

Objective Bayes Model Selection in Probit Models

Luis Leon-Novelo*
University of Florida

Elías Moreno†
University of Granada

George Casella‡
University of Florida

January 20, 2011

Abstract

We describe a variable selection procedure for categorical responses where the candidate models are all probit regression models having a potential set of k covariates. 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 k 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. Lastly, we assess, through simulations, the accuracy of the procedure, and apply the variable selector to a gene expression data set, where the response is whether a patient exhibits pneumonia.

Key Words: Intrinsic Priors, Linear Models, Bayes Factors, Model Selection, Probit Models, Stochastic Search.

*Postdoctoral Associate, Department of Statistics, University Florida, 102 Griffin-Floyd Hall, Gainesville, FL 32611. Supported by NIH 1R01GM081704. Email: luis@stat.ufl.edu

†Professor, Department of Statistics, University of Granada, 18071, Granada, Spain. Supported by Ministerio de Ciencia y Tecnología, Grant MTM2010-16087, and Junta de Andalucía Grant SEJ-02814. Email: emoreno@ugr.es

‡Distinguished Professor, Department of Statistics, University of Florida, Gainesville, FL 32611. Supported by National Science Foundation Grants DMS-0631632 and SES-0631588, and NIH 1R01GM081704. Email: casella@ufl.edu.

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 k 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 k covariates those which have some influence in the observed responses. Thus, we have a model selection problem in the class of 2^k models.

In this paper we describe a variable selection procedure applicable to dichotomous responses which 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 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 a posterior probability over a given threshold. Thus, the output of our model selection procedure will be a ranking of the class of models according with their posterior probabilities.

The literature on variable selection in logistic or probit models is not large. Variable selection in logistic regression has been done with “lasso” type procedures. For instance, Meier et al. [1] used a group lasso for logistic regression, while Kyung et al. [2] used the latent variable probit model for Bayesian lasso variable selection. The Bayesian approaches tend to dominate, and the main difficulties in its formulation, the choice of the prior for the models involved and for the model parameters, are typically solved by using subjective priors. Swartz and Shete [3] did a simulation study of 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. They developed subjective priors, and selected the models based on their posterior probabilities. The evaluation of the procedure was carried out through examples and simulations. Correlated data was also considered by Kinney and Dunson [5], who selected both fixed and random effects using a fully Bayesian approach. Sha et al. [6] also used Bayesian variable selection, and were particularly interested in the case where k is very large. They used conjugate priors and cross-validation to check their predictions.

Our proposal is to use objective priors for model and for model parameters in the Bayesian variable selection problem. The justification for doing so is that since we are interested in variable selection, it seems that the experimenter will not have prior information about the influential covari-

ates, 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. There, clinical variables (age, health) are always included in the model, and the variable selector is run on the set of genes.

We will use the uniform prior distribution for the discrete set of models, and intrinsic priors for the regression coefficients. These intrinsic priors are known to satisfy many reasonable properties for model selection (Consonni and Rocca [7], Casella and Moreno [8]), and provide, under wide conditions, a consistent variable selection procedure (Casella et al. [9]).

Although the use of automatic objective priors allows us to circumvent the serious difficulty of eliciting priors for the regression coefficients, we need to implement the objective intrinsic priors in probit models. 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, or impossible, and hence so is the computation of the intrinsic priors for models comparison. However, we are able to 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, \dots, y_n) and compute the marginals under the models for all the subsets of the k potential set of regressors. Once this is done, 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, \dots, z_n) .

In addition, we develop a new controlled-dimension stochastic search algorithm, which allows us to search through all feasible subsets of covariates containing no more than q covariates, where q can be set by the experimenter. If the number of covariates k is greater than n , the model posterior probabilities cannot be computed. Thus, the search must be restricted to these lower dimensional models. We note that the incorporation of subjective prior information allows for selecting models with more than n covariates, or even without any sample information, but this is not the case with an objective analysis where there is no subjective prior input, as occurs with intrinsic priors.

The remainder of the paper is organized as follows. In Section 2 we briefly summarize for completeness the standard Bayesian model selection framework. We give the definition of intrinsic priors for the model parameters. We also explain more in 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, \dots, z_n) . In

Section 4 we derive a stochastic search algorithm that explores the entire model space, but restricts the search to models with a number of covariates specified by the user, and smaller than the sample size n , and in Section 5 we perform a simulation study to assess the accuracy of the procedure. The proposed model selection criterion is also applied to a real pneumonia data set with patients for which their gene expression 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 as the incomplete observation of a continuous variable, a latent variable hierarchical model (see, for example Albert and Chib [10]).

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 $\mathcal{M} = \{M_j, j = 1, \dots, N\}$. Let $p(\mathbf{z}|M_j) = \int p(\mathbf{z}|\theta_j, M_j)\pi(\theta_j|M_j)d\theta_j$ 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

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

where

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

is the Bayes factor to compare models M_j and M_1 , where M_1 is a particular fixed model. In regression analysis with a potential set of k regressors, $N = 2^k$ 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 value 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 (Girón et al. [11], Casella et al. [9]).

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 (see Section 2.3 for details).

2.2 Bayes Factors for Intrinsic Priors

Consider two general models

$$M_1 = \{p(y|\alpha), \pi^N(\alpha)\} \quad \text{and} \quad M_2 = \{p(y|\beta), \pi^N(\beta)\},$$

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 [12] addressed this problem by creating the “intrinsic Bayes factor”, a useable pseudo-Bayes factor constructed from the Bayes factor for the above improper priors as follows.

Given a sample $\mathbf{y} = (y_1, \dots, y_n)$, they defined 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 $\int p(\mathbf{y}_T|\alpha)\pi^N(\alpha)d\alpha$ and $\int p(\mathbf{y}_T|\beta)\pi^N(\beta)d\beta$ are positive and finite and there is no subsample of \mathbf{y}_T satisfying these conditions.

For a mTS \mathbf{y}_T consider the posterior of the parameters

$$\pi^N(\alpha|\mathbf{y}_T) \propto p(\mathbf{y}_T|\alpha)\pi^N(\alpha) \quad \text{and} \quad \pi^N(\beta|\mathbf{y}_T) \propto p(\mathbf{y}_T|\beta)\pi^N(\beta).$$

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

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

where $\mathbf{y}_{-T} = \mathbf{y} \setminus \mathbf{y}_T$. 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 to compare M_2 and M_1 for the whole sample \mathbf{y} , and $BF_{12}^N(\mathbf{y}_T)$ is the Bayes factor to compare M_1 and M_2 for the mTS \mathbf{y}_T . Both Bayes factors use improper default priors under both models.

To try to eliminate the dependence of the partial Bayes factor on \mathbf{y}_T , Berger and Pericchi [12] 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} . They called this arithmetic mean the Intrinsic Bayes factor $BF_{21}^{AI}(\mathbf{y})$, and it is given by

$$BF_{21}^{AI}(\mathbf{y}) = BF_{21}^N(\mathbf{y}) \times \text{mean}_{\mathbf{y}_T}[BF_{12}^N(\mathbf{y}_T)].$$

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 (Berger and Pericchi [12]). Later, Moreno et al. [13] proposed a ‘‘Limiting Intrinsic Procedure’’ for nested models (M_1 is nested in M_2 if for every α there is a β_α such that $p(y|\alpha) = p(y|\beta_\alpha)$) for defining the intrinsic priors $(\pi^N(\alpha), \pi^I(\beta))$. They suggested considering the Bayesian model for these priors, that is

$$M_1 = \{p(y|\alpha), \pi^N(\alpha)\} \quad \text{and} \quad M_2 = \{p(y|\beta), \pi^I(\beta)\}, \quad (2)$$

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

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

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 $\pi^I(\beta) = \pi^N(\beta) E_{\mathbf{y}_T|\beta}^{M_2} BF_{12}^N(\mathbf{y}_T)$. The Bayes factor for the intrinsic prior is then given by

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

We note that the Bayes factor (4) does not depend on the data set, but only on the sampling models. Moreno et al. [13] showed that it is a limit of Bayes factors for proper priors, and 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 $\mathbf{z} = (z_1, \dots, z_n)$, where $z_i, i = 1, \dots, n$, 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 ,

$j \leq k$. That is, this probit model M_j has the form

$$p(z_i|\theta_i, M_j) = \text{Bernoulli}(z_i|\theta_i) \quad \text{and} \quad \theta_i|M_j = \Phi(x_i'\beta_j), \quad (5)$$

where Φ denotes the standard normal distribution function, and β_j is a vector of dimension j . 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 $k + 1$.

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

For the sample $\mathbf{y} = (y_1, \dots, y_n)'$ 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} . For a generic model M_j with $j + 1$ regressors the alternative model is

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

where the design matrix \mathbf{X}_j has dimensions $n \times (j + 1)$. We aim for an automatic specification of the priors $\pi(\alpha)$ and $\pi(\beta)$, and hence we use the intrinsic methodology, 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 distribution for the sample \mathbf{y} under the null and the alternatives models for the intrinsic priors are formally written as

$$m_1(\mathbf{y}) = \int N_n(\mathbf{y}|\alpha\mathbf{1}_n, \mathbf{I}_n)\pi^N(\alpha)d\alpha, \quad (6)$$

$$m_j(\mathbf{y}) = \int \int N_n(\mathbf{y}|\mathbf{X}_j\beta_j, \mathbf{I}_n)\pi^I(\beta|\alpha)\pi^N(\alpha)d\alpha d\beta.$$

Since model M_1 is nested in M_j for any j , the Bayes factor for the intrinsic priors $BF_{j1}^{IP}(\mathbf{y}) = m_j(\mathbf{y})/m_1(\mathbf{y})$ provides a consistent model selection procedure; that is, provided that the true model is one of the 2^k regression models the procedure chooses this true model when the sample size grows to infinity (Casella et al. [9]).

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

$$m_j(\mathbf{z}) = \int_{A_1 \times \dots \times A_n} m_j(\mathbf{y}) d\mathbf{y}, \quad j = 1, \dots, 2^k, \quad (7)$$

where

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

and the required Bayes factors for intrinsic priors are $BF_{j1}^{IP}(\mathbf{z}) = m_j(\mathbf{z})/m_1(\mathbf{z})$, $j = 1, \dots, 2^k$.

3 Computing the Bayes Factor

To compute the Bayes factor for the observable sample \mathbf{z} we proceed by finding first the analytic expressions of both the intrinsic priors for the regression model and the marginal probabilities of \mathbf{y} given in (6). Later we give an algorithm to compute the Bayes factor for the responses \mathbf{z} defined in (7) based on the computation of multivariate normal distribution probabilities.

3.1 Intrinsic Priors for Normal Regression Models

Let \mathbf{Z}_T be the design matrix of a 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

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

Therefore, it follows that the mTS size is $j + 1$. We assume that \mathbf{Z}_T is standardized¹, 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, since the priors for α and β priors are proportional to 1, the intrinsic prior for comparing M_j versus M_1 , given in formula (3), becomes after some simplifications

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

¹Although this assumption is not necessary, it is typically good practice and stabilizes the numerics.

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

However, the matrix $\mathbf{Z}'_j \mathbf{Z}_j$ is unknown since it is a theoretical design matrix corresponding to the training sample y_T , although it can be estimated with the average of all possible submatrices of the design matrix \mathbf{X} with $j + 1$ rows (Girón et al. [14]). This average turns out to be (see Appendix A) $(j + 1)/(2n)$ $(\mathbf{X}'\mathbf{X})$, and therefore,

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

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

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

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

where $\Sigma_j = \mathbf{I}_n + 2n/(j+1) \mathbf{X}_j(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'_j$. 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

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

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

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

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

where $n s_y^2 = \sum_{i=1}^n (y_i - \bar{y})^2$ 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

Based on the observed sample \mathbf{z} we now compute the marginals $m_j(\mathbf{z})$ in (7) for $j = 1, \dots, 2^k$. From the expression (7) and (8) we have

$$\begin{aligned} m_j(\mathbf{z}) &= \int_{A_1 \times \dots \times A_n} m_j(\mathbf{y}) d\mathbf{y} = \int_{A_1 \times \dots \times A_n} \int_{-\infty}^{\infty} m_j(\mathbf{y}|\alpha) d\alpha d\mathbf{y} \\ &= \int_{-\infty}^{\infty} \int_{A_1 \times \dots \times A_n} N_n(\mathbf{y}|\alpha \mathbf{1}, \Sigma_j) d\mathbf{y} d\alpha. \end{aligned} \quad (10)$$

The integral over $A = A_1 \times \dots \times A_n$ is the probability of the hypercube A assigned by the a n -variate normal distribution. Genz and Bretz [15] 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 Genz et al. [16]. In other words, this function is able to compute $f_j(A|\alpha) = \int_A N_n(\mathbf{y}|\alpha\mathbf{1}, \Sigma_j) d\mathbf{y}$. The problem then reduces to the computation of

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

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 to compare M_j versus M_1 for the intrinsic priors and data \mathbf{z} can be computed by

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

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 MLE 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

We look for the model with maximum posterior probability (see Subsection 2.1) defined by $m_j(\mathbf{z})/m_0(\mathbf{z})$, or, equivalently, with maximum $m_j(\mathbf{z})$. As mentioned in Subsection 3.1, this quantity only can 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 application below, we have significantly more genes per patient than patients. Thus, we propose a random walk through the space of models with $q \leq n - 1$ covariates. The value of q is selected by the researcher having in mind that the smaller the value of q is the smaller the space for the search is, making the search algorithm more efficient.

We identify the models with a vector $\gamma \in \{0, 1\}^k$, where \mathcal{M}_γ includes the covariate j only if $\gamma_j = 1$. Since the intercept is always included in the model it is not considered in γ explicitly. For example, for $\gamma = (0, 1, 1, 0, \dots, 0)$, \mathcal{M}_γ is the model that includes only the intercept and the second and third covariates.

In theory, the vector γ could be any k -dimensional vector of 0's and 1's. There are 2^k such models, and we denote this model space by $\mathcal{M}_{k:k}$. However, the feasible search space is the set of

all models taken from $\mathcal{M}_{k:k}$ but having no more than $n - 1$ total covariates. In fact our search works for any $q \leq n - 1$ chosen by the researcher. There are $\sum_{j=0}^q \binom{k}{j}$ such models. We denote this model space by $\mathcal{M}_{k:q}$, and now describe a random walk with stationary distribution proportional to $m_\gamma(\mathbf{z})$ for $\gamma \in \mathcal{M}_{k:q}$.

We start by defining three vectors of indicator functions.

- $\delta \in \mathcal{M}_{k:k}$: This is a vector with 0 – 1 entries of the covariates in a latent model, where $\delta_j = 1$ indicates that the j^{th} covariate is in the latent model. Note that δ can choose models having more than q coefficients.
- $A = (a_1, \dots, a_k)$: A vector with 0 – 1 entries indicating the *active covariates* in the model, where $a_j = 1$ indicates that the j^{th} 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 $\sum_{j=1}^k a_j = q$.
- $\gamma \in \mathcal{M}_{k:q}$: This is a vector with 0 – 1 entries where $\gamma_j = 1$ indicates that the j^{th} of the covariates is in the model. The model has no more than q covariates, so $\sum_{j=1}^k \gamma_j \leq q$.

The initial point of the random walk is any $(A^{(0)}, \delta^{(0)}, \gamma^{(0)})$ such that $\sum_j a_j^{(0)} = q$, $\delta^{(0)} \in \mathcal{M}_{k:k}$, and $\gamma^{(0)} \in \mathcal{M}_{k:q}$. 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 probability r , $0 \leq r \leq 1$ is set by the user, 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 with coefficients

$$\delta'_{A,j} = \begin{cases} 0 & \text{with probability } 1/2 \\ 1 & \text{with probability } 1/2, \end{cases}$$

set $\delta' = (\delta'_A, \delta_{A^c}^{(t)})$ and construct the candidate $\gamma' = \delta' \star A^{(t)}$. 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 $0 \rightarrow 1$ or $1 \rightarrow 0$, whichever applies, to create δ'_A , set $\delta' = (\delta'_A, \delta_{A^c}^{(t)})$ and construct the candidate $\gamma' = \delta' \star A^{(t)}$.

Calculate the Metropolis-Hastings ratio $\alpha_1 = \min\{1, m_{\gamma'}(\mathbf{z})/m_{\gamma^{(t)}}(\mathbf{z})\}$ and set

$$\delta^{(t+1)} = \begin{cases} \delta' & \text{with probability } \alpha_1 \\ \delta^{(t)} & \text{with probability } 1 - \alpha_1 . \end{cases}$$

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

2. Update $A^{(t+1)}$: Choose one coefficient of $A^{(t+1)}$ at random, and swap it from $0 \rightarrow 1$ or $1 \rightarrow 0$, whichever applies, to create the candidate A' . Set $\gamma'' = \delta^{(t+1)} \star A'$

Calculate the Metropolis-Hastings ratio $\alpha_2 = \min\{1, m_{\gamma''}(\mathbf{z})/m_{\gamma^{(t+1)}}(\mathbf{z})\}$ and set

$$A^{(t+2)} = \begin{cases} A' & \text{with probability } \alpha_2 \\ A^{(t+1)} & \text{with probability } 1 - \alpha_2 . \end{cases}$$

Set $\delta^{(t+2)} = \delta^{(t+1)}$ and $\gamma^{(t+2)} = \delta^{(t+2)} \star A^{(t+2)}$.

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 $\mathcal{M}_{k;q}$ of feasible models, and has stationary distribution proportional to $m_\gamma(\mathbf{z})$, or equivalently, to the 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} .

A question that arises is whether our objective BFIP procedure should be compared with the variable selection procedure of Schwarz [17], the Bayesian information criterion (BIC). We will not do it here given that there have been extensive comparisons for normal linear models. Indeed, Casella et al. [9] note that for large sample sizes the variable selection procedure based on the BFIP is equivalent to that based on the Schwarz approximation, and for small sample sizes the Schwarz approximation has a very high Type I error (as high as 75%), which soon becomes very small as the sample size increases. Thus, the Schwarz approximation will be biased away from the null model for small n , or, more generally, in the cases where j is close to n . As n increases, the Type I error goes rapidly to 0, and the Schwarz approximation will then be biased toward the null model. In contrast, the intrinsic procedure has a less variable Type I error, being smaller than that of the Schwarz approximation for small n and somewhat larger for large n .

5.1 Simulation Study

Here we simulate the original normal data y 's to which the probit transformation z 's is applied, and examine the performance of the BFIP for both data sets, the y 's and the z 's. 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" when dichotomizing.

We analyze the results of the proposed variable selection procedure for moderate sample sizes and under four simulation scenarios. We first generate a sample $\mathbf{y} = (y_1, \dots, y_n)$ from the normal regression model $N_n(\mathbf{y}|\mathbf{X}\beta, \mathbf{I}_n)$ for $n = 20$, where we have 6 regressors. All covariate values were sampled from the uniform $U(0, 6)$. The probit sample was obtained 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 's, and according to the values of $m_j(\mathbf{y})$ in (9) when using the information in the y 's. We repeat the simulation 200 times and count the number of times that the method puts the true model in the first place of the ranking, and the number of times it is in the top three of the ranking.

The four scenarios are the same except for the value of the regression coefficients (including the intercept) of the true model that are given in the first column of Table 1. The second and third columns in Table 1 give the number of times the top model is the true one when using the sample y 's and the sample z 's, respectively, and the fourth and fifth columns the number of times the true model is in the top three model of the ranking.

From the second and third columns of Table 1, for the original data y 's the BFIP ranks the true model first in 74% of the simulations, while this percentage drops to 31% for the probit transformation z 's. The difference decreases as the number of covariates increases; for instance, the percentage of times the BFIP chooses the true model $(-1, 1, -1, 1, 0, 0)$ based on the y 's is of 97%, and this percentage is of 51% for the probit transformation z 's. These percentages grow considerably when we only require that the true model be in the top three of the ranking.

An overall conclusion is that the effectiveness of the model selector BFIP is smaller when using the z 's than when using the original data y 's, 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.

Table 1: PERFORMANCE OF THE BFIP MODEL SELECTION PROCEDURE, WHEN 200 SAMPLES OF SIZE 20 ARE SIMULATED FROM A NORMAL REGRESSION MODEL. THE LEFTMOST COLUMN GIVES THE TRUE SIMULATION MODEL, AND THE NEXT TWO COLUMNS GIVE THE PERCENTAGE OF TIMES THAT THE BFIP PROCEDURE RANKED THE TRUE MODEL FIRST, OR IN THE TOP THREE, USING BOTH THE y 'S AND THE z 'S.

True Model Coefficients	Top Choice %		Top Three %	
	BFIP (y)	BFIP (z)	BFIP (y)	BFIP (z)
-1,0,0,0,0,0	74	31	86	45
-1,1,0,0,0,0	83	61	99	82
-1,1,-1,0,0,0	92	70	98	92
-1,1,-1,1,0,0	97	51	100	86

5.2 Application in Association Genetics

This work was first motivated by the following association genetics problem. The data corresponds to 47 patients in an intensive care unit following trauma surgery. The physicians are concerned with how to better manage post-operative 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, gene expression of 296 genes was measured in peripheral blood, along with three clinical covariates: age, gender and abbreviated injury score (AIS). 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 is 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). Since we have a small sample size, we expect that the models with the highest BFIP will have few covariates, and we focus our search in the models with at most 10 genes (that is, considering the intercept and the three patient-level covariates, with at most $q = 14$ covariates).

Table 2: . THE 10 MODELS WITH THE HIGHEST BFIP FOUND BY THE STOCHASTIC SEARCH ALGORITHM.

Rank	Number	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

We ran 10,000 iterations of the stochastic search. The 10 models with the highest Bayes factors found by this search are shown in Table 2. Genes ERICH1 (non annotated), OLFM1 (“its abundant expression in the brain suggests that it may have an essential role in nerve tissue”, genebank) 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 or the most convenient.

As an illustration of possible use of this information, we look at some of the most frequently appearing genes in Table 2 and select, for example, the genes listed in the third row of the table. Assuming a multivariate normal prior for the regression coefficients with covariance matrix $100 \times \mathbf{I}$, Table 3 shows the 95% highest posterior density (HPD) credible intervals 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 consider a future patient with the same covariate values and computed the probability (and its HPD credible interval) of exhibiting the disease. The results are shown in the left panel of Figure 1, where we see that all of these future patients have a high 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 1 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

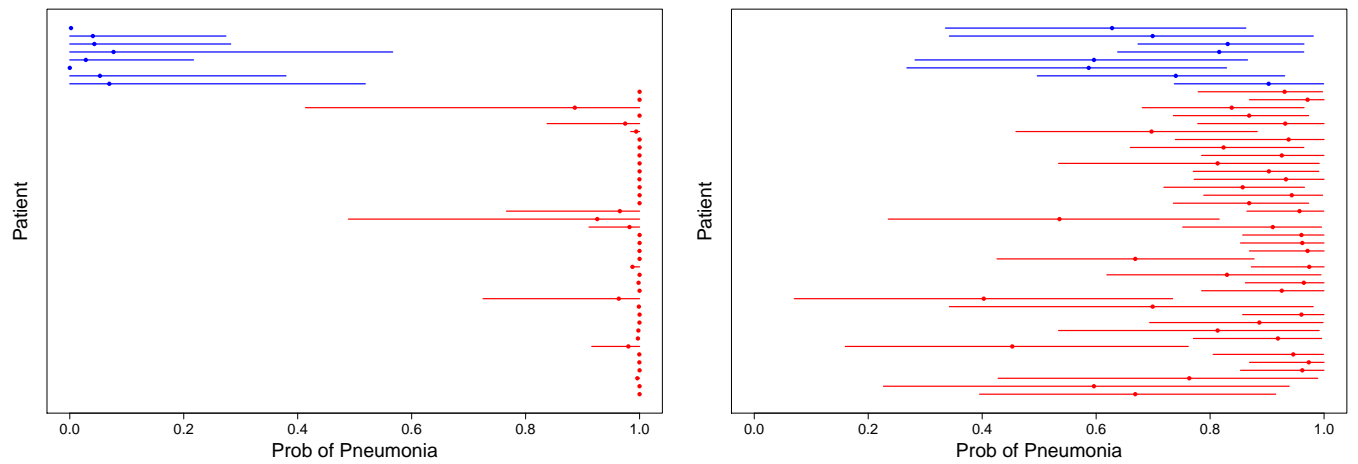


Figure 1: EACH LINE REPRESENTS THE 95% HPD CREDIBLE INTERVALS FOR THE PROBABILITY OF EXHIBITING PNEUMONIA FOR A (FUTURE) PATIENT WITH THE SAME COVARIATE VALUES AS ONE IN THE SAMPLE. THE DOT REPRESENTS THE MEAN. THE LINE IS RED WHEN THE PATIENT IN THE SAMPLE WITH THESE COVARIATE VALUES EXHIBITED THE DISEASE AND BLUE OTHERWISE. THE LEFT PANEL CORRESPONDS TO THE MODEL WITH 3 GENES FOUND BY USING THE PROPOSED MODEL SELECTION ALGORITHM AND RIGHT PANEL CORRESPONDS TO THE MODEL WITH ONLY CLINICAL VARIABLES AND WITHOUT GENETIC INFORMATION.

Table 3: 95%-HPD CREDIBLE INTERVAL AND MEAN FOR REGRESSION COEFFICIENTS IN MODEL IN THE SECOND ROW OF TABLE 2. PRIOR COVARIANCE MATRIX FOR COEFFICIENTS $100 \times \mathbf{I}$.

	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

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 and 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 subjective assigned by the researcher, so that 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.

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 p covariates out of a total of k possible covariates, then the prior for it is $\binom{k}{p}^{-1}$, and the prior for p is k^{-1} , $p = 1, \dots, k$. Under this prior the models with few (and many) covariates are assigned higher probability than that assigned to the models with approximately $k/2$ covariates. The use of this prior in a simulation study (results not shown) turned out to not be a desirable performer.

The proposed procedure does not use subjective prior information and, consequently, 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 so that it can be used to search models not driven by the posterior probability of models, but some different criteria. This search is a reliable alternative to the 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 taking into account the uncertainty introduced by the model selection procedure. More research needs to be done to make an accurate inference in the presence of model uncertainty.

References

- [1] Meier, L, Geer, SVD, Bühlmann, P, Zürich, ETH. The group lasso for logistic regression. *Journal of the Royal Statistical Society, Series B* 2008; .
- [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, Dey, DK. Variable selection for multivariate logistic regression models. *Journal of Statistical Planning and Inference* 2003; **111**:37–55.
- [5] Kinney, S, Dunson, D. 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. URL <http://www.jstor.org/stable/3695405>.
- [7] Consonni, G, Rocca, LL. Tests based on intrinsic priors for the equality of two correlated proportions. *Journal of the American Statistical Association* 2008; **103**:1260–1269.
- [8] Casella, G, Moreno, E. Objective bayes variable selection. *Journal of the American Statistical Association* 2006; **101**:157–167.
- [9] Casella, G, Girón, FJ, Martínez, ML, Moreno, E. Consistency of Bayesian procedures for variable selection. *The Annals of Statistics* 2009; **37**:1207– 1228.
- [10] Albert, J, Chib, S. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* 1993; **88**:669–679.
- [11] Girón, FJ, Moreno, E, Martínez, ML. *An Objective Bayesian Procedure for Variable Selection in Regression*. Birkhäuser Boston, 2006b.
- [12] Berger, JO, Pericchi, LR. The intrinsic bayes factor for model selection and prediction. *Journal of the American Statistical Association* 1996; **91**(433):109–122. URL <http://www.jstor.org/stable/2291387>.
- [13] 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. URL <http://www.jstor.org/stable/2670059>.
- [14] Girón, FJ, Martínez, M, Moreno, E, Torres, F. Objective testing procedures in linear models: Calibration of the p-values. *Scandinavian Journal of Statistics* 2006a; **33**(4):765–784.
- [15] Genz, A, Bretz, F. *Computation of Multivariate Normal and t Probabilities*. Lecture Notes in Statistics. Springer-Verlag, Heidelberg, 2009.
- [16] Genz, A, Bretz, F, Miwa, T, Mi, X, Leisch, F, Scheipl, F, Hothorn, T. *mvtnorm: Multivariate Normal and t Distributions*, 2010. URL <http://CRAN.R-project.org/package=mvtnorm>, r package version 0.9-92.
- [17] Schwarz, G. Estimating the dimension of a model. *Ann. Statist* 1978; **6**:461–464.

A Estimating $\mathbf{Z}_T^t \mathbf{Z}_T$

The following result is asserted in Girón *et al.* (2006), but the proof only appears in a technical report. We reproduce it here for completeness. We estimate $\mathbf{Z}_T^t \mathbf{Z}_T$ (\mathbf{Z}_T is the expected design matrix of the mTS) by averaging the design matrix of all possible minimal training samples. 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 , $l = 1, \dots, L$, 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,

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

Therefore,

$$\widehat{\mathbf{Z}_T^t \mathbf{Z}_T} = \frac{1}{L} \sum_l \mathbf{X}(l)^t \mathbf{X}(l) = \frac{j+1}{n} \mathbf{X}^t \mathbf{X}$$