Research Article

Statistics in Medicine

Received 24 January 2011, Accepted 2 August 2011 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4406

Objective Bayes model selection in probit models

Luis Leon-Novelo,[a] Elías Moreno[b] and George Casella[a*†]

We describe a new variable selection procedure for categorical responses where the candidate models are all probit regression models. The procedure uses objective intrinsic priors for the model parameters, which do not depend on tuning parameters, and ranks the models for the different subsets of covariates according to their model posterior probabilities. When the number of covariates is moderate or large, the number of potential models can be very large, and for those cases, we derive a new stochastic search algorithm that explores the potential sets of models driven by their model posterior probabilities. The algorithm allows the user to control the dimension of the candidate models and thus can handle situations when the number of covariates exceed the number of observations. We assess, through simulations, the performance of the procedure and apply the variable selector to a gene expression data set, where the response is whether a patient exhibits pneumonia. Software needed to run the procedures is available in the `R` package `varselectIP`.

Keywords: intrinsic priors; linear models; Bayes factors; model selection; probit models; stochastic search

1. Introduction

Categorical responses often appear in the analysis of the effect of a medical treatment, which is applied to a collection of patients. These patients are typically characterized by a set of p potential covariates, and a first important statistical problem in the presence of covariates is to reduce the dimension of the model by retaining from the original p covariates those that have some influence in the observed responses. Thus, we have a model selection problem in the class of 2^p models.

In this paper, we describe a new variable selection procedure applicable to dichotomous responses that are modeled with a probit regression. Although, under the $0-1$ loss function, an optimal solution is the model having the highest posterior probability, the $0-1$ loss function might not be realistic for many applications. One potential shortcoming is that submodels different from the true one are assigned the same loss regardless of how far or close to the true one they are. Therefore, it seems appropriate to not choose a single 'optimal' model but a subset of them having posterior probabilities over a given threshold. Thus, the output of our model selection procedure will be a ranking of the models according to their posterior probabilities.

The literature on variable selection in logistic or probit models is not large. Researchers have carried out variable selection in logistic regression with 'lasso'-type procedures. For instance, Meier *et al.* [1] used a group lasso for logistic regression, whereas Kyung *et al.* [2] used the latent variable probit model for Bayesian lasso variable selection. The Bayesian approaches tend to dominate, and researchers typically solve the main difficulties in its formulation, the choice of the prior for the models involved and for the model parameters, by using subjective priors. Swartz and Shete [3] did a simulation study of the five types of variable selection in case–control logistic regression and found that a Bayesian stochastic search based on a 'spike-and-slab' prior was the best performer. Chen and Dey [4] also used a Bayesian approach and were particularly concerned with correlated binary responses in multivariate logistic regression. In their study, the researchers developed subjective priors and selected models on the

[a] *Department of Statistics, University of Florida , Gainesville, FL 32611, U.S.A.*
[b] *Department of Statistics, University of Granada , 18071 Granada, Spain*
**Correspondence to: George Casella, Department of Statistics, University of Florida, 408 McCarty Hall, Building C, Gainesville, FL 32611, U.S.A.*
[†] *E-mail: casella@ufl.edu*

basis of their posterior probabilities. They carried out the evaluation of the procedure through examples and simulations. Kinney and Dunson [5], who selected both fixed and random effects by using a fully Bayesian approach, also considered correlated data. Sha *et al.* [6] also used Bayesian variable selection and were particularly interested in the case where p is very large. Their analysis used conjugate priors, and the researchers checked predictions with cross-validation.

Our proposal is to use objective priors for both the models, and their model parameters, in the Bayesian variable selection problem. The justification for doing so is that because we are interested in variable selection, it seems that the experimenter will not have prior information about the influential covariates, and hence, little subjective prior information on the distribution of the regression coefficients can be expected. Of course, if there is specific prior information, that can also be accommodated. Moreover, it may also be the case that certain covariates are always to be included; our genomics example is such a case. In this case, clinical variables (age and health) are always included in the model, and the variable selector is run on the set of genes.

Although the use of automatic objective priors allows us to circumvent the difficulty of eliciting priors for the regression coefficients, we need to implement objective intrinsic priors in probit models, which has not yet been carried out. The main difficulty comes from the fact that even the analytic expression of the Jeffreys prior for the probit regression model is very difficult to obtain, and hence, so is the computation of the intrinsic priors for model comparison. However, we can cope with this by considering the probit model to be a normal regression model with incomplete information, or equivalently, we look at our dichotomous data z as a $0-1$ thresholding transformation of a latent normal regression random variable y. We proceed as follows. We first apply the standard intrinsic prior formulation to the latent normal regression variables $(y_1, \ldots, y_n)$ and compute the marginals under the models for all the subsets of the p potential set of regressors. Once this is carried out, we transform these marginals into marginals for the dichotomous data z, our actual observations. This is carried out via integration on the cosets of the $0-1$ observations $(z_1, \ldots, z_n)$.

The proposed approach handles the large p small n problem through a new controlled-dimension stochastic search in the space of models containing no more than q covariates, where q can be set by the experimenter. If the number of covariates, p, is greater than n, the model posterior probabilities cannot be computed. Thus, the search must be restricted to these lower dimensional models, that is, $q \leqslant n$. We note that, in the large p small n case, the incorporation of subjective prior information allows for selecting models with more than n covariates, or even without any sample information. This is not the case when using intrinsic priors, which produce an objective analysis where there is no subjective prior input.

We organize the remainder of the paper as follows. In Section 2, we briefly summarize, for completeness, the standard Bayesian model selection framework and give the definition of intrinsic priors for the model parameters. We also explain, in more detail, the basic idea of our approach. In Section 3, we compute the intrinsic priors for the underlying hidden regression model and describe a numerical approach to compute the Bayes factors for the dichotomous responses $(z_1, \ldots, z_n)$. Section 4 contains details for a stochastic search algorithm that explores the entire model space but allows the user to specify an upper bound on the number of covariates. In Section 5, there is a simulation study to both assess the accuracy of the procedure and compare it with other model selectors, notably the model selection procedure of Hu and Johnson [7]. We also apply the proposed model selection criterion to a genomics data set, for which the response is the incidence of pneumonia in patients in a hospital intensive care unit and their gene expression levels are the covariates of the model. Section 6 contains some final remarks, and there is a small technical appendix.

2. Using intrinsic priors

In this section, we first summarize a standard selection procedure based on Bayes factors, briefly describe intrinsic priors, and finally show how a dichotomous observation can be thought of as the incomplete observation of a continuous variable, a latent variable hierarchical model (see, e.g., [8]). The dichotomous observation is an indicator of the sign of this continuous latent random variable modeled through a regression model with known variance.

2.1. Model selection and Bayes factors

Let $p(\mathbf{z}|\theta_j, M_j)$ be the distribution of the sample $\mathbf{z}$ under a generic regression model M_j, where θ_j represents the parameters under model M_j and M_j belongs to a finite set of models $\mathcal{M} = \{M_j, j =$

$1, \ldots, N\}$. Let $p(\mathbf{z}|M_j) = \int p(\mathbf{z}|\theta_j, M_j)\pi(\theta_j|M_j)\mathrm{d}\theta_j$ be the marginal distribution of the sample $\mathbf{z}$ under model M_j, where $\pi(\theta_j|M_j)$ denotes the prior distribution for the model parameters θ_j, and $p(\mathbf{z}) = \sum_{j=1}^{N} p(\mathbf{z}|M_j)\pi(M_j)$ is the marginal distribution of $\mathbf{z}$, where $\pi(M_j)$ denotes the prior probability of model M_j.

In this setting, the posterior probability of model M_j is given by

$$\pi(M_j|\mathbf{z}) = \frac{p(\mathbf{z}|M_j)\pi(M_j)}{p(\mathbf{z})} = \frac{BF_{j1}(\mathbf{z})\ \pi(M_j)/\pi(M_1)}{1 + \sum_{j=2}^{N} BF_{j1}(\mathbf{z})\ \pi(M_j)/\pi(M_1)}, \tag{1}$$

where

$$BF_{j1}(\mathbf{z}) = \frac{p(\mathbf{z}|M_j)}{p(\mathbf{z}|M_1)} = \frac{\int p(\mathbf{z}|\theta_j, M_j)\pi(\theta_j|M_j)\mathrm{d}\theta_j}{\int p(\mathbf{z}|\theta_1, M_1)\pi(\theta_1|M_1)\mathrm{d}\theta_1}$$

is the Bayes factor for comparing models M_j and M_1, where M_1 is a particular fixed model. In regression analysis with a potential set of p regressors, $N = 2^p$, and M_1 is typically the intercept-only model.

Our Bayesian model selection procedure searches for models with high posterior probability, and from expression (1), it follows that this is equivalent to searching for models with high values of $BF_{j1}(\mathbf{z})$ $\pi(M_j)$. We remark that for intrinsic priors $\pi^I(\theta_j|M_j)$, the Bayesian variable selection procedure for normal regression models have excellent properties. In particular, they are consistent model selectors and have moderate type I and type II errors for finite sample sizes [9, 10].

Here, for the probit model, our variable selection procedure transforms classes of marginal densities for normal regression variables into marginal densities for probit regression variables, so that the variable selection procedure for probit models enjoys the original properties that are invariant under the probit transformation. For instance, in the normal regression setting, consistency means that the posterior probability of the true model tends to one as the sample size grows. Now, in probit regression, the sample we observe is a probit transformation of the sample from a normal regression. Thus, the true probit model is contained in the image of a class of normal models that contains the true one. Consistency properties of the procedure should now be understood in this setting. It is also the case that the probit transformation of a normal sample entails a notable loss of sampling information (Section 2.3 for details).

2.2. *Bayes factors for intrinsic priors*

Consider two general models:

$$M_1 : \left\{p_1(y|\alpha), \pi_1^N(\alpha)\right\} \text{ and } M_2 : \left\{p_2(y|\beta), \pi_2^N(\beta)\right\},$$

where $\pi^N(\alpha)$ and $\pi^N(\beta)$ are default priors, for example, Jeffreys priors or reference priors. Frequently, these priors are not integrable and thus are not suitable for testing. Berger and Pericchi [11] addressed this problem by creating the 'intrinsic Bayes factor', a useable pseudo-Bayes factor constructed from the Bayes factor for the aforementioned improper priors as follows.

Given a sample $\mathbf{y} = (y_1, \ldots, y_n)$, define a minimal training sample (mTS) as any subsample of minimal size such that the posterior distributions under both models are integrable. Formally, a subsample $\mathbf{y}_T$ of the observed sample $\mathbf{y}$ is an mTS if both $\int p_1(\mathbf{y}_T|\alpha)\pi_1^N(\alpha)d\alpha$ and $\int p_2(\mathbf{y}_T|\beta)\pi_2^N(\beta)d\beta$ are positive and finite and if there is no subsample of $\mathbf{y}_T$ satisfying these conditions.

For an mTS $\mathbf{y}_T$, consider the posterior of the parameters

$$\pi_1^N(\alpha|\mathbf{y}_T) \propto p_1(\mathbf{y}_T|\alpha)\pi_1^N(\alpha) \text{ and } \pi_2^N(\beta|\mathbf{y}_T) \propto p_2(\mathbf{y}_T|\beta)\pi_2^N(\beta).$$

The partial Bayes factor $BF_{21}^P(\mathbf{y}_T)$ is defined for the sample $\mathbf{y}_T$ as

$$BF_{21}^P(\mathbf{y}_T) = \frac{\int p_2(\mathbf{y}_{-T}|\beta)\pi_2^N(\beta|\mathbf{y}_T)d\beta}{\int p_1(\mathbf{y}_{-T}|\alpha)\pi_1^N(\alpha|\mathbf{y}_T)d\alpha},$$

where $\mathbf{y}_{-T} = \mathbf{y}\backslash\mathbf{y}_T$. It can be easily shown that $BF_{21}^P(\mathbf{y}_T) = BF_{21}^N(\mathbf{y})BF_{12}^N(\mathbf{y}_T)$, where $BF_{21}^N(\mathbf{y})$ is the Bayes factor for comparing M_2 with M_1 for the entire sample $\mathbf{y}$ and $BF_{12}^N(\mathbf{y}_T)$ is the Bayes factor for comparing M_1 with M_2 for the mTS $\mathbf{y}_T$. Both Bayes factors use improper default priors under both models.

To try to lessen the dependence of the partial Bayes factor on $\mathbf{y}_T$, Berger and Pericchi [11] introduced the average of the partial Bayes factors $BF_{21}^P(\mathbf{y}_T)$ over the existing mTS $\mathbf{y}_T$ in the sample $\mathbf{y}$. This arithmetic mean was called the arithmetic intrinsic Bayes factor $BF_{21}^{AI}(\mathbf{y})$, given by

$$BF_{21}^{AI}(\mathbf{y}) = BF_{21}^N(\mathbf{y}) \times mean_{\mathbf{y}_T}\left[BF_{12}^N(\mathbf{y}_T)\right].$$

We note that $BF_{21}^{AI}(\mathbf{y})$ is not a Bayes factor. In particular, it does not satisfy the symmetric property of the Bayes factors: $BF_{21}^{AI}(\mathbf{y}) \neq 1/BF_{12}^{AI}(\mathbf{y})$. However, it is asymptotically equivalent to a Bayes factor for the so-called intrinsic priors [11]. Later, Moreno *et al.* [12] proposed an 'intrinsic limiting procedure' for nested models (M_1 is nested in M_2 if for every α there is a β_α such that $p_1(y|\alpha) = p_2(y|\beta_\alpha)$) for defining the intrinsic priors $(\pi_1^N(\alpha), \pi^I(\beta))$. The suggestion was to consider the Bayesian model for these priors, that is,

$$M_1 : \left\{p_1(y|\alpha), \pi_1^N(\alpha)\right\} \ and \ M_2 : \left\{p_2(y|\beta), \pi^I(\beta)\right\}, \tag{2}$$

where $\pi^I(\beta) = \int \pi^I(\beta|\alpha)\pi_1^N(\alpha)d\alpha$. The intrinsic prior for β is obtained from the intrinsic prior for β conditional on α,

$$\pi^I(\beta|\alpha) = \pi_2^N(\beta)\ E_{\mathbf{y}_T|\beta}^{M_2}\left[\frac{p_1(\mathbf{y}_T|\alpha)}{\int p_2(\mathbf{y}_T|\beta)\pi_2^N(\beta)d\beta}\right]. \tag{3}$$

In this latter expression, the expectation is taken with respect to the distribution of mTS $\mathbf{y}_T$ under the larger model M_2. Equivalently, we can write $\pi^I(\beta) = \pi_2^N(\beta)\ E_{\mathbf{y}_T|\beta}^{M_2} BF_{12}^N(\mathbf{y}_T)$. The Bayes factor for the intrinsic prior (BFIP) is then given by

$$BF_{21}^{IP}(\mathbf{y}) = \frac{\int p_2(\mathbf{y}|\beta)\pi^I(\beta)d\beta}{\int p_1(\mathbf{y}|\alpha)\pi_1^N(\alpha)d\alpha}. \tag{4}$$

We note that the Bayes factor (4) does not depend on the data set but only on the sampling models. Moreno *et al.* [12] showed that it is a limit of Bayes factors for proper priors and that it satisfies the properties of a Bayes factor. In this paper, we will compute the Bayes factor for intrinsic priors in our variable selection problem.

2.3. *Probit models and intrinsic Bayes factors*

Consider a sample $\mathbf{z} = (z_1, \ldots, z_n)$, where z_i, $i = 1, \ldots, n$, is a $0-1$ random variable such that, under model M_j, it follows a probit regression model with a $j+1$ dimensional vector of covariates x_j, $j \leqslant p$. That is, this probit model M_j has the form

$$z_i|\theta_i, M_j \sim \text{Bernoulli}(z_i|\theta_i) \text{ with } \theta_i|M_j = \Phi(x_i'\beta_j), \tag{5}$$

where Φ denotes the standard normal cumulative distribution function and β_j is a vector of dimension $j+1$. The first component of the vector x_i is set equal to one so that when considering models of the form (5), the intercept is in any submodel. The maximum length of the vector of covariates is $p+1$.

The probit model (5) can be thought of as a regression model with incomplete sampling information. Indeed, consider a random variable y_i following a normal regression model, but only the sign of y_i is observed. More specifically, we observe the variable $z_i = 1(y_i > 0)$, and on the basis of the information provided by the sample $z = (z_1, \ldots, z_n)$, we want to compare the regression models M_j having j covariates, $j \in \{1, \ldots, p\}$, with the intercept-only model M_1.

For the sample $\mathbf{y} = (y_1, \ldots, y_n)'$, the null normal model is

$$M_1 : \{N_n(\mathbf{y}|\alpha\mathbf{1}_n, \mathbf{I}_n),\ \pi(\alpha)\},$$

where $N_n(\mathbf{y}|\mu, \Lambda)$ denotes the n-variate normal density with mean μ and covariance matrix Λ evaluated at the vector $\mathbf{y}$ and $\mathbf{1}_n$ denotes a vector of ones. For a generic model M_j with $j1$ regressors, the alternative model is

$$M_j : \left\{N_n(\mathbf{y}|\mathbf{X}_j\beta_j, \mathbf{I}_n),\ \pi(\boldsymbol{\beta}_j)\right\},$$

where the design matrix $\mathbf{X}_j$ has dimensions $n \times (j+1)$. Researchers have developed intrinsic methodology for the linear model, starting with Berger and Pericchi [13], which was further developed in [14] by

using the methods of Moreno *et al.* [12]. This intrinsic methodology gives us an automatic specification of the priors $\pi(\alpha)$ and $\pi(\beta)$, starting with the reference priors $\pi^N(\alpha)$ and $\pi^N(\beta)$ for α and β, which are both improper and proportional to 1. We again note that if there is a set of covariates that are to be kept in every model, that set becomes the null model M_1, and we proceed in the same way.

The marginal distributions for the sample $\mathbf{y}$ under the null model, and under the alternative model with intrinsic prior, are formally written as

$$\begin{aligned} m_1(\mathbf{y}) &= \int N_n(\mathbf{y}|\alpha\mathbf{1}_n, \mathbf{I}_n)\pi^N(\alpha)\mathrm{d}\alpha, \\ m_j(\mathbf{y}) &= \iint N_n(\mathbf{y}|\mathbf{X}_j\beta_j, \mathbf{I}_n)\pi^I(\beta|\alpha)\pi^N(\alpha)\mathrm{d}\alpha\mathrm{d}\beta. \end{aligned} \tag{6}$$

Because model M_1 is nested in M_j for any j, the BFIPs $BF_{j1}^{IP}(\mathbf{y}) = m_j(\mathbf{y})/m_1(\mathbf{y})$ provide a consistent model selection procedure; that is, provided that the true model is one of the 2^p regression models, the procedure chooses this true model with probability one when the sample size grows to infinity [10].

However, these are marginals of the sample $\mathbf{y}$, but our selection procedure requires us to compute the Bayes factor of model M_j versus the reference model M_1 for the sample $\mathbf{z} = (z_1, \ldots, z_n)$. Then, we transform the marginal $m_j(\mathbf{y})$ into the marginal $m_j(\mathbf{z})$ by using the probit transformations $z_i = 1(y_i > 0)$, $i = 1, \ldots, n$. These latter marginals are given by

$$m_j(\mathbf{z}) = \int_{A_1 \times \cdots \times A_n} m_j(\mathbf{y})\mathrm{d}\mathbf{y}, \tag{7}$$

where

$$A_i = \begin{cases} (0, \infty) & \text{if} \quad z_i = 1, \\ (-\infty, 0) & \text{if} \quad z_i = 0, \end{cases}$$

and the required Bayes factor based on the intrinsic prior is $BF_{j1}^{IP}(\mathbf{z}) = m_j(\mathbf{z})/m_1(\mathbf{z})$.

3. Computing the Bayes factor

To compute the Bayes factor for the observable sample $\mathbf{z}$, we proceed by first finding the analytic expressions of both the intrinsic priors for the regression model and the marginal probabilities of $\mathbf{y}$ given in (6). Then, we give an algorithm to compute the Bayes factor for the responses $\mathbf{z}$ defined above on the basis of the computation of multivariate normal distribution probabilities.

3.1. Intrinsic priors for normal regression models

Let $\mathbf{Z}_T$ be the design matrix of an mTS of a normal regression model M_j for the variable y that includes j covariates plus the intercept. Then, if $j+1$ is the dimension of β_j, we have

$$\int N_{j+1}(\mathbf{y}_T|\mathbf{Z}_T\beta, \mathbf{I}_{j+1})\mathrm{d}\beta = \begin{cases} |\mathbf{Z}_T'\mathbf{Z}_T|^{-1/2} & \text{if rank } (\mathbf{Z}_T) \geqslant j+1 \\ \infty & \text{otherwise} \end{cases}.$$

Therefore, it follows that the mTS size is $j+1$. We assume that the $(j+1) \times (j+1)$ square matrix $\mathbf{Z}_T$ is standardized[‡]; that is, all columns have mean zero and variance 1, except the first column, which has all its entries equal to one.

In our context, because the priors for α and β are proportional to 1, the intrinsic prior for comparing M_j versus M_1, given in formula (3), becomes, after some simplification,

$$\pi^I(\beta|\alpha) = N_{j+1}\left(\beta|\alpha e_1, 2(\mathbf{Z}_T'\mathbf{Z}_T)^{-1}\right),$$

where e_1 is a vector with the first component equal to 1 and the others equal to zero and $\mathbf{Z}_T'$ has $j+1$ columns corresponding to j covariates and an intercept.

[‡]*Although this assumption is not necessary, it is typically good practice and stabilizes the numerics.*

However, the matrix $\mathbf{Z}_T'\mathbf{Z}_T$ is unknown because it is a theoretical design matrix corresponding to the training sample $\mathbf{y}_T$. It can be estimated by averaging over all submatrices, containing $j+1$ rows, of the $n\times(j+1)$ design matrix $\mathbf{X}_j$ [15]. This average turns out to be (Appendix A) $\frac{j+1}{n}(\mathbf{X}_j'\mathbf{X}_j)$, and therefore,

$$\pi^I(\beta|\alpha) = N_{j+1}\left(\beta|\alpha e_1, \frac{2\,n}{j+1}(\mathbf{X}_j'\mathbf{X}_j)^{-1}\right).$$

Because we require $\mathbf{X}_j'\mathbf{X}_j$ to be invertible, we need $j+1 \leqslant n$. In other words, we will be able to compute the intrinsic prior when the number of covariates, including the intercept, is smaller than or equal to the sample size n.

The marginal of the sample $\mathbf{y} = (y_1, \ldots, y_n)'$ under model M_j, conditional on α, is

$$m_j(\mathbf{y}|\alpha) = \int N_n(\mathbf{y}|\mathbf{X}_j\beta, \mathbf{I}_n)\; N_{j+1}\left(\beta|\alpha e_1, \frac{2n}{j+1}(\mathbf{X}_j'\mathbf{X}_j)^{-1}\right)\mathrm{d}\beta = N_n(\mathbf{y}|\alpha\mathbf{1}, \Sigma_j), \tag{8}$$

where $\Sigma_j = \mathbf{I}_n + 2\,[n/(j+1)]\,\mathbf{X}_j(\mathbf{X}_j'\mathbf{X}_j)^{-1}\mathbf{X}_j'$. Integrating out the parameter α in expression (8) with respect to the reference prior $\pi^N(\alpha) = c$ (c is an arbitrary positive constant), we obtain

$$m_j(\mathbf{y}) = \frac{c}{(2\pi)^{(n-1)/2}|\mathbf{1}'\Sigma_j^{-1}\mathbf{1}|^{1/2}|\Sigma_j|^{1/2}}\exp\left\{-\frac{1}{2}\mathbf{y}'\Lambda_j\mathbf{y}\right\}, \tag{9}$$

where $\Lambda_j = \Sigma_j^{-1} - \Sigma_j^{-1}\mathbf{1}(\mathbf{1}'\Sigma_j^{-1}\mathbf{1})^{-1}\mathbf{1}'\Sigma_j^{-1}$ and has rank $n-1$.

Similarly, the marginal of the sample $\mathbf{y} = (y_1, \ldots, y_n)'$ under model M_1 is

$$m_1(\mathbf{y}) = \frac{c}{n^{1/2}(2\pi)^{(n-1)/2}}\exp\left\{-\frac{1}{2}ns_y^2\right\},$$

where $ns_y^2 = \sum_{i=1}^n(y_i - \bar{y})^2$ and $\bar{y} = \sum_{i=1}^n y_i/n$. We note that both marginals $m_j(\mathbf{y})$ and $m_1(\mathbf{y})$ depend on the arbitrary positive constant c that appears in $\pi^N(\alpha)$.

3.2. *Bayes factors for probit models*

On the basis of the observed sample $\mathbf{z}$, we now compute the marginals $m_j(\mathbf{z})$ in (7). From expressions (7) and (8), we have

$$\begin{aligned} m_j(\mathbf{z}) &= \int_A m_j(\mathbf{y})\mathrm{d}\mathbf{y} = \int_A\int_{-\infty}^{\infty} m_j(\mathbf{y}|\alpha)\mathrm{d}\alpha\;\mathrm{d}\mathbf{y} \\ &= \int_{-\infty}^{\infty}\int_A N_n(\mathbf{y}|\alpha 1, \Sigma_j)\mathrm{d}\mathbf{y}\;\mathrm{d}\alpha. \end{aligned} \tag{10}$$

The integral over $A = A_1 \times \cdots \times A_n$ is the probability of the hypercube A assigned by the n-variate normal distribution. Genz and Bretz [16] described an algorithm to efficiently and accurately compute this probability possibly having $\pm\infty$ as the extreme points of its edges. The algorithm is implemented in the R function `pmvnorm` in the `R` package `mvtnorm` by [17]. In other words, this function is able to compute $f_j(A|\alpha) = \int_A N_n(\mathbf{y}|\alpha 1, \Sigma_j)\mathrm{d}\mathbf{y}$. The problem then reduces to the computation of

$$m_j(\mathbf{z}) = \int_{-\infty}^{\infty} f_j(A|\alpha)\mathrm{d}\alpha.$$

The marginal $m_1(\mathbf{z})$ under M_1 is obtained by replacing Σ_j in (10) by the identity matrix $\mathbf{I}_n$. Therefore, the Bayes factor for comparing M_j with M_1 for the intrinsic priors and data $\mathbf{z}$ can be computed as

$$BF_{j1}^{IP}(\mathbf{z}) = \frac{\int_{-\infty}^{\infty} f_j(A|\alpha)\mathrm{d}\alpha}{\int_{-\infty}^{\infty} f_1(A|\alpha)\mathrm{d}\alpha}. \tag{11}$$

We observe that $f_h(A|\alpha), h = 1, j$, as a function of α, is very close to zero outside of the interval $(\hat{\alpha}-6, \hat{\alpha}+6)$ where $\hat{\alpha} = \Phi^{-1}(\sum_{i=1}^n z_i/n)$ is the maximum likelihood estimator of α under M_1, so that the integral over the real line can be approximated by the integral over this interval.

4. A controlled-dimension stochastic search

Referring to the discussion following (1), we search for models with high values of $BF_{j1}(\mathbf{z})\pi(M_j)$, which is equivalent to searching for models with maximum posterior probability (Section 2.1). For the variable selection problem, we find that using uniform prior weights on the models works well, so our search is only driven by the Bayes factor (11). As mentioned in Section 3.1, this quantity can only be calculated when the number of covariates (including the intercept) considered by the model is fewer than the sample size, which is not always the case. For example, in the succeeding application, we have significantly more genes per patient than patients.

Thus, we propose a random walk through the space of models with $q \leqslant n-1$ covariates (recall that the intercept is always included). The researcher selects the value of q, keeping in mind that the smaller the value of q, the smaller is the space for the search, making the search algorithm more efficient. We identify the models with a vector $\gamma \in \{0,1\}^p$, where M_γ includes the covariate j only if $\gamma_j = 1$. Because the intercept is always included in the model, it is not considered in γ explicitly. For example, for $\gamma = (0,1,1,0,\ldots,0)$, M_γ is the model that includes only the intercept and the second and third covariates.

In theory, the vector γ could be any p-dimensional vector of 0's and 1's. There are 2^p such models, and we denote this model space by $\mathcal{M}_{p:p}$. However, the feasible search space is the set of all models taken from $\mathcal{M}_{p:p}$ but having no more than $n-1$ total covariates. In fact, our search works for any $q \leqslant n-1$ chosen by the researcher. There are $\sum_{j=0}^{q}\binom{p}{j}$ such models. We denote this model space by $\mathcal{M}_{p:q}$ and now describe a random walk with stationary distribution proportional to $m_\gamma(\mathbf{z})$ for $\gamma \in \mathcal{M}_{p:q}$.

We start by defining three vectors of indicator functions.

- $\delta \in \mathcal{M}_{p:p}$: This is a vector with $0-1$ entries of the covariates in a latent model, where $\delta_j = 1$ indicates that the jth covariate is in the latent model. Note that δ can choose models having more than q coefficients.
- $A = (a_1,\ldots,a_p)$: A vector with $0-1$ entries indicating the *active covariates* in the model, where $a_j = 1$ indicates that the jth covariate is in the active set. This is the current subset of q covariates from which we will use to iterate the stochastic search. We require that $\sum_{j=1}^{p} a_j = q$.
- $\gamma \in \mathcal{M}_{p:q}$: This is a vector with $0-1$ entries where $\gamma_j = 1$ indicates that the jth covariate is in the model. The model has no more than q covariates, so $\sum_{j=1}^{p}\gamma_j \leqslant q$.

The initial point of the random walk is any $(A^{(0)},\delta^{(0)},\gamma^{(0)})$ such that $\sum_j a_j^{(0)} = q$, $\delta^{(0)} \in \mathcal{M}_{p:p}$ and $\gamma^{(0)} \in \mathcal{M}_{p:q}$. The random walk consists of two Metropolis–Hasting steps, one for δ and one for A, from which we construct γ. Define $\delta \star A$ as the componentwise multiplication of δ and A. The user sets the probability r, $0 \leqslant r \leqslant 1$, and at iteration t, starting from $(A^{(t)},\delta^{(t)},\gamma^{(t)})$, we have the following:

1. Update $\delta^{(t)}$: We only update components of $\delta^{(t)}$ that are in the active set. Write $\delta^{(t)} = (\delta_A^{(t)},\delta_{A^c}^{(t)})$, where $\delta_A^{(t)} = \{\delta_j^{(t)} : a_j^{(t)} = 1\}$ contains the *active* coefficients and $\delta_{A^c}^{(t)} = \{\delta_j^{(t)} : a_j^{(t)} = 0\}$ contains the *inactive* coefficients.
   (a) With probability r: Replace $\delta_A^{(t)}$ with the candidate δ_A^{cand} with coefficients
   $$\delta_{A,j}^{\text{cand}} = \begin{cases} 0 & \text{with probability } 1/2 \\ 1 & \text{with probability } 1/2\,, \end{cases}$$
   set $\delta^{\text{cand}} = (\delta_A^{\text{cand}},\delta_{A^c}^{(t)})$, and construct the candidate $\gamma^{\text{cand}} = \delta^{\text{cand}} \star A^{(t)}$. This is a vector of 0's and 1's of length p with no more than q $1's$.
   (b) With probability $1-r$: Choose one coefficient of $\delta_A^{(t)}$ at random, and change $0 \to 1$ or $1 \to 0$, whichever applies, to create δ_A^{cand}. Set $\delta^{\text{cand}} = (\delta_A^{\text{cand}},\delta_{A^c}^{(t)})$ and construct the candidate $\gamma^{\text{cand}} = \delta^{\text{cand}} \star A^{(t)}$.
   Calculate the Metropolis–Hastings ratio $\alpha_1 = \min\{1, m_{\gamma^{\text{cand}}}(\mathbf{z})/m_{\gamma^{(t)}}(\mathbf{z})\}$ and set
   $$\delta^{(t+1)} = \begin{cases} \delta^{\text{cand}} & \text{with probability } \alpha_1 \\ \delta^{(t)} & \text{with probability } 1-\alpha_1\,. \end{cases}$$
   We now create a second γ candidate, $\gamma^{\text{cand}\,2} = \delta^{(t+1)} \star A^{(t)}$.
2. Update $A^{(t+1)}$: Choose one coefficient of $A^{(t)}$ at random, and swap it from $0 \to 1$ or $1 \to 0$, whichever applies, to create the candidate A^{cand}. Create a third γ candidate $\gamma^{\text{cand}3} = \delta^{(t+1)} \star A^{\text{cand}}$.

Calculate the Metropolis-Hastings ratio $\alpha_2 = \min\{1, m_{\gamma^{\text{cand }3}}(\mathbf{z})/m_{\gamma^{\text{cand }2}}(\mathbf{z})\}$ and set

$$A^{(t+1)} = \begin{cases} A^{\text{cand}} & \text{with probability } \alpha_2 \\ A^{(t)} & \text{with probability } 1-\alpha_2 \end{cases},$$

and set $\gamma^{(t+1)} = \delta^{(t+1)} \star A^{(t+1)}$.

The model represented by γ contains the covariates that are both active and in the latent model, formally, $\gamma = \delta \star A$. The stochastic search is a random walk on the set $\mathcal{M}_{p:q}$ of feasible models and has stationary distribution proportional to $m_\gamma(\mathbf{z})$ or, equivalently, to the Bayes factor.

5. Simulations and applications

In this section, we evaluate the performance of the variable selection procedure through simulations and then apply the procedure to the problem that originally motivated this work, a problem of selecting candidate genes associated with outcomes in an intensive care unit based on the information provided by a probit sample $\mathbf{z}$.

5.1. *Comparison with the Hu and Johnson model selection algorithm*

To compare the intrinsic procedure with that of Hu and Johnson (H&J) [7], we ran the simulation scenario in Section 4 of that paper, evaluating all $2^{15} = 32,768$ models. H&J generated the covariates as independent and identically distributed standard normal random variables, and as they did, we simulated $n = 30$ dichotomous observations from model (5) where the true simulated values are $\boldsymbol{\beta}_i = 0.5i$ for $i = 0, \ldots, 6$ and $\boldsymbol{\beta}_i = 0$ for $i = 7, \ldots, 15$. We used the hyperparameter values given in Section 4 of H&J, and even though the true intercept parameter is equal to zero, it was included in every model.

We then looked at the top 50 models that each procedure chose (highest posterior probability). We note that, although the true model contains six nonzero coefficients, this turns out to be a bad model in the sense that if $\beta_3 - \beta_6$ are in the model, no selection procedure would then include β_1 and β_2, as the amount of additional variability that they explain cannot overcome any dimension penalty (even the naive F-to-enter has p-value equal to 0.18). This observation underlies an emerging strategy in selection problems, especially with a large number of covariates. As statisticians, we have typically been trained to look for the 'true' model; but sometimes, that is not feasible and may not be the right goal. Perhaps a more suitable goal is to find good, meaningful models that explain a sufficient amount of variability.

Figure 1 compares the two procedures. In the left panel, we see that the intrinsic procedure consistently selects more true coefficients, with the H&J procedure showing a number of dips in the number of true coefficients selected. The right panel shows that the intrinsic procedure will select slightly more false coefficients in the lower-ranking models but selects a comparable, and sometimes fewer, number of false coefficients in the first 20 models (which, in practice, may be more than any experimenter is going to examine closely).

5.2. *Loss of information from dichotomizing*

Here, we simulate the original normal data y, to which the probit transformation z is applied, and examine the performance of the BFIP for both data sets, the y and the z. By doing so, we illustrate the behavior of BFIP for probit samples and also compare the behavior of the BFIP for the original samples with the behavior for probit samples, which give us an idea of the 'loss of information' from dichotomizing.

In our next comparisons, we have decided to include not only the H&J procedure but also the BIC, the variable selection procedure of Schwarz [18]. Although much is known about the relationship of BIC and BFIP in the linear model setting [10, 19], less is known in the probit regression case. We typically expect BIC to be biased toward small models, and we want to see if this is the case in probit regression. BIC also provides a default procedure to which we can calibrate any improvement.

We analyze the results of the proposed variable selection procedure for moderate sample sizes and under four simulation scenarios. We first generate a sample $\mathbf{y} = (y_1, \ldots, y_n)$ from the normal regression model $N_n(\mathbf{y}|\mathbf{X}\beta, \mathbf{I}_n)$ for $n = 20$, where we have six regressors. We sampled all covariate values from the uniform $U(0, 6)$. We obtained the probit sample by setting $z_i = 1_{(y_i>0)}$.

We assign a uniform prior to the models M_j, and thus, we rank the models according to the values of $m_j(z)$ in (10) when using the information in the z data and according to the values of $m_j(\mathbf{y})$ in (9)

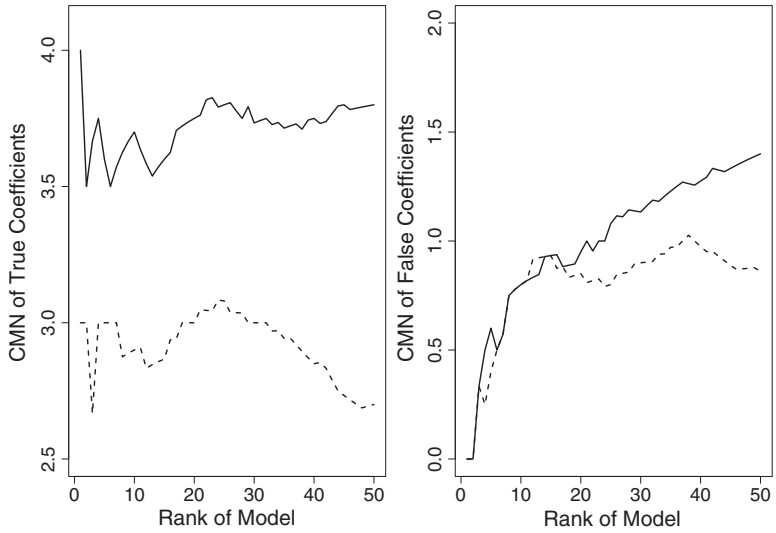

Figure 1. Comparison of the intrinsic procedure (solid lines) with the Hu and Johnson procedure (dashed lines). For each procedure, we took the top 50 models that it selected. The left panel shows the cumulative mean number (CMN) of true coefficients selected, and the right panel shows the CMN of false coefficients selected.

when using the information in the y data. We repeat the simulation 200 times and count the number of times that the method ranks the true model in first place and the number of times it is in the top three of the ranking.

The four scenarios are the same except for the values of the regression coefficients (the intercept is always included) of the true model that are given in the first column of Table I. The H&J procedure is the winner in the first row of Table I, where there is no signal in the data and the correct model is the intercept-only model. This behavior is similar to what was seen in Section 5.1, where the H&J procedure tended to select models with smaller numbers of coefficients. Throughout the rest of Table I, with respect to the y data, the BFIP and the H&J procedure are competitive; and the BIC also does reasonably well. However, the situation changes drastically when we look at selection using the z data. In this case, the BFIP procedure clearly dominates the other two. This suggests that the loss of information in going from the y to the z data affects BFIP much less than the other two procedures.

An overall conclusion is that the effectiveness of all of the model selectors is lessened when using the z data than when using the original data y, which is certainly reasonable. However, the BFIP (z) still provides valuable knowledge on the structure of the problem, more so as the true model contains more covariates.

5.3. Application in association genetics

The following association genetics problem first motivated this work. The data corresponds to 47 patients in an intensive care unit following trauma surgery. The physicians are concerned with how to better

Table I. Comparison of performance of the proposed model selection procedure, the BIC and the H&J criterion for model selection.

	Top choice						Top three					
Model		y			z			y			z	
True coefficients	BFIP	BIC	H&J	BFIP	BIC	H&J	BFIP	BIC	H&J	BFIP	BIC	H&J
0, 0, 0, 0, 0	144	132	156	62	82	107	170	173	187	89	121	138
1, 0, 0, 0, 0	166	148	169	121	68	83	197	185	193	164	82	134
−1, −1, 0, 0, 0	183	153	171	139	87	90	196	191	196	183	127	153
1, −1, 1, 0, 0	193	167	184	102	44	40	200	197	197	172	148	147

$M = 200$ samples of size $n = 20$ with $p = 5$ (plus intercept) covariates were simulated. From left to right: coefficients of true model; number of times this model ranked first by using the information in y and z data, respectively, by the Bayes factor for the intrinsic prior (BFIP), the BIC, and the Hu and Johnson (H&J) criteria; number of times the model ranked in the top three by using the y and z data, respectively.

manage postoperative sepsis (infection) and are interested to see if there is association with any subset of genes. Here, we consider the $0-1$ endpoint 'pneumonia'; of the 47 patients, 39 of them exhibited pneumonia. We employ our model selection algorithm to select the variables that are most highly associated with the response.

For each patient, we measured gene expression of 296 genes in peripheral blood, along with three clinical covariates: age, gender, and abbreviated injury score. These clinical covariates are always in the model, and hence, they constitute the null model. We look for the set of genes that better explains the response pneumonia after taking into consideration the clinical covariates. In other words, our goal was to get the best model that includes the three clinical covariates and relevant genes.

We applied the proposed variable selection procedure to our data set of gene expression and search for the model with the highest value of m_j in (10). Because we have a small sample size, we expect that the models with the highest BFIP will have few covariates, and we focus our search on the models with at most 10 genes (i.e., considering the intercept and the three patient-level covariates, with at most $q = 14$ covariates).

We ran 10,000 iterations of the stochastic search. We show in Table II the 10 models with the highest Bayes factors found by this search. Genes ERICH1 (non-annotated), OLFM1 ('its abundant expression in the brain suggests that it may have an essential role in nerve tissue', GenBank), and BCL3 (related with leukemia/lymphoma) are frequent in these models (in fact, ERICH1 appeared in 16 of the top 20 models). The clinician must choose the model with the best biological interpretation.

As an illustration of the possible use of this information, we look at some of the most frequently appearing genes in Table II and select, for example, the genes listed in the third row of the table. With the assumption of a multivariate normal prior for the regression coefficients with covariance matrix $100 \times \mathbf{I}$, Table III shows the 95% highest posterior density CIs for the regression coefficients. As expected, the value zero is not included in any gene effect intervals, and the values in the table tell us that the lower is the gene expression the more likely the patient will exhibit pneumonia (coded as 'success' or $z_i = 1$).

Furthermore, for each patient in the sample, we considered a future patient with the same covariate values and computed the probability (and its highest posterior density CI) of exhibiting the disease. We show the results in the left panel of Figure 2, where we see that all of these future patients have a high

Table II. The 10 models with the highest Bayes factor for the intrinsic prior found by the stochastic search algorithm.

Rank	Number of genes	Genes		
1	3	ARL10	ERICH1	OR4D1
2	3	GCLM	OLFM1	TEP1
3	3	ERICH1	OLFM1	TEP1
4	3	BCL3	ERICH1	TMEM56
5	3	C8orf34	ERICH1	WDR26
6	3	ARPC5	ERICH1	ITGB1
7	2	ERICH1	PCNX	
8	2	ERICH1	MLLT6	
9	3	C8orf34	ERICH1	MLLT6
10	3	ERICH1	SETD4	TRIO

Table III. 95% highest posterior density CIs and means for regression coefficients from the model in the third row of Table II.

	Lower	Upper	Mean
(Intercept)	−9.26	14.02	2.87
Age	−0.10	0.56	0.20
Gender	−13.66	2.91	−5.35
AIS	−2.04	3.82	0.67
ERICH1	−22.75	−1.46	−10.91
OLFM1	−24.27	−2.57	−13.65
TEP1	−23.34	−3.50	−13.64

The prior covariance matrix for the coefficients is $100 \times \mathbf{I}$.
AIS, abbreviated injury score.

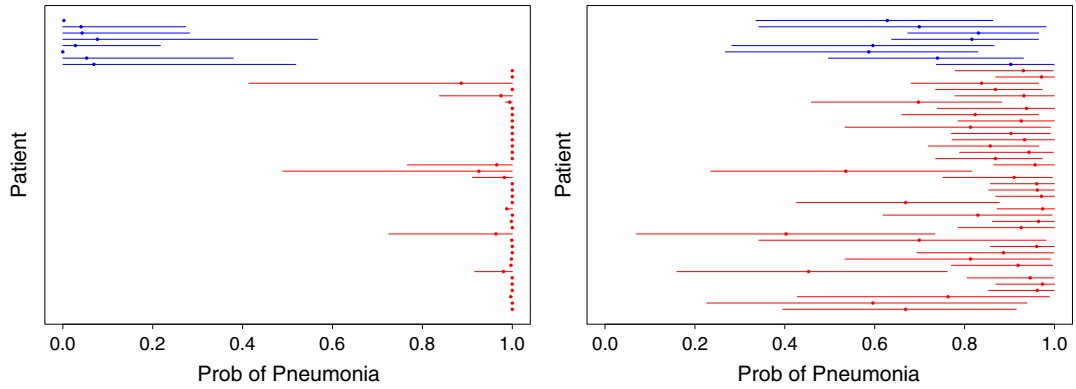

Figure 2. Each line represents the 95% highest posterior density CIs for the probability of exhibiting pneumonia for a (future) patient with the same covariate values as one in the sample. The dot represents the mean. The line is red when the patient in the sample with these covariate values exhibited the disease and blue otherwise. The left panel corresponds to the model with clinical variable plus the three genes found by using the proposed model selection algorithm. The right panel corresponds to the model with only clinical variables (no genomic information).

probability of matching the disease status of their in-sample counterpart. Many of these probabilities of matching are close to one, which is not the case if we only consider the clinical variables. To see this, the right panel in Figure 2 is the analogous plot to the one on the left but only including the clinical covariates in the model, thus showing the relevance of the genetic information.

6. Discussion

We have proposed a new variable selection procedure for probit models. This procedure is embedded in the intrinsic prior framework, and hence, the priors for the parameters of each competing model are objective resulting in automatic priors that only depend on the structure of the sampling regression models. Further, these priors do not depend on any hyperparameter whose values need to be subjectively assigned by the researcher; thus, no subjective prior elicitation is required.

We avoid the need to work with the extremely complex objective priors for probit models, in particular the intrinsic priors, by working instead with the intrinsic priors for the normal regression models, the underlying models of the latent random variables that define the probit models. This allows us to use the much simpler intrinsic priors for the regression parameters in normal models for computing the marginals of the latent variables. Then, these marginals of the latent variables are transformed into marginals of the observable probit data. That is all we need for doing variable selection in probit models. We also note that this methodology can be generalized to the case of multiple responses, for example, to the case of an ordered probit response or to the case of normal latent variables with unknown variances.

The use of the latent normal model leads us to probit regression, as opposed to logistic regression, perhaps the most popular model for analyzing dichotomous data. However, this is really not a drawback, as the end result from the models is quite similar. Indeed, Caffo and Griswold [20] noted that 'in typical settings, data cannot distinguish between probit and logit conditional link functions' and further argue in favor of the probit model. The typically small sample sizes in genetic studies makes distinguishing these models even more difficult.

Here, we assumed *a priori* that all possible regression models are equally likely. Other priors for models can be used in (1), and in fact, we also explored the use of a uniform prior for models conditional on a given number of covariates and a uniform prior for the number of covariates. Specifically, if a model includes k covariates out of a total of p possible covariates, then the prior for it is $\binom{p}{k}^{-1}$, and the prior for k is p^{-1}, $k = 1, \ldots, p$. Under this prior, the models with few (and many) covariates are assigned higher probability than that assigned to the models with approximately $p/2$ covariates. The use of this prior in a simulation study (results not shown) showed that it was not a desirable performer.

Because the proposed procedure does not use subjective prior information, it can only consider models with dimension smaller than the sample size. This led us to construct a random search through the space of models having a limited number of covariates, which is fixed by the researcher. This random

search is, to our knowledge, new, and it is not specific to the intrinsic prior methodology. Thus, it can be used for search models using other criteria, not just the posterior probability of models. This search is a reliable alternative to step forward or backward searches.

We have applied the proposed model selection procedure to the detection of a subset of genes that, in light of the data, have a high impact in the dichotomous response. As a by-product, we also provided some inference conditional on the selected model, although we are aware that the inference does not take into account the uncertainty introduced by the model selection procedure. More research needs to be carried out to make an accurate inference in the presence of model uncertainty.

Lastly, all software needed to run the procedures in the paper is available free in the `R` package `varselectIP` [21].

APPENDIX A. An estimating $\mathbf{Z}_T'\mathbf{Z}_T$

The following result is asserted in [15], but the proof only appears in a technical report. We reproduce it here for completeness. We estimate $\mathbf{Z}_T'\mathbf{Z}_T$ ($\mathbf{Z}_T$ is the expected design matrix of the mTS) by averaging the design matrix of all possible mTS. In our notation model, M_j has j covariates plus an intercept, so a submatrix $\mathbf{X}_j$ has $j+1$ columns. The total number of different training samples in the sample is $L = \binom{n}{j+1}$. Index with l, $l = 1, \ldots, L$, each one of these samples and denote by $\mathbf{X}(l)$ the $j+1 \times j+1$ submatrix of the design matrix $\mathbf{X}$ corresponding to the subsample l. Using the fact that each row of $\mathbf{X}$ is in exactly $\binom{n-1}{j}$ subsamples, we have

$$\begin{aligned}\left(\sum_l \mathbf{X}(l)'\mathbf{X}(l)\right)_{ij} &= \sum_l \sum_{h=1}^{j+1} (\mathbf{X}(l)')_{ih}(\mathbf{X}(l))_{hj} \\ &= \sum_{h=1}^{j+1} \sum_l (\mathbf{X}(l))_{hi}(\mathbf{X}(l))_{hj} \\ &= \sum_{h=1}^{j+1} \binom{n-1}{j} (\mathbf{X})_{hi}(\mathbf{X})_{hj} \\ &= \binom{n-1}{j} (\mathbf{X}'\mathbf{X})_{ij}.\end{aligned}$$

Therefore,

$$\widehat{\mathbf{Z}_T'\mathbf{Z}_T} = \frac{1}{L}\sum_l \mathbf{X}(l)'\mathbf{X}(l) = \frac{j+1}{n}\mathbf{X}'\mathbf{X}. \tag{12}$$

Acknowledgement

We thank a reviewer for pointing out that, in a spirit similar to (12), Berger and Pericchi [22] used the Cauchy–Binet theorem (which relates the determinant of a matrix product to the sum of products of determinants of submatrices) to obtain weights for training samples. However, the matrix identity (12) does not appear there.

NIH 1R01GM081704, Ministerio de Ciencia y Tecnología MTM2010-16087, Junta de Andalucía SEJ-02814, National Science Foundation DMS-0631632 and SES-0631588.

References

1. Meier L, Van De Geer S, Bühlmann P. The group lasso for logistic regression. *Journal of the Royal Statistical Society, Series B* 2008; **70**:53–71.
2. Kyung M, Gill J, Ghosh M, Casella G. Fixed and random effects selection in linear and logistic models. *Bayesian Analysis* 2010; **5**:369–411.
3. Swartz MD, Shete S. Finding factors influencing risk: comparing Bayesian stochastic search and standard variable selection methods applied to logistic regression models of cases and controls. *Statistics in Medicine* 2008; **27**:6158–6174.
4. Chen M-H, Dey DK. Variable selection for multivariate logistic regression models. *Journal of Statistical Planning and Inference* 2003; **111**:37–55.
5. Kinney S, Dunson DB. Fixed and random effects selection in linear and logistic models. *Biometrics* 2007; **63**:690–698.
6. Sha N, Vannucci M, Tadesse MG, Brown PJ, Dragoni I, Davies N, Roberts TC, Contestabile A, Salmon M, Buckley C, Falciani F. Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage. *Biometrics* 2004; **60**(3):812–819. http://www.jstor.org/stable/3695405.

7. Hu J, Johnson VE. Bayesian model selection using test statistics. *Journal of the Royal Statistical Society Series B* 2009; **71**(1):143–158. http://ideas.repec.org/a/bla/jorssb/v71y2009i1p143-158.html.
8. Albert J, Chib S. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* 1993; **88**:669–679.
9. Girón FJ, Moreno E, Martínez ML. An objective Bayesian procedure for variable selection in regression. In *Advances on Distribution Theory, Order Statistics and Inference*, Balakrishnan N, Castillo E, Sarabia JM (eds). Birkhäuser: Boston, 2006b; 389–404.
10. Casella G, Girón FJ, Martínez ML, Moreno E. Consistency of Bayesian procedures for variable selection. *The Annals of Statistics* 2009; **37**:1207–1228.
11. Berger JO, Pericchi LR. The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association* 1996; **91**(433):109–122. http://www.jstor.org/stable/2291387.
12. Moreno E, Bertolino F, Racugno W. An intrinsic limiting procedure for model selection and hypotheses testing. *Journal of the American Statistical Association* 1998; **93**(444):1451–1460. http://www.jstor.org/stable/2670059.
13. Berger JO, Pericchi LR. The intrinsic Bayes factor for linear models (with discussion). In *Bayesian Statistics 5*, Bernardo JM, Berger JO, Dawid AP, Smith AFM (eds). Oxford University Press, 1996; 23–42.
14. Casella G, Moreno E. Objective Bayesian variable selection. *Journal of the American Statistical Association* 2006; **101**(473):157–167.
15. Girón FJ, Martínez ML, Moreno E, Torres F. Objective testing procedures in linear models: calibration of the p-values. *Scandinavian Journal of Statistics* 2006; **33**(4):765–784. http://ideas.repec.org/a/bla/scjsta/v33y2006i4p765-784.html.
16. Genz A, Bretz F. *Computation of Multivariate Normal and t Probabilities*, Lecture Notes in Statistics. Springer-Verlag: Heidelberg, 2009.
17. Genz A, Bretz F, Miwa T, Mi X, Leisch F, Scheipl F, Hothorn T. mvtnorm: Multivariate Normal and t Distributions, 2010. http://CRAN.R-project.org/package=mvtnorm, r package version 0.9-92.
18. Schwarz G. Estimating the dimension of a model. *Annals of Statistics* 1978; **6**:461–464.
19. Moreno E, Girón FJ, Casella G. Consistency of objective Bayes factors as the model dimension grows. *The Annals of Statistics* 2010; **38**:1937–1952.
20. Caffo B, Griswold M. A user-friendly introduction to link-probit-normal models. *The American Statistician* 2006; **60**(2):139–145.
21. Gopal V, Leon-Novelo L, Casella G. varSelectIP: Objective Bayes Model Selection in Linear Regression and Probit models, 2011. http://CRAN.R-project.org/package=varSelectIP, r package version 0.1-4.
22. Berger JO, Pericchi LR. Training samples in objective Bayesian model selection. *The Annals of Statistics* 2004; **32**(3):841–869. http://www.jstor.org/stable/3448577.