Bayesian Variable Selection in Qualitative Models by Kullback–Leibler projections

J. A. Dupuis[1] and C. P. Robert[2]

[1] *Laboratoire de Statistique et Probabilité, Université Paul Sabatier, 118 Route de Narbonne, 31062 Toulouse, Cedex, France, email:* dupuis@cict.fr

[2] *Laboratoire de Statistique, CREST–INSEE, Timbre J340 75675 Paris cedex 14, France, email:* robert@ensae.fr

Abstract

The Bayesian variable selection method proposed in the paper is based on the evaluation of the Kullback-Leibler distance between the full (or encompassing) model and the submodels. The implementation of the method does not require a separate prior modeling on the submodels since the corresponding parameters for the submodels are defined as the Kullback-Leibler projections of the full model parameters. The result of the selection procedure is the submodel with the smallest number of covariates which is at an acceptable distance of the full model. We introduce the notion of explanatory power of a model and scale the maximal acceptable distance in terms of the explanatory power of the full model. Moreover, an additivity property between embedded submodels shows that our selection procedure is equivalent to select the submodel with the smallest number of covariates which has a sufficient explanatory power. We illustrate the performances of this method on a breast cancer dataset, where they appear to be quadratic (as opposed to exponential) in the number of covariates.

Keywords: Additivity property; Kullback-Leibler distance; logit model; transitivity; entropy; explanatory power.

AMS Subject Classification (1991): 62A15, 62F15, 62J02, 62H15.

1. Introduction

Bayesian model choice is usually based on the premise that posterior probabilities, Bayes factors or related quantities should be compared, according to various scales. As noted by Gelfand and Dey (1994), there is no agreement on the Bayesian course of action in this setup as the problem can be stated in many different formats. For instance, consider the ongoing debate on the alternative extensions to noninformative setups with the solutions of Aitkin (1991), O'Hagan (1995), or Berger and Perrichi (1996), among others,

---

This paper is dedicated to the memory of Costas Goutis, who died in an accident in July 96. He has been influential in the development of the projection approach to testing.

and the difficulty to use improper priors in this particular setup. (See Gelfand and Dey (1994), George and McCullogh (1994), Carlin and Chib (1995), Kass and Raftery (1995), or Raftery, Madigan and Volinsky (1996) for recent perspectives on the whole issue of Bayesian model choice.) Most of these different approaches require prior specifications for each possible submodel, with at least a proper prior on each (sub)set of parameters and often a prior weight attached to each submodel. The complexity of this modeling is at odds with the initial parsimony requirement inherent to variable selection and it creates difficulties and adhocqueries even in moderately informative settings, as discussed in Goutis and Robert (1997). For instance, usual prior modeling rules imply that the weights depend on the number of submodels which are considered, notwithstanding the prior information and the tree structure of the submodels. Automated prior selection methods as in Bernardo (1979) and McCullogh and Rossi (1993) also encounter difficulties, on either implementation or theoretical grounds.

The alternative of model averaging advocated by Phillips and Smith (1996) and Madigan and Raftery (1994) encounters similar difficulties, while formally falling outside the model choice category since this procedure does not propose a particular (sub)model as its output and, similarly, fails the parsimony requirement.

The variable selection strategy we advocate in this paper and illustrate for various qualitative models has already been defended in Goutis and Robert (1994) for testing in generalized linear models. Contrary to existing methods, it only requires a (possibly improper) prior distribution on the full model. The submodels under consideration are projections of the full model, namely the closest submodels in the sense of the Kullback-Leibler distance to the full model. This definition thus avoids measure theoretic difficulties of defining a prior distribution on a set of prior measure zero, as in McCullogh and Rossi (1993) who consider the distribution of the projected parameters. Besides addressing variable selection issues, our extension of Goutis and Robert (1994) puts additional emphasis on the inferential interpretation of the Kullback-Leibler distance, which allows us to solve the scaling problem.

The general principle of the method is presented in Section 2, while the derivation of the Kullback-Leibler projections is detailed in Sections 3 (in the discrete case) and 4 (in the logit and polylogit cases). Section 5 examines the important issue of scaling the Kullback-Leibler distance and of deriving the proper bound which determines the answer to the variable selection problem. Section 6 proposes an algorithmic implementation of the method, including an excursion path in the submodel tree, which is illustrated in Section 7 on a breast cancer dataset studied in Raftery and Richardson (1996). Although there is no theoretical guarantee, our tree descent method appears to give the selected submodel in a quadratic time, thus avoiding the combinatoric explosion observed in usual variable selection techniques.

2. Variable selection

2.1. Distance between models.

We consider $p+1$ random variables y, $x^1, \ldots, x^k, \ldots, x^p$, where y is a qualitative variable which represents the phenomenon under scrutiny and the x^k's are either discrete or continuous covariates. As understood in this paper, the goal of variable selection is to reduce as much as possible the dimension of the covariates vector from the original p while preserving enough of the *explanatory power* of the full model, this notion being rigorously defined below. The decision to select a particular submodel is based on n i.i.d. replications of the random vector (y, x), with $x = (x^1, \ldots, x^p)$. The i-th random variable is denoted (y_i, x_i) with $x_i = (x_i^1, \ldots, x_i^k, \ldots, x_i^p)$.

First, consider the case where all covariates are discrete. We denote by $\alpha_j(x)$ the probability $P(y_i = j | x_i = x)$ $(j = 1, \ldots, J)$. When p is small (in practice, $p \leq 4$), as in usual contingency tables or in some discriminant analyses, the suppression of a given covariate, say x^1, is generally associated with the point null hypothesis

$$\mathrm{H}_0: \quad \forall j = \{1, \ldots, J\}, \quad \forall (u, v) \in \mathcal{X}_1^2, \quad \alpha_j(u; s_2, \ldots, s_p) - \alpha_j(v; s_2, \ldots, s_p) = 0, \tag{2.1}$$

where $(s_2, \ldots, s_p)$ represents any value of $(x^2, \ldots, x^p)$. See for instance Santner and Duffy (1989).

When some covariates are continuous, or when they are discrete but p is large, a parameterized covariate dependent model can be considered instead, as for instance in a generalized linear model,

$$P(y_i = j | x_i, \alpha) = \Phi_j(x_i^t \alpha), \qquad j = 1, \ldots, J, \ i = 1, \ldots, n, \ \alpha \in \mathbb{R}^p . \tag{2.2}$$

In this setup, the statistical issues related with variable selection are usually expressed in terms of null hypotheses on the parameters of (2.2). For instance, null hypotheses are of the form

$$\mathrm{H}_0: \quad \alpha_{i_1} = \ldots = \alpha_{i_q} = 0 \tag{2.3}$$

for a subset $\{i_1, \ldots, i_q\}$ of $\{1, \ldots, p\}$. In both discrete and continuous cases, an approach to variable selection based on such null hypotheses represents a drastic simplification and mostly a misrepresentation of the genuine purposes of the experimenter. As such, it does induce a substantial bias in the subsequent inference (see, e.g., Dupuis, 1997; Goutis and Robert, 1997). In particular, while the goal of the experimenter is to preserve most of the explanatory power of his/her model at a lower cost in terms of number of covariates, the exact nullity of the coefficients in (2.1) or (2.3) is generally meaningless. Indeed, the discontinuities at $\alpha_{i_j} = 0$ are not duplicated by the predictive performances of the corresponding models, i.e. do not induce major changes in the explanatory power of the model.

In the remainder of the paper, we refer to the setting associated with discrete covariates, small values of p and the null hypothesis (2.1) as the 'discrete case', and consider the second setting, associated with model (2.2), only for the logit and polylogit models, although our developments easily extend to other qualitative generalized linear models.

As mentioned in the Introduction, we consider the model selection alternative derived from Goutis and Robert (1994). (See also Mengersen and Robert (1996) for a first use in the setup of mixtures, and Dupuis (1994, 1997) for the use in a Bayesian test of homogeneity for Markov chains.) The principal motivation for this approach is based on the above argument that only a major reduction in the explanatory power of a submodel (when compared with the full model) must lead to the rejection for this submodel. The loss of *explanatory power* is defined through the evaluation of a comprehensive distance between the full model and the submodel of interest.

Several distances are acceptable candidates for this global perspective but the choice usually settles on the *Kullback-Leibler distance*,

$$d(f,g) = \int \log\left(\frac{f(z)}{g(z)}\right) f(z)dz \quad \text{or} \quad d(f,g) = \sum_z \log\left(\frac{f(z)}{g(z)}\right) f(z), \tag{2.4}$$

for information theoretic and intrinsic considerations (see Bernardo and Smith, 1994), as well as computational reasons. Moreover, this distance enjoys appreciable properties such as transitivity and additivity, in the settings of this paper, and it relates to the theory of generalized linear models, as shown below. (See also McCullogh and Rossi, 1993, or Kass and Wasserman, 1996, for other arguments.)

We want to stress at this point the fact that the choice of a functional distance between two densities f and g as an expression of the differences between the corresponding models appears as a sufficiently comprehensive summarizer to encompass all possible effects of a change from f to g. This representation of the explanatory power (or loss of) of a model is robust (or generic) enough to address the different uses of variable selection and thus to cover the possible uses of this selection by an arbitrary practitioner. The lack of symmetry of the Kullback-Leibler distance is far from being a deterrent in the model selection framework, since the formulation of the test is actually intrinsically asymmetric (as pointed out in Dupuis, 1997): indeed, in this context, only the full model matters, the submodels being approximations to the full model, whether they are acceptable or not. Note that the *I-divergence* of Csiszár (1975), which is defined in the opposite way, namely in terms of the projected density g, is less compelling from this point of view.

2.2. Selection by projections.

Given a functional distance like (2.4), we rephrase the null hypothesis that a subset of the covariates gives an acceptable approximation of the full model (with density f), as

$$\mathrm{H}_0 : d(f, f^{\perp}) \leq \epsilon, \tag{2.5}$$

where $f^{\perp}$ denotes the projection of f onto the corresponding subspace (i.e. the closest element of this subspace in terms of distance d) and ϵ is the maximum acceptable distance or, in our terminology, the maximum loss in explanatory power.

The main issue in this approach is scaling, i.e. the choice of the bound ϵ. First, it can be determined by selecting an appropriate difference for simple distributions. For instance, in the case of contingency tables, an appropriate scaling distribution is the binomial $\mathcal{B}(n, p)$ distribution, to compare with the binomial $\mathcal{B}(n, 0.5)$ distribution. McCullogh and Rossi (1989) and Goutis and Robert (1994) provide some developments about the choice of ϵ for binomial, Poisson and normal distributions. An alternative is proposed in Dupuis (1994), by derivation of an upper bound on the maximum Kullback-Leibler distance $d(f, f^{\perp})$, and it could apply in the present setup, as shown in Section 5, in the sense that ϵ is a percentage of this upper bound. We actually opt for a third scaling approach which seems closer to modeling purposes and relates to the loss in explanatory power compared with the full model. This loss is derived from the distance between the full model and the covariate free model, by choosing ϵ as a percentage ρ (say, $\rho = 5\%$ or $\rho = 10\%$) of this explanatory power. We justify this choice by a deeper analysis in Section 5. Moreover we stress that, by virtue of an additivity property between embedded models (see Propositions 3.2 and 4.1), this scaling approach is equivalent to impose that submodels have an *explanatory power* which is at least $(1 - \rho)$ % of the *explanatory power* of the full model (see Section 5 for details).

2.3. Conditional issues.

Consider the density $f(x, y)$ of the full model, denoted M_g, for the qualitative variable y and the covariates. Denote by $\theta = (\alpha, \xi)$ the parameter of the joint distribution, where ξ

is the parameter associated with the density of x, $f(x|\xi)$, and α is the parameter associated with the conditional density $f(y|x,\alpha)$. (Note that α has different meanings in the discrete and logit settings.) For $\mathcal{A} \subset \{1,\ldots,p\}$ and $x_{\mathcal{A}}$ the associated subset of the covariate set x, $\mathcal{M}_{\mathcal{A}}$ denotes the class of submodels such that the density of (y,x) is

$$g(x,y) = g(y|x_{\mathcal{A}},\alpha) f(x|\xi).$$

The parameter of the Kullback-Leibler projection of $f(x,y)$ on $\mathcal{M}_{\mathcal{A}}$, $\theta^{\perp} = (\alpha^{\perp},\xi^{\perp})$, then satisfies

$$\begin{aligned} d(f(x,y|\theta), g(x,y|\theta^{\perp}) &= d(f(x|\xi), f(x|\xi^{\perp})) + \mathbb{E}_x[d(f(y|x,\alpha), g(y|x_{\mathcal{A}},\alpha^{\perp}))] \\ &= \mathbb{E}_x[d(f(y|x,\alpha), g(y|x_{\mathcal{A}},\alpha^{\perp}))] \end{aligned} \tag{2.6}$$

since $\xi^{\perp} = \xi$. Therefore, only the second term, $\mathbb{E}_x[d(f(y|x,\alpha), g(y|x_{\mathcal{A}},\alpha^{\perp}))]$, is relevant for the variable selection procedure.

Although this formulation of (2.5) involves the conditional distribution of y, it requires evaluating $\mathbb{E}_x[d(f(y|x,\alpha), g(y|x_{\mathcal{A}},\alpha^{\perp}))]$ for the joint density of (x,y). In some discrete cases, the term $\mathbb{E}_x[d(f(y|x,\alpha), g(y|x_{\mathcal{A}},\alpha^{\perp}))]$ can be computed (see Section 3), but, in the continuous case, this joint distribution is most often unknown and we propose to use instead the approximation

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_y\left[\log\left(\frac{f(y|x_i,\alpha)}{g(y|x_{\mathcal{A}},\alpha^{\perp})}\right)\middle| x_i\right],$$

which a.s. converges to $\mathbb{E}_x[d(f(y|x,\alpha), g(y|x,\alpha^{\perp}))]$ as n goes to infinity.

The test is then implemented in a conditional version for continuous covariates. For instance, if $\overline{x}_{\mathcal{A}}$ denotes the complement covariates to $x_{\mathcal{A}}$, the component $\overline{x}_{\mathcal{A}}$ of x is thus eliminated when

$$\frac{1}{n}\sum_{i=1}^{n} d(f(y|x_i,\alpha), g(y|x_i,\alpha^{\perp})) < \epsilon \tag{2.7}$$

where $g(y|x_i,\beta)$ is such that $\Phi_j(x_i^t\beta) = \Phi_j(\sum_{v\in\mathcal{A}}\beta_v x_{i,v})$ and $\alpha^{\perp}$ is the Kullback-Leibler projection of $\alpha \in \mathbb{R}^p$ in the space where $\alpha_l = 0$ for all $l \notin \mathcal{B}$.

In a Bayesian framework, condition (2.7) can be assessed by evaluating the posterior expectation of $d(f(y|x_i,\alpha), g(y|x_i,\alpha^{\perp}))$ and comparing it to the upper bound ϵ, rather than by computing the posterior probability of (2.5), which requires the determination of the acceptance probabilities, given the embedded nature of the test. Since this expectation can rarely be computed in closed form in continuous setups, as shown in Section 4, it has to approximated by an MCMC algorithm (see Section 6.1).

2.4. *Operational principle.*

If $d(\mathrm{M}_g, \mathcal{M}_\mathcal{A})$ denotes the distance between the full model M_g and its projection on the class $\mathcal{M}_\mathcal{A}$,

$$d(\mathrm{M}_g, \mathcal{M}_\mathcal{A}) = \mathbb{E}_x[d(f(y|x,\alpha), g(y|x_\mathcal{A}, \alpha^\perp))],$$

our selection principle is therefore articulated as follows:

Among all subsets $\mathcal{A}$ of covariates which are acceptable in the sense that

$$d(\mathrm{M}_g, \mathcal{M}_\mathcal{A}) < \epsilon$$

is satisfied, select the submodel with the smallest cardinal. In case of ex-æquos in terms of numbers of covariates, select the submodel which is closest to the full model for the distance (2.6).

This selection principle obviously follows from *Occam's razor* rule, since it selects the most parsimonious submodel among those which are compatible with the full complex model. As most variable choice techniques, this method is partially exhaustive in the sense that most cases have to be examined in descending, ascending or mixed trees, the embedded (smaller) submodels only being eliminated by the rejection of embedding (larger) submodels. Nonetheless, in Section 6, we propose an algorithmic implementation which appears to be quadratic (rather than exponential) in the number of covariates.

3. **Projections in the discrete case.**

Starting from $\mathcal{A} \subset \{1, \ldots, p\}$, we denote by $\beta(x_\mathcal{A})$ the vector of the $\beta_j(x_\mathcal{A}) = \Pr(y = j|x_\mathcal{A})$. For a subset $\mathcal{B}$ of $\mathcal{A}$, the density of (x, y) in the class $\mathcal{M}_\mathcal{B}$ is

$$h(x,y) = h(y|x_\mathcal{B})f(x|\xi) = \gamma_y(x_\mathcal{B})f(x|\xi).$$

An advantage of the discrete case, when compared with the general setup addressed in Section 4, is that the projection $(\beta^\perp(x_\mathcal{B}), \xi)$ of model $\mathrm{M}_\mathcal{A}$ on the class $\mathcal{M}_\mathcal{B}$ can be derived in closed form.

Proposition 3.1. – *The minimization program*

$$\operatorname{Arg}\min_{\gamma(x_\mathcal{B})} \mathbb{E}_{x_\mathcal{A}}\left[d(g(y|x_\mathcal{A}), h(y|x_\mathcal{B}))\right]$$

has a unique solution

$$\beta^\perp(x_\mathcal{B}) = (\mathbb{E}[\beta_j(x_\mathcal{A})|x_\mathcal{B}])_{j=1,\ldots,J}\,.$$

Proof.– Since

$$\mathbb{E}_x[d(f(y|x_{\mathcal{A}}), g(y|x_{\mathcal{B}})] = \sum_x f(x) \sum_j \beta_j(x_{\mathcal{A}}) \log\left(\frac{\beta_j(x_{\mathcal{A}})}{\gamma_j(x_{\mathcal{B}})}\right)$$

we have to solve the following minimization program for a given $x_{\mathcal{B}}$:

$$\operatorname{Arg}\min_{\gamma(x_{\mathcal{B}})} \sum_x f(x) \sum_{j=1}^{J} \beta_j(x_{\mathcal{A}}) \log[1/\gamma_j(x_{\mathcal{B}})] \tag{3.1}$$

under the constraint $\sum_j \gamma_j(x_{\mathcal{B}}) = 1$.

Let $\overline{x}_{\mathcal{B}}$ be the complement covariates to $x_{\mathcal{B}}$ in $x_{\mathcal{A}}$. The solution to (3.1) is given by

$$\begin{aligned}\beta_j^{\perp}(x_{\mathcal{B}}) &= \frac{\sum_{\overline{x}_{\mathcal{B}}} f(x_{\mathcal{B}}, \overline{x}_{\mathcal{B}})\beta_j(x_{\mathcal{B}}, \overline{x}_{\mathcal{B}})}{\sum_j \sum_{\overline{x}_{\mathcal{B}}} f(x_{\mathcal{B}}, \overline{x}_{\mathcal{B}})\beta_j(x_{\mathcal{B}}, \overline{x}_{\mathcal{B}})} = \sum_{\overline{x}_{\mathcal{B}}} f(\overline{x}_{\mathcal{B}}|x_{\mathcal{B}})\beta_j(x_{\mathcal{B}}, \overline{x}_{\mathcal{B}}) \\ &= \mathbb{E}_{\overline{x}_{\mathcal{B}}|x_{\mathcal{B}}}[\beta_j(x_{\mathcal{A}})] = \mathbb{E}[\beta_j(x_{\mathcal{A}})|x_{\mathcal{B}}]\ .\end{aligned}$$

■■

The solution $\beta^{\perp}(x_{\mathcal{B}})$ is thus the conditional expectation of $\beta(x_{\mathcal{A}})$ given $x_{\mathcal{B}}$. This means that $\beta^{\perp}(x_{\mathcal{B}})$ is the projection (for the L_2 norm) of $\beta(x_{\mathcal{A}})$ on the σ-algebra induced by $x_{\mathcal{B}}$. This remark implies that the usual properties of projections (transitivity, multiple projection theorem) apply. In addition, a remarkable additivity property on the distances between the projected models can be exhibited.

Proposition 3.2. –*Distances between embedded submodels are additive, in the sense that, when $\mathcal{C} \subset \mathcal{B} \subset \mathcal{A} \subseteq \{1, \ldots, p\}$,*

$$d(M_a, \mathcal{M}_{\mathcal{C}}) = d(M_a, \mathcal{M}_{\mathcal{B}}) + d(M_b, \mathcal{M}_{\mathcal{C}})\ ,$$

where M_a and M_b denote the projections of M_g on $\mathcal{M}_{\mathcal{A}}$ and on $\mathcal{M}_{\mathcal{B}}$, respectively.

Proof.– Omitting ξ for commodity, we denote by $\beta^{\perp}(x_{\mathcal{A}})$, $\beta^{\perp}(x_{\mathcal{B}})$ and $\beta^{\perp}(x_{\mathcal{C}})$ the parameters of the projection of the full model on the classes $\mathcal{M}_{\mathcal{A}}$, $\mathcal{M}_{\mathcal{B}}$ and $\mathcal{M}_{\mathcal{C}}$, respectively. By virtue of the transitivity property, $\beta^{\perp}(x_{\mathcal{C}})$ is simultaneously the projection of $\beta^{\perp}(x_{\mathcal{A}})$ on $\mathcal{M}_{\mathcal{C}}$ and the projection of $\beta^{\perp}(x_{\mathcal{B}})$ on $\mathcal{M}_{\mathcal{C}}$, and $\beta^{\perp}(x_{\mathcal{B}})$ is also the projection of $\beta^{\perp}(x_{\mathcal{A}})$ on $\mathcal{M}_{\mathcal{B}}$. Taking into account those remarks, it is easy to show the additive property is equivalent to

$$\mathbb{E}_{x_{\mathcal{A}}}\left[\sum_j \beta_j^{\perp}(x_{\mathcal{A}}) \log\left(\frac{\beta_j^{\perp}(x_{\mathcal{B}})}{\beta_j^{\perp}(x_{\mathcal{C}})}\right)\right] = \mathbb{E}_{x_{\mathcal{B}}}\left[\sum_j \beta_j^{\perp}(x_{\mathcal{B}}) \log\left(\frac{\beta_j^{\perp}(x_{\mathcal{B}})}{\beta_j^{\perp}(x_{\mathcal{C}})}\right)\right]\ .$$

Now, by virtue of Proposition 3.1, we have, for all j,

$$\begin{aligned}\mathbb{E}_{x_\mathcal{B}}\left[\beta_j^\perp(x_\mathcal{B})\log\left(\frac{\beta_j^\perp(x_\mathcal{B})}{\beta_j^\perp(x_\mathcal{C})}\right)\right] &= \mathbb{E}_{x_\mathcal{B}}\left[\mathbb{E}_{\overline{x}_\mathcal{B}|x_\mathcal{B}}[\beta_j^\perp(x_\mathcal{A})]\log\left(\frac{\beta_j^\perp(x_\mathcal{B})}{\beta_j^\perp(x_\mathcal{C})}\right)\right]\\ &= \mathbb{E}_{x_\mathcal{B}}\,\mathbb{E}_{\overline{x}_\mathcal{B}|x_\mathcal{B}}\left[\beta_j^\perp(x_\mathcal{A})\log\left(\frac{\beta_j^\perp(x_\mathcal{B})}{\beta_j^\perp(x_\mathcal{C})}\right)\right]\\ &= \mathbb{E}_{x_\mathcal{A}}\left[\beta_j^\perp(x_\mathcal{A})\log\left(\frac{\beta_j^\perp(x_\mathcal{B})}{\beta_j^\perp(x_\mathcal{C})}\right)\right]\,.\end{aligned}$$

■■

4. Dichotomous and polychotomous regression models

4.1. Logit model.

If $y_i \in \{0,1\}$ and $x_i \in \mathbb{R}^k$ are related by a logit model

$$P(y_i=1|x_i,\alpha) = 1 - P(y_i=0|x_i,\alpha) = \frac{\exp(\alpha^t x_i)}{1+\exp(\alpha^t x_i)}\,, \tag{4.1}$$

the projection $\alpha^\perp$ on the subspace corresponding to the covariates z_i is associated with β, solution of the minimization program

$$\min_\beta \sum_{i=1}^n \left\{(\alpha^t x_i - \beta^t z_i)\frac{\exp\alpha^t x_i}{1+\exp\alpha^t x_i} - \log\left(\frac{1+\exp\alpha^t x_i}{1+\exp\beta^t z_i}\right)\right\}\,,$$

i.e. of the implicit equations (in β)

$$\sum_{i=1}^n \frac{\exp\beta^t z_i}{1+\exp\beta^t z_i} z_i = \sum_{i=1}^n \frac{\exp\alpha^t x_i}{1+\exp\alpha^t x_i} z_i. \tag{4.2}$$

As already noticed in Goutis and Robert (1994), these equations are formally equivalent to the MLE equations for the logit model,

$$\sum_{i=1}^n \frac{\exp\beta^t z_i}{1+\exp\beta^t z_i} z_i = \sum_{i=1}^n y_i z_i,$$

with $\exp\alpha^t x_i/\{1+\exp\alpha^t x_i\}$ playing the role of the y_i's. This formal equivalence has practical relevance, since it guarantees the existence of the projection $\alpha^\perp$ of α as well as the availability of standard computing softwares for the practical derivation of this projection (or simple Newton-Raphson procedures, see McCullagh and Nelder, 1989, or Jensen *et al.*, 1991).

An additional interesting feature of the logit model pertains to variable choice since, in this case too, *Kullback-Leibler projections are transitive.* Indeed, if ω_i is a subvector of z_i and if the corresponding parameter is γ, the projection of β is solution of

$$\sum_{i=1}^{n} \frac{\exp \gamma^t \omega_i}{1+\exp \gamma^t \omega_i} \omega_i = \sum_{i=1}^{n} \frac{\exp \beta^t z_i}{1+\exp \beta^t z_i} \omega_i . \tag{4.3}$$

Since the l.h.s. of (4.3) is a subvector of the r.h.s. of (4.2), the solution of (4.3) is also solution to

$$\sum_{i=1}^{n} \frac{\exp \gamma^t \omega_i}{1+\exp \gamma^t \omega_i} \omega_i = \sum_{i=1}^{n} \frac{\exp \alpha^t x_i}{1+\exp \alpha^t x_i} \omega_i \tag{4.4}$$

if β is solution of (4.2). This property is sufficient to establish that our selection procedure is coherent, in both senses that the order of selection of the eliminated covariates is not relevant and that embedded models are farther away from the full model than embedding models, as shown by the additive property below.

Note that the distance between the full model and a reduced model also involves several constant terms, since it is equal to

$$\frac{1}{n} \sum_i \frac{\exp \alpha^t x_i}{1+\exp \alpha^t x_i} (\alpha^t x_i - \beta^t z_i) - \log \left(\frac{1+\exp \alpha^t x_i}{1+\exp \beta^t z_i} \right) =$$

$$\mathrm{Ent}\,(\alpha) + \frac{1}{n} \sum_{i=1}^{n} \log \left(1+\exp \beta^t z_i\right) - \beta^t \sum_{i=1}^{n} \frac{\exp \alpha^t x_i}{1+\exp \alpha^t x_i} z_i ,$$

where $\mathrm{Ent}\,(\alpha)$ denotes the entropy associated with the full model, namely

$$\frac{1}{n} \sum_i \left\{ \frac{\exp \alpha^t x_i}{1+\exp \alpha^t x_i} \log \left(\frac{\exp \alpha^t x_i}{1+\exp \alpha^t x_i} \right) + \frac{1}{1+\exp \alpha^t x_i} \log \left(\frac{1}{1+\exp \alpha^t x_i} \right) \right\} .$$

Therefore, the distance to a reduced model only requires the computation of

$$\sum_{i=1}^{n} \frac{\exp \alpha^t x_i}{1+\exp \alpha^t x_i} \log \left(1+\exp \beta^t z_i\right) .$$

As in the discrete case, additivity holds in this setup.

PROPOSITION 4.1. –*Distances between embedded submodels are additive, in the sense that, when $\mathcal{C} \subset \mathcal{B} \subset \mathcal{A}$,*

$$d(M_a, \mathcal{M}_{\mathcal{C}}) = d(M_a, \mathcal{M}_{\mathcal{B}}) + d(M_b, \mathcal{M}_{\mathcal{C}}) .$$

Proof.– let $\beta^{\perp}(x_{\mathcal{U}})$ denotes the solution of (4.2), for subset $\mathcal{U}$, as in the proof of Proposition 3.2. Then $d(\mathrm{M}_a, \mathcal{M}_{\mathcal{C}})$ is equal to

$$
\begin{aligned}
&= \sum_i \frac{\exp \beta^{\perp}(x_{\mathcal{A}})^t x_{\mathcal{A}i}}{1+\exp \beta^{\perp}(x_{\mathcal{A}})^t x_{\mathcal{A}i}} (\beta^{\perp}(x_{\mathcal{A}})^t x_{\mathcal{A}i} - \beta^{\perp}(x_{\mathcal{C}})^t x_{\mathcal{C}i}) - \log\left(\frac{1+\exp \beta^{\perp}(x_{\mathcal{A}})^t x_{\mathcal{A}i}}{1+\exp \beta^{\perp}(x_{\mathcal{C}})^t x_{\mathcal{C}i}}\right) \\
&= \sum_i \frac{\exp \beta^{\perp}(x_{\mathcal{A}})^t x_{\mathcal{A}i}}{1+\exp \beta^{\perp}(x_{\mathcal{A}})^t x_{\mathcal{A}i}} (\beta^{\perp}(x_{\mathcal{A}})^t x_{\mathcal{A}i} - \beta^{\perp}(x_{\mathcal{B}})^t x_{\mathcal{B}i}) - \log\left(\frac{1+\exp \beta^{\perp}(x_{\mathcal{A}})^t x_{\mathcal{A}i}}{1+\exp \beta^{\perp}(x_{\mathcal{B}})^t x_{\mathcal{B}i}}\right) \\
&\quad + \sum_i \frac{\exp \beta^{\perp}(x_{\mathcal{A}})^t x_{\mathcal{A}i}}{1+\exp \beta^{\perp}(x_{\mathcal{A}})^t x_{\mathcal{A}i}} (\beta^{\perp}(x_{\mathcal{B}})^t x_{\mathcal{B}i} - \beta^{\perp}(x_{\mathcal{C}})^t x_{\mathcal{C}i}) - \log\left(\frac{1+\exp \beta^{\perp}(x_{\mathcal{B}})^t x_{\mathcal{B}i}}{1+\exp \beta^{\perp}(x_{\mathcal{C}})^t x_{\mathcal{C}i}}\right) \\
&= \sum_i \frac{\exp \beta^{\perp}(x_{\mathcal{A}})^t x_{\mathcal{A}i}}{1+\exp \beta^{\perp}(x_{\mathcal{A}})^t x_{\mathcal{A}i}} (\beta^{\perp}(x_{\mathcal{A}})^t x_{\mathcal{A}i} - \beta^{\perp}(x_{\mathcal{B}})^t x_{\mathcal{B}i}) - \log\left(\frac{1+\exp \beta^{\perp}(x_{\mathcal{A}})^t x_{\mathcal{A}i}}{1+\exp \beta^{\perp}(x_{\mathcal{B}})^t x_{\mathcal{B}i}}\right) \\
&\quad + \sum_i \frac{\exp \beta^{\perp}(x_{\mathcal{B}})^t x_{\mathcal{B}i}}{1+\exp \beta^{\perp}(x_{\mathcal{B}})^t x_{\mathcal{B}i}} (\beta^{\perp}(x_{\mathcal{B}})^t x_{\mathcal{B}i} - \beta^{\perp}(x_{\mathcal{C}})^t x_{\mathcal{C}i}) - \log\left(\frac{1+\exp \beta^{\perp}(x_{\mathcal{B}})^t x_{\mathcal{B}i}}{1+\exp \beta^{\perp}(x_{\mathcal{C}})^t x_{\mathcal{C}i}}\right)
\end{aligned}
$$

by virtue of (4.3). ■■

4.2. Polylogit regression.

The above section only applies for dichotomous variables y_i. However, the framework relevant for most population studies is often multinomial. The natural extension to the previous section is to use a polylogit modeling, namely to impose that

$$\frac{P(y_i = k)}{P(y_i = K)} = \exp(\alpha_k^t x_i), \qquad k \neq K.$$

In this case, the Kullback-Leibler distance between the $f(y_i|x_i, \alpha)$'s and the $g(y_i|x_i, \beta)$'s is given by

$$\frac{1}{n} \sum_{i=1}^{n} \left\{ \sum_{k=1}^{K-1} \frac{\exp(\alpha_k^t x_i)}{1+\sum_{\ell=1}^{K-1} \exp(\alpha_\ell^t x_i)} (\alpha_k - \beta_k)^t x_i \right\} + \log\left(\frac{1+\sum_{\ell=1}^{K-1} \exp(\beta_\ell^t x_i)}{1+\sum_{\ell=1}^{K-1} \exp(\alpha_\ell^t x_i)}\right),$$

where $\alpha = (\alpha_1, \ldots, \alpha_{K-1})$. If β is a subvector of α, associated with the subvector z_i of x_i, the projection $\alpha^{\perp}$ is defined by the equations (in β) $(1 \leq k \leq K-1)$

$$\sum_{i=1}^{n} \frac{\exp(\alpha_k^t x_i)}{1+\sum_{\ell=1}^{K-1} \exp(\alpha_\ell^t x_i)} z_i = \sum_{i=1}^{n} \frac{\exp(\beta_k^t z_i)}{1+\sum_{\ell=1}^{K-1} \exp(\beta_\ell^t z_i)} z_i$$

which are again equivalent to the polylogit MLE equations and can thus be solved using softwares such as GLIM. Moreover, this property shows that the transitivity and additivity results of the previous section extend to this case.

5. Scaling the threshold.

In this section, the p covariates are either discrete or continuous. The first step of the selection procedure outlined in Section 2.4 is to eliminate the subsets $\mathcal{A}$ such that

$$d(\mathrm{M}_g, \mathcal{M}_{\mathcal{A}}) > \epsilon \tag{5.1}$$

before selecting the subsets with smallest cardinal. This procedure thus requires a scaling of the threshold ϵ. As in Dupuis (1994), the solution takes advantage of the fact that $d(\mathrm{M}_g, \mathcal{M}_0)$ is bounded (see Proposition 5.1), where the class $\mathcal{M}_0$ contains the covariate free submodels of M_g, that is those such that the density of (x, y) satisfies

$$g(x, y) = g(y|\alpha)f(x|\xi).$$

For instance, in the setup of Section 4, $\mathcal{M}_0$ corresponds to the submodels with only an intercept. A selection procedure can then be based on the choice of $\rho\ \in]0, 1[$ and the derivation of ϵ as $\rho\, d(\mathrm{M}_g, \mathcal{M}_0)$. Note that, due to the property of additivity of the distance d (see Propositions 3.2 and 4.1), we have

$$d(\mathrm{M}_g, \mathcal{M}_{\mathcal{A}}) > \rho\, d(\mathrm{M}_g, \mathcal{M}_0) \qquad \text{if and only if} \qquad d(\mathrm{M}_a, \mathcal{M}_0) < (1-\rho) d(\mathrm{M}_g, \mathcal{M}_0). \tag{5.2}$$

where M_a is the projection of M_g on the class $\mathcal{M}_{\mathcal{A}}$. As a consequence of (5.2), the selection procedure only considers submodels M_a such that their *relative explanatory power*,

$$\mathcal{P}_r(\mathrm{M}_a) = d(\mathrm{M}_a, \mathcal{M}_0)/d(\mathrm{M}_g, \mathcal{M}_0),$$

is greater than $1-\rho$. (See Proposition 5.2 for a justification of the interpretation of those distances as *explanatory power* of the corresponding models.)

PROPOSITION 5.1. – *The explanatory power $d(M_g, \mathcal{M}_0)$ is bounded from above by* $\log J$.

Proof.– The conditional entropy of y given x is

$$H(y|x) = -\sum_y f(y|x,\alpha) \log f(y|x,\alpha) \geq 0.$$

Therefore,

$$\begin{aligned} d(\mathrm{M}_g, \mathcal{M}_0) &= \mathbb{E}_x\left[d(f(y|x,\alpha), g(y|\alpha^{\perp}))\right] \\ &= -\mathbb{E}_x\left[H(y|x)\right] - \mathbb{E}_x\left[\mathbb{E}_y\left[\log g(y|\alpha^{\perp})|x\right]\right] \\ &\leq \mathbb{E}_x\left[\mathbb{E}_y\left[\log \frac{1}{g(y|\alpha^{\perp})}\middle| x\right]\right] \end{aligned} \tag{5.3}$$

where

$$\mathbb{E}_y\left[\log g(y|\alpha^\perp)|x\right] = \sum_y f(y|x,\alpha)\log g(y|\alpha^\perp).$$

We now establish that $\mathbb{E}_x\mathbb{E}_y[\log \frac{1}{g(y|\alpha^\perp)}]$ is bounded by $\log J$.

Since

$$\mathbb{E}_{(x,y)}\left[\log g(y|\alpha^\perp)\right] = \mathbb{E}_x\left[\sum_j \alpha_j(x)\log\alpha_j^\perp\right],$$

and Proposition 3.1 has shown that $\alpha_j^\perp = \mathbb{E}_x[\alpha_j(x)]$,

$$\mathbb{E}_x\left[\sum_j \alpha_j(x)\log\alpha_j^\perp\right] = \sum_j \mathbb{E}_x\left[\alpha_j(x)\right]\log\mathbb{E}_x\left[\alpha_j(x)\right] \ .$$

(Note that Proposition 3.1 also applies when the covariates are continuous, since $\mathcal{M}_0$ is made of the covariate free submodels.) Taking into account $\sum_j \mathbb{E}_x[\alpha_j(x)] = 1$ we have,

$$\mathbb{E}_{(x,y)}\left[\log\frac{1}{g(y|\alpha^\perp)}\right] = \sum_j \mathbb{E}_x[\alpha_j(x)]\log(1/\mathbb{E}_x[\alpha_j(x)]) \le \log J \ , \tag{5.4}$$

since this sum over j represents the entropy of a multinomial distribution with J cells. ■■

Note that this upper bound only depends on the number of modalities of y and is independent of the dimension of $\mathcal{X}$. Moreover, it is equal to the maximum entropy of y (since y is multinomial with J cells). An analogous result has been established in Dupuis (1994) in a longitudinal setup, which shows the importance of the inherent multinomial structure of the problem.

The above result can be refined by a determination of the conditions under which the upper bound of $d(\mathrm{M}_g,\mathcal{M}_0)$ is exactly $\log J$. Consider thus the subset of Θ,

$$\Theta^* = \{\theta\in\Theta : d(\mathrm{M}_g,\mathcal{M}_0) = \log J\} \ .$$

Proposition 5.2. – *Assume that $f(x)>0$. Then $d(\mathcal{M}_g, M_0) = \log J$ if and only if, for every x and j, $H(y|x) = 0$ and $\alpha_j^\perp = 1/J$.*

Proof.– We have the following equivalences:

$$\begin{aligned}
\theta\in\Theta^* &\iff d(\mathrm{M}_g,\mathcal{M}_0) = \log J \\
&\iff \mathbb{E}_x\left[H(y|x)\right] = 0 \quad \text{and} \quad \mathbb{E}_{(x,y)}\left[\log\frac{1}{g(y|\alpha^\perp)}\right] = \log J \qquad (5.5)\\
&\iff \forall\, x\in\mathcal{X}, \quad H(y|x) = 0 \quad \text{and} \quad \forall\, j = 1,\dots,J \quad \alpha_j^\perp = 1/J \ . \qquad (5.6)
\end{aligned}$$

Equivalence (5.5) is deduced from (5.3) and from (5.4). The first part of (5.6) follows from the assumption $f(x) > 0$. The second part is due to the fact that $\mathbb{E}_{(x,y)}[\log g(y|\alpha^{\perp})] = \sum_j \alpha_j^{\perp} \log \alpha_j^{\perp}$, which is the negentropy of a multinomial distribution. And, for the multinomial case with J cells, the upper bound of the entropy, is obtained when all the probabilities (the $\alpha_j^{\perp}$'s) are equal to $1/J$. ■■

It turns out that a process (x, y) whose explanatory power $d(\mathrm{M}_g, \mathcal{M}_0)$ reaches the upper bound $\log J$ is such that, given x, the variable y is deterministic. In other words, the variable y is entirely explained by the covariate vector x. This result reinforces our interpretation of the quantity $d(\mathrm{M}_g, \mathcal{M}_0)$ as a measure of the explanatory power of M_g. In addition, this result justifies, a posteriori, the expression *loss of explanatory power* we have used to call the quantity $d(\mathrm{M}_g, \mathcal{M}_{\mathcal{A}})$ since, it is actually equal to a difference of *explanatory powers* (namely $d(\mathrm{M}_g, \mathcal{M}_0) - d(\mathrm{M}_a, \mathcal{M}_0)$), by virtue of the additivity property.

Note, however, that, for a given set of covariates and a given value J, Θ^* can be empty, i.e. that the bound $\log J$ on $d(\mathrm{M}_g, \mathcal{M}_0)$ may be too large. In the discrete case, we can exhibit sufficient and necessary conditions on J, $\mathrm{N}=\prod_k \mathrm{N}_k$, where N_k is the number of modalities of the covariate x^k, and on the rank of the matrix A of the $\alpha_j(x)$'s, for this bound to be tight.

The problem is to solve (in $f(x)$ and A) the system

$$\forall\, j = 1, \ldots, J, \quad \sum_{x \in \mathcal{X}} f(x)\, \alpha_j(x) = \frac{1}{J} \quad \text{and} \quad \forall\, x \in \mathcal{X}, \quad \sum_{j=1}^{J} \alpha_j(x) \log \alpha_j(x) = 0 \qquad (5.5)$$

under the constraints

$$\forall\, x \in \mathcal{X}, \quad \sum_{j=1}^{J} \alpha_j(x) = 1 \qquad \text{and} \qquad \sum_{x \in \mathcal{X}} f(x) = 1.$$

Note that the nullity of the negentropy $\sum_j \alpha_j(x) \log \alpha_j(x)$ implies that there exists a unique j_x such that $\alpha_{j_x}(x) = 1$ and $\alpha_j(x) = 0$ for $j \neq j_x$. In addition, the l.h.s. of (5.5) can be written under the matricial representation

$$Af = \frac{1}{J}\mathbf{1}$$

where A is the matrix of the $\alpha_j(x)$'s, f is the vector of the $f(x)$'s and $\mathbf{1} = (1, \ldots, 1)^t \in \mathbb{R}^J$.

Consider the three exclusive cases: $J < N$, $J = N$ and $J > N$. In the first case, there is no solution since it leads to the contradiction $1/J = 0$. When $J = N$, for any collection of j_x such that the matrix A is invertible (this condition is satisfied if there is no more than one zero per row), the system has a same and unique solution, namely $f(x) = 1/J$ uniformly in x. When $J > N$, the system has an infinity of solutions as long as the rank of A is larger than J. Note again that a result analogous to Proposition 5.2 has been obtained by Dupuis (1994), when dealing with a Bayesian test of homogeneity for Markov chains, since the space Θ^* contains Markov chains whose entropy is zero.

As a side remark, note that Proposition 5.2 can also be used for an absolute scaling of the explanatory power of the model by comparison with the bound $\log J$.

6. Implementation issues

6.1. MCMC implementation.

As mentioned in the introduction, the method is forcibly distinct from a testing approach. From a Bayesian perspective, this signifies that the focus is on estimating the posterior distance between the (embedding) full model and some submodels. In the setup of Section 4, given a sample of α's produced from the posterior distribution for the full model by an MCMC algorithm (see Albert and Chib, 1993, Gilks *et al.*, 1996, or Robert, 1996), it is then possible to compute the projected samples for the submodels through the MLE equations (4.1) and to derive the distances to the full model by averaging (4.4) over the α's.

Instead of using the data augmentation steps of Albert and Chib (1993), we generate the MCMC sample of the α's via a random walk Hastings-Metropolis normal step, using the second order approximation of the posterior distribution, namely

$$L(\alpha) = \prod_{i=1}^{n} \frac{\exp(\alpha^t x_i)}{1 + \exp(\alpha^t x_i)}$$

in the case of a flat prior on α, to construct the variance. We also introduce a scale factor τ in the normal variance in order to control the acceptance rate, following the recommendations of Gelman *et al.* (1996). The Hastings-Metropolis proposal for $\alpha^{(t+1)}$ is thus generated as a normal $\mathcal{N}(\alpha^{(t)}, \tau^2(X^t \mathrm{diag}(\nabla L) X)^{-1})$.

6.2. Excursions in the submodel tree.

The projections of each point of the MCMC sample are derived via a standard Newton-Raphson algorithm implemented in `C` (programs are available from the authors upon request) with a ridge type stabilizing of the second derivative. Since the procedure selects the smallest submodel which is at a distance less than ϵ from the full model, it seems to call for an complete exploration of the submodel tree and is thus almost exhaustive in the computation of the estimators of the parameters of the submodels. The "almost" is due to the elimination of the farthest branches (submodels) of the tree by rejection of one of their ancestors. In order to minimize the number of submodels to be considered, we suggest below a particular type of excursion in the tree.

The basic remark underlying this exploration path is that, since the submodel to be selected is the model with the smallest number k_0 of covariates which is at an acceptable distance from the full model, there is no reason to consider a larger number of covariates if an upper bound on k_0 can be found. The two first steps of the algorithm provide such an upper bound. The first approach is a *downward* excursion which starts from the full model M_g and removes the covariate which reduces the less the explanatory power in terms of distance (2.4). The resulting submodel with $p-1$ covariates is M_{p-1}. Covariates are then successively removed, as previously, from models M_{p-j} $(j \geq 1)$ till the bound ϵ is exceeded. The last accepted submodel M_{k_1} gives a first upper bound k_1 on k_0. The second step in the algorithm proceeds symmetrically, being an *upward* step which starts from the constant model M_0 by adding successively the most explanatory covariate till the resulting submodel M_{k_2} is accepted, thus producing a second upper bound on k_0. In this upward step, covariates are ranked by order of (decreasing) importance through their contribution to the explanatory power of the full model.

The two next steps of the algorithm determine (a) whether any submodel with less than $\min(k_1, k_2)$ covariates is acceptable and, if not, (b) whether any submodel with $k_0 = \min(k_1, k_2)$ covariates is at a closer distance of the full model than the model obtained in the downward and upward steps. These two last steps are more time consuming. In particular, the last step (b) involves at most

$$\binom{p}{\min(k_1, k_2)}$$

submodels, since some of these have already been excluded through one of their ancestors. However, it appears in practice that both downward and upward steps always lead to the same submodel which is furthermore the best submodel (for the projection criterion) obtained after the four steps above. We thus suggest to run only the downward and upward steps to check for coincidence when time is at stake.

Note that the submodels considered in the above steps are not necessarily all the submodels embedded in the previous model and containing the previous model respectively, since some of these submodels may be removed in a previous step. Indeed, if one submodel is not acceptable in the downward step, all its descendents are not acceptable and should not be considered in the subsequent steps. Moreover, in the upward step, submodels whose ancestors have been rejected in the downward step do not need to be examined.

7. An illustration.

The projection method developed in this paper is used to select covariates in a logistic regression model for an epidemiological study already considered by Richardson *et al.* (1989) from a classical point of view and by Raftery and Richardson (1996), who were using Bayes factors and transformed variables using the ACE (Alternating Conditional Expectations) of Madigan and Raftery (1994).

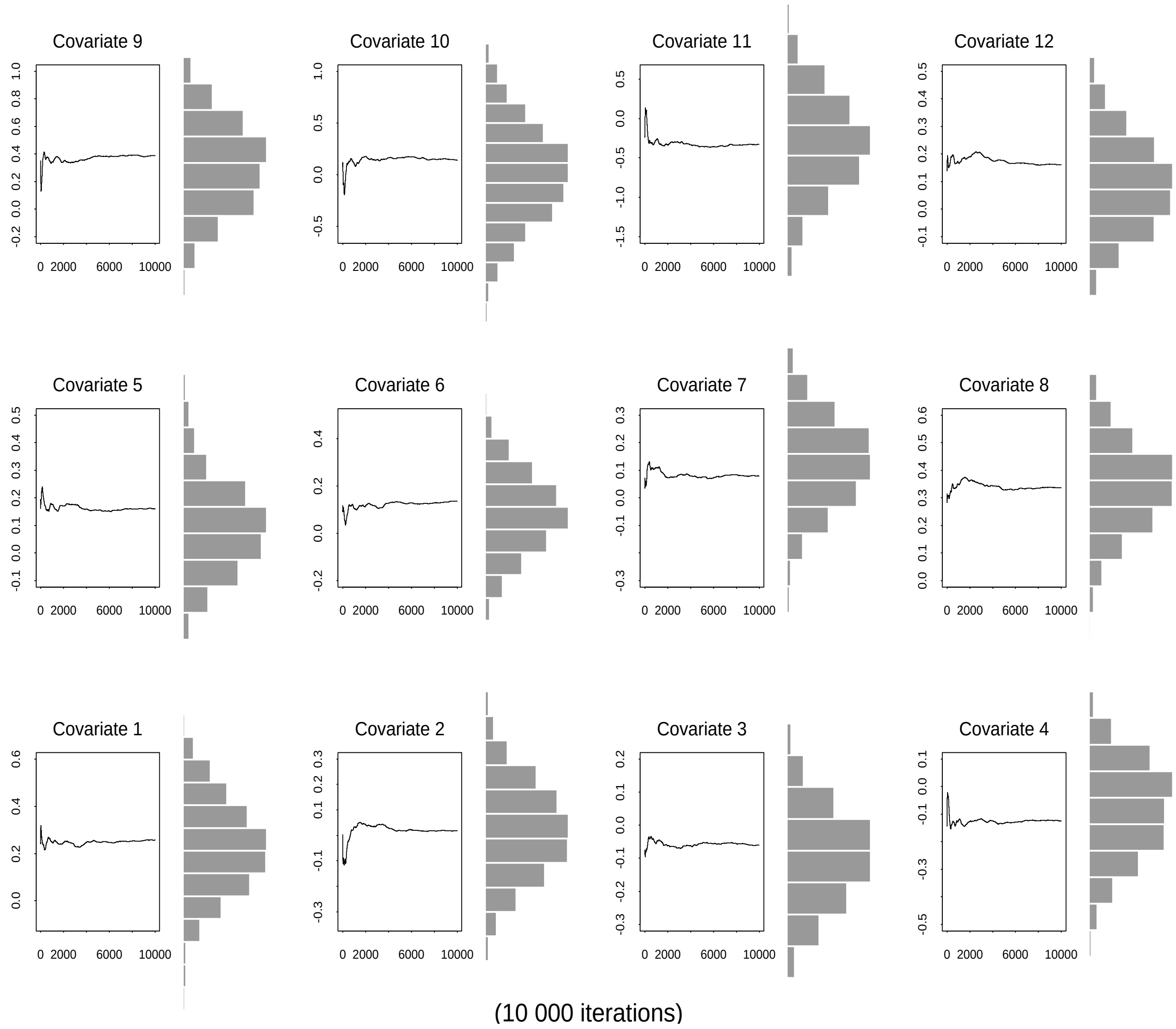

Figure 7.1 – Convergence curves and histograms for the Gibbs samples of the coefficients of the covariates in the logit model for the breast cancer dataset from Richardson *et al.* (1989).

The study evaluates the role of dietary factors on breast cancer and consists of 740 women from Montpelier (France) (after removal of the missing values). The 13 risk factors registered in the study were age, menauposal age, age at menarche, parity, family history of breast cancer, age of first pregnancy, age at the end of studies, Quetelet's body mass index (weight/squared height), alcohol consumption (number of drinks per week), fat intake (total and saturated), and history of benign breast disease. (See Raftery and Richardson (1996) for more details.) Some of these factors are categorical while others are continuous, and we standardized all the factors by subtracting the means and dividing by the standard deviation, since all submodels contain an intercept. Note also that we do not consider interactions, following Raftery and Richardson's (1996) findings.

The results of the selection procedure are provided in Table 7.1, which gives the sequence of submodels examined in the downward and upward steps, as well as the sequence of the submodels with the same number of covariates which are evaluated in the last step (the other being directly eliminated by rejection of an ancestor in the model tree). The first noteworthy feature of this sequence of submodels is that both the downward and the upward steps provide the correct submodel. As mentioned in the previous section, this phenomenon occurred in all cases we examined and we conjecture that it should be true in all cases except for pathological features of the covariate matrix.

The second fact of interest is, of course, the resulting submodel which excludes menauposal age, age at menarche, and both fat intakes, a result which somehow coincides with the findings of Raftery and Richardson (1996), although these authors transformed the covariates. For comparison purposes, note that the estimated distance between the full model and the constant probability model is 0.042. Since Raftery and Richardson (1996) imposed that the classical risk factors (age, menauposal status, age at menarche, parity, familial background, age at the end of studies, Quetelet's index) must be part of the model, we also ran the method under this constraint. The selected variable among alcohol, total fat and saturated fat, is again alcohol, as shown by Table 7.2. This second experiment illustrates the freedom allowed by our projection method. In models with small numbers of covariates, each possible submodel can be evaluated in terms of its explanatory power, and the comparison can be led in a more qualitative way, rather than follows from the reference to a strict level. For instance, in Table 7.1, Step `2.`, it may be preferable to choose the model `100111011001` with relative explanatory power 0.88, which is only slightly smaller than the selected 0.91, because it eliminates an additional covariate, namely age at the end of studies.

step	subset $\mathcal{A}$	$d(\mathrm{M}_g, \mathcal{M}_{\mathcal{A}})$ ($\times 740$)	$\mathcal{P}_r(\mathrm{M}_a)$
1.	101111111111	0.508	0.98
	101111111011	1.146	0.96
	100111111011	1.800	0.94
	100111111001	**2.726**	**0.91**
2.	000000010000	21.78	0.29
	000010010000	16.97	0.45
	100010010000	13.81	0.55
	100010011000	10.61	0.66
	100010011001	7.601	0.75
	100011011001	5.224	0.83
	100111011001	3.736	0.88
	100111111001	**2.726**	**0.91**
3.	111111110000	8.170	0.73
	111111001010	13.72	0.55
	111100111010	8.349	0.73
	110011111010	5.988	0.81
	001111111010	9.215	0.70
	111110011001	4.542	0.85
	111101011001	4.761	0.85
	111011011001	3.91	0.87
	110111011001	3.265	0.89
	101111011001	3.017	0.90
	011111011001	5.895	0.81
	100111111001	**2.726**	**0.91**
	100111011101	3.109	0.899
	100011111101	3.826	0.88
	111011010011	5.284	0.83
	110110110011	6.04	0.80
	101101110011	5.9	0.81
	101011011011	3.576	0.88
	100111011011	2.77	0.91
	101010111011	5.08	0.84
	011001111011	9.346	0.70
	100110011111	4.151	0.87
	100101011111	4.224	0.86
	100011011111	3.787	0.88
4.	101111011001	3.017	0.90
	100111111001	**2.726**	**0.91**
	100111011011	2.77	0.91

Table 7.1 – Successive steps of the selection method for the study in Raftery and Richardson (1996). (The submodel is represented by the indicators of the covariates, see text.) The selected submodel is in bold. The computations involve 5000 Gibbs iterations and the upper bound is a 10% loss of the explanatory power, 0.042, or equivalently, $\mathcal{P}_r(\mathrm{M}_a) > 0.9$. For each set of covariates, we indicate the distance to the full model and the relative explanatory power.

alcohol	total fat	saturated fat	distance($\times 740$)
0	0	0	5.346
1	0	0	1.334
0	1	0	2.950
0	0	1	2.745
1	0	1	0.623

Table 7.2 – Comparison of the effects of the factors alcohol, total fat and saturated fat on the distance to the full model.

Acknowledgments

The authors are grateful to Sylvia Richardson for sharing and discussing her dataset, as well as to Peter Müller and Mauro Pacifico for helpful discussions.

References

Aitkin, M. (1991) Posterior Bayes factors (with discussion). *J. Royal Statistical Society* B **53**, 111–142.

Albert, J.H. and Chib, S. (1993) Bayesian analysis of binary and polychotomous response data. *J. American Statistical Association* **88**, 669–679.

Berger, J.O. and Perrichi, L.R. (1996) The intrinsic Bayes factor for model selection and prediction. *J. American Statistical Association* **91**, 109–122.

Bernardo, J.M. (1979) Reference posterior distributions for Bayesian inference (with discussion). *J. Royal Statistical Society* B **41**, 113–147.

Carlin, B.P. and Chib, S. (1995) Bayesian model choice through Markov-Chain Monte-Carlo. *J. Royal Statist. Soc.* (Ser. B), **57**, 473–484.

Csiszár, I. (1975) *I*-divergence geometry of probability distributions and minimization problems. *Ann. Probability* **3**, 146–158.

Dupuis, J.A. (1994). Bayesian test of homogeneity for Markov chains with missing data by Kullback proximity. Tech. Report 9457, CREST, Paris.

Dupuis, J.A. (1997). Bayesian test of homogeneity for Markov chains. *Statistics and Probability Letters* **31**, 333-338.

George, E.I. and McCullogh, R.E. (1994) Variable selection via Gibbs sampling. *J. American Statistical Association* **89**, 881–889.

Gelfand, A. and Dey, D. (1994) Bayesian model choice: asymptotics and exact calculations. *J. Royal Statistical Society* B **56**, 501–514.

Gelman, A., Gilks, W.R. and Roberts, G.O. (1996) Efficient Metropolis jumping rules. In *Bayesian Statistics 5*, J.O. Berger, J.M. Bernardo, A.P. Dawid, D.V. Lindley and A.F.M. Smith (Eds.). Oxford University Press, Oxford, 599–608.

Gilks, W.R., Richardson, S. and Spiegelhalter, D.I. (1996) *Markov Chain Monte Carlo in Practice.* Chapman and Hall, London.

Goutis, C. and Robert, C.P. (1994) Model choice in generalized linear models: a Bayesian approach via Kullback–Leibler projections. *Biometrika* (to appear).

Goutis, C. and Robert, C.P. (1997) Choice among hypotheses using estimation criteria. *Ann. Eco. Statist.* (in press).

Jensen, S.T., Johansen, S. and Lauritzen, S.L. (1991) Globally convergent algorithms for maximizing a likelihood function. *Biometrika* **78**(4), 867-877.

Kass, R.E. and Raftery, A.E. (1995) Bayes factors. *J. American Statistical Association* **90**, 773–795.

Kass, R.E. and Wasserman, L. (1996) Discussion of "Posterior predictive assessment of model fitness via realized discrepancies" by Gelman, Meng and Stern. *Statistica Sinica* **6**, 774–779.

O'Hagan, A. (1995) Fractional Bayes factors for model comparison (with discussion). *J. Royal Statistical Society* B **57**, 99–118.

Madigan, D. and Raftery, A.E. (1994) Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. American Statistical Association* **89**, 1535–1546.

McCullagh, P. and Nelder, J. (1989) *Generalized Linear Models.* Chapman and Hall, London.

McCulloch, R. E. and Rossi, P. R. (1993) Bayes factors for nonlinear hypotheses and likelihood distributions. *Biometrika* **79**, 663–76.

Mengersen, K. and Robert, C. (1996) Testing for mixtures: a Bayesian entropic approach. In *Bayesian Statistics 5*, J.O. Berger, J.M. Bernardo, A.P. Dawid, D.V. Lindley and A.F.M. Smith (Eds.). Oxford University Press, Oxford, 255–276.

Phillips, D.B. and Smith, A.F.M. (1996) Bayesian model comparison via jump diffusions. In *Markov chain Monte-Carlo in Practice* (Ed. W.R. Gilks, S.T. Richardson and D.J. Spiegelhalter), 215–240. Chapman and Hall, London.

Raftery, A.E., Madigan, D. and Volinsky, C.T. (1996) Accounting for model uncertainty in survival analysis improves predictive performance. In *Bayesian Statistics 5*, J.O. Berger, J.M. Bernardo, A.P. Dawid, D.V. Lindley and A.F.M. Smith (Eds.). Oxford University Press, Oxford, 323–349.

Raftery, A.E. and Richardson, S. (1996) Model selection for generalized linear models via GLIB: application to nutrition and breast cancer. In *Bayesian Biostatistics*, D.A. Berry and D.K. Stangl (eds.), 321–353.. Marcel Dekker, New York.

Richardson, S., de Vincenzi, I., Gerber, M., and Pujol, H. (1989) Alcohol consumption in a case-control study of breast cancer in southern France. *International Journal of Cancer* **44**, 84–89.

Robert, C.P. (1996) *Méthodes de Monte Carlo par chaînes de Markov.* Economica, Paris.

Santner, T.J. and Duffy, D. (1989) *The Statistical Analysis of Discrete Data.* Springer-Verlag, New York.