View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by UCL Discovery

HIERARCHICAL BAYESIAN VARIABLE SELECTION IN THE PROBIT MODEL WITH MIXTURE OF NOMINAL AND ORDINAL RESPONSES

Eleftheria Kotti[*]*, Ioanna Manolopoulou and Tom Fearn*

Department of Statistical Science, University College London, United Kingdom
e.kotti.12@ucl.ac.uk, i.manolopoulou@ucl.ac.uk, t.fearn@ucl.ac.uk

ABSTRACT

Multi-class classification problems have been studied for pure nominal and pure ordinal responses. However, there are some cases where the multi-class responses are a mixture of nominal and ordinal. To address this problem we build a hierarchical multinomial probit model with a mixture of both types of responses using latent variables. The nominal responses are each associated to distinct latent variables whereas the ordinal responses have a single latent variable. Our approach first treats the ordinal responses as a single nominal category and then separates the ordinal responses within this category. We introduce sparsity into the model using Bayesian variable selection (BVS) within the regression in order to improve variable selection classification accuracy. Two indicator vectors (indicating presence of the covariate) are used, one for nominal and one for ordinal responses. We develop efficient posterior sampling. Using simulated data, we compare the classification accuracy of our method to existing ones.

Index Terms— Hierarchical probit model, Bayesian variable selection, classification, nominal and ordinal responses, latent variables

1. INTRODUCTION

The motivation for this research was an application to Barretts oesophagus, where progression from healthy through three stages of the disease can be viewed as a continuum and thus may benefit from being treated as an ordinal sequence, whilst a possible progression to cancer is qualitatively different and not part of this continuum.

When analyzing multi-class responses, it is important to note whether each one of the responses is nominal or ordinal. Some types of models are appropriate only for ordinal responses and others for nominal. For example, support vector machines (SVMs) have been introduced for pure nominal [1] and, recently, pure ordinal [2]. In frequentist statistics, the four most common models are multinomial probit model for pure nominal/ordinal responses and multinomial logit model for pure nominal/ordinal responses [3]. The Bayesian approach for the multinomial probit models uses latent variables [4]. Treating the responses as entirely nominal when in fact they are a mixture of nominal and ordinal leads to loss of information which may be important.

Including all the predictors in the model is inefficient, as most of them do not inform the response variable. Introducing sparsity in the model can improve prediction accuracy, especially in the context of a large number of predictors [5]. Variable selection has been implemented, for example, via classification trees (CT) and random forests (RF) using pure nominal [1] or ordinal responses [6], [7]. In frequentist statistics, least absolute shrinkage and selection operator (Lasso) [8] is one of the best known techniques for variable selection. In Bayesian statistics, penalized methods have a Bayesian interpretation that can take advantage of prior knowledge. Priors that offer penalisation in variable selection are usually a mixture of two distributions, known as a spike and slab prior [9]. A classical choice is a spike at zero and a normal slab. We focus on BVS in multi-class classification problems, where the literature is limited. References in multi-class BVS have studied either just the (nominal) multinomial probit model [10] or just the ordinal multinomial probit model [11].

In the current study we build a hierarchical model for a mixture of nominal and ordinal responses with BVS for the latent regression, combining [10] and [11] under different assumptions. The advantage of our model is that it harnesses the structural feature of both nominal and ordinal responses. The proposed variable selection approach gives high classification accuracy by allowing the model to incorporate the additional information of the ordinal structure, as well as using a different set of important variables for the ordinal and nominal pieces. The added flexibility can accommodate a wider range of features observed in the data which can be learnt efficiently through our tailored posterior sampler. The new method can be applied, for example, in finance (loan applications, income), bioinformatics (microarrays) and social sciences (nursery application, students' grades).

[*]E. Kotti would like to thank the 'Foundation for Education and European Culture' for the financial support.

2. HIERARCHICAL MODEL FOR MIXTURE OF RESPONSES

We denote the observed data by $\mathbf{X}$ and $\mathbf{Y}$, where $\mathbf{X}$ is an $n \times p$ design matrix (n is the number of samples, p is the number of variables) and $\mathbf{Y}$ is an $n \times 1$ response vector. We assume that from the total of M responses that the response vector can take, $|\mathbf{t}|$ are ordinal, $\mathbf{t} = (t_0, \ldots, t_{|\mathbf{t}|-1})$. Then the remaining $M - |\mathbf{t}|$ responses are nominal. Assuming that zero is the 'baseline' category, we proposed a hierarchical multinomial probit model using latent variables for the mixture of nominal and ordinal responses. We use a latent variable representation of the response classes, where subsets of continuous latent variables correspond to the different response classes.

2.1. Nominal approach: treat ordinal responses as one nominal

At the first step, we treat all ordinal responses as one nominal response (one extra group), and so the total number of nominal responses is $M - |\mathbf{t}| + 1$. Then, we built a probit model with latent variables that takes into account the nominal responses (including the one group of ordinal responses). Since zero is the 'baseline', we need $s = M - |\mathbf{t}|$ latent variables. We introduce $\mathbf{Z}$ as the $n \times s$ matrix of latent variables and assume matrix normal (MN) distribution

$$\mathbf{Z} - \mathbf{1}_n \boldsymbol{\alpha}^{'} - \mathbf{XB} \sim MN(\mathbf{I}_n, \boldsymbol{\Sigma}), \tag{1}$$

where $\mathbf{1}_n$ is a n dimensional column vector of ones, $\boldsymbol{\alpha}$ is the $s \times 1$ vector of intercepts (prime denotes transposition) and $\mathbf{B}$ is a $p \times s$ matrix of regression coefficients. $\boldsymbol{\Sigma}$ is the $s \times s$ covariance matrix that can be unknown [10] or known $\boldsymbol{\Sigma} = \sigma_r^2 \mathbf{I}_s, r = 1, \ldots, s$. We denote by $Z_i^{\star} = \max_{1 \le r \le s}\{Z_{i,r}\}$, $i = 1, \ldots, n$. The relation between nominal responses and latent variables is given by

$$Y_i = \begin{cases} 0, & \text{if } Z_i^{\star} \le 0 \\ r, & \text{if } Z_i^{\star} > 0 \text{ and } Z_{i,r} = Z_i^{\star}. \end{cases} \tag{2}$$

In order to perform variable selection for the model of $M - |\mathbf{t}| + 1$ nominal responses, we use a common indicator vector $\boldsymbol{\gamma}$ across different latent variables,

$$\gamma_j = \begin{cases} 1, & \text{if } B_{j,r} \neq 0 \text{ for all } r, \\ 0, & \text{if } B_{j,r} = 0 \text{ for all } r, \end{cases}$$

where $B_{j,r}$ is the entry in the j-th row and r-th column of $\mathbf{B}$, for $j = 1, \ldots, p$ and $r = 1, \ldots, s$. Selection of the j-th variable corresponds to $\gamma_j = 1$.

2.2. Treat ordinal responses

At the second hierarchical step, ordinal responses have a single latent variable and are specified via a boundary vector $\mathbf{k} = (k_0 = -\infty, k_1 = 0, k_2, \ldots, k_{|\mathbf{t}|} = +\infty)$ on the latent variables. Let us denote by $\underline{\mathbf{Z}}$ the $n \times 1$ vector of latent variables that is distributed as multivariate normal (MVN) with common variance across different responses

$$\underline{\mathbf{Z}} - \mathbf{1}_n \lambda + \mathbf{X}\boldsymbol{\beta} \sim MVN(0, \sigma^2 \mathbf{I}_n), \tag{3}$$

where the scale λ is the intercept, $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients, σ^2 is the variance which may be known ($\sigma^2 = 1$) [11] or unknown. The relationship between the latent variable and ordinal responses, according to [4], is the following

$$Y_i = t_g, \text{ if } k_g < Z_i \le k_{g+1}, \tag{4}$$

where $g = 0, \ldots, |\mathbf{t}| - 1$. Note that in the ordinal case $\underline{\mathbf{Z}}$ is a latent vector in contrast to nominal where $\mathbf{Z}$ is a matrix.

In this case we use an indicator vector $\boldsymbol{\xi}$, distinct from $\boldsymbol{\gamma}$, to indicate the inclusion or exclusion of the coefficient β_j of ordinal response.

2.3. Model summary

In summary, the hierarchical multinomial probit model with mixture of nominal and ordinal responses using latent variables is given via (1) and (3) under two different settings: covariance matrix $\boldsymbol{\Sigma}$ and variance σ^2 are known or unknown.

3. HIERARCHICAL BVS FOR MIXTURE OF RESPONSES

3.1. Prior distributions

For the model in Sec. 2.1 the priors are: $\boldsymbol{\alpha}^{'} - \boldsymbol{\alpha}_0^{'} \sim MN(h, \boldsymbol{\Sigma})$, $\mathbf{B}_{\boldsymbol{\gamma}} - \mathbf{B}_{0\boldsymbol{\gamma}} \sim MN(\mathbf{H}_{\boldsymbol{\gamma}}, \boldsymbol{\Sigma})$ which corresponds to the non-zero coefficients, $\gamma_j \sim \text{Bernoulli}(w_{(\text{nom})})$ and $\boldsymbol{\Sigma} \sim \text{InverseWishart}(\delta; \mathbf{Q})$, where $\delta = n - s + 1$ [10] or we consider that the covariance matrix is fixed.

For the model in Sec. 2.2 the priors are: $\lambda \sim N(\lambda_0, \sigma^2 h)$, for non-zero coefficients $\boldsymbol{\beta}_{\boldsymbol{\xi}} \sim MVN(\boldsymbol{\beta}_{0\boldsymbol{\xi}}, \sigma^2 \mathbf{F}_{\boldsymbol{\xi}})$, $\xi_j \sim \text{Bernoulli}(w_{(\text{ord})})$ [11] and we consider a conjugate prior $\sigma^2 \sim \text{InverseGamma}(d_1, d_2)$. Finally, $k_2, k_3, \ldots, k_{|\mathbf{t}|-1}$ are uniformly distributed on the interval $(0, +\infty)$ subject to the constrain that $k_2 < k_3 < \ldots < k_{|\mathbf{t}|-1}$.

In both cases sparsity in the models is considered by assigning spike and slab priors on the coefficients, with the spike at zero. In addition, the probability of success of a Bernoulli distribution corresponds to the probability of including variables in the model a-priori.

3.2. Posterior inference

If the covariance matrix and variance of latent variables are unknown, posterior inference for the model in Sec. 2.1 is done in [10], but we need to derive the posteriors for the model in Sec. 2.2 (ordinal responses). Setting $\lambda_0 = 0$ and $\boldsymbol{\beta}_{0\boldsymbol{\xi}} = \mathbf{0}$ we derive the conditional distribution of the vector of latent

variable, which is a multivariate Student distribution (MVT) [12],

$$\underline{\mathbf{Z}}|\boldsymbol{\xi},\mathbf{k},\mathbf{X},\mathbf{Y} \sim MVT\left(2d_1;\mathbf{0},\frac{d_2}{d_1}\mathbf{P}_{\boldsymbol{\xi}}\right)\prod_{i=1}^{n}\mathbb{1}(Z_i \in A_i),$$

where $\mathbf{P}_{\boldsymbol{\xi}} = \mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n^{'} + \mathbf{X}_{\boldsymbol{\xi}}\mathbf{F}_{\boldsymbol{\xi}}\mathbf{X}_{\boldsymbol{\xi}}^{'}$ and $\mathbb{1}(.)$ is the indicator function of the set $A_i = \{Z_i : k_g < Z_i \leq k_{g+1}\}$ if $Y_i = t_g$ according to (4). The conditional distribution of boundaries is uniform [11].

On the other hand, if the covariance matrix and variance are known, we can derive the posterior inference for the model in Sec. 2.1 (nominal responses), similarly to [11]. Setting $\boldsymbol{\alpha}_0 = \mathbf{0}$ and $\mathbf{B}_{0\boldsymbol{\gamma}} = \mathbf{0}$, we can sample from latent matrix $\mathbf{Z}$ according to

$$\mathbf{Z}|\boldsymbol{\gamma},\mathbf{X},\mathbf{Y} \sim MN(\mathbf{P}_{\boldsymbol{\gamma}},\boldsymbol{\Sigma})\prod_{i=1}^{n}\mathbb{1}(\mathbf{Z}_i \in R_i),$$

where $\mathbf{P}_{\boldsymbol{\gamma}} = \mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n^{'} + \mathbf{X}_{\boldsymbol{\gamma}}\mathbf{H}_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}}^{'}$ and the set R_i can be calculated according to (2).

Since the full conditional distributions of both indicator vectors $\boldsymbol{\gamma}$ and $\boldsymbol{\xi}$ do not have a closed form solution, the Metropolis algorithm [13] is applied within the Gibbs step, using a symmetric proposal, with probability of 0.5 to add or delete a variable and probability 0.5 to swap two variables. In both cases, to speed up the process of sampling $\boldsymbol{\gamma}$'s and $\boldsymbol{\xi}$'s we apply QR-decomposition [14].

3.3. Method

We would like to build a variable selection algorithm based on the hierarchical mixture model of both types of responses that is proposed in the previous section. We construct an algorithm that consist of the following two parts. BVS approach is implemented for nominal (included the one group of ordinal) responses with unknown parameters $\mathbf{Z}$ and $\boldsymbol{\gamma}$ (see part A of Algorithm 1). BVS approach is implemented for ordinal responses for the unknown parameters $\underline{\mathbf{Z}}$, $\boldsymbol{\xi}$ and $\mathbf{k}$ (see part B of Algorithm 1). The two models may have some variables in common but the models that are most frequently selected by the approach may be different for those two parts. Taking the combined results of the two parts into account we can identify the best models and the important variables for the mixture of nominal and ordinal responses.

Let us denote the sample of j_A-th iteration of the part A with the upper index (j_A), and the sample of j_B-th iteration of the part B with the upper index (j_B). We construct the Gibbs steps as summarized in Algorithm 1. This consists of two parts, A and B, and the final conclusion, which is the combination of both. Since the two parts are independent, for the training procedure the order of the parts it does not matter.

Algorithm 1 Hierarchical BVS: mixture of nominal and ordinal responses

Part A: BVS on $M - |\mathbf{t}| + 1$ nominal responses ($M - |\mathbf{t}|$: nominal responses and all ordinal responses are treated as one nominal response)

1: Initialize values $\boldsymbol{\gamma}^{(0)}$ and $\mathbf{Z}^{(0)}$
2: Draw $\boldsymbol{\gamma}^{(j_A)}$ from $p(\boldsymbol{\gamma}|\mathbf{Z}^{(j_A-1)},\mathbf{X},\mathbf{Y})$
3: Draw $\mathbf{Z}^{(j_A)}$ from $p(\mathbf{Z}|\boldsymbol{\gamma}^{(j_A)},\mathbf{X},\mathbf{Y})$
4: Repeat steps 2 and 3 until the number of iterations achieved and stop (Results: VS_A and MS_A, see footnote 1 for abbreviations)

Part B: BVS on $|\mathbf{t}|$ ordinal responses

1: Initialize values $\boldsymbol{\xi}^{(0)}$, $\underline{\mathbf{Z}}^{(0)}$ and $\mathbf{k}^{(0)}$
2: Draw $\boldsymbol{\xi}^{(j_B)}$ from $p(\boldsymbol{\xi}|\underline{\mathbf{Z}}^{(j_B-1)},\mathbf{X},\mathbf{Y})$
3: Draw $\mathbf{k}^{(j_B)}$ from $p(\mathbf{k}|\underline{\mathbf{Z}}^{(j_B-1)},\mathbf{X},\mathbf{Y})$
4: Draw $\underline{\mathbf{Z}}^{(j_B)}$ from $p(\underline{\mathbf{Z}}|\boldsymbol{\xi}^{(j_B)},\mathbf{k}^{(j_B)},\mathbf{X},\mathbf{Y})$
5: Repeat steps 2, 3 and 4 until the number of iterations achieved and stop (Results: VS_B and MS_B)

Combine parts A **and** B
$VS = VS_A \cup VS_B$ and $MS = MS_A \cup MS_B$

3.4. Classification and prediction

The classification procedure for a new sample is done according to the following process: First, we find the best model (the model with the highest posterior probability) for nominal responses and we do predictions according to (2). If the predicted response is nominal, then we finish the prediction. If the predicted response corresponds to the group of ordinal responses that are treated as one nominal case, then we find the best model for ordinal responses and we do predictions according to (4).

4. RESULTS

The experimental study was performed using simulated data from the proposed probit model with multi-class nominal and ordinal responses. Simulations are created according to the two step approach.

We ran two different simulations to cover the following scenarios: (i) $\boldsymbol{\Sigma}$ and σ^2 are unknown and (ii) $\boldsymbol{\Sigma}$ and σ^2 are known. In both cases, for generating simulated data we set $n = 100$, $p = 200$, $M = 5$, $\mathbf{t} = [1,2,3]$ and the error terms of latent variables distributed as standard normal. The majority of $\mathbf{B}$'s entries (related to the nominal responses) are zero except for $B_{[3,8],1} = [0.85,-0.81]$ and $B_{[3,8],2} = [-0.83,-0.62]$. In addition, the majority of $\boldsymbol{\beta}$'s

[1] VS_A (VS_B): the set of selected variables for nominal (ordinal) responses using marginal probabilities, VS: the final set of selected variables (nominal and ordinal responses jointly), MS: the corresponding set of variables in the most probable model.

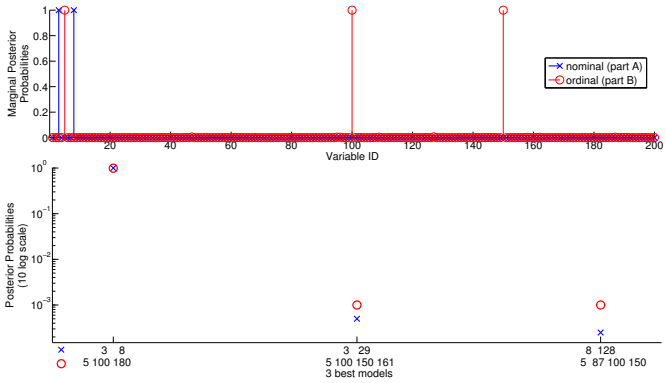

Fig. 1. Marginal posterior probabilities (top) and posterior probabilities in log. scale (bottom) of average of chains for scenario (i).

entries (related to the ordinal responses) are zero except for $\beta_5 = -1.4$, $\beta_{100} = 1.2$ and $\beta_{150} = 1.3$.

For scenario (i), in the variable selection approach we set the values for the hyperparameters. In part A, the hyperparameters of the unknown covariance matrix $\boldsymbol{\Sigma}$ are $\delta = 3$, $\mathbf{Q} = \mathbf{I}_s$ [10]. In addition, we set $h = 10^6$ (flat prior for the intercept), $\mathbf{H}_{\boldsymbol{\gamma}} = c_1\mathbf{I}_{p_{\boldsymbol{\gamma}}}$ (easy for calibration), where $p_{\boldsymbol{\gamma}} = \sum_{j=1}^{p}\gamma_j$, $c_1 = 5$ [10] and $w_{(\text{nom})} = 2/200$. We initialize the $\boldsymbol{\gamma}^{(0)}$ selecting randomly two variables. Then we initialize $\mathbf{Z}^{(0)}$. We ran four different chains with 5000 iterations after 2000 burn-in iterations. In part B, the hyperparameters of the variance σ^2 are $d_1 = \delta/2 = 1.5$ and $d_2 = 0.5$ (inverse Gamma is the univariate case of inverse Wishart). In addition, we select $h = 10^6$, $\mathbf{F}_{\boldsymbol{\xi}} = c_2\mathbf{I}_{p_{\boldsymbol{\xi}}}$, $c_2 = 5$ and $w_{(\text{ord})} = 3/200$. We initialize the $\boldsymbol{\xi}^{(0)}$ selecting randomly three variables. Then, we initialize $\underline{\mathbf{Z}}^{(0)}$ and $\mathbf{k}^{(0)}$. Fig. 1 contains the results of variable and model selection of two parts, for the average of the chains. Our proposed algorithm correctly identifies the individual important variables 3, 8 (part A) and 5, 100, 150 (part B). The remaining variables have marginal posterior probabilities close to zero. In addition, the combination of variables 3 and 8 is the best model for part A (bottom of Fig.1: first line of $x-$axis) and the model with the variable 5, 100, 150 is the best model for part B (bottom of Fig.1: second line of $x-$axis), with posterior probability much higher than the the second, third, etc. following best models respectively.

For scenario (ii), in the variable selection approach we set $\boldsymbol{\Sigma}$ equal to $\mathbf{I}_s$ ($\sigma^2 = 1$, known). The remaining parameters and hyperparameters are the same as in the scenario (i). The figure is similar to the Fig. 1.

In order to do prediction of a new (future) sample, we generate new data (a hundred samples that will be referred to as the test set) according to the parameters of each specific scenario. We pick from the test design matrix only the variables that had been selected after applying Algorithm 1 and we do prediction. Then, we repeat one hundred times the process of generating test sets. Based on the inherent amount of error that the simulated data have, the highest classification accuracy that we can achieve is on average 66.02%. The proposed method achieves on average a 61.55% classification accuracy for the test set, which is very close to the highest possible.

Table 1. The comparison of classification accuracy for the test set after applying variable selection approaches with $\boldsymbol{\Sigma}$ to be unknown.

	Accuracy (%)	
	Nominal	Ordinal
'Highest' accuracy	66	
Our proposed method	60	
BVS	46	23
Lasso	50	47
CT	41	29
RF	48	31
SVMs	36	35

In order to compare the proposed method with existing methods, we select one test set (out of a hundred). For scenario (i), the highest classification accuracy for this test set is 66% and our method achieves classification accuracy 60%. Table 1 contains the results of the comparison with other methods that were proposed for pure nominal or pure ordinal responses. The BVS with the wrong latent structure (nominal) seems to be able to identify a satisfactory amount of the structure by using latent variables, here in a higher 4-dimensional space, but is beaten by our proposed BVS method. The poor performance of the BVS with the wrong latent structure (ordinal) is not surprising in a situation where the simulated data have a structure that cannot be modelled properly. When the responses are all treated as ordinal, the other methods cope better with the misspecification, but are beaten by our proposed BVS with the correct latent structure. For scenario (ii), our method achieves 61% accuracy (66% is the 'highest' possible) in both cases beating existing methods.

5. CONCLUSIONS

We propose a hierarchical Bayesian probit model, that is appropriate for mixtures of nominal and ordinal responses, using latent variables. Then we proposed a hierarchical BVS method for model selection using mixture of nominal and ordinal responses. The hierarchical approach consist of two parts: ordinal responses treated as one nominal response and apply BVS for nominal responses and afterwards apply BVS just for ordinal responses. We use two indicator vectors (one for each hierarchical part) to represent the presence of absence of a predictor in the regression. The hierarchical proposed algorithm for variable selection is simple and computationally efficient because the nominal and ordinal parts are decoupled and can be performed in parallel. The novel method does automated classification via the model and achieves better classification accuracy compared to other methods.

6. REFERENCES

[1] K.P. Murphy, *Machine learning: a probabilistic perspective*, MIT press, 2012.

[2] W. Chu and S.S. Keerthi, "Support vector ordinal regression," *Neural computation*, vol. 19, no. 3, pp. 792–815, 2007.

[3] P. McCullagh and J.A. Nelder, *Generalized linear models*, vol. 37, CRC press, 1989.

[4] J.H. Albert and S. Chib, "Bayesian analysis of binary and polychotomous response data," *Journal of the American statistical Association*, vol. 88, no. 422, pp. 669–679, 1993.

[5] R. Tibshirani T. Hastie and M. Wainwright, "Statistical learning with sparsity," 2015.

[6] R. Piccarreta, "Classification trees for ordinal variables," *Computational Statistics*, vol. 23, no. 3, pp. 407–427, 2008.

[7] S. Janitza, G. Tutz, and A.L. Boulesteix, "Random forests for ordinal response data: prediction and variable selection," Tech. Rep., Ludwig-Maximilians-University Munich, Department of Statistics, 2014.

[8] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of statistical software*, vol. 33, no. 1, pp. 1, 2010.

[9] E.I. George and R.E. McCulloch, "Variable selection via Gibbs sampling," *Journal of the American Statistical Association*, vol. 88, no. 423, pp. 881–889, 1993.

[10] N. Sha, M. Vannucci, M.G. Tadesse, P.J. Brown, I. Dragoni, N. Davies, T.C. Roberts, A. Contestabile, M. Salmon, C. Buckley, and F. Falcian, "Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage," *Biometrics*, vol. 60, no. 3, pp. 812–819, 2004.

[11] D. Kwon, M.G. Tadesse, N. Sha, R.M. Pfeiffer, and M. Vannucci, "Identifying biomarkers from mass spectrometry data with ordinal outcome," *Cancer informatics*, vol. 3, pp. 19, 2007.

[12] J. Geweke, "Efficient simulation from the multivariate normal and Student-t distributions subject to linear constraints and the evaluation of constraint probabilities," in *Computing science and statistics: Proceedings of the 23rd symposium on the interface*. Citeseer, 1991, pp. 571–578.

[13] P.J. Brown, M. Vannucci, and T. Fearn, "Bayesian wavelength selection in multicomponent analysis," *Journal of Chemometrics*, vol. 12, no. 3, pp. 173–182, 1998.

[14] P.J. Brown, M. Vannucci, and T. Fearn, "Bayes model averaging with selection of regressors," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 64, no. 3, pp. 519–536, 2002.