Supplement to

Variable Selection in the presence of factors: a model selection perspective

Contents

This paper contains additional material that has pedagogical value to better understand the content of the main article. References with a prefix S, e.g. (S.1), correspond to this supplement; all others, e.g. (1), relate to the actual article.

S.1 The role of parameterization in model selection methods

To understand the main ideas, it suffices to focus on the case where we only have one factor with ℓ levels in the set of explanatory variables. Then, we have to quantify the evidence in favor of the factor being a relevant explanatory variable for the response.

In this situation, model (1) reduces to

$$\boldsymbol{y} = \mathbf{1}\beta_0 + \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{S.1}$$

with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_\ell)^\top$, $\boldsymbol{Z}^\top = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$, $\boldsymbol{Z}_i = (z_{i1}, \ldots, z_{i\ell})^\top$, with $z_{ij} = 1$ if individual i belongs to the j-th level of the factor, and 0 otherwise, $j = 1, \ldots, \ell$, $i = 1, \ldots, n$. The null model is obtained by setting $\boldsymbol{Z} = \mathbf{0}$ in (S.1) and states that the factor is irrelevant as an explanatory variable. Any submodel that is obtained by removing columns from (S.1) can be identified with a parameter vector $\boldsymbol{\delta} \in \{0, 1\}^\ell$ that indicates which levels of the factor are present. We refer to each model either by $M_{\boldsymbol{\delta}}$ or simply by $\boldsymbol{\delta}$. The full model corresponds to $\boldsymbol{\delta} = \mathbf{1}$, the null model to $\boldsymbol{\delta} = \mathbf{0}$.

Suppose that $\ell = 3$. The design matrix $[\mathbf{1} \mid \boldsymbol{Z}]$ (which has 4 columns) is rank deficient, which means that the full model is not identifiable, that it is overparametrized. This issue is traditionally dealt with by imposing a full-rank parametrization. In fact, in the `R` software if one of the variables in a linear model is a factor, by default a treatment parametrization is imposed on the model without any interference of the user. This results in an associated design matrix $[\mathbf{1} \mid \boldsymbol{Z}^\star]$ (a 3-column, full rank matrix). Suppose that we use the treatment parameterization where the first level is treated as the reference; hence, $\beta_1 = 0$ and $\boldsymbol{Z}^\star$ results from removing the first column from $\boldsymbol{Z}$. If, as it is customary in variable selection, we produce the class of competing models by deleting columns of $\boldsymbol{Z}^\star$, we obtain a class containing 4 models, namely $M^{b=1}_{(0,0)}$, $M^{b=1}_{(0,1)}$, $M^{b=1}_{(1,0)}$ and $M^{b=1}_{(1,1)}$, that we have collected in Table S.1 (left column). The superscript stands for the fact that the first level is the baseline level. Now repeat the exercise but choosing the second level as reference (right column on Table S.1). Comparing both model spaces, we note that the null models coincide. [Notice also that the interpretation of β_0 is not the same as in the full model. Although in the full model β_0 is the average response for units in the reference level, in the null model it is the average response obtained if the factor had no effect on the response, so that it means the same regardless of the reference level chosen.] Likewise, $M^{b=1}_{(0,1)}$ and $M^{b=2}_{(0,1)}$ are equivalent since one is a reparameterization of the other, and similarly for the full models. However, $M^{b=1}_{(1,0)}$ and $M^{b=2}_{(1,0)}$ are distinct models. Furthermore, $M^{b=1}_{(1,0)}$ does not exist in the second list and similarly $M^{b=2}_{(1,0)}$ does not appear in the first model space. This phenomenon was previously reported in Fernández et al. (2002), specifically in their section 5.1. There, they entertain the example of the categorical variable "month of the year". To reproduce here their reasoning, let December, January and February be level 1, 2 and 3 of the factor (the factor has obviously 12 levels, but that is irrelevant for this discussion). Then they state that, when using December as a reference level (the model space is the one in the left column in Table S.1) one would not be able to capture the situation where "January has the same effect as [. . .] February yet not the same as December" (because such situation corresponds to model $M^{b=2}_{(1,0)}$ which is in not in the list of entertained models but in the right column of the table). This phenomenon has an unavoidable impact on the variable selection results, of particular importance if we consider the parameterization that results in a set of competing models that does not include the true data generating process. Obviously, in this situation the true model cannot be chosen (independently of the sample size). Since the null model is close to the missed model (particularly if the size of the signal is not very strong) and is a quite parsimonious representation of the data, hence being favored by Ockham's razor, the evidence we collect from the data will likely underestimate the importance of the factor.

Next, we use two examples to illustrate the impact of the described dependency on the parametrization, as measured by the posterior probabilities of the factor, i.e., the sum of the posterior probabilities of all the models under which the factor is an explanatory variable. The details on how we compute posterior model probabilities are described in Section 3; note that here we use a constant prior on the model space. Notice also that the problem would similarly affect *any* model selection-based approach (and subsequent model average estimates) simply because regardless of how one measures evidence, the class of competing models is not exhaustive if one imposes a full-rank parametrization on the full model.

Example S.1. *For the model* (S.1) *with* $\ell = 3$ *we have simulated* $n = 100$ *observations with* $\beta_0 = 1$*,* $\beta_1 = 0.5$*,* $\beta_2 = \beta_3 = 0$*. The observations were randomly assigned to each of the groups. The standard deviation of the errors is fixed at* $\sigma = 1$*.*

In the true data generating process, the factor is relevant (with only the first level being active) so we should expect support in our results in favor of the factor. Nevertheless, if the first level is used as baseline (so, as we discussed above, the true model is missed), we obtain a posterior probability of the factor of 0.41 (concluding that the factor is not relevant) while if the second level is used as baseline (the list of considered models contains the true model) the probability of the factor is quite more sensible and becomes 0.77. For the BSGS (Bayesian Sparse Group Selection by Lee and Chen, 2015, implemented with the R package `BSGS`*) the results are 0.28 in the first case and 0.54 in the second.*

In the case where the first level is used as baseline, what happens is that the null model is simple and "close" to the missed true model; hence, it is favored by the Bayes factor over all the remaining models. If the second level is used as baseline, the results are different and more sensible since the true model is considered.

As the signal increases, the null model becomes less and less appealing. If we repeat the same experiment but with $\beta_1 = 0.75$*, we obtain a posterior probability of the factor of 0.92 if first level is used as baseline, and 0.99 if level 2 is used as baseline. These numbers with BSGS are 0.33 and 0.92 respectively.*

●

Example S.2. *Our second example concerns a study about child obesity in Spain (Zurriaga et al., 2011), which we introduce in Section S.5. In this research, one goal* *was to identify the relevance (if any) of practicing sports (the factor) on the body mass index y of $n = 1002$ children between 2 and 14 years of age. This factor has $\ell = 6$ levels, ranging from no practice to very intense practice of sports .*

If the first level of the factor (no practice of sports) is used as baseline, the posterior probability of the factor is 0.26, while if the second level is used as baseline this probability increases to 0.98 (a remarkable change). With BSGS similar results are obtained, with 0.26 in the first case and 0.91 in the second.

Since this is a real example, the truth is unknown, but intuition tells us that the factor ought to be influential on the response. The model with the first level of the factor present and the others not (modeling the two situations of sendentary vs. active lifestyle) is arguably a good model that simply disappears if this parameterization is used as baseline.

●

One could think that the problem is restricted to parameterizations where one of the levels is used as a baseline. Nevertheless, alternative full rank parameterizations will always have associated a model space with 4 models, so again certainly one model would be missing. Consider for instance the sum to zero parameterization $\beta_3 = -\beta_1 - \beta_2$. The set of models obtained when deleting columns from the design matrix are listed in Table S.2 (superscript "sum=0" stands for the competing models in this parameterization). In this model space, obviously the null and full models are equivalent to their counterparts in the parameterizations above, but for instance $M^{b=1}_{(0,1)}$ (third level of the factor has a significant impact, which is different from the first two levels) is not in this new list of models.

The conclusion is then that the full-column rank formulation of the model does not allow us to span the whole class of models by deleting columns of $\boldsymbol{Z}^{\star}$, and that the set of models that we miss is parameterization-dependent. This observation invalidates the tempting approach of just feeding a (model-selection-based) variable selection package (like those compared in Forte et al., 2018) with the design matrix of the full model (which is now full rank) and using the results without taking into consideration the fact that the potential regressor is a factor. In Section 3, we show that the number of models that disappear may be quite large when the number of

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

Table S.1: Refer to model (S.1). List of entertained models for the problem with $\ell = 3$ using two different reference levels using the treatment parametrization. $M^{b=1}_{\boldsymbol{\delta}}$ stands for the models where the first level is used as reference; similarly for $M^{b=2}_{\boldsymbol{\delta}}$. Here, μ_{ij} stands for the expectation of y_i under the corresponding model, with unit i belonging to the j-th level of the factor.

Sum to zero constraint
$M^{\text{sum}=0}_{(0,0)} : \mu_{ij} = \beta_0$
$M^{\text{sum}=0}_{(0,1)} : \mu_{i1} = \beta_0, \mu_{i2} = \beta_0 + \beta_2, \mu_{i3} = \beta_0 - \beta_2$
$M^{\text{sum}=0}_{(1,0)} : \mu_{i1} = \beta_0 + \beta_1, \mu_{i2} = \beta_0, \mu_{i3} = \beta_0 - \beta_1$
$M^{\text{sum}=0}_{(1,1)} : \mu_{i1} = \beta_0 + \beta_1, \mu_{i2} = \beta_0 + \beta_2, \mu_{i3} = \beta_0 - \beta_1 - \beta_2$

Table S.2: Refer to model (S.1). List of entertained models for the problem with $\ell = 3$ using the sum to zero parameterization: $\beta_3 = -\beta_1 - \beta_2$. Notation for μ_{ij} is similar to that in Table S.1.

variables, factors or levels within factor is even moderate.

The approach that we propose, detailed in Section 3, which is obviously independent of the parameterization, is to use the original (rank deficient) matrix $\boldsymbol{Z}$ and the models we obtain as we delete columns from it. This solution is what Fernández et al. (2002) name “free reference level”. In the case with $\ell = 3$, we formally obtain 8 models, enumerated in Table S.1, and it’s clear that all the models that were missing are now present. In particular, $M^{b=1}_{(1,0)}$ (that was missed in the second parameterization) is equivalent to $M_{(0,1,0)}$ while $M^{b=2}_{(1,0)}$ (that, recall, did not appear in first parameterization) is the same as $M_{(1,0,0)}$. A drawback of this approach is that, upon inspection, we realize that we have in fact 5 unique models since 4 of them (the last four) are simply reparameterizations of each other, all representing the same model.

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

Table S.3: Refer to model (S.1). List of entertained models for the problem with $\ell = 3$ using the original parameterization. Models below second line are different representations of the same model. Column Rank contains the rank of the design matrix in each model. Notation for μ_{ij} is similar to that in Table S.1. The convention that we follow when defining the set of competing models $\mathcal{M}$ is to keep the overparametrized model, $M_{(1,1,1)}$ in this case.

The issue of repeated models has of course a simple solution: removing all but one of the repeated models. As long as the criteria used to evaluate the single models is invariant to full rank parameterizations then it is irrelevant which model is kept. The Bayesian approach that we develop in detail in Section 3 is based on the conventional Bayes factors that are invariant in this sense (see details in Bayarri and García-Donato, 2007). The convention that we follow when defining the set of competing models $\mathcal{M}$ is to keep the overparametrized model, $M_{(1,1,1)}$ in this case.

For illustrative purposes, and still with the constant prior over the model space (so each of the unique models has a prior probability of 1/5) we have computed the posterior probability of the factor for the simulated Example S.1 obtaining 0.78 (when $\beta_1 = 0.5$) and 0.99 (when $\beta_1 = 0.75$) which are quite sensible. Similarly, for the Example S.2 we obtain that the probability of the factor “practice of sports” being significant is 0.99.

S.2 Enumeration of models for the case $k = 1$, $p = 2$ $(\ell_1 = 2, \ell_2 = 3)$

Table S.4 contains the values for the vectors $\boldsymbol{\gamma}$, $\boldsymbol{\delta}$ and $\boldsymbol{\tau}$ for all the models spanned as we delete columns from $[\boldsymbol{X} \mid \boldsymbol{Z}]$ in (1) of a problem with one covariate ($k = 1$) and two factors ($p = 2$) with $\ell_1 = 2$ and $\ell_2 = 3$. We also indicate whether each model is a member of $\mathcal{M}$.

S.3 Proof of Theorem 1

Proof. Since, for each factor, what is not permitted is to remove only one column, it is straightforward to obtain that the cardinality of $\mathcal{M}$ is given by (2). What is more laborious is to prove that the models in $\mathcal{M}$ are all unique under the stated conditions.

First notice that two linear models differ when the spaces spanned by the columns of their corresponding design matrices, $\mathcal{C}(\cdot)$, do not coincide. Additionally, it is well known that two vector spaces generated by certain vectors differ if and only if each vector in one space can be expressed as a linear combination of the vectors in the other space (and conversely). If $\boldsymbol{A}$ is a matrix and $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ are two of its column vectors, we denote by $\boldsymbol{A} - \{\boldsymbol{a}_1, \boldsymbol{a}_2\}$ the matrix that results from removing $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ from $\boldsymbol{A}$.

Denote the vector columns in $\mathbb{X} \equiv [\mathbf{1} \mid \boldsymbol{X} \mid \boldsymbol{Z}]$ as in

$$[\,\mathbf{1} \mid \boldsymbol{x}_1 \mid \cdots \mid \boldsymbol{x}_k \mid \boldsymbol{z}_{11} \mid \cdots \mid \boldsymbol{z}_{1\ell_1} \mid \cdots \mid \boldsymbol{z}_{p1} \mid \cdots \mid \boldsymbol{z}_{p\ell_p}\,], \tag{S.2}$$

and recall that dummies associated with factors must satisfy, for any $j = 1, \ldots, p$,

$$\sum_{r=1}^{\ell_j} \boldsymbol{z}_{ir} = \mathbf{1}\ . \tag{S.3}$$

The proof follows by *reductio ad absurdum*: we take two models M_1 and M_2 in $\mathcal{M}$ (so two different set of columns of (S.2)), assume they define the same linear model, and conclude that the rank of $\mathbb{X}$ should be smaller than the assumed $1 + k + L - p$.

If there is a numeric variable (say x_1) in M_1 but not in M_2 (or vice versa), that would imply that the column vector defined by this variable can be expressed as a linear combination of other column vectors in $\mathbb{X}$ so

$$\text{rank}\big(\mathbb{X}\big) = \text{rank}\big(\mathbb{X} - \{\boldsymbol{x}_1\}\big) = \text{rank}\big(\mathbb{X} - \{\boldsymbol{x}_1, \boldsymbol{z}_{11}, \ldots, \boldsymbol{z}_{p1}\}\big) \leq k + L - p,$$

(the last equality holds true because of (S.3)), which contradicts the hypothesis.

If there is a factor (say Λ_1) that is present in M_1 but not in M_2 (or vice versa) then one dummy vector column in this factor (say $\boldsymbol{z}_{11}$) can be expressed as a linear combination of other column vectors in $\mathbb{X}$ (different from $\{\boldsymbol{z}_{12}, \ldots, \boldsymbol{z}_{1\ell_1}\}$) and similarly for $\boldsymbol{z}_{12}$ (due to (S.3)). Then, just as before,

$$\text{rank}\big(\mathbb{X}\big) = \text{rank}\big(\mathbb{X} - \{\boldsymbol{z}_{11}, \boldsymbol{z}_{12}\}\big) = \text{rank}\big(\mathbb{X} - \{\boldsymbol{z}_{11}, \boldsymbol{z}_{12}, \boldsymbol{z}_{21}, \ldots, \boldsymbol{z}_{p1}\}\big) \leq k + L - p,$$

contradicting the hypothesis.

Finally, suppose that M_1 and M_2 have the same active numerical variables and the same factors. There should be (at least) one factor (say Λ_1) for which the choice of dummies differ. Then we have

$$M_1 = [\mathbf{1} \mid \boldsymbol{z}_{1i_1} \mid \ldots \mid \boldsymbol{z}_{1i_a} \mid \cdots], \quad M_2 = [\mathbf{1} \mid \boldsymbol{z}_{1j_1} \mid \ldots \mid \boldsymbol{z}_{1j_b} \mid \cdots].$$

Without loss of generality we assume that $a \leq b$ so at least one dummy in M_2 is not in M_1 and we suppose it is $\boldsymbol{z}_{1j_1}$. Finally, due to the definition of $\mathcal{M}$ then $a \neq \ell_1 - 1$ and $b \neq \ell_1 - 1$. If M_1 and M_2 would define the same linear model we must consider two possibilities

- Λ_1 is not oversaturated in M_2 (so $b \leq \ell_1 - 2$). Then $\boldsymbol{z}_{1j_1}$ can be expressed as linear combination of the variables in M_1 of the type

$$\boldsymbol{z}_{1j_1} = c_0\mathbf{1} + c_{i_1}\boldsymbol{z}_{1i_1} + \cdots + c_{i_a}\boldsymbol{z}_{1i_a} + \cdots \tag{S.4}$$

Additionally, there is another dummy variable that is not in M_2 (say $\boldsymbol{z}_{11}$) that can be expressed as the linear combination given by (S.3) that is linearly independent of (S.4) (simply because it involves the rest of $\ell_1 - 1$ with coefficients different from zero and (S.4) uses $a \leq \ell_1 - 2$.)

Then, as before,

$$\text{rank}\big(\mathbb{X}\big) = \text{rank}\big(\mathbb{X} - \{\boldsymbol{z}_{11}, \boldsymbol{z}_{12}\}\big) = \text{rank}\big(\mathbb{X} - \{\boldsymbol{z}_{11}, \boldsymbol{z}_{12}, \boldsymbol{z}_{21}, \ldots, \boldsymbol{z}_{p1}\}\big) \leq k + L - p,$$

contradicting the hypothesis.

- Λ_1 is oversaturated in M_2 (so $b = \ell_1$). Here, M_2 has, apart from z_{1j_1}, (at least) another dummy (say $\boldsymbol{z}_{1j_2}$) not included in M_1 corresponding to Λ_1. Both can be expressed as independent linear combinations of the vectors in M_2 and again:

$$\text{rank}\big(\mathbb{X}\big) = \text{rank}\big(\mathbb{X} - \{\boldsymbol{z}_{1j_1}, \boldsymbol{z}_{1j_2}\}\big) = \text{rank}\big(\mathbb{X} - \{\boldsymbol{z}_{1j_1}, \boldsymbol{z}_{1j_2}, \boldsymbol{z}_{21}, \ldots, \boldsymbol{z}_{p1}\}\big) \leq k + L - p,$$

contradicting the hypothesis.

□

S.4 Results about $\mathcal{M}$

Definition S.1. *For $\ell \geq j \geq 0$, define*

$$\left(\!\!\binom{\ell}{j}\!\!\right) = \begin{cases} 1 & \text{if } \ \ell - 1 \leq j \leq \ell \\ \binom{\ell}{j} & \text{if } \quad j < \ell - 1 \end{cases}$$

The following result establishes several properties of the rank of models that will be useful in specifying the model prior probabilities.

Result S.1. *In the conditions of Theorem 1, let $\kappa(\boldsymbol{\gamma}, \boldsymbol{\delta})$ (dimension of a model) denote the rank of the design matrix of each model $(\boldsymbol{\gamma}, \boldsymbol{\delta}) \in \mathcal{M}$ minus one and denote by $\mathcal{F}_k^{\ell_1,\ldots,\ell_p}(r)$ the number of models with the same dimension r in a problem with k numerical variables and p factors with number of levels $\ell_1, \ldots, \ell_p$. Then* *i)*

$$\kappa(\boldsymbol{\gamma},\boldsymbol{\delta}) = \mathbf{1}^\top\boldsymbol{\gamma} + \sum_{j=1}^{p} \min\{\mathbf{1}^\top\boldsymbol{\delta}_j, \ell_j - 1\}.$$

ii) The dimension of models in $\mathcal{M}$ varies in the interval:

$$0 \leq \kappa(\boldsymbol{\gamma},\boldsymbol{\delta}) \leq L - p + k,$$

where the dimension of the null model is $\kappa(\mathbf{0},\mathbf{0}) = 0$ and the dimension of the full model is $\kappa(\mathbf{1},\mathbf{1}) = L - p + k$.

iii)

$$\mathcal{F}_k^{\ell_1,\ldots,\ell_p}(r) = \sum_{0\leq i\leq k,\, 0\leq j_1\leq \ell_1-1,\, \cdots,\, 0\leq j_p\leq \ell_p-1,\, 1+j_1+\cdots+j_p=r} \binom{k}{i} \left(\!\!\binom{\ell_1}{j_1}\!\!\right) \cdots \left(\!\!\binom{\ell_p}{j_p}\!\!\right),$$

where $0 \leq r \leq L - p + k$ and by convention $\binom{0}{0} = 1$.

Proof. Results i) and ii) easily follows using similar arguments as those in the proof of Theorem 1 and now noticing that $\kappa(\boldsymbol{\gamma},\boldsymbol{\delta})+1$ is the number of linearly independent vectors in

$$\{\ \mathbf{1} \mid \gamma_1\boldsymbol{x}_1 \mid \cdots \mid \gamma_k\boldsymbol{x}_k \mid \delta_{11}\boldsymbol{z}_{11} \mid \cdots \mid \delta_{1\ell_1}\boldsymbol{z}_{1\ell_1} \mid \cdots \mid \delta_{p1}\boldsymbol{z}_{p1} \mid \cdots \mid \delta_{p\ell_p}\boldsymbol{z}_{p\ell_p}\ \}.$$

Result iii) is a standard multi-binomial expression that defines the number of different ways to choose i elements from a set of k, j_1 from a set of ℓ_1 and so on. The singularity here is that, when $j_h \geq \ell_h - 1$ there is only one possible choice (the rest are repetitions of the same model).

□

Result S.2. *In the conditions of Theorem 1, suppose the active factors in $\mathcal{M}(\boldsymbol{\gamma},\boldsymbol{\tau})$ have indices $\{j_1,\ldots,j_{m_2}\}$ (note $m_2 = \mathbf{1}^\top\boldsymbol{\tau}$) and denote $\mathcal{G}^{\ell_{j_1},\ldots,\ell_{j_{m_2}}}(r)$ the number of models in $\mathcal{M}(\boldsymbol{\gamma},\boldsymbol{\tau})$ with the same dimension $r + \mathbf{1}^\top\boldsymbol{\gamma}$. Then:*

i) the cardinality of $\mathcal{M}(\boldsymbol{\gamma},\boldsymbol{\tau})$ is:

$$\#\mathcal{M}(\boldsymbol{\gamma},\boldsymbol{\tau}) = \prod_{j=1}^{p} \left(2^{\tau_j\ell_j} - \tau_j\ell_j - \tau_j\right).$$

ii) The dimension of models $(\boldsymbol{\gamma}, \boldsymbol{\delta}) \in \mathcal{M}(\boldsymbol{\gamma}, \boldsymbol{\tau})$ *varies in the interval*

$$\kappa(\boldsymbol{\gamma}, \boldsymbol{\delta}) \in R(\boldsymbol{\gamma}, \boldsymbol{\tau}) \equiv [\mathbf{1}^\top \boldsymbol{\gamma} + \mathbf{1}^\top \boldsymbol{\tau},\ \mathbf{1}^\top \boldsymbol{\gamma} + \sum_{j=1}^{p} \tau_j(\ell_j - 1)]. \tag{S.5}$$

iii)

$$\mathcal{G}^{\ell_{j_1},\ldots,\ell_{j_{m_2}}}(r) = \sum_{1 \leq j_1 \leq \ell_{j_1}-1, \cdots, 1 \leq j_{m_2} \leq \ell_{j_{m_2}}-1,\, j_1+\cdots+j_{m_2}=r} \left(\!\!\binom{\ell_1}{j_1}\!\!\right) \cdots \left(\!\!\binom{\ell_{m_2}}{j_{m_2}}\!\!\right),$$

where $m_2 \leq r \leq \ell_{j_1} + \cdots + \ell_{j_{m_2}} - m_2$.

Proof. The proof is straightforward and follows using similar arguments as those in the proofs of Theorem 1 and Result S.2. □

Theorem S.1. *With* (C)*, the prior inclusion probability of any numerical variable* (x_i) *is* $p(\gamma_i = 1) = 1/2$*, while the inclusion probability of a factor* (Λ_j) *is*

$$p(\tau_j = 1) = 1 - \frac{1}{2^{\ell_j} - \ell_j}\ .$$

With (SB)*, the inclusion probability of a variable is*

$$p(\gamma_i = 1) = 1 - \frac{1}{L - p + k + 1} \sum_{r=0}^{L-p+k-1} \frac{\mathcal{F}_{k-1}^{\ell_1,\ldots,\ell_p}(r)}{\mathcal{F}_{k}^{\ell_1,\ldots,\ell_p}(r)};$$

and, for a factor, it is

$$p(\tau_j = 1) = 1 - \frac{1}{L - p + k + 1} \sum_{r=0}^{L-\ell_j-(p-1)+k} \frac{\mathcal{F}_{k}^{\ell_1,\ldots,\ell_{j-1},\ell_{j+1},\ell_p}(r)}{\mathcal{F}_{k}^{\ell_1,\ldots,\ell_p}(r)}.$$

Proof. About the prior (C). The result for numerical variables follows simply noting that half of the models in $\mathcal{M}$ contain x_i. About a factor it's easier to work with the complement. The factor Λ_j is not included if none of its levels is and there are a total of

$$2^k \times \prod_{r=1, r\neq j}^{p} [2^{\ell_r} - \ell_r]$$

and the result follows.

About the prior (SB) we work with the complement in both cases. Recall that

$$\mathcal{F}_k^{\ell_1,\ldots,\ell_p}(r)$$

is, for a problem with k numerical variables and p factors (levels $\ell_1,\ldots,\ell_p$) the number of models of dimension r, where r is a natural number that ranges in the interval $[0, L-p+k]$ (see ii) in Result S.1).

Of the models in $\mathcal{M}$, there are $\mathcal{F}_{k-1}^{\ell_1,\ldots,\ell_p}(r)$ models that do not contain x_i when $r \leq L-p+k-1$ and the full model (with dimension $r = L-p+k$) necessarily contains x_i. Hence the probability that the variable x_i is not included corresponds to the sum:

$$\frac{1}{L-p+k+1} \sum_{r=0}^{L-p+k-1} \frac{\mathcal{F}_{k-1}^{\ell_1,\ldots,\ell_p}(r)}{\mathcal{F}_k^{\ell_1,\ldots,\ell_p}(r)}$$

and the result follows.

The proof for a factor is similar, but now the number of models that do not contain a given factor Λ_j is

$$\mathcal{F}_{k-1}^{\ell_1,\ldots\ell_{j-1},\ell_{j+1},\ell_p}(r),$$

and the dimension varies in the interval $[0, L-\ell_j-(p-1)+k]$. □

S.5 Childhood obesity study in Spain: a real application

An epidemiological study was conducted in Spain (Zurriaga et al., 2011) to determine the association strength between dietary behavior, sedentary habits and childhood obesity.

Data were collected on children (2-14 years old) based on questionnaires filled in by pediatricians and families. Our analysis of the data has mainly illustrative purposes and we focus solely on determining which variables have an influence on the body mass index of children. Childhood obesity is known to be caused by an interplay of causes of quite different nature: medical history, dietary and behavioral habits and social environment. Of these groups, we consider $k = 4$ numerical variables and $p = 13$ factors described in Table S.5,where we also define the associated labels. For our analysis, we considered the $n = 925$ children that didn't have any missing values in the variables.

The results we obtained applying our methodology are summarized in Table S.6 in the form of posterior inclusion probabilities, the median inclusion probability and the highest posterior probability model. The results clearly corroborate the well-known importance of the medical history variables that are endorsed by very high inclusion probabilities. The case of Sex is an exception and it is uncertain if it has an effect on the body mass index; it has a posterior inclusion probability close to 1/2. Among the dietary habits, Meals5 and AfternoonSnack have a strong effect on the response while the others do not seem to have influence. Within the group of behavioral habits, the amount of time devoted to screens and the consumption of candy have an important effect on BMI, followed by sports activity. Finally, body mass index of children is not associated with any of the variables informing about the social environment.

S.6 Simulation regarding grouping

Experiment S.1. *For each value of* $\beta_{11} \in \{0.20, 0.50, 0.75, 1.00, 1.25, 1.50\}$ *we generated 10 simulated datasets from the model* (1) *with no covariates (*$k = 0$*), only one factor* $p = 1$ *with effect:* $\boldsymbol{\beta}^\top = (\beta_{11}, \dots, \beta_{11}, 0)$*. To clarify,* $\boldsymbol{\beta}$ *has the first* $\ell - 1$ *components equal to* β_{11} *and the last is zero. In our simulation, we have* $n = 100$ *sampled units of which the first* 80 *were randomly assigned to the first* $\ell - 1$ *levels of the factor and the remaining* 20 *to the last level. Here, the intercept was fixed at zero and the standard deviation at one.*

In the true generating model, the effect of the first $\ell - 1$ *levels can be collapsed into a single one. Hence the true model can be simply expressed as:*

$$M : y_{i1} = \beta_0 + \beta_{11} + \epsilon_{i1}, \;\; \textit{for} \;\; i = 1, \dots, 80, \;\; y_{i2} = \beta_0 + \epsilon_{i2}, \;\; \textit{for} \;\; i = 81, \dots, 100.$$

We have considered three values of the number of levels, namely, $\ell \in \{3, 5, 10\}$*. We now compute the posterior probability of* M *above versus the null model (with*

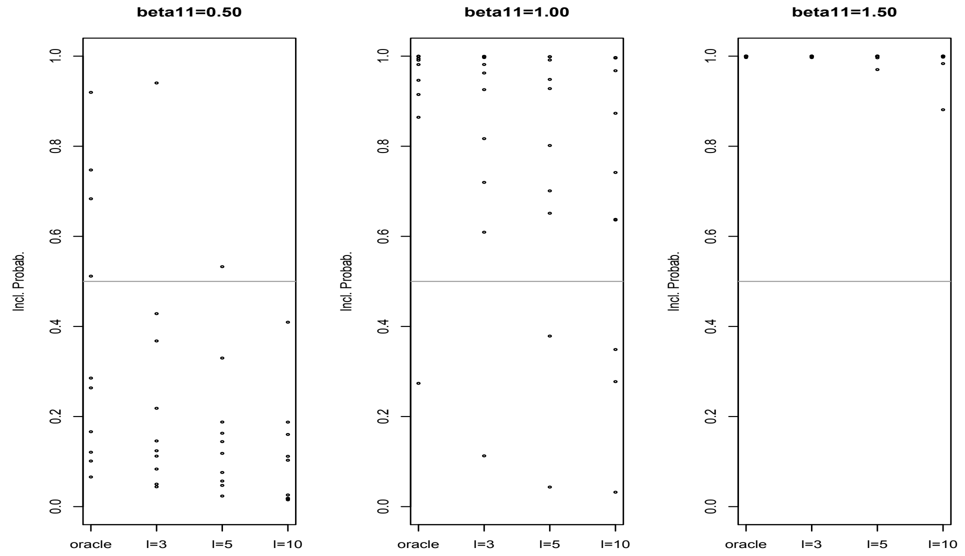

Figure S.1: For experiment S.1: posterior probability of the data generating process (oracle) and posterior inclusion probability using our methodology.

only the intercept). This probability, that only considers two models, can be easily computed with standard software (Forte et al., 2018). Note that this probability is the best evidence in favor of the factor that we could obtain as it is computed knowing exactly the distribution of the coincidental levels of the factor. We call it "oracle" and it is, for the different datasets, represented in Figure S.1 (in the interest of space, only three values of β_{11} *are shown; these are representative). Using our methodology, we have obtained the inclusion probability of the factor and the results are also collected in Figure S.1 in the form of points.*

The "loss of power" effect is visible in the form of a smaller evidence (compared to the oracle) in favor of the factor. As expected, the loss increases as the number of levels increase, but remarkably the increase is very smooth. In many cases, the evidence reported is remarkably comparable with the "oracle". Furthermore, in terms of reporting whether the factor is relevant or not (probability larger than 0.5) our method and the oracle coincide, regardless of whether the effect of the factor is small

$\beta_{11} = 0.2$ *or large* $\beta_{11} \geq 1.25$ *(even when* $\ell = 10$*) and has a quite satisfactory behavior in the rest of the cases (the worst scenario being* $\beta_{11} = 0.5$ *where the oracle detects the factor 4 times out of ten while our methodology detects the presence of the factor 1 time in* $\ell = 3$ *or* $\ell = 5$ *and none in* $\ell = 10$*).*

•

The example above serves to illustrate that, when the effect of certain levels of factors is equal, the loss of power in summarizing the evidence of the factor is far from being dramatic. This is because, given the model selection nature of our approach, many submodels are considered that provide a reasonable fit at a reduced cost in terms of the number parameters used and that's why the factor preserves a large proportion of the evidence. In our opinion, these are very satisfactory results, given the simplicity of our methodology and its completely absence of any tuning parameters.

References

Bayarri, M. J. and G. García-Donato (2007). Extending conventional priors for testing general hypotheses in linear models. *Biometrika 94*(1), 135–152.

Fernández, C., E. Ley, and M. Steel (2002). Bayesian modeling of catch in a northwest Atlantic fishery. *Applied Statistics 51*, 257–280.

Forte, A., G. Garcia-Donato, and M. F. Steel (2018). Methods and tools for Bayesian variable selection and model averaging in normal linear models. *International Statistical Review 86*(2), 237–258.

Lee, K.-J. and R.-B. Chen (2015). BSGS: Bayesian sparse group selection. *The R journal 7*(2), 122–133.

Zurriaga, O., J. Perez-Panades, J. Izquiero, M. Gil, Y. Anes, C. Quiñones, M. Margolles, A. Lopez-Maside, A. T. Vega-Alonso, and M. T. Miralles (2011). Factors associated with childhood obesity in spain. the obice study: a case–control study based on sentinel networks. *Public Health Nutrition 14*(6), 1105–113.

	$\boldsymbol{\delta}_1$		$\boldsymbol{\delta}_2$								$\boldsymbol{\delta}_1$		$\boldsymbol{\delta}_2$						
γ_1	δ_{11}	δ_{12}	δ_{21}	δ_{22}	δ_{23}	$\in \mathcal{M}$?	τ_1	τ_2	$\kappa(\boldsymbol{\gamma},\boldsymbol{\delta})$	γ_1	δ_{11}	δ_{12}	δ_{21}	δ_{22}	δ_{23}	$\in \mathcal{M}$?	τ_1	τ_2	$\kappa(\boldsymbol{\gamma},\boldsymbol{\delta})$
0	0	0	0	0	0	y	0	0	0	1	0	0	0	0	0	y	0	0	1
0	0	0	0	0	1	y	0	1	1	1	0	0	0	0	1	y	0	1	2
0	0	0	0	1	0	y	0	1	1	1	0	0	0	1	0	y	0	1	2
0	0	0	0	1	1	n	-	-	-	1	0	0	0	1	1	n	-	-	-
0	0	0	1	0	0	y	0	1	1	1	0	0	1	0	0	y	0	1	2
0	0	0	1	0	1	n	-	-	-	1	0	0	1	0	1	n	-	-	-
0	0	0	1	1	0	n	-	-	-	1	0	0	1	1	0	n	-	-	-
0	0	0	1	1	1	y	0	1	2	1	0	0	1	1	1	y	0	1	3
0	0	1	0	0	0	n	-	-	-	1	0	1	0	0	0	n	-	-	-
0	0	1	0	0	1	n	-	-	-	1	0	1	0	0	1	n	-	-	-
0	0	1	0	1	0	n	-	-	-	1	0	1	0	1	0	n	-	-	-
0	0	1	0	1	1	n	-	-	-	1	0	1	0	1	1	n	-	-	-
0	0	1	1	0	0	n	-	-	-	1	0	1	1	0	0	n	-	-	-
0	0	1	1	0	1	n	-	-	-	1	0	1	1	0	1	n	-	-	-
0	0	1	1	1	0	n	-	-	-	1	0	1	1	1	0	n	-	-	-
0	0	1	1	1	1	n	-	-	-	1	0	1	1	1	1	n	-	-	-
0	1	0	0	0	0	n	-	-	-	1	1	0	0	0	0	n	-	-	-
0	1	0	0	0	1	n	-	-	-	1	1	0	0	0	1	n	-	-	-
0	1	0	0	1	0	n	-	-	-	1	1	0	0	1	0	n	-	-	-
0	1	0	0	1	1	n	-	-	-	1	1	0	0	1	1	n	-	-	-
0	1	0	1	0	0	n	-	-	-	1	1	0	1	0	0	n	-	-	-
0	1	0	1	0	1	n	-	-	-	1	1	0	1	0	1	n	-	-	-
0	1	0	1	1	0	n	-	-	-	1	1	0	1	1	0	n	-	-	-
0	1	0	1	1	1	n	-	-	-	1	1	0	1	1	1	n	-	-	-
0	1	1	0	0	0	y	1	0	1	1	1	1	0	0	0	y	1	0	2
0	1	1	0	0	1	y	1	1	2	1	1	1	0	0	1	y	1	1	3
0	1	1	0	1	0	y	1	1	2	1	1	1	0	1	0	y	1	1	3
0	1	1	0	1	1	n	-	-	-	1	1	1	0	1	1	n	-	-	-
0	1	1	1	0	0	y	1	1	2	1	1	1	1	0	0	y	1	1	3
0	1	1	1	0	1	n	-	-	-	1	1	1	1	0	1	n	-	-	-
0	1	1	1	1	0	n	-	-	-	1	1	1	1	1	0	n	-	-	-
0	1	1	1	1	1	y	1	1	3	1	1	1	1	1	1	y	1	1	4

Table S.4: List of models for the problem with one covariate ($k = 1$) and two factors ($p = 2$) with $\ell_1 = 2$ and $\ell_2 = 3$, respectively. It is specified which models define $\mathcal{M}$, the model space without repetitions. Here $\#\mathcal{M} = 20$, with dimension of models ranging between 0 and 4. $\mathcal{M}$ can be partitioned in 8 sets of cardinality $\#\mathcal{M}(\gamma_1 = 0, \tau_1 = 0, \tau_2 = 0) = 1$, $\#\mathcal{M}(0,0,1) = 4$, $\#\mathcal{M}(0,1,0) = 1$, $\#\mathcal{M}(0,1,1) = 4$, $\#\mathcal{M}(1,0,0) = 1$, $\#\mathcal{M}(1,0,1) = 4$, $\#\mathcal{M}(1,1,0) = 1$ and $\#\mathcal{M}(1,1,1) = 4$.

Description	Code	Factor	ℓ
Medical History			
Weight at birth	WeightBorn	N	-
Height at birth	HeightBorn	N	-
Is the father obese?	FaObese	Y	2
Is the mother obese?	MoObese	Y	2
Age	Age	N	-
Sex	Sex	Y	2
Dietary habits			
Place having lunch	Wherelunch	Y	3
Does he/she have 5 meals regularly?	Meals5	Y	2
Does he/she eat vegetables regularly?	Vegeta	Y	2
Does he/she eat fruit regularly?	Fruit	Y	2
Does he/she eat afternoon snack regularly?	AfternoonSnack	Y	2
Behavioral habits			
Was he/she breastfed?	Breastfeed	Y	2
Hours (daily) for screens	HrsScrDay	N	-
Sports activity	Sports	Y	6
Consumption of candies	Candies	Y	6
Social environment			
Social class of child	SocialClass	Y	4
Studies of mother	MoStudies	Y	4

Table S.5: Description of potential explanatory variables in the childhood obesity dataset

Variable	Inclusion Probs.	MPM	HPM
Medical History			
WeightBorn	1.00 ■	■	■
HeightBorn	1.00 ■	■	■
FaObese	1.00 ■	■	■
MoObese	1.00 ■	■	■
Age	1.00 ■	■	■
Sex	0.52 ■	■	
Dietary habits			
Wherelunch	0.29 ■		
Meals5	0.99 ■	■	■
Vegeta	0.31 ■		
Fruit	0.30 ■		
AfternoonSnack	0.96 ■	■	■
Behavioral habits			
Breastfeed	0.37 ■		
HrsScreen	0.91 ■	■	■
Sports	0.78 ■	■	
Candies	0.86 ■	■	■
Social environment			
SocialClass	0.49 ■		
MoStudies	0.28 ■		

Table S.6: For childhood obesity dataset, posterior inclusion probabilities (both numerical values and gray code); variables in the Median Inclusion probability model and in the Highest Posterior probability model.