal{N}_n(\boldsymbol{y}^* \mid c\,(\boldsymbol{X}_0\boldsymbol{\beta}_0 + \boldsymbol{X}_i\boldsymbol{\beta}_i) + \boldsymbol{X}_0\boldsymbol{b},\, (c\,\sigma)^2\boldsymbol{I}).$$

This density can also be expressed as

$$\mathcal{N}_n(\boldsymbol{y}^* \mid \boldsymbol{X}_0\boldsymbol{\beta}_0^* + \boldsymbol{X}_i\boldsymbol{\beta}_i^*,\, (\sigma^*)^2\boldsymbol{I})$$

where $(\boldsymbol{\beta}_0^*,\, \boldsymbol{\beta}_i^*,\, \sigma^*) = (c\,\boldsymbol{\beta}_0 + \boldsymbol{b},\, c\,\boldsymbol{\beta}_i,\, c\,\sigma)$. Therefore, taking $\boldsymbol{\beta}_0^* = c\,\boldsymbol{\beta}_0 + \boldsymbol{b}$ and $\sigma^* = c\sigma$ the integrated likelihood can be expressed as

$$m_i(\boldsymbol{y}^* \mid \boldsymbol{\beta}_0^*,\, \sigma^*) = \int_{\mathbb{R}^{k_i}} \mathcal{N}_n(\boldsymbol{y}^* \mid \boldsymbol{X}_0\boldsymbol{\beta}_0^* + \boldsymbol{X}_i(c\,\boldsymbol{\beta}_i),\, \sigma^{*2}\boldsymbol{I}) \cdot \\ \cdot \frac{c^{k_i}}{\sigma^{*k_i}} f_i\left(\frac{c\,\boldsymbol{\beta}_i}{\sigma^*}\right) d\boldsymbol{\beta}_i,$$

and making the change of variables $\boldsymbol{\beta}_i^* = c\,\boldsymbol{\beta}_i$ gives the desired result.

□

E.3 Proof of proposition 3.4

Proposition. *For any $(a,\, b,\, \rho_i) \in \mathcal{A}$, (where $\mathcal{A}$ is the parametric espace in (3.7)) and $n \geq k_i + k_0$, the prior predictive distribution for $\boldsymbol{y}$ under M_i using the Conventional Robust prior is:*

$$m_i^R(\boldsymbol{y}) = m_0^R(\boldsymbol{y})\, Q_{i0}^{-\frac{n-k_0}{2}} \frac{2a}{k_i + 2a} \left(\rho_i\,(n+b)\right)^{-\frac{k_i}{2}}\, AP_{i0},$$

where

$$m_0^R(\boldsymbol{y}) = \frac{1}{2}\,\pi^{-\frac{n-k_0}{2}}\, |\boldsymbol{X}_0^t\boldsymbol{X}_0|^{-\frac{1}{2}}\, \Gamma\left[\frac{n-k_0}{2}\right]\, SSE_0^{-\frac{n-k_0}{2}},$$

and AP_{i0} is a hypergeometric function of two variables or Appell hypergeometric function:

$$AP_{i0} = F_1\Big[a + \frac{k_i}{2};\, \frac{k_0 + k_i - n}{2},\, \frac{n-k_0}{2};\\ a + 1 + \frac{k_i}{2};\, \frac{(b-1)}{\rho_i\,(b+n)};\, \frac{b - Q_{i0}^{-1}}{\rho_i\,(b+n)}\Big].$$

Recall that Q_{i0} is the ratio of residual sum of squares under each model SSE_i/SSE_0 and was defined in Section 2.3.2.

Note that for computing $m_i(\boldsymbol{y})$ a sample of size $n \geq k_i + k_0$ is needed. Indeed, if we want to compute $m_i(\boldsymbol{y})$ for every model M_i with $i = 0, \ldots, 2^p - 1$ we need a sample of size $n \geq p + k_0$.

First of all let us introduce some results we need for this proof.

Result E.1. *Let $\boldsymbol{X}$ be any $n \times k$ matrix, then*

$$\int_{\mathbb{R}^k} (2\pi\sigma^2)^{-n/2} \exp\{-\frac{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^t (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}{2\sigma^2}\} d\boldsymbol{\beta} =$$

$$= |\boldsymbol{X}^t\boldsymbol{X}|^{-1/2} \exp\{-\frac{1}{2\sigma^2} SSE\} (2\pi\sigma^2)^{(k-n)/2}.$$

where $SSE = \boldsymbol{y}^t (I_n - \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t)^{-1} \boldsymbol{y}$.

Proof. See García-Donato (2003). □

Result E.2. *Given any model M_i in its orthogonal parameterization*

$$M_i : \mathcal{N}_n(\boldsymbol{y} \mid \boldsymbol{X}_0\boldsymbol{\gamma} + \boldsymbol{V}_i\boldsymbol{\beta}_i, \sigma^2\boldsymbol{I}_n),$$

then for every positive constant, c,

$$SSE_0 - (1+c)^{-1} \boldsymbol{y}^t (\boldsymbol{V}_i(\boldsymbol{V}_i^t\boldsymbol{V}_i)^{-1}\boldsymbol{V}_i^t) \boldsymbol{y} = (1+c)^{-1} (SSE_i + c\, SSE_0),$$

where SSE_i is the residual sum of squares under model M_i.

Proof. See García-Donato (2003). □

Result E.3.

$$\int_0^\infty \sigma^{-(a+1)} \exp\left\{-\frac{b}{\sigma^2}\right\} d\sigma = \frac{\Gamma(a/2)}{2\, b^{a/2}}.$$

Proof. See García-Donato (2003). □

Now we can proceed with the proof:

Proof. First of all, let's proof that $m_i^R(\boldsymbol{y})$ does not change whether the model is in its orthogonal parameterization or not.

The orthogonal reparameterization is:

$$(\boldsymbol{\beta}_i,\, \boldsymbol{\beta}_0,\, \sigma) \to (\boldsymbol{\beta}_i,\, \boldsymbol{\beta}_0 + (\boldsymbol{X}_0^t \boldsymbol{X}_0)^{-1} \boldsymbol{X}_0^t \boldsymbol{X}_i\, \boldsymbol{\beta}_i,\, \sigma).$$

The Jacobian of this change of variables is 1:

$$\mathcal{J} = \det \begin{bmatrix} \boldsymbol{I}_n & (\boldsymbol{X}_0^t \boldsymbol{X}_0)^{-1} \boldsymbol{X}_0^t \boldsymbol{X}_i & 0 \\ \boldsymbol{0} & \boldsymbol{I}_n & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & 1 \end{bmatrix} = 1.$$

Hence, the likelihood does not change.

On the other hand, the Conventional Robust prior is invariant under any parameterization that leaves $\boldsymbol{\beta}_i$ unchanged, that is:

$$(\boldsymbol{\beta}_0,\, \boldsymbol{\beta}_i,\, \sigma) \to (\boldsymbol{\gamma},\, \boldsymbol{\beta}_i,\, \sigma)$$

with $\boldsymbol{\gamma} = \boldsymbol{L}\boldsymbol{\beta}_0$ and $\boldsymbol{L}$ any $n \times k_0$ matrix, including the orthogonal parameterization. Then, as neither the likelihood nor the prior change, the resulting $m_i^R(\boldsymbol{y})$ does not change either. For simplicity in calculations we assume here the orthogonal parameterization.

Given any sample size n, the prior predictive distribution under M_0 is

$$m_0^R(\boldsymbol{y}) = \int_{\mathbb{R}^{k_0}} \int_0^\infty \mathcal{N}_n(\boldsymbol{y} \mid \boldsymbol{X}_0\boldsymbol{\gamma},\, \sigma^2 \boldsymbol{I}_n)\, \frac{1}{\sigma}\, d\boldsymbol{\gamma}_0\, d\sigma,$$

and using Result E.1 for integrating out $\boldsymbol{\beta}_0$

$$\begin{aligned} m_0^R(\boldsymbol{y}) &= (2\pi)^{-\frac{n-k_0}{2}}\, |\boldsymbol{X}_0^t \boldsymbol{X}_0|^{-\frac{1}{2}} \int_0^\infty \sigma^{-(n-k_0+1)}\, \exp(-\frac{SSE_0}{2\sigma^2})\, d\sigma \\ &= \frac{1}{2} \pi^{-\frac{n-k_0}{2}}\, |\boldsymbol{X}_0^t \boldsymbol{X}_0|^{-\frac{1}{2}}\, \Gamma\left[\frac{n-k_0}{2}\right]\, \mathrm{SSE}_0^{-\frac{n-k_0}{2}}. \end{aligned}$$

Also, the prior predictive distribution under M_i is

$$m_i^R(\boldsymbol{y}) = \int_\Theta \mathcal{N}_n(\boldsymbol{y} \mid \boldsymbol{X}_0\boldsymbol{\gamma} + \boldsymbol{V}_i\boldsymbol{\beta}_i,\, \sigma^2\boldsymbol{I}_n)\cdot$$
$$\cdot\mathcal{N}_{k_i}(\boldsymbol{\beta}_i \mid 0, \boldsymbol{B}(\lambda))\; a\;\lambda^{a-1}\;\sigma^{-1}\;d(\boldsymbol{\gamma},\,\boldsymbol{\beta}_i,\,\sigma,\,\lambda),$$

with $\Theta = \mathbb{R}^{k_0} \times \mathbb{R}^{k_i} \times [0,\infty) \times [0,1]$.

First we integrate out $\boldsymbol{\beta}_i$ and $\boldsymbol{\gamma}$ using Result E.1, giving

$$m_i^R(\boldsymbol{y}) = (2\pi)^{-\frac{n-k_0}{2}}\,|\boldsymbol{X}_0^t\boldsymbol{X}_0|^{-\frac{1}{2}}$$
$$\int \exp\left\{-\frac{1}{2\sigma^2}\left[SSE_0 - \left(1+\frac{\lambda}{\rho_i(b+n)-b\lambda}\right)^{-1}\boldsymbol{y}^t\boldsymbol{V}_i(\boldsymbol{V}_i^t\boldsymbol{V}_i)^{-1}\boldsymbol{V}_i^t\boldsymbol{y}\right]\right\}$$
$$a\,\lambda^{a+\frac{k_i}{2}-1}\sigma^{-(n-k_0+1)}(\rho_i(b+n)-(b-1)\lambda)^{-\frac{k_i}{2}}\,d(\sigma,\lambda).$$

Now, using Result E.2 and integrating out σ through Result E.3, we finally obtain

$$m_i^R(\boldsymbol{y}) = \frac{1}{2}\pi^{-\frac{n-k_0}{2}}\,|\boldsymbol{X}_0^t\boldsymbol{X}_0|^{-\frac{1}{2}}\,\Gamma\left[\frac{n-k_0}{2}\right]$$
$$\int_0^1 a\,\lambda^{a+\frac{k_i}{2}-1}\,(\rho_i\,(b+n)-(b-1)\,\lambda)^{\frac{n-k_i-k_0}{2}} \qquad \text{(E.3)}$$
$$(\mathrm{SSE}_i\,(\rho_i\,(b+n)-b\,\lambda)+\lambda\,\mathrm{SSE}_0)^{-\frac{n-k_0}{2}}\;d\lambda\,.$$

This expression can be rewritten as:

$$m_i^R(\boldsymbol{y}) = a\,Q_{i0}^{-\frac{n-k_0}{2}}\,(\rho_i\,(n+b))^{-k_i/2}\,m_0(\boldsymbol{y})$$
$$\int_0^1 \lambda^{a+\frac{k_i}{2}-1}\left(1-\frac{b-1}{\rho_i(b+n)}\lambda\right)^{\frac{n-k_i-k_0}{2}}\left(1-\frac{b-Q_{i0}^{-1}}{\rho_i(b+n)}\lambda\right)^{-\frac{n-k_0}{2}}\,d\lambda\,,$$

and using the expression of the Appell function in Appendix C the previous expression can be written as

$$m_i^R(\boldsymbol{y}) = m_0(\boldsymbol{y})\, Q_{i0}^{-\frac{n-k_0}{2}} (n\rho_i + b\rho_i)^{-k_i/2} \frac{2a}{k_i + 2a}\ \mathrm{AP}_{i0},$$

where AP_{i0} is the hypergeometric function of two variables or Appell hypergeometric function.

$$\begin{aligned} \mathrm{A}P_{i0} = \mathrm{Appell}_2 F_1 \Big[a + \frac{k_i}{2};\ \frac{k_0 + k_i - n}{2},\ \frac{n - k_0}{2};& \\ a + 1 + \frac{k_i}{2};\ \frac{(b-1)}{\rho_i\,(b+n)};\ \frac{b - Q_{i0}^{-1}}{\rho_i\,(b+n)} \Big]&. \end{aligned}$$

□

E.4 Proof of Proposition 3.5

Proposition. *Given a model, M_i with k_i extra covariates, for any sample $\boldsymbol{y}^*$ of size $n^* = k_i + k_0$ we have*

$$m_i^R(\boldsymbol{y}^*) = m_0^R(\boldsymbol{y}^*) = \frac{1}{2}\pi^{-\frac{k_i}{2}}\, |\boldsymbol{X}_0^t \boldsymbol{X}_0|^{-\frac{1}{2}}\, \Gamma\left[\frac{k_i}{2}\right]\ SSE_0^{-\frac{k_i}{2}},$$

which only depends on SSE_0, $\boldsymbol{X}_0$ and k_i.

Proof. The proof follows immediately from next Result and equation (E.3) in Appendix E.3.

Result E.4. *For a specific model M_i with k_i extra variables if the sample size is $n = k_i + k_0$ the corresponding residual sum of squares is $SSE_i = 0$*

Proof. For a sample of size $n = k_i + k_0$ the design matrix, $\boldsymbol{X} = [\boldsymbol{X}_0 \mid \boldsymbol{X}_i]$, is a square, full rank matrix and

$$\boldsymbol{X}\left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t = \boldsymbol{I}_{k_i+k_0}. \tag{E.4}$$

Now, recall $SSE_i = \boldsymbol{y}^t(\boldsymbol{I}_n - \boldsymbol{P}_i)\boldsymbol{y}$ with $\boldsymbol{P}_i = \boldsymbol{X}\left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t$, then, if $n = k_i + k_0$ it is easy to see that $SSE_i = 0$ □

□

Appendix F

Proofs of results in Chapter 4

F.1 Convergence Theorems.

Theorem F.1 (Monotone convergence.)**.** *Let $(\mathcal{S}, \Sigma, \mu)$ be a measure space. Let f_n be a pointwise non-decreasing sequence of positive functions. If $\exists\ M > 0$ such that $\int_{\mathcal{S}} f_n(x)\, dx \leq M$ and $\exists\ f = \lim_{n\to\infty} f_n$ almost every where, then f is integrable and*

$$\int_{\mathcal{S}} f\, d\mu = \lim_{n\to\infty} \int_{\mathcal{S}} f_n\, d\mu.$$

Corollary F.1. *Let $f_n(x)$ denote a sequence of real-valued measurable functions on a measure space $(\mathcal{S}, \Sigma, \mu)$. Assume that $\exists\ f = \lim_{n\to\infty} f_n$ almost every where and that f is not measurable in $\mathcal{S}$, then:*

$$\lim_{n\to\infty} \int_{\mathcal{S}} f_n\, d\mu = \infty.$$

Theorem F.2 (Dominated convergence)**.** *Let f_n be a sequence of real-valued measurable functions on a measure space $(\mathcal{S}, \Sigma, \mu)$. Assume that the sequence converges pointwise to a function f, $\lim_{m\to\infty} f_m = f$, and*

it is dominated by some integrable function g

$$|f_n(x)| \leq g(x); \;\; \forall\, x \in \mathcal{S}$$

Then the limiting function f is integrable and

$$\int_{\mathcal{S}} f\, d\mu = \lim_{m\to\infty} \int_{\mathcal{S}} f_m\, d\mu$$

F.2 Proof of proposition 4.3

Proposition. *The Conventional Robust Bayes factor is information consistent if and only if* $n \geq k_i + k_0 + 2a$.

Proof. Conventional Robust Bayes factor can be written as

$$\begin{aligned} B_{i0}^R =& a\,(\rho_i(n+b))^{-\frac{k_i}{2}} (Q_{i0})^{-\frac{n-k_0}{2}} \int_0^1 \lambda^{a+\frac{k_i}{2}-1} \\ & \left[1 - \frac{b-1}{\rho_i(b+n)}\lambda\right]^{\frac{n-k_i-k_0}{2}} \left[1 - \frac{b - Q_{i0}^{-1}}{\rho_i(b+n)}\lambda\right]^{-\frac{n-k_0}{2}} d\lambda = \qquad \text{(F.1)} \\ =& a\,(\rho_i(n+b))^{-\frac{k_i}{2}} \int_0^1 \lambda^{a+\frac{k_i}{2}-1} \\ & \left[1 - \frac{b-1}{\rho_i(b+n)}\lambda\right]^{\frac{n-k_i-k_0}{2}} \left[Q_{i0}\left(1 - \frac{b\lambda}{\rho_i(b+n)}\right) + \frac{\lambda}{\rho_i(b+n)}\right]^{-\frac{n-k_0}{2}} d\lambda. \end{aligned}$$

Let $\{q_m\}$ be an arbitrary decreasing sequence of real numbers such that $\lim_{m\to\infty} q_m = 0$ and define $f_m(\lambda, a, b, n)$ as

$$\begin{aligned} f_m(\lambda, a, b, n) = \lambda^{a+\frac{k_i}{2}-1} & \left[1 - \frac{b-1}{\rho_i(b+n)}\lambda\right]^{\frac{n-k_i-k_0}{2}} \\ & \left[q_m\left(1 - \frac{b\lambda}{\rho_i(b+n)}\right) + \frac{\lambda}{\rho_i(b+n)}\right]^{-\frac{n-k_0}{2}} d\lambda. \end{aligned}$$

It is easy to proof that $f_m(\lambda, a, b, n)$ is also an increasing sequence of functions with m such that $|f_m| < h$, with

$$h(\lambda) = \lambda^{a-1-\frac{n-k_0-k_i}{2}} \left(1 - \frac{b-1}{\rho_i(b+n)}\lambda\right)^{\frac{n-k_i-k_0}{2}} (\rho_i(b+n))^{\frac{n-k_0}{2}}, \qquad \text{(F.2)}$$

and

$$\lim_{m\to\infty} f_m(\lambda, a, b, n) = h(\lambda).$$

- **Case 1:** $n < 2a + k_i + k_0$ or equivalently $a - 1 - \frac{n-k_0-k_i}{2} > -1$. Since (F.2) is integrable for $a - 1 - \frac{n-k_0-k_i}{2} \leq -1$, it follows from the dominated convergence theorem that

$$\lim_{m\to\infty} \int_0^1 f_m(\lambda, a, b, n)\, d\lambda < \infty.$$

- **Case 2:** $n \geq 2a + k_i + k_0$ or equivalently $a - 1 - \frac{n-k_0-k_i}{2} \leq -1$. Since (F.2) is not integrable for $a - 1 - \frac{n-k_0-k_i}{2} \leq -1$, it follows from Corollary F.1 that

$$\lim_{m\to\infty} \int_0^1 f_m(\lambda, a, b, n)\, d\lambda = \infty.$$

Since this holds for every increasing sequence such that $q_m \to 0$ it also holds for $Q_{i0} \to 0$ as desired. □

F.3 Proof of proposition 4.4

Proposition. *For any n, B_{i0}^R is bounded above by a constant for $Q_{i0} \to 1$. This constant is smaller than 1, and depends only on k_i and a. Specifically,*

$$\lim_{Q_{i0}\to 1} B_{i0}^R \;\leq\; \left[1 + \frac{k_i}{2a}\right]^{-1} < 1.$$

Proof. It can be seen that for Q_{i0} close to 1, the integrand in (F.1) is bounded by

$$g(\lambda, a, b, \rho_i) = K \times \lambda^{a+\frac{k_i}{2}-1} \left[1 - \frac{b-1}{\rho_i(b+n)}\lambda\right]^{-\frac{k_i}{2}},$$

which is an integrable function. So by the dominated convergence theorem we can exchange limit and integral to get

$$\lim_{Q_{i0}\to 1} B^R_{i0} = a\left(\rho_i(n+b)\right)^{-\frac{k_i}{2}} \int_0^1 \lambda^{a+\frac{k_i}{2}-1} \left[1 - \frac{b-1}{\rho_i(b+n)}\lambda\right]^{-\frac{k_i}{2}} d\lambda.$$

We next proof that this function is bounded by 1. Define the constant B^0_{i0} as

$$B^0_{i0} = \int_0^1 a\,\lambda^{a+\frac{k_i}{2}-1} \left(\rho_i(b+n) - (b-1)\lambda\right)^{-\frac{k_i}{2}} d\lambda. \tag{F.3}$$

Since $\rho_i > \frac{b}{b+n}$ we have that $\rho_i(b+n) > b$. Moreover

$$\left(\rho_i(b+n) - (b-1)\lambda\right)^{-\frac{k_i}{2}} \leqslant \left(\rho_i(b+n) - b + 1\right)^{-\frac{k_i}{2}}.$$

It then follows that

$$B^0_{i0} \leqslant \int_0^1 a\,\lambda^{a+\frac{k_i}{2}-1}\, d\lambda = \left[1 + \frac{k_i}{2a}\right]^{-1},$$

and, since $k_i \geqslant 1$ and $a < \infty$, we finally get

$$B^0_{i0} \leqslant \frac{1}{1+\frac{1}{2a}} < 1,$$

as desired. □

F.4 Proof of proposition 4.2

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

Proof. First of all recall that the Conventional Robust prior distribution can be expressed as a scale mixture of normals as shown in Proposition 3.1.

Liang et al. (2008) showed that model consistency holds for Zellner-Siow prior, the Hyper-g prior, Hyper-g/n prior and, in general, for any "scale mixture of normals" prior, if

$$\int_0^\infty (1+g)^{-\frac{k_i}{2}}\, h_n(g)\, dg,$$

vanishes as n grows to infinity.

In Liang et al. (2008) this condition is shown to hold for Zellner-Siow and Hyper-g/n prior. For the Conventional Robust prior mixing function we get

$$\int_0^\infty (1+g)^{-\frac{k_i}{2}}\, h_n^R(g)\, dg = \int_{\rho_i(b+n)-b}^\infty (1+g)^{-\frac{k_i}{2}}\, \frac{a\,(\rho_i(b+n))^a}{(g+b)^{(a+1)}}\, dg.$$

Now, making the change of variables: $z = g - (\rho_i(b+n) - b)$,

$$\int_0^\infty (1+g)^{-\frac{k_i}{2}}\, h_n^R(g)\, dg = \\ = \int_0^\infty \frac{a(\rho_i(b+n))^a}{(z+\rho_i(b+n))^{(a+1)}\,(1+z+\rho_i(b+n)-b)^{\frac{k_i}{2}}}\, dg.$$

It is now easy to see that, if $\rho_i(b+n)$ goes to ∞ with n, this integral vanishes as $n \to \infty$ as desired. □

Appendix G

Proofs of results in Chapter 5

G.1 Proof of Proposition 5.1

Proposition. *Let* $\boldsymbol{y}$ *be a sample of size* $n = k_i + k_0 + 1$ *with* $Q_{i0} \geq (k_i + 1)^{-1}$. *For* $a = 1/2$ *and* $b = 1$, B_{i0}^R *is decreasing with* ρ_i.

Moreover, for the distribution of Q_{i0} *under* M_0 *the region* $Q_{i0} \geq (k_i+1)^{-1}$ *accumulates at least at* 30% *of probability and contains the mean (that is,* $(k_i + 1)^{-1}$ *is not larger than the* 70% *percentile).*

Proof. Suppose that $a = 1/2$, $b = 1$ and $Q_{i0} \geq (k_i + 1)^{-1}$

B_{i0}^R is decreasing with ρ_i

Next lemma identify the conditions under which the Conventional Robust Bayes factor is decreasing with ρ_i:

Lemma G.1. *Let's* A, *and* C *be positive constants; then:*

$$g(x) = x^{-\frac{k}{2}} \int_0^1 \lambda^{\frac{k-1}{2}} \left(1 + \frac{C\lambda}{x}\right)^{-A} d\lambda, \tag{G.1}$$

is decreasing with x if

$$\int_0^1 \lambda^{\frac{k-1}{2}} \left(1+\frac{C\lambda}{x}\right)^{-A} d\lambda < 2\left(1+\frac{C}{x}\right)^{-A}. \tag{G.2}$$

Proof. First, let us compute the derivative of g:

$$\frac{\partial}{\partial x} g(x) \propto -\frac{k}{2}\int_0^1 \lambda^{\frac{k-1}{2}} \left(1+\frac{C\lambda}{x}\right)^{-A} d\lambda + \\ + x\int_0^1 \lambda^{\frac{k-1}{2}+1}(-A)\left(-\frac{C\lambda}{x^2}\right)\left(1+\frac{C\lambda}{x}\right)^{-(A+1)} d\lambda\,.$$

Integrating by parts, the second term of this expression becomes:

$$\begin{aligned} & \int_0^1 \lambda^{\frac{k-1}{2}+1}(-A)\left(-\frac{C\lambda}{x}\right)\left(1+\frac{C\lambda}{x}\right)^{-(A+1)} d\lambda \\ = & \left\{\lambda^{\frac{k-1}{2}+1}\left[-\left(1+\frac{C\lambda}{x}\right)^{-A}\right]\right\}_0^1 + \frac{k+1}{2}\int_0^1 \lambda^{\frac{k-1}{2}}\left(1+\frac{C\lambda}{x}\right)^{-A} d\lambda \\ = & \frac{k+1}{2}\int_0^1 \lambda^{\frac{k-1}{2}}\left(1+\frac{C\lambda}{x}\right)^{-A} d\lambda - \left(1+\frac{C}{x}\right)^{-A}. \end{aligned}$$

Then we finally have:

$$\frac{\partial}{\partial x} g(x) \propto \frac{1}{2}\int_0^1 \lambda^{\frac{k-1}{2}}\left(1+\frac{C\lambda}{x}\right)^{-A} d\lambda - \left(1+\frac{C}{x}\right)^{-A}.$$

Hence, for $g(x)$ to be decreasing with x, we need the inequality in (G.2) □

Note that B^R_{i0} for $a = 1/2$ and $b = 1$ can be seen to be proportional to the expression in Lemma G.1 by taking $A = \frac{n-k_0}{2}$, $C = Q_{i0}^{-1} - 1$ and $x = \rho_i(n+1)$. Then if we want the Bayes factor to be decreasing with ρ, we need the inequality in (G.2) to hold, which is equivalent to:

$$\int_0^1 \lambda^{\frac{k_i-1}{2}}\left(\frac{\rho_i(n+1)+Q_{i0}^{-1}-1}{\rho_i(n+1)+(Q_{i0}^{-1}-1)\lambda}\right)^{\frac{n-k_0}{2}} d\lambda < 2.$$

When $n = k_i + k_0 + 1$, the left hand side is:

$$\begin{aligned}
&\int_0^1 \lambda^{\frac{k_i-1}{2}} \left(\frac{\rho_i(k_i+k_0+2) + Q_{i0}^{-1} - 1}{\rho_i(k_i+k_0+2) + (Q_{i0}^{-1}-1)\lambda} \right)^{\frac{k_i+1}{2}} d\lambda \\
= &\int_0^1 \lambda^{-1} \left(\frac{\rho_i(k_i+k_0+2)\lambda + (Q_{i0}^{-1}-1)\lambda}{\rho_i(k_i+k_0+2) + (Q_{i0}^{-1}-1)\lambda} \right)^{\frac{k_i+1}{2}} d\lambda \\
= &\int_0^1 \lambda^{-1} \left(1 - \frac{1-\lambda}{1 + \frac{(Q_{i0}^{-1}-1)}{\rho_i(k_i+k_0+2)}\lambda} \right)^{\frac{k_i+1}{2}} d\lambda.
\end{aligned}$$

Since $\rho_i \geq \frac{1}{1+n}$, we have $\rho_i(k_i + k_0 + 2) > 1$ thus the expression above is bounded by:

$$\int_0^1 \lambda^{-1} \left(1 - \frac{1-\lambda}{1 + (Q_{i0}^{-1}-1)\lambda} \right)^{\frac{k_i+1}{2}} d\lambda. \tag{G.3}$$

The idea is to find a value Q_{i0}^* for which the expression (G.3) is smaller than 2. Then, since this expression is decreasing with Q_{i0} for all values of $Q_{i0} \geq Q_{i0}^*$, the inequality is still true. Making the change of variables

$$t = \frac{1-\lambda}{1 + (Q_{i0}^{-1}-1)\lambda},$$

the expression in (G.3) can be rewritten as:

$$\frac{1}{Q_{i0}} \int_0^1 (1-t)^{\frac{k_i-1}{2}} \left(1 + (Q_{i0}^{-1}-1)t\right)^{-1} d\lambda. \tag{G.4}$$

We can bound this expression with:

$$\frac{1}{Q_{i0}} \int_0^1 (1-t)^{\frac{k_i-1}{2}} d\lambda = \frac{2}{(k_i+1)Q_{i0}}.$$

Then for (G.3) to be smaller than 2, we just need to take

$$Q_{i0}^* = (k_i + 1)^{-1}.$$

The value of ρ_i maximizing B^R_{i0} in this scenario is then the minimal value for ρ_i in $\mathcal{A}$ (see (3.7)). For $b = 1$ and $n = k_i + k_0 + 1$, this value is $\rho_i = 1/(k_i + k_0 + 2)$, but since n can be $n = k_i + k_0$, the maximum value we can reach is at $\rho_i = 1/(k_i + k_0 + 1)$.

$P(Q_{i0} \geq (k_i + 1)^{-1} \mid M_0) \approx 0.3$**.**

As we show in Section 2.3.2, under M_0 Q_{i0} follows a beta distribution $\mathcal{B}e(\frac{n-k_i-k_0}{2}, \frac{k_i}{2})$. In particular, for a sample of size $n = k_i + k_0 + 1$

$$Q_{i0} \mid M_0 \sim \mathcal{B}e(\frac{1}{2}, \frac{k_i}{2}).$$

It is easy to proof that under this distribution the region $Q_{i0} \geq (k_i+1)^{-1}$ represents 30% of the values under the null. This region clearly includes the mean $\mathrm{E}[Q_{i0} \mid M_0] = (k_i + 1)^{-1}$, and of course $Q_{i0} = 1$, the usual values of Q_{i0} indicating "compatibility" with M_0.

□

G.2 Proof of Proposition 5.2

Proposition. *Let* $\boldsymbol{y}$ *be a sample of size* $n = k_i + k_0 + 1$ *with* $Q_{i0} \geq (k_i + 1)^{-1}$*. For* $a = 1/2$*,* $b = 1$ *and* $\rho_i = 1/(k_i + k_0 + 1)$ *the maximum value of* B^R_{i0} *is always less than 1.*

Proof. We show next that $B^R_{i0} < 1$ for $\rho_i = 1/(k_i + k_0 + 2)$.

As shown in the proof of proposition 5.1, for $a = 1/2$, $b = 1$ and $n = k_i + k_0 + 1$

$$\int_0^1 \lambda^{\frac{k_i-1}{2}} \left(1 + \frac{(Q_{i0}^{-1} - 1)\lambda}{\rho_i(n+1)}\right)^{-\frac{n-k_0}{2}} d\lambda < 2\left(1 + \frac{(Q_{i0}^{-1} - 1)}{\rho_i(n+1)}\right)^{-\frac{n-k_0}{2}}.$$

Taking $\rho_i = 1/(k_i + k_0 + 2)$ this inequality turns out to be

$$\int_0^1 \lambda^{\frac{k_i-1}{2}} \left(1 + (Q_{i0}^{-1} - 1)\lambda\right)^{-\frac{k_i+1}{2}} d\lambda < 2(Q_{i0})^{\frac{k_i+1}{2}}.$$

Multiplying both sides by $0.5\,(Q_{i0})^{-\frac{k_i+1}{2}}$, we finally get: $B_{i0}^R < 1$.

Hence, as $1/(k_i+k_0+1) > 1/(k_i+k_0+2)$ and B_{i0}^R is a decreasing function of ρ_i, B_{i0}^R with $\rho_i = 1/(k_i + k_0 + 1)$ is also less than 1. □

Proposition G.1

Proposition G.1. *Let $\boldsymbol{y}$ be a sample of size $n = k_i + k_0 + 1$ with $Q_{i0} \geq (k_i+1)^{-1}$. For $a = 1/2$, $b = 1$ and $\rho_i = 1/(k_i+k_0+2)$, B_{i0}^R is a decreasing function of k_i.*

Proof. The Conventional Robust Bayes factor for $n = k_i+k_0+1$, $a = 1/2$, $b = 1$, and $\rho_i = 1/(k_i + k_0 + 2)$ is:

$$\begin{aligned} B_{i0} &= \frac{1}{2}(Q_{i0})^{-\frac{k_i+1}{2}} \int_0^1 \lambda^{\frac{k_i-1}{2}} \left(1 + (Q_{i0}^{-1} - 1)\lambda\right)^{-\frac{k_i+1}{2}} d\lambda \\ &= \frac{1}{2}\int_0^1 \lambda^{-1} \left(Q_{i0}\lambda^{-1} + (1 - Q_{i0})\right)^{-\frac{k_i+1}{2}} d\lambda \\ &= \frac{1}{2}\int_0^1 \lambda^{-1} \left(Q_{i0}(\lambda^{-1} - 1) + 1\right)^{-\frac{k_i+1}{2}} d\lambda. \end{aligned}$$

Since $[Q_{i0}(\lambda^{-1} - 1) + 1] > 1 \;\forall\; \lambda$, this function is clearly a decreasing function of k_i. □

Bibliography

Abramowitz, M. and Stegun, I.A. (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* New York: Dover.

Appell, P. (1925). *Sur les Fonctions Hypergéométriques de Plusieurs Variables.* Paris: Gauthier-Villars.

Barbieri, M.M. and Berger, J.O. (2004). Optimal Predictive Model Selection. *The Annals of Statistics*, **32**(3): pp. 870–897.

Bayarri, M.J. and Berger, J.O. (1998). Quantifying Surprise in the Data and Model Verification. In J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, eds., *Bayesian Statistics 6*, pp. 53–82. Oxford University Press.

Bayarri, M.J. and Berger, J.O. (2000). P Values for Composite Null Models. *Journal of the American Statistical Association*, **95**(452): pp. 1127–1142.

Bayarri, M.J. and Castellanos, M.E. (2007). Bayesian Checking of the Second Levels of Hierarchical Models. *Statistical Science*, **22**(3): pp. 322–343.

Bayarri, M.J. and García-Donato, G. (2007). Extending Conventional Priors for Testing General Hypotheses in Linear Models. *Biometrika*, **94**(1): pp. 135–152.

Bayarri, M.J. and García-Donato, G. (2008). Generalization of Jeffreys Divergence-Based Priors for Bayesian Hypothesis Testing. *Journal of the Royal Statistical Society: Series B*, **70**(5): pp. 981–1003.

Bayarri, M.J. and Morales, J. (2003). Bayesian Measures of Surprise for Outlier Detection. *Journal of Statistical Planning and Inference*, **111**(1-2): pp. 3 – 22.

Berger, J.O. (1976). Admissible Minimax Estimation of a Multivariate Normal Mean with Arbitrary Quadratic Loss. *The Annals of Statistics*, **4**(1): pp. 223–226.

Berger, J.O. (1980). A Robust Generalized Bayes Estimator and Confidence Region for a Multivariate Normal Mean. *The Annals of Statistics*, **8**(4): pp. 716–761.

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer, 2nd edition.

Berger, J.O. (2006). The Case for Objective Bayesian Analysis. *Bayesian Analysis*, **1**(3): pp. 385–402.

Berger, J.O., Bayarri, M.J., and et al. (2010a). Generalization of BIC. Working Document, Statistical and Applied Mathematical Sciences Institute (SAMSI).

Berger, J.O., Bayarri, M.J., and Pericchi, L.R. (2010b). The Effective Sample Size. Working Document, Statistical and Applied Mathematical Sciences Institute (SAMSI).

Berger, J.O., Bernardo, J.M., and Sun, D. (2009). The Formal Definition of Reference Priors. *The Annals of Statistics*, **37**(2): pp. 905–938.

Berger, J.O., Ghosh, J.K., and Mukhopadhyay, N. (2003). Approximations and Consistency of Bayes Factors as Model Dimension Grows. *Journal of Statistical Planning and Inference*, **112**(1-2): pp. 241 – 258.

Berger, J.O. and Molina, G. (2005). Posterior Model Probabilities Via Path-Based Pairwise Priors. *Statistica Neerlandica*, **59**(1): pp. 3–15.

Berger, J.O. and Pericchi, L.R. (1996). The Intrinsic Bayes Factor for Model Selection and Prediction. *Journal of the American Statistical Association*, **94**: pp. 542–554.

Berger, J.O. and Pericchi, L.R. (2001). Objective Bayesian Methods for Model Selection: Introduction and Comparison. *Lecture Notes-Monograph Series*, **38**(3): pp. 135–207.

Berger, J.O., Pericchi, L.R., and Varshavsky, J.A. (1998). Bayes Factors and Marginal Distributions in Invariant Situations. *Sankhya: The Indian Journal of Statistics, Series A*, **60**(3): pp. 307–321.

Bernardo, J.M. and Smith, A.F.M. (1994). *Bayesian Theory*. John Wiley and Sons, ltd.

Breiman, L. and Friedman, J.H. (1985). Estimating Optimal Transformations for Multiple Regression and Correlation. *Journal of the American Statistical Association*, **80**(391): pp. 580–598.

Carlin, B.P. and Chib, S. (1995). Bayesian Model Choice via Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**(3): pp. 473–484.

Casella, G., Girón, F., Martínez, M., and Moreno, E. (2009). Consistency of Bayesian Procedures for Variable Selection. *The Annals of Statistics*, **37**(3): pp. 1207–1228.

Casella, G. and Moreno, E. (2006). Objective Bayesian Variable Selection. *Journal of the American Statistical Association*, **101**(473).

Chattamvelli, R. (1995). On the Doubly Non-Central F Distribution. *Computational Statistics and Data Analysis*, **20**(5): pp. 481 – 489.

Dmochowski, J. (1996). Intrinsic Priors Via Kullbakc-Leibler Geometry. In J.M. Bernardo, M. DeGroot, D. Lindley, and A.F.M. Smith, eds., *Bayesian Statistics 5*, pp. 543–549. London: Oxford University Press.

Eaton, M. (1989). *Group Invariance Applications in Statistics.*, volume 1 of *Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics, Hayward, Ca.

Ehrlich, I. (1973). Participation in Illegitimate Activities: a Theoretical and Empirical Investigation. *Journal of Political Economics*, **81**(3).

Fernández, C., Ley, E., and Steel, M. (2001). Benchmark Priors for Bayesian Model Averaging. *Journal of Political Economics*, **100**: pp. 381–427.

Galassi, M., Davies, J., Theiler, J., Gough, B., Jungman, G., Alken, P., Booth, M., and Rossi, F. (2009). *GNU Scientific Library Reference Manual*, 3rd edition.
URL `http://www.gnu.org/software/gsl`

García-Donato, G. (2003). *Factores Bayes y Factores Bayes Convencionales: Algunos Aspectos Relevantes*. Ph.D. thesis, Universidad de Valencia.

García-Donato, G. and Martínez-Beneyto, M. (2010). Variable Selection with Gibbs Samplers and Zellner-Siow Priors. Working Document, Centro Superior de Investigaciones en Salud Publica (CSISP).

Garthwaite, P.H., Kadane, J.B., and O'Hagan, A. (2005). Statistical Methods for Eliciting Probablity Distributions. *Journal of the American Statistical Association*, **100**(470): pp. 680–701.

Gelfand, A.E. and Ghosh, S.K. (1998). Model Choice: A Minimum Posterior Predictive Loss Approach. *Biometrika*, **85**(1): pp. 1–11.

George, E.I. (2000). The Variable Selection Problem. *Journal of the American Statistical Association*, **95**(452): pp. 1304–1308.

George, E.I. and McCulloch, R.E. (1993). Variable Selection Via Gibbs Sampling. *Journal of the American Statistical Association*, **88**(423): pp. 881–889.

George, E.I. and McCulloch, R.E. (1997). Approaches for Bayesian Variable Selection. *Statistica Sinica*, **7**(2): pp. 339–373.

Ghosh, J.K. and Samanta, T. (2002). Nonsubjective Bayes Testing: An Overview. *Journal of Statistical Planning and Inference*, **103**(1-2): pp. 205–223.

Goldstein, M. (2006). Subjective Bayesian Analysis: Principles and Practice. *Bayesian Analysis*, **1**(3): pp. 403–420.

Goutis, C. and Robert, C.P. (1998). Model Choice in Generalised Linear Models: A Bayesian Approach Via Kullback-Leibler Projections. *Biometrika*, **85**(1): pp. 29–37.

Guo, R. and Speckman, P.L. (2009). Bayes Factors Consistency in Linear Models. Presented in O'Bayes 09 conference.

Guttman, I. (1982). *Linear Models. An Introduction.* Wiley Series in Probability and Mathematical Statistics. Wiley, John and Sons inc.

Hald, A. (1952). *Statistical Theory with Engineering Applications.* New York: Wiley.

Hoeting, J.A., Madigan, D., Raftery, A.E., and Volinsky, C.T. (1999). Bayesian Model Averaging: A Tutorial. *Statistical Science*, **14**(4): pp. 382–401.

Hsiao, C.K. (1997). Approximate Bayes Factors When a Mode Occurs on the Boundary. *Journal of the American Statistical Association*, **92**(438): pp. 656–663.

Ibrahim, J.G., Chen, M.H., and Sinha, D. (2001). Criterion-Based Methods for Bayesian Model Assessment. *Statistica Sinica*, **11**(2): pp. 419–443.

Ibrahim, J.G. and Laud, P.W. (1994). A Predictive Approach to the Analysis of Designed Experiments. *Journal of the American Statistical Association*, **89**(425): pp. 309–319.

Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press, 3rd edition.

Kadane, J.B., Dickey, J.M., Winkler, R.L., Smith, W.S., and Peters, S.C. (1980). Interactive Elicitation of Opinion for a Normal Linear Model. *Journal of the American Statistical Association*, **75**(372): pp. 845–854.

Kass, R.E. (1993). Bayes Factors in Practice. *Journal of the Royal Statistical Society: Series D*, **42**: pp. 551–560.

Kass, R.E. and Greenhouse, J. (1989). Comment on Investigating Therapies of Potentially Great Benefit: ECMO by Ware(1989). *Statistical Science*, **4**: pp. 310–317.

Kass, R.E. and Raftery, A.E. (1995). Bayes Factors. *Journal of the American Statistical Association*, **90**(430): pp. 773–795.

Kass, R.E. and Vaidyanathan, S.K. (1992). Approximate Bayes Factors and Orthogonal Parameters, with Application to Testing Equality of Two Binomial Proportions. *Journal of the Royal Statistical Society. Series B (Methodological)*, **54**(1): pp. 129–144.

Kullback, S. (1999). *Information Theory and Statistics*. New York: Dover.

Kuo, L. and Mallick, B. (1998). Variable Selection for Regression Models. *Sankhya: The Indian Journal of Statistics*, **60**(1): pp. 65–81.

Laud, P.W. and Ibrahim, J.G. (1995). Predictive Model Selection. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**(1): pp. 247–262.

Leamer, E.E. (1978). *Specification Searches: ad hoc Inference with Nonexperimental Data*. New York: Wiley.

Liang, F., Paulo, R., Molina, G., Clyde, M.A., and Berger, J.O. (2008). Mixtures of g Priors for Bayesian Variable Selection. *Journal of the American Statistical Association*, **103**(481): pp. 410–423.

Miller, A. (2001). *Subset Selection in Regression*. New York: Chapman and Hall.

Moreno, E., Giron, F., and Casella, G. (2009). Consistency of Objective Bayes Tests as the Model Dimensions Increases. Presented in O'Bayes 09 conference.

O'Hagan, A. (1988). *Probability: Methods and Measurements*. Chapman and Hall.

O'Hagan, A. (1994). *Kendall's Advanced Theory of Statistics. Volume 2B: Bayesian Inference*. London: Edward Arnold.

O'Hagan, A. (1995). Fractional Bayes Factors for Model Comparison. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**(1): pp. 99–138.

O'Hagan, A. (2003). HSSS Model Criticism (with discussion). In P.J. Green, N.L. Hjort, and S.T. Richardson, eds., *Highly Structured Stochastic Systems*, pp. 423–453. Oxford University Press, Oxford, UK.

Pérez, J.M. (1998). *Development of Expected Posterior Prior Dstributions for Model Comparisons*. Ph.D. thesis, Purdue University.

Press, S.J. (2003). *Subjective and Objective Bayesian Statistics: Principles, Models and Applications*. Wiley, 2nd edition.

Raftery, A.E., Madigan, D., and Hoeting, J.A. (1997). Bayesian Model Averaging for Linear Regression Models. *Journal of American Statistical Association*, **92**: pp. 179–191.

Rao, C. (1965). *Linear Statistical Inference and its Applications*. New York: Wiley.

Robert, C.P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. New York: Springer-Verlag.

Scott, J.G. and Berger, J.O. (2010). Bayes and Empirical-Bayes Multiplicity Adjustment in the Variable-Selection Problem. *The Annals of Statistics*, **38**(5): pp. 2587–2619.

Spiegelhalter, D.J. and Smith, A.F.M. (1982). Bayes Factors for Linear and Log-Linear Models with Vague Prior Information. *Journal of the Royal Statistical Society. Series B (Methodological)*, **44**(3): pp. 377–387.

Strawderman, W.E. (1971). Proper Bayes Minimax Estimators of the Multivariate Normal Mean. *The Annals of Mathematical Statistics*, **42**(1): pp. 385–388.

Strawderman, W.E. (1973). Proper Bayes Minimax Estimators of the Multivariate Normal Mean Vector for the Case of Common Unknown Variances. *The Annals of Statistics*, **1**(6): pp. 1189–1194.

Suzuki, Y. (1983). On Bayesian Approach to Model Selection. In *Proceedings of the International Statistical Institute*, pp. 288–291. Voorburg, ISI Publications.

Tierney, L., Rossini, A.J., and Li, N. (2007). Simple Parallel Statistical Computing in R. *Journal of Computational and Graphical Statistics*, **16**(2): pp. 399–420.

Vandaele, W. (1978). Participation in Illegitimate Activities: Ehrlich Revisited. In *Deterrence and Incapacitation*, pp. 270–335. US National Academy of Sciences.

Wasserman, L. (2000). Bayesian Model Selection and Model Averaging. *Journal Mathematical Psychology*, **44**(1): pp. 92–107.

Wolfram (2010a).
URL `http://functions.wolfram.com/HypergeometricFunctions/AppellF1/`

Wolfram (2010b).
URL `http://functions.wolfram.com/HypergeometricFunctions/Hypergeometric2F1/`

Woods, H., Steinour, H., and Starke, H. (1932). Effect of Composition of Porland Cement on Heat Evolved During Hardening. *Industrial and Engineering Chemistry Research*, **24**: pp. 1207–1214.

Zellner, A. (1986). On Assessing Prior Distributions and Bayesian Regression Analysis with g-prior Distributions. In A. Zellner, ed., *Bayesian Inference and Decision techniques: Essays in Honor of Bruno de Finetti*, pp. 389–399. Edward Elgar Publishing Limited.

Zellner, A. and Siow, A. (1980). Posterior Odds Ratio for Selected Regression Hypotheses. In J.M. Bernardo, M. DeGroot, D. Lindley, and A.F.M. Smith, eds., *Bayesian Statistics 1*, pp. 585–603. Valencia: Univeristy Press.

Zellner, A. and Siow, A. (1984). *Basic Issues in Econometrics*. Chicago: University of Chicago Press.