Bayesian variable selection for probit models with an application to clinical diagnosis

by
Eleftheria Kotti

A Thesis Submitted for the Degree of
Doctor of Philosophy

in the
Faculty of Mathematical and Physical Sciences
Department of Statistical Science
University College London

April 2017

Declaration of Authorship

I, Eleftheria Kotti, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signature: .....

Abstract

My research focuses on the development of a probabilistic model for the classification using spectral measurements of tissue from different stages in the progression of Barrett's oesophagus (BE) and the implementation in this model of variable selection with the aim of improving the classification accuracy.

In Chapter 1, a brief introduction to the BE disease, to spectroscopy, and to importance of variable selection is presented. Chapter 2 focuses on (penalized) likelihood methods for variable selection, including also evaluation measures for the performance of prediction models. Chapter 3 introduces Bayesian variable selection (BVS) using a probit model with binary responses. Then, BVS is studied under different prior assumptions for the coefficients and for the indicator vector (indicating if the variable is important). The next chapter contains the results of applying these different assumptions either on real or on simulated binary datasets.

The remaining chapters regard the extension of BVS from binary to multi-class responses. Multi-class classification problems have been studied for pure nominal and pure ordinal responses (Chapter 5). However, there are cases with both types of responses, e.g. BE disease. We develop a BVS approach for which the stages of the disease are a mixture of nominal and ordinal responses. To address this problem we build three probit models based on latent variables: (i) a decomposed approach using two indicator vectors, one for nominal and one for ordinal responses (Chapter 6), (ii) BVS approach using a common indicator vector (Chapter 7), and (iii) BVS approach using an indicator matrix, which is a collection of indicator vectors (Chapter 8). Finally, Chapter 9 contains the results of applying the proposed methods to BE for clinical diagnosis and comparing with existing methods. The last chapter contains the conclusions and suggestions for future directions.

Acknowledgements

First and foremost, I would like to express my special appreciation and thanks to my primary supervisor Prof. Tom Fearn, for encouraging my research, providing priceless research advice, and for always being happy to help me. Also, many thanks to Dr. Ioanna Manolopoulou, who was an active secondary supervisor, providing valuable advice and feedback.

Thanks to the multidisciplinary team of the UCL Department of Structural and Molecular Biology, the specialists in infrared spectroscopy, and the UCL Hospital Gastrointestinal and Histopathology unit that provided the data of the Barrett's oesophagus (BE) disease for the application part of this thesis. Special thanks to Dr. Liberty Foreman who helped me understand the biological concepts of BE.

I gratefully acknowledge the funding sources, consisting of UCL and the Foundation for Education and European Culture, that made my Ph.D. study possible by covering my tuition fees as well as my living and travel expenses.

I am very grateful to all my friends in London and abroad who supported me during my Ph.D. studies and helped me to achieve a good work-life balance, for example via playing board games and doing sports. Special thanks to my partner who encouraged me, listened to my concerns, and helped me to overcome the difficulties during my studies.

Many thanks to my parents for their continued support and encouragement. Also, I would like to thank my family in the UK, my sister and her husband, who with their experience helped me along the Ph.D. journey, as well as their little son with whom I had many good times playing and having fun.

Contents

List of Figures

List of Tables

Acronyms

AIC	Akaike information criterion
APD	Adjacent-paired dataset
AUC	Area under the curve
BE	Barrett's oesophagus (according to the US "esophagus")
BVS	Bayesian variable selection
BIC	Bayesian information criterion
CDF	Cumulative distribution function
FTIR	Fourier transform infrared
HGD	High grade dysplasia
IMD	Intercepted-matched dataset
IR	Infrared
k-NN	k-nearest neighbors
LASSO	Least absolute shrinkage and selection operator
LDA	Linear discriminant analysis
LGD	Low grade dysplasia
LOOCV	Leave-one-out cross validation
MCMC	Markov chain Monte Carlo
MH	Metropolis-Hastings
NDBE	Non-dysplastic Barrett's oesophagus
OAC	Oesophageal adenocarcinoma
PCA	Principal component analysis
PDF	Probability distribution function
ROC	Receiver operating characteristic
SFS	Sequential forward selection
SQ	Squamous
SVM	Support vector machine
VS	Variable selection

Subscript notation for matrices

All matrices are denoted with bold upper case letters. When two subscripts are used, then the first corresponds to a row of the matrix and the second to a column. On the other hand, when only one subscript is used then the subscript is an indicator vector, which is how we select rows and columns.

Related publications

Part of the work included in this thesis has been presented in international peer-reviewed conferences and workshops:

- Kotti, E., Manolopoulou, I., and Fearn, T. (2016c). Hierarchical Bayesian variable selection in the probit model with mixture of nominal and ordinal responses. *In 2016 IEEE Workshop on Statistical Signal Processing (SSP)* pp. 576–580. IEEE.
- Kotti, E., Manolopoulou, I., and Fearn, T. (2016b). Bayesian variable selection for mixture of nominal and ordinal responses via an indicator matrix. In *2nd UCL Conference on the Theory of Big Data.*
- Kotti, E., Manolopoulou, I., and Fearn, T. (2016a). Bayesian variable selection for a mixture of nominal and ordinal responses. In *13th International Society for Bayesian Analysis World Meeting (ISBA).*
- Kotti, E., Manolopoulou, I., and Fearn, T. (2015). Bayesian variable selection in the probit model with mixture of nominal and ordinal responses. In *Workshop Autonomous Citizens: Algorithms for Tomorrow's Society.*

Chapter 1

Introduction

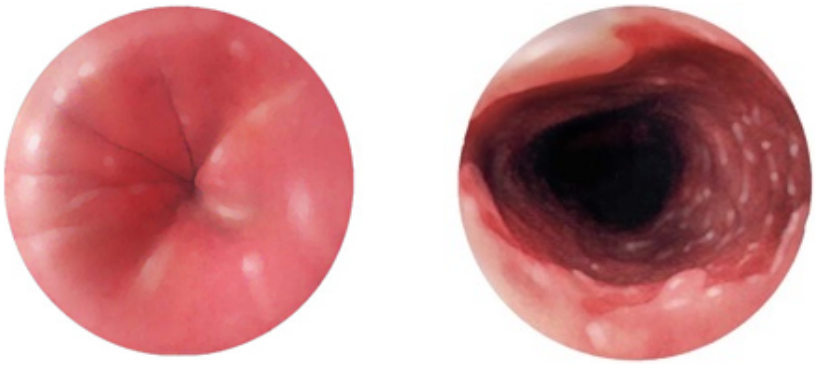
The theme of this thesis is the application of Bayesian variable selection methods to high-dimensional spectroscopic data arising from a study aiming to use such measurements to classify the stages of Barrett's oesophagus. As an introduction to the topic, this chapter has three parts: some background to Barrett's oesophagus, a brief introduction to spectroscopy applied to tissue samples and an introduction to variable selection from a statistical standpoint, which includes the motivation for variable selection.

1.1 Barrett's oesophagus

Background

The oesophagus connects the mouth to the stomach via a food pipe whose normal lining is made up of squamous epithelium cells. In Barrett's oesophagus (BE), also known as Barrett syndrome or columnar epithelium lined lower oesophagus, the cells of the food pipe have started to be replaced by another cell type normally found lower in the gut. This is called metaplasia. Metaplasia can be illustrated via on endoscopic photograph. A photograph taken with an endoscopic camera in the oesophagus demonstrates the difference between squamous epithelium (light pink) and metaplastic epithelium (dark pink), Figure 1.1. This distinction between healthy tissue and BE can be seen easily. However, if we want a more detailed diagnosis we need a more sophisticated technique like spectroscopy.

Many people with BE may have metaplasia, but not have cells that are growing abnormally, which is called dysplasia. Before oesophageal adenocarcinoma (OAC), also known as cancer, occurs, there are three different stages of BE (according to the UK classification): non-dysplastic Barrett's oesophagus (NDBE), low grade dysplasia (LGD) and high grade dysplasia



(a) Normal oesophagus. (b) BE.

Figure 1.1: Normal versus Barrett's oesophagus. Light pink corresponds to squamous epithelium (healthy part) and dark pink to metaplastic epithelium (diseased part) of the oesophagus. By permission of Mayo Foundation for Medical Education and Research. All rights reserved.

(HGD). The grade of a cell is what it looks like under a microscope. The less normal the cells look, the higher the grade is: cells with NDBE are slightly abnormal, those with LGD are mildly abnormal and cells with HGD are very abnormal. The 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 in any classification model, whilst the progression to cancer where is occurs is qualitatively different and it may be advantageous not to treat it as part of this continuum. Some physiological features of each BE stage and of squamous (SQ), or healthy, tissue are represented in Figure 1.2.

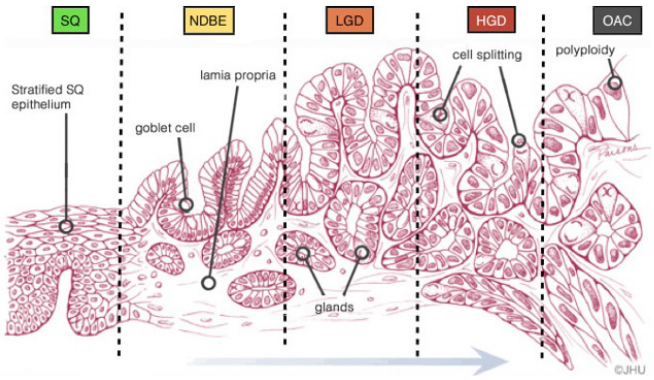


Figure 1.2: BE progression from SQ to OAC. Permission to reproduce this image has been granted by Johns Hopkins University.

BE epidemiology

Oesophageal cancer is the 6th biggest killer and the 8th most common cancer in the world. It is most common in white males between 40 to 60 years old (Zhang, 2013). Patients with BE have a 30 fold increased risk of OAC development, but even so only 0.5% will turn into OAC (Old et al., 2015).

There is an increased risk of OAC development as BE advances. Specifically, there is 15-20% risk of LGD patients and 40-60% risk of HGD of progressing to OAC (Conteduca et al., 2012). If a patient is diagnosed with OAC, there is an 85% probability of mortality within five years (de Jonge et al., 2013; Foreman, 2016).

In addition, some patients may suffer from gastro-oesophageal reflux disease, which can damage the lining of the oesophagus, leading to BE. The cell changes in the food pipe are caused by stomach juices coming back up through the valve at the top of the stomach (acid reflux). The acid in the juices irritates the lining of the food pipe, which is able to change the cell types to abnormal. Studies have noted that the risk of having acid reflux is higher if the patient is overweight, a smoker or drinks large amounts of alcohol (Zagari et al., 2008).

Biomarkers for BE

The benefit of identifying biomarkers is two-fold: they facilitate the early detection of BE disease and support the selection of target treatment. Goblet cells present in NDBE and gland cells present in LGD are obvious biomarkers for the presence of BE (Figure 1.2) and are part of standard histopathology.

In addition, experimental and clinical researches are carried out with the aim of identifying biomarkers of BE. As BE advances, DNA changes can occur — common ones are aneuploidy, tetraploidy and loss of heterozygosity. An aneuploid cell contains an abnormal number of chromosomes (compared with 46 chromosomes in a normal cell), a tetraploid cell contains double the amount of chromosomes and loss of heterozygosity is the loss of an entire gene and the surrounding chromosomal region. Unstable tetraploid cells can evolve into tumorigenic aneuploid cells. In addition, the tumour protein p53 is a significant biomarker for the progression to OAC. 77% of patients who are progressing to cancer had p53 tetraploidy and loss of heterozygosity (Rabinovitch et al., 2001). Biomarker p53 combined with biomarker Ki67 may reduce the inter-observer agreement and hence help in accurate grading of BE disease (McManus et al., 2004). Another biomarker that increases during tumorigenesis is Mcm2 (Lao-Sirieix et al., 2006).

Although there is extensive literature identifying BE biomarkers, due to clinical cost, time and poor performance measurement of trials, it is difficult to clinically try all possible combinations of the potential biomarkers in order to predict the stage of BE disease.

Diagnosis of BE

The usual way to diagnose BE is from histopathology of a biopsy taken via endoscopy. An endoscope, a long, flexible tube that has a light source and a

camera at the end, is inserted from the mouth down to the oesophagus of the patient in order to take one or more biopsies which are the biological samples of cells or tissues for examination. The resulting biopsy is processed by paraffin embedding, sectioning with a microtome at the thickness of 4μm, deparaffinisation, staining with hematoxylin and eosin, and microscopic analysis of the physiological appearance of many thick sections. The latest analysis is called Vienna classification (Conteduca et al., 2012). According to that classification there are five stages of the BE and the BE histopathological representation of their sections is given in Figure 1.3. Some physiological features are given in Figure 1.2. However, the classification process remains somewhat subjective, since it depends on the histopathologist's training and experience. Kerkhof et al. (2007) recommend that at least two histopathologists evaluate the BE biopsy, and, when indicated, consult a third histopathologist to establish a final diagnosis.

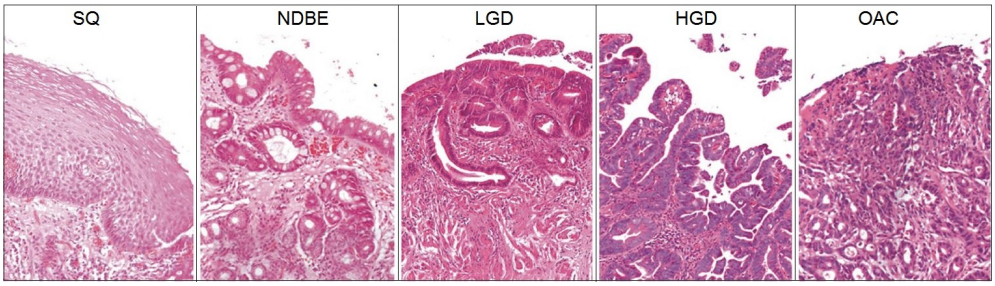


Figure 1.3: BE histopathological images for different stages of BE in the Vienna classification representation. Permission to reproduce this image has been granted by Baishideng Publishing Group Inc.

Usually there is a low inter-observer agreement between histopathologists at the dysplastic stages of BE. The inter-observer agreement of two histopathologists is measured using the Cohen (κ) statistic. A κ value of one means the histopathologists are in complete agreement, and the other ranges are: poor, any negative value to 0; slight, 0 to 0.2; fair, 0.2 to 0.4; moderate, 0.4 to 0.6; substantial, 0.6 to 0.8; and almost perfect, 0.8 to 1.0 (Montgomery, 2005). According to Kerkhof et al. (2007) the agreement between histopathologists when analysing two groups HGD+OAC versus all the others (in the case study of 920 patients in total), was substantial, $\kappa = 0.61$. On the other hand, κ value of 0.25 for the same study using four groups, SQ, NDBE+LGD, HGD and OAC, shows relatively poor inter-observer agreement between histopathologists. This is because distinguishing between the middle stages (NDBE, LGD, HGD) and cancer is more difficult than the discrimination between healthy and cancer biopsies. So, it is important to identify less subjective biomarkers in order to determine the stage of the BE

disease.

Symptoms of BE

Like in many cancer types, many people do not have any symptoms and the cell changes are found when tests are carried out for something else. However, people with BE may also have gastro-oesophageal reflux disease of which long term burning indigestion is the most common symptom.

Treatment of BE

Treatment of BE is based on the stage of the disease and can also be affected by other factors such as the patient's overall health. The general rule is that the earlier cancer is diagnosed, the better chance of successful treatment the patient has. Treatment aims to lower the amount of acid reflux and to remove any damaged areas of the oesophagus.

An early, well known treatment includes medicines (patients take tablets until they control the symptoms, and then reduce the dose). If the symptoms are not well controlled by medicines, patients may have surgery in order to strengthen the valve at the lower end of the oesophagus.

Most people have surgery through an endoscope, where the doctor puts a flexible tube called an endoscope down to the throat. The endoscope contains a camera so the doctor can see inside the food pipe. During surgery a lower part of the food pipe is removed to stop a cancer from developing. The doctor may suggest to patients alternative endoscopic treatments, such as radiofrequency ablation or photodynamic therapy, if the cells are very abnormal, in order to destroy them. Radiofrequency ablation involves administering a high frequency radio wave, which generates heat and treats the tissue. This treatment has been effective in over 90% of patients with LGD. In more advanced OAC, photodynamic therapy can be used where the patient is administered an intravenous non-toxic photosensitiser (light-sensitive drug) which is activated by a laser light in order to kill cancer cells, but the patient becomes photosensitive and needs to stay out of direct sun light for at least 24 hours. The British Society of Gastroenterology (Fitzgerald et al., 2014), National Institute for Health and Care Excellence (NICE, 2014), American College of Gastroenterology (Shaheen et al., 2016), and other organizations provide detailed guidelines for the treatment of BE.

1.2 Spectroscopy of biological samples

As mentioned above, diagnosis of BE requires endoscopy, where via a tube light is shone into the oesophagus. Using the same tube to measure spectra in-situ would be the ideal, but the current stage of the art is to measure spectra

on biopsies. The advantage of the spectroscopy is that it can characterize biopsies based on how they interact with light; molecules in different types of tissue absorb light at different wavelengths (distance between two adjacent peaks or troughs). Below we study how these biopsies produce spectra which are the input of our study.

Molecular vibrations

A molecule is made up of two or more atoms, for example water (H_2O) consists of compound molecule made up of 2 hydrogen atoms and 1 oxygen atom. A molecule has three possible types of motion that can occur in any combination, translational (whole atom or molecule changes its location in three dimensional space), rotational (whole molecule spins around an axis in three dimensional space) and vibrational (motion that changes the shape of the molecule) transitions. Absorption of light quanta and inelastic scattering of photos can both provoke vibrational transitions. In particular, when the energy difference between the ground state and the final vibrational state matches the energy of a photon (Foreman, 2016). Vibrational transitions are responsible for absorption in the region around 4000 cm^{-1}, which corresponds to infrared (IR) light (Figure 1.4), where cm^{-1} is the unit of measurement of the wavenumber (the reciprocal of wavelength). The absorption of IR radiation causes excitation of vibrations of the atoms of a molecule or the crystal lattice and causes bands in the spectra. IR light is (roughly) divided into three ranges: near infrared includes light of wavelengths of $14000 - 4000$ cm^{-1}, mid-IR includes light lengths of $4000 - 400$ cm^{-1} and far-infrared $400 - 10$ cm^{-1}. In general, the name IR spectroscopy conventionally refers to the mid-IR region (Pasquini, 2003).

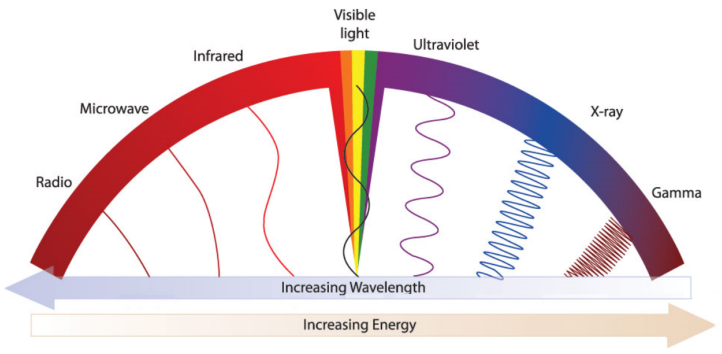


Figure 1.4: Electromagnetic spectrum. The energy of the waves increases as the wavelengths decrease. University of Waikato. All Rights Reserved. www.sciencelearn.org.nz

There are six types of fundamental vibrational modes (Figure 1.5), but not all of them absorb IR light. For example, oxygen (O_2) is a symmetrical diatomic molecule, which has only one bond and one vibrational state, which

is symmetrical and so this mode is not IR active, since it does not have a changing dipole, because the net dipole moments are in opposite directions and as a result, they cancel each other. On the other hand, the asymmetric stretching and twisting are IR active modes, because the bonds move in opposite direction and they do not cancel each other. For example, the asymmetric stretching mode of water (H_2O) is IR active.

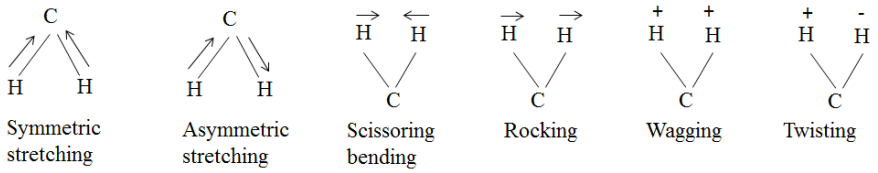


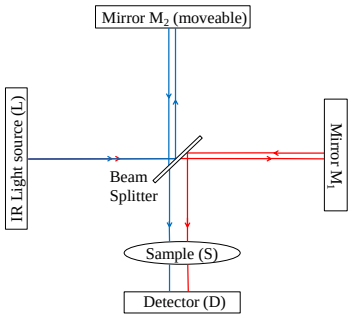
Figure 1.5: Six types of molecular vibration modes (for example of methylene group CH_2).

Fourier transform IR spectroscopy

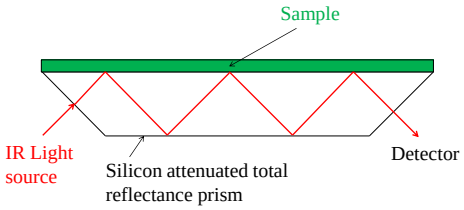
Fourier transform infrared (FTIR) spectroscopy uses an interferometer to create an interferogram, which allows simultaneous recording at all frequencies. In this case, the Michelson interferometer produces the interferogram. Then, to create a plot of intensity (power) versus wavenumber, a Fourier transform is applied to the interferogram (Foreman, 2016).

The Michelson interferometer (Figure 1.6a) has a beam splitter which is placed between the fixed and moveable mirrors. The beam splitter divides the IR light into two parts: half of the light is reflected to a fixed mirror and the other half is transmitted to a moveable mirror. When the two beams meet again at the splitter, they recombine. The recombined beam is aimed at the sample and recorded by the detector. The function of the transmitted light intensity versus the moveable mirror position produces an interferogram. Then, a Fourier transformation applied to the interferogram provides an IR power spectrum (which is a function of transmitted light intensity versus wavenumber (Trevisan et al., 2012)). Multiple interferograms are then averaged to achieve a sufficient signal to noise ratio. The absorbance spectrum is calculated according to the Beer-Lambert law $A = \log_{10}(I_0/I)$, where I_0 represents the intensity, also known as power spectra, of the incident light beam (reference intensity) and I represents the intensity of the light coming out of the sample (Pasquini, 2003). Both intensities are measured in the same units but because the Beer-Lambert law uses a transformation of their ratio, the absorbance spectrum is properly unitless, and usually reported as absorbance units. The entire process can be automated by using the OPUS 6.5 software (Bruker spectrometer).

Attenuated total reflectance is a technique that, coupled with FTIR, can



(a) Michelson interferometer. Two paths: L → M_1 → S → D (red) and L → M_2 → S → D (blue).

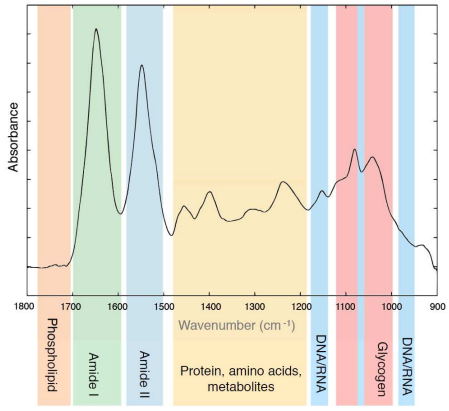


(b) Attenuated total reflectance prism.

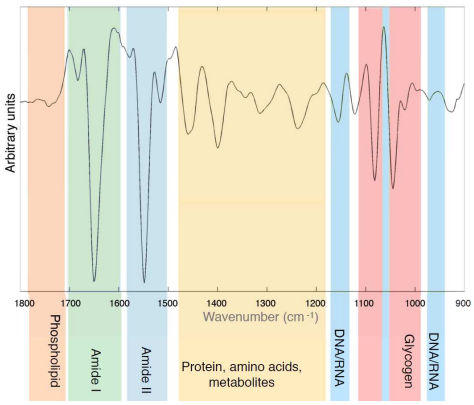
Figure 1.6: Tools to produce IR spectra.

combats the most challenging aspects of IR analyses, namely sample preparation and spectral reproducibility. These aspects greatly speed sample analysis. In our case, before the IR beam is aimed to the detector, it is internally reflected three times within an IR transmitting prism (Figure 1.6b).

The result that the interferometer records, after applying the Fourier transformation, is a spectrum (Figure 1.7). The fingerprint region, between $1800 - 900$ cm^{-1} is an important region, since most of the biologically important cellular compounds such as DNA/RNA and protein, absorb in this region, see Figure 1.7 (Foreman, 2016).



(a) Absorbance spectra.



(b) Second derivative spectra.

Figure 1.7: Absorbance and second derivative spectra: peaks are colour coded and labelled where compounds are known to produce a band. Permission to include this figure has been granted by Dr. Liberty Foreman.

1.3 Nature of spectral data and biological interpretation

The interpretation of peaks in the fingerprint region is complicated by the large number of different vibrations that occur here.

The simplest way to identify a compound would be to compare the spectrum to a library of spectra from known compounds. Advances in computer retrieval techniques can be used to rapidly compare a spectrum from an unknown compound to a library of known compounds. However, the algorithm will try to match the library to the unknown compound and name the known compounds that best fit the peaks. The more complex the mixture of unknown compounds, the harder it will be to match it to the known library.

This technique is of limited use when interpreting biological data as it is most effective when dealing with either a single or a small number of compounds. A biological sample typically consists of a complex mixture of many different types of bio-compounds, such as, tissue, blood, glycogen, DNA/RNA and many more which are not known. Each of these components has a different and unknown concentration and each has their own distinct spectral signature. Further, the system is in general neither linear nor additive. The resulting spectrum is complex and cannot be simply compared to a library of known compound spectra.

Instead, the most effective way of interpreting biological spectra in the context of diagnosis is to find the regions within the spectra that change as a disease progresses. Then we compare these peaks to those within a library of known biological compounds, which were first identified as important from the literature. It is unlikely that all the changes seen will be due to a single compound. Therefore, several compounds are visually compared simultaneously and by process of elimination, it is possible to predict which compounds are likely to be changing. This is difficult to reproduce computationally because peaks are completely dependent on the type of tissue and the disease being analysed.

For example a tissue spectrum containing glycogen will have many bands that are characteristic to changes in glycogen, but we know that there is little else in tissue that absorbs in the 1081 at cm^{-1} (CO stretch) region. Therefore, we can conclude that changes in this particular region of the spectrum are most likely to be from glycogen. Another example is the 964 cm^{-1} band. We know that little else in a tissue sample absorbs here apart from DNA, and therefore changes in this region can be attributed to DNA (Foreman, 2016).

The disease studied within this thesis is BE. There has been some previous

research into the spectral changes occurring as BE progresses. Most of the differences observed between the diseased and the healthy stage were in the 1170-1000 cm^{-1}, region, particularly at the wavenumbers 1168, 1154, 1116, 1066 and 1022 cm^{-1} which are related to DNA/RNA and glycogen. As BE progresses from NDBE to HGD the region that has been reported to separate them is 1610-1530 cm^{-1} and this region is attributed to Amide II. The spectral differences between NDBE and HGD/OAC can be seen in the regions 1290-1210 cm^{-1} and 1130-870 cm^{-1} which correspond to protein and DNA/RNA (Foreman, 2016). In addition, the differences seen between SQ and NDBE spectra are most likely to be related to either glycoproteins (mucin) or DNA/RNA (Quaroni and Casson, 2009).

Some information about the spectra of the main components of the tissue is available. Figure 1.8 represents four of the compounds: DNA, glycogen, blood and mucin. From the biological point of view, the DNA should increase as BE progresses, the glycogen is expected to change as dysplasia progresses, the blood is important because all biopsies will contain different amounts of blood and for the mucin it is expected to see changes in the signal for all stages (apart from SQ versus NDBE) as the disease progresses.

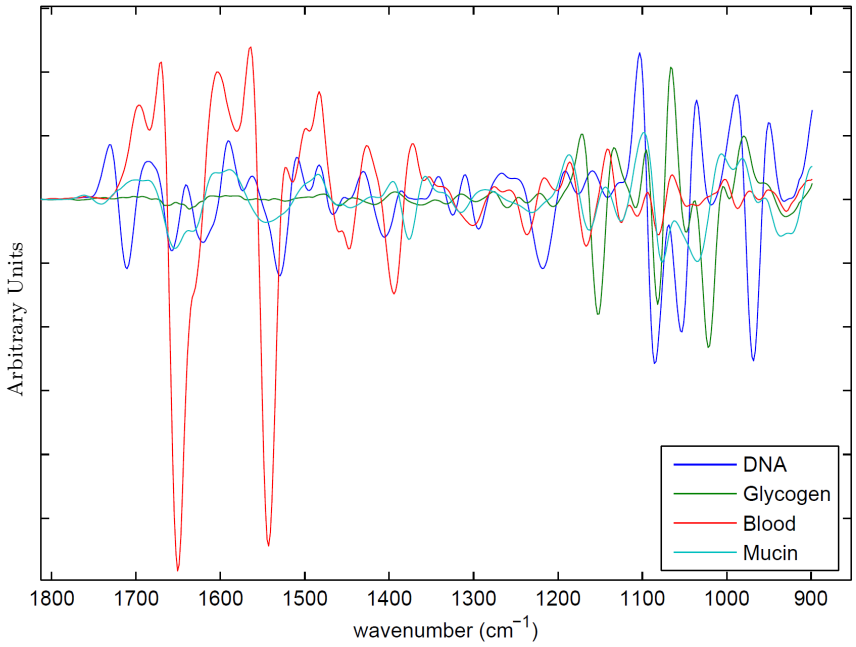


Figure 1.8: Spectra of four main compounds: DNA, glycogen, blood and mucin.

Even though we know something about the compounds, their spectra are very complex (Figure 1.8). Each compound has many peaks in it and which peaks are useful for diagnosis will depend not only of the presence of a compound but what other compounds are there and how the relevant

position of these peaks of the compound compared with this one. Given the amount of uncertainty, we will choose to put vague priors on the coefficients in our Bayesian variable selection, rather than trying to force the selection of particular peaks. Afterwards we try to interpret the results. By contrast, our prior for the inclusion probability of variables will be informative, with a small probability chosen because we want a sparse solution.

1.4 Spectral data of BE

This study is motivated by the use of FTIR spectra of biopsies on patients to diagnose BE disease via spectroscopy. Our diagnostic data are a discretized version of a functional vector of absorbance over a range of wavenumbers. The discretization depends on the resolution of the interferometer, here absorbances at 676 wavenumbers (number of variables) ranging from 2200 cm^{-1} to 900 cm^{-1} are recorded and the resulting spectrum presented in the second derivative form. Taking into account all patients' spectra (samples) we can construct the design matrix. Two histopathologists classify the samples to one of the five possible stages of the BE disease, which is the response vector. From this high-dimensional data we are interested in selecting only a small subset of wavenumbers that carry information about the stages of the BE disease with the aim of improving the classification accuracy. This subset of wavenumbers may not only contain wavenumbers that correspond to peaks or troughs in the spectrum but also wavenumbers that are not obvious from the plot (Figure 1.7).

As mentioned in the previous paragraph, our data are a discretized version of a continuous spectrum. The smoothness of the underlying spectrum implies that there will be strong correlations between absorbances at nearby wavenumbers.

A more subtle collinearity problem is the following. Peaks are absorbances arising from particular molecular bonds. Many molecular bonds will absorb at several wavelengths. So, there are distant peaks that are strongly correlated. For example, peaks that present absorbances at wavenumbers around 1050 cm^{-1} and 950 cm^{-1} seem to carry similar information about DNA/RNA (Figure 1.7).

1.5 Variable selection and prediction

FTIR spectrometers typically record absorbance at a very large number of wavenumbers that depend on the resolution that they use. At the same time,

a relatively small number of biopsies are available from patients. One way to deal with high-dimensional problems ($p \gg n$, p is the number of variables, here measurements, and n is the number of observations, here spectra), known as large p, small n problems, as well as the case of highly correlated variables with $p < n$, is the selection of the most important variables.

When there is collinearity among predictors, regressions become unstable. One solution to handle collinearity is to use ridge regression estimators which stabilize the least squares estimation. Another is to remove the 'redundant' variables (those that are linear combinations of existing absorbances at different wavenumbers) using variable selection methods. We need to be aware though that there may be multiple solutions because of the highly correlated variables may substitute for each other.

The goal of variable selection is to identify a small subset of variables which together give accurate predictions. To achieve accurate prediction given a model, overfitting and underfitting has to be avoided. If the model is too simple, then the model has high bias/low variance and conversely, if the model is too complex, then the model has low bias/high variance (Figure 1.9).

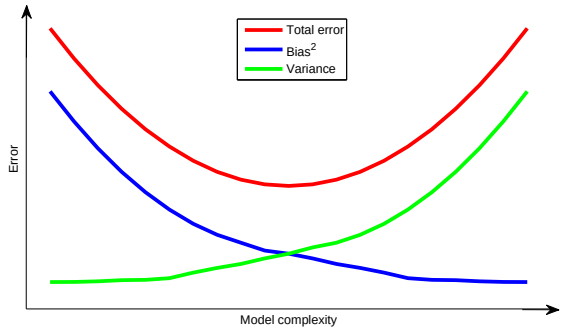


Figure 1.9: Bias-variance tradeoff.

The aim is to choose the model that has the best balance of goodness of fit against model complexity. This report is focused on methods that find the important variables so that we can predict the outcome or find the best model. However, variables that are important individually may not be a part of the best overall predictive model and vice versa. The results of the variable selection approach can summarized under two different settings: to identify individually variables that are potential biomarkers and to identify the best model (combination of variables) that can be used for predictions.

The predictions we are interested in are out-of-sample predictions. By out-of-sample we mean the prediction of future measurements given hitherto unobserved explanatory variables. There is a tougher challenge than

in-sample prediction, where the samples to be predicted are the same as, or very similar to, the ones in the training set.

In the Bayesian framework there are some preferences concerning how to select the prior distributions for the coefficients for each of the two tasks. Because our focus is on out-of-sample prediction we should avoid priors like the g-prior that effectively reinforce the pattern in the training set. We will select priors that induce a more general shrinkage, with the aim of producing robust predictors.

1.6 Reasons for variable selection

Variable selection is, in general, useful for interpretational reasons: simpler models are easier to interpret than complex ones. Variable selection removes non significant effects, giving us a chance to interpret the reduced model.

Variable selection is also useful for reducing noise in the dataset and consequently achieving a good class separation. This may be important when working with biological data, because they are often noisy and it is complex to understand them. As a result, variable selection tends to improve the prediction performance. Irrelevant variables in the input data may decrease the classification performance. In this study variable selection not only improves the classification performance but may help researchers to understand the cell changes involved in the disease.

In addition, variable selection reduces the difficulty of handling and visualising the data. After the process of variable selection, parameter estimation will be stable, data visualization is easier to do and efficient storage is feasible.

Variable selection is particularly attractive for high-dimensional data, especially when p approaches or exceeds n. In these cases several problems can arise: overfitting of the model to the sample, collinearity among the independent variables and computational difficulties.

A very efficient way to explore the high-dimensional predictor space is to apply Bayesian variable selection approaches. The probability-driven stochastic search is able to visit many different combinations of predictors. The result of the search is a rich one, providing both joint and marginal inclusion probabilities for predictors. For example, the most visited model suggests the best combination of wavenumbers that altogether contribute the most to predicting the progression of the BE disease. For biologists it is also important to identify some possible biomarkers, via the marginal inclusion probabilities, which indicate wavenumbers that appear quite often in good models.

One issue in identifying important variables is the number of possibilities

and the risk of spurious associations. In the frequentist framework one way to check if each variable is important would be to do hypothesis testing, but the probability of false positive results rapidly increases as the number of variables increases. For example, if we perform a single test using a significance level of 1% and there truly is no effect of the factor being tested, there is only a 1% chance of a false positive result. However, if we perform 10000 independent tests using the same significance level, we expect $10000 \times 0.01 = 100$ of the tests to have p-value less that 0.01, so that 100 of the tests would be falsely significant.

A good way to pick individual variables is by controlling the false discovery rate (Muller et al., 2006). In the Bayesian set-up, the idea is to define a threshold for the marginal posterior probability of inclusion with respect to a specified false discovery rate level. This process (Saadi et al., 2016) can be implemented in the following way: (i) For a given threshold, let R be the number of predictors with marginal posterior probability bigger than the threshold (the idea is similar to if the p-value is less than or equal to a significance level then you reject the null hypothesis). (ii) Repeatedly permute the class labels and using the same threshold estimate the false discovery rate as the ratio of the average of the number of false positives to R. (iii) Choose the threshold so that the empirical false discovery rate (as calculated at step ii) is not greater that a specified level (usually 0.05).

1.7 Motivation and contributions

The motivation of this work is to analyse spectral data related to the Barrett's oesophagus (BE) disease, where the stages of the disease can be described as a mixture of nominal and ordinal variables. The aim of the work is to provide a good model for clinical diagnosis and fully to identify some interesting regions of the high-dimensional spectra. To the best of the author's knowledge, this is the first attempt to build a Bayesian variable selection method using both nominal and ordinal variables.

The principal contributions of this thesis are:

- Chapter 2: A summary of the most famous penalised regression methods.
- Chapter 2: A summary of the most important measures to evaluate the classifier.
- Chapter 3: A summary of the most important methods to perform

Bayesian variable selection (BVS) using a probit model with binary responses.

- Chapter 4: An application of BVS approaches to some datasets.
- Chapter 5: A detailed explanation of BVS using pure nominal and pure ordinal responses, including also the similarities and differences between the two approaches.
- Chapter 6: The building of a decomposed probit model for mixture of nominal and ordinal responses.
- Chapter 6: A proposed algorithm for decomposed BVS.
- Chapter 7: The building of a probit model for mixture of nominal and ordinal responses.
- Chapter 7: The construction of an algorithm that implements BVS using a common indicator vector for all the latent variables.
- Chapter 8: The construction of an algorithm that implements BVS using both types of responses and different indicator vectors across different latent variables.
- Chapter 9: The extension of the proposed approach of decomposed variable selection in existing methods.
- Chapter 9: The application of the three proposed methods in BE for clinical diagnosis.
- Chapter 9: The comparison of the three proposed methods with existing ones.

Chapter 2

Classification methods for high-dimensional data

This chapter starts with a general description of the classification problem and how to evaluate the performance of a classification method. Afterwards we study statistical models for categorical responses and methods to assess different models (different variables included in the model via variable selection). In the last part, we focus on the study of variable selection for the case of high-dimensional data, where standard approaches are not applicable or are not efficient. Variable extraction methods are presented briefly as an alternative way for dimensionality reduction.

2.1 The classification problem

Classification problems study how to learn a rule or a model to classify observations into a given set of classes on the basis of an observed feature vector. In order to learn the rule or the model a set of training data is required and the task is to produce a method that will generalise to new observations.

A classical choice, which for two groups projects the data on to a line in order that samples from the different classes are well separated, is Fisher's linear discriminant analysis (LDA). In the classification context the LDA classifier uses the criterion of maximizing the ratio of between-to within-class variance to construct variables that are linear combinations of the original ones. Those extracted variables may be used directly or as input for another classifier. Other algorithmic based classification methods are k-nearest neighbours (k-NN), classification trees, and random forest (Murphy, 2012).

On the other hand, model based methods rely on parametric assumptions about the data. Various models that are appropriate for categorical responses

can be used, for example in the probit model the errors are assumed to follow a normal distribution. Interestingly, some methods like LDA can be derived not only as algorithmic methods for classification but also as model based methods. For example, LDA can also be derived by assuming that the two classes are Gaussian distributed with a common covariance matrix (Hastie et al., 2001).

2.1.1 Assessing the performance of a classification method

For models predicting continuous variables, classical choices to assess the predictions are via mean square error or mean absolute error. However, the best way to evaluate binary predictions is not so clear, and it becomes even more complicated for multi-class predictive responses.

Evaluation measures

Particularly in medical diagnostic studies, we are often interested in predicting if the result is positive (patient has the disease) or negative (patient does not have the disease). In different contexts the idea of positive and negative results can be adapted. Four different outcomes are possible when a binary case is classified. A true positive (TP) is a correctly predicted positive example. A true negative (TN) is a correctly predicted negative example. A false positive (FP) is an incorrect prediction that an example was positive, when in fact was negative. This is also known as a type I error. The last one, false negative (FN) is an incorrect prediction that an example was negative, when in fact was positive. This is also known as a type II error. The four outcomes can be summarized in the confusion matrix C (Table 2.1). High values of the diagonal elements TP and TN mean that the classifier is very good. On the other hand, high values on the off-diagonal elements mean that mistakes were often made. Although the confusion matrix itself is simple to understand, there is a lot of additional terminology that needs to be explained.

Table 2.1: Confusion matrix for a binary classification problem.

		Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

Based on these definitions several new functions which measure different aspects of classification performance can be defined. Here are some of them. The fraction of actual positives that are correctly predicted as positives is

called sensitivity. This is also known as recall or true positive rate and is given by

$$\text{Sensitivity: } \rho = \frac{TP}{TP + FN}.$$

Sensitivity increases as the number of FN decreases.

The fraction of actual negatives that are correctly predicted as negatives is called specificity. This is also known as the true negative rate and is given by

$$\text{Specificity} = \frac{TN}{TN + FP}.$$

Specificity increases as the number of FP reduces. The ideal classifier has sensitivity and specificity equal to unity. In general, however this cannot be achieved: as sensitivity increases, specificity decreases.

The fraction of predicted positives that are true positives is called precision. This is also known as positive predictive value and it is given by

$$\text{Precision: } \pi = \frac{TP}{TP + FP}.$$

Accuracy counts the number of correct classifications as a proportion of all the cases

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

and the error is simply calculated by

$$\text{Error} = 1 - \text{Accuracy}.$$

In classification, the main goal is to maximize the accuracy of the classifier or equivalently to minimize the error rate. However, all misclassification cannot be equally considered (Pazzani et al., 1994; Fearn, 2012). For example, for a medical diagnosis problem, the cost of diagnosing a healthy patient as diseased may not be the same as the cost of diagnosing a diseased patient as healthy. In some medical contexts for example a cost of 2 may assigned to misclassifying a negative as positive. This means that it is 2 times more important to correctly classify a negative as negative, than it is avoid to misclassifying a positive as negative. Based on this idea, it may be of interest to minimize an expected cost or some other criterion taking cost into account instead of minimizing the classification error rate.

If the sensitivity (ρ) has large value but the precision (π) has small value,

then the F_β measure may be used to select a best classifier

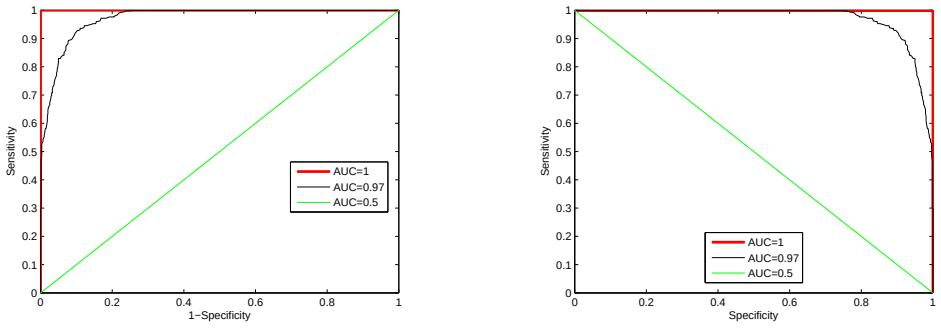
$$F_\beta = \frac{(\beta^2+1)\pi\rho}{\beta^2\pi+\rho} = \frac{(\beta^2+1)TP}{(\beta^2+1)TP+\beta^2FP+FN}, \tag{2.1}$$

where the scalar β controls the trade-off between the precision and the sensitivity. In practice, where the recall and the precision are equally important ($\beta = 1$), the F_1 micro-averaged-measure is given by

$$F_{1-micro} = \frac{2\pi\rho}{\pi+\rho}, \tag{2.2}$$

which is the weighted harmonic mean of the precision and sensitivity.

A widely used evaluation measure that is based on the definitions of sensitivity and specificity is the receiver operating characteristic (ROC) curve. This is produced for a given classifier by varying a tuning parameter such as a threshold to give pairs of sensitivity/specificity values. It is usually represented as a graph of sensitivity versus 1-specificity. An example of a ROC curve is shown in Figure 2.1a. As we change the threshold the number of FP decreases, while the number of FN increases. Maximizing sensitivity corresponds to a large value on $y-$axis and minimizing the $1-$specificity corresponds to a small value on $x-$axis on the ROC curve. Thus, if the line of the graph is close to the top left, then the classifier is good.



(a) Sensitivity versus $1-$specificity.

(b) Sensitivity versus specificity.

Figure 2.1: ROC curves and the corresponding AUC values.

Instead of studying the ROC curve itself as an evaluation measure, the area under the ROC curve (AUC) is a well known measure for classification performance. AUC can be calculated via numerical approximation methods for example using trapezoidal rule. AUC equal to 0.5 corresponds to no discrimination between classes (random classifier), which means that the ROC curve is the diagonal line that starts from the origin (Figure 2.1a). On the other hand, if the ideal classifier has AUC equal to unity this means that the

ROC curve consists of two straight lines that have common point the upper left corner. So, AUC is typically between 0.5 and 1.

An alternative, but not so commonly used, representation of the ROC curve is as a graph of sensitivity versus specificity. In that case, maximizing specificity (instead of minimizing the 1−specificity) corresponds to a large value on $x-$axis on the ROC curve (Figure 2.1b). So, if the line of the graph is close to the top right (instead to the top left), then the classifier is good. With respect to the AUC, the area under the curve is located differently for graphical representation of sensitivity versus 1−specificity and sensitivity versus specificity but the values of AUC are the same in both cases. Finally, graphical measures to quantify the performance of the classifier, which are not so well known, are lift chart, precision-recall curves, and cost curves. More details are given in Japkowicz and Shah (2011).

The majority of the aforementioned evaluation measures for binary responses easily extend to the multi-class responses. Measures to evaluate the classifier for binary and multi-class cases are summarized in Table 2.2. For the multi-class case, more details are given below.

In a multi-class classification problem, in the first instance we may be interested in predicting if the result is member of the class m ($m = 0, \ldots, M-1$, where M is the total number of classes) or not. Four different outcomes are possible when a multi-class case is classified: TP_m is the number of correctly predicted examples of a member of the class m, FP_m are examples that are not members of the class m but are predicted as members of class m, FN_m are examples that are members of the class m but are predicted as not members of class m, and TN_m is the number of correctly predicted examples of not be a member of the class m. Those four possible outcomes can be summarized in the collapsed confusion matrix (Table 2.3).

Table 2.3: Collapsed confusion matrix for a multi-class classification problem.

		Predicted	
		Be a member of class m	Not be a member of class m
Actual	Be a member of class m	TP_m	FN_m
	Not be a member of class m	FP_m	TN_m

We can calculate those possible outcomes from the $M \times M$ confusion

Table 2.2: Summary of the most important measures to evaluate the classifier for binary and multi-class case.

Measure	Binary study (2 classes)	Multi-class study (M classes)	
		Average (macro)	Overall (micro)
Accuracy	$\frac{TP+TN}{TP+FP+FN+TN}$	$\frac{\sum_{m=0}^{M-1} \frac{TP_m+TN_m}{TP_m+FP_m+FN_m+TN_m}}{M}$	$\frac{\sum_{m=0}^{M-1} TP_m+TN_m}{\sum_{m=0}^{M-1}(TP_m+FP_m+FN_m+TN_m)}$
Error rate	$1 -$ Accuracy	$1 -$ Accuracy	$1 -$ Accuracy
Precision	$\frac{TP}{TP+FP}$	$\frac{\sum_{m=0}^{M-1} \frac{TP_m}{TP_m+FP_m}}{M}$	$\frac{\sum_{m=0}^{M-1} TP_m}{\sum_{m=0}^{M-1}(TP_m+FP_m)}$
Sensitivity	$\frac{TP}{TP+FN}$	$\frac{\sum_{m=0}^{M-1} \frac{TP_m}{TP_m+FN_m}}{M}$	$\frac{\sum_{m=0}^{M-1} TP_m}{\sum_{m=0}^{M-1}(TP_m+FN_m)}$
Specificity	$\frac{TN}{TN+FP}$	$\frac{\sum_{m=0}^{M-1} \frac{TN_m}{TN_m+FP_m}}{M}$	$\frac{\sum_{m=0}^{M-1} TN_m}{\sum_{m=0}^{M-1}(TN_m+FP_m)}$
F_β	$\frac{(\beta^2+1)\text{Precision}\cdot\text{Sensitivity}}{\beta^2\text{Precision}+\text{Sensitivity}}$	$\frac{\sum_{m=0}^{M-1} \frac{(\beta^2+1)\text{Precision}_m\cdot\text{Sensitivity}_m}{\beta^2\text{Precision}_m+\text{Sensitivity}_m}}{M}$	$\frac{(\beta^2+1)\sum_{m=0}^{M-1} TP_m}{\beta^2\sum_{m=0}^{M-1}(2TP_m+FP_m+FN_m)}$
AUC	AUC numerical approx.	$\frac{\sum_{m=0}^{M-1} \text{AUC}_m}{M}$	$\frac{2}{M(M-1)}\sum_{\forall m,m':m\neq m'} \text{AUC}(m,m')$

matrix (C) according to the following equations

$$TP_m = C(m,m), FN_m = \sum_{m'=0, m\neq m'}^{M-1} C(m,m'),$$
$$FP_m = \sum_{m'=0, m\neq m'}^{M-1} C(m',m), TN_m = \sum_{m'=0, m\neq m'}^{M-1} C(m',m'),$$

where M denotes the number of classes. Sensitivity, specificity and precision of each class m are defined according to the last three equations and those are

$$\text{Sensitivity}_m\text{: } \rho_m = \frac{TP_m}{TP_m + FN_m}, \text{Specificity}_m = \frac{TN_m}{TN_m + FP_m},$$
$$\text{Precision}_m\text{: } \pi_m = \frac{TP_m}{TP_m + FP_m}$$

respectively.

With respect to the F_β measure, Equation (2.1) can also be used in the multi-class case and a typical choice in this case is $\beta = 1$. The F_1 micro-averaged-measure for the multi-class case is given again by Equation (2.2), where the overall sensitivity and precision are given by

$$\text{Sensitivity: } \rho = \frac{\sum_{m=0}^{M-1} TP_m}{\sum_{m=0}^{M-1}(TP_m + FN_m)}, \text{Precision: } \pi = \frac{\sum_{m=0}^{M-1} TP_m}{\sum_{m=0}^{M-1}(TP_m + FP_m)}.$$

The F_1 macro-averaged-measure of Equation (2.1) in the multi-class case is

$$F_{1-macro} = \frac{\sum_{m=0}^{M-1} F_{1m}}{M}, F_{1m} = \frac{2\pi_m\rho_m}{\pi_m + \rho_m}.$$

ROC curves were originally developed for binary problems, but they have also been generalized for multi-class problems (Hand and Till, 2001),

$$AUC_{HT} = \frac{2}{M(M-1)} \sum_{\forall m,m': m\neq m'} AUC(m,m'),$$

where $AUC(m,m')$ is the area under the ROC curve involving the pair of classes m and m'.

Data splitting

The experimental estimation of the performance of a classifier may be based on random partition into training and test parts using various methods, but the idea of data splitting is common: to build a classifier using the training

set and evaluate it using the test set. In cases where an additional set (sometimes called the tuning set) is needed to tune one or more parameters the data are split into three sets. In that case, we build a model on the training set, we tune the parameters on the tuning set and estimate the classification performance on the test set. This triple splitting is important because if you use the test set for both tuning and prediction, this leads to optimistic estimates of classification performance, since the tuning optimises the classifier for the particular test set. In addition, sometimes the researchers instead of reporting the classification performance on the test set report the classification performance on the training set, which can be very over-optimistic.

One way to implement the random partition is by using the hold-out method. It uses two separate datasets, training set (for example 2/3) and test set (for example 1/3), with repeated splits. This is useful for medium sized datasets. Another idea is to implement the random partition by using k-fold cross validation, which means that the dataset is randomly split into k equal size sub-samples, we use $k-1$ sub-samples as training data and the remaining sub-sample as test data and repeat this k times. Extensive experiments have shown that 10-fold cross validation is a good choice to get an accurate estimate. However, 5-fold cross validation is also popular, Hastie et al. (2001). Variance can be reduced by using the mean of multiple cross-validations as an estimate of, for example, the accuracy. Note that if a large number of folds is selected, then the bias of the true error rate estimator will be small, but the variance will be large. In addition, the computational time will be very large. On the other hand, if a small number of folds is selected, the variance of the estimator will be small and the computation time will be reduced, but the bias of the estimator will be large. So, in practice, the choice for k depends on the size of the dataset. For large datasets, even 3-fold cross validation will be quite good.

A special case of cross validation is the leave-one-out cross validation method, LOOCV ($k = n$). Train the model with $n-1$ observations and predict the one that was left out and repeat this process n times. Using LOOCV the variance of the error estimates is high (Hastie et al., 2001). In addition, it is quite computationally expensive to repeat the process n times. For large n the computational expense becomes prohibitive, but for small size data, it is not so slow. For very sparse datasets, LOOCV may be used in order to train with as many samples as possible.

Within the context of variable selection (VS), mistakes in cross validation are very common. The choice of which variables are important for the model needs to be evaluated as part of the cross validation. We have to take care and

put every supervised method inside the training part of the cross validation step, not do VS using global information, and then do cross validation.

2.1.2 Unbalanced dataset

In some sciences, such as Genetics and Medicine, it is important to balance unbalanced dataset (for example 95% of observations are healthy, only 5% are diseased). In an unbalanced dataset the minority class (one class is represented by only a small number of observations) may not have a sufficient number of observations to learn the data and so deriving a good rule may be a difficult task. Usually, the classification accuracy seems very high. However, this is not the case since it is only reflecting the majority class (one class is represented by only a large number of observations). In the binary classification problem one way to handle an unbalanced dataset is to build a balanced training set, use it for classifier training in two steps: first, randomly select the desired number of minority class and then add an equal number of randomly selected majority class via re-sampling. Another way to handle an unbalanced dataset is to build a balanced test set and use this balanced set to test the classification performance. There is an extensive literature on how to balance unbalanced dataset for supervised learning, see for example Ganganwar (2012), but it is not discussed here since it is beyond the scope of this study.

2.2 Model based methods

In this section we study statistical models for discrete outcomes criteria, often based on likelihood, to assess the different models. The statistical models belong to the generalized linear model with Bernoulli or multinomial (discrete) distributions, because of the discrete outcome.

2.2.1 Generalized linear models for categorical responses

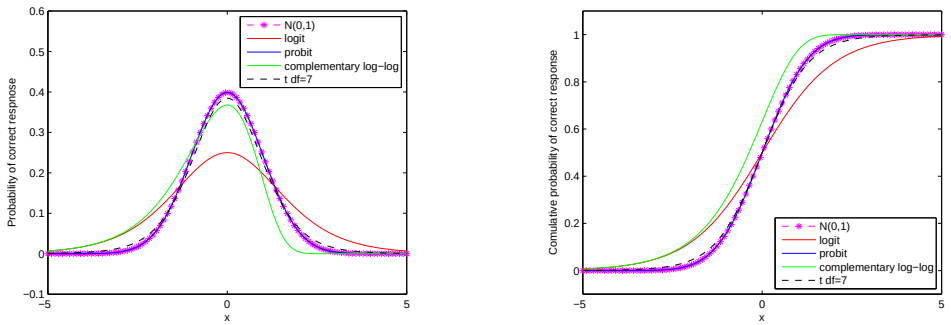
The relationship between the responses y_i and potential explanatory variables $\mathbf{X}_{i,:}$ (row vector of data matrix $\mathbf{X}$ with p variables, $i = 1, \ldots, n$) is described via a generalized linear model, which is made up of a linear predictor $\eta_i = \mathbf{X}_{i,:}\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is the vector of coefficients, and the link function, which describes how the mean, $E(Y_i|\mathbf{X}_{i,:}) = \mu_i$ depends on the linear predictor via $g(\mu_i) = \eta_i$. For a classification problem, where the responses are categorical, the three most common link functions (Table 2.4) are logit, probit

and complementary log-log (Hastie et al., 2001). Responses are distributed as Bernoulli or binomial for the binary case and can be generalized to multinomial or categorical for the multi-class case.

Table 2.4: Link functions for generalized linear model with categorical responses.

Logit link	Probit link	Complementary log-log link
$\eta_i = \log \frac{\mu_i}{1-\mu_i}$	$\eta_i = \Phi^{-1}(\mu_i)$	$\eta_i = \log[-\log(1-\mu_i)]$

The logit link can be interpreted as modelling log odds. Logit is computationally easier than working with normal distributions through the probit link function. Logit has slightly heavier tails, which means that a probit curve approaches the axes more quickly than a logit curve (Figure 2.2). However, probit is preferred when it seems plausible that there is an underlying vector of dependent variables which is normally distributed (Finney, 1947). In addition, probit is preferred for Bayesian models where it has the advantage that we can assign conjugate priors to the regression parameters. Albert and Chib (1993) propose instead of using the inverse cumulative density function (CDF) of the probit link to use the Student t inverse CDF as the link function in a Bayesian setting, where the degrees of freedom parameter becomes part of the estimation problem. The logistic function is similar to the normal except in the tails, where it is heavier, resembling a Student $t-$distribution. In Figure 2.2, the complementary log-log link is not symmetric. Due to this asymmetry, this link is usually only used for problems when μ_i is small, and then a complementary log-log is close to a logit model.



(a) Link functions mapping probabilities to the real line.

(b) Link functions mapping cumulative probabilities to the real line.

Figure 2.2: Visualization of common link functions with binary responses, compared to the standard normal and a Student $t-$distribution.

2.2.2 Model assessment criteria

Given a model, its evaluation is carried out according to some criteria that are usually based on the loss function. The basic idea is that, based on the criteria, we can compare the full model (includes all the variables) with other models that have fewer variables in order to find the best model.

The model based methods have a loss function $L(y_i, \hat{y}_i)$, where $\hat{y}_i$ denotes the prediction of the i-th sample given the observed data, $i = 1, \ldots, n$. For categorical response variables a typical choice of loss function is the $0 - 1$ loss,

$$L(y_i, \hat{y}_i) = \begin{cases} 0, & \text{if } y_i = \hat{y}_i \\ 1, & \text{if } y_i \neq \hat{y}_i. \end{cases}$$

Another option is to select as a loss function the negative log likelihood and to minimize that. If this loss function does not have a penalty term, then the full model will always be the best model. This penalty term should depend on the number of parameters, since a flexible model with many parameters has more potential for overfitting than a simple one. We will study different criteria that are based on the idea of minimizing the negative log likelihood plus a penalty term.

Akaike's information criterion (AIC) was introduced by Akaike (1998), and it is a simple way to compare all possible models via a penalty on the number of estimated parameters. The AIC is defined as

$$\text{AIC} = -2\log[L(\hat{\boldsymbol{\beta}}; \mathbf{y})] + 2p,$$

where $\hat{\boldsymbol{\beta}}$ denotes the maximum likelihood estimate of the model parameter and p is the dimension of $\boldsymbol{\beta}$. After calculating AIC for all possible models, the best model is the one with the smallest AIC. AIC was derived as an approximation to the estimated risk, based on expected Kullback-Leibler information, of predicting using the regression model. For small sample sizes, n, AIC underestimates the risk, and has been improved (Peruggia, 2003) to

$$\text{AICc} = -2\log[L(\hat{\boldsymbol{\beta}}; \mathbf{y})] + 2p + \frac{2p(p+1)}{(n-p-1)},$$

which is correct to $O(1/n)$ for normal linear regression. This stronger penalty corrects the tendency of AIC to choose models that are too complex when n is small. Various authors have derived corrections to this bias in AIC that are correct to $O(1/n)$ for generalized linear models, see for example Imori et al. (2011). Unfortunately these corrections, which use derivatives of the log likelihood, depend on both the probability distribution and the link function

in the generalized linear model, so that the simplicity of AIC is lost.

The deviance information criterion can be considered as a generalization of AIC and is defined by

$$\text{DIC} = D(\bar{\boldsymbol{\beta}}) + 2p_D = \bar{D} + p_D,$$

where $D(\bar{\boldsymbol{\beta}}) = -2\log[L(\bar{\boldsymbol{\beta}}; \mathbf{y})]$ is the deviance, $\bar{\boldsymbol{\beta}}$ is the posterior expectation of $\boldsymbol{\beta}$ (for AIC posterior mean is substituted by the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$), $p_D = \bar{D} - D(\bar{\boldsymbol{\beta}})$ and $\bar{D} = E[-2\log[L(\bar{\boldsymbol{\beta}}; \mathbf{y})]]$ is the posterior mean. The p_D measures the complexity of the model. A smaller deviance information criterion indicates a better fit to the dataset.

An alternative to AIC is the Bayesian information criterion, (BIC) or Swartz criterion, (Dayhoff and Schwartz, 1978)

$$\text{BIC} = -2\log[L(\hat{\boldsymbol{\beta}}; \mathbf{y})] + \log_e(n)p,$$

which penalizes model complexity more heavily, using a penalty term dependent on sample size. Generalizations of Schwarz's derivation are presented by Cavanaugh and Neath (1999).

Model selection by AIC and BIC aims to indicate the best model, in the sense of having the smallest error when the model is used for prediction. AIC and BIC can be used in place of hypothesis testing in stepwise model selection: the model with the lowest AIC/BIC score always being the one selected at each stage. Unlike hypothesis testing methods, AIC and BIC can be used to compare models that are not nested.

Comparing the two methods, BIC selects simpler models than AIC, and for large sample sizes can select simpler models than hypothesis testing based methods as well. AIC is asymptotically efficient yet not consistent and BIC is consistent yet not asymptotically efficient. AIC is commonly used for small sample sizes and performs better for complex models when BIC performs better for simpler models.

Nevertheless, BIC is motivated as an approximation to the Bayes Factor, which can be used to compare two models, M_1 and M_2, in a Bayesian setting, using the ratio of their posterior probabilities

$$\frac{P(M_1|\mathbf{X}, \mathbf{y})}{P(M_2|\mathbf{X}, \mathbf{y})} = \frac{P(M_1)}{P(M_2)} \times \frac{P(\mathbf{X}, \mathbf{y}|M_1)}{P(\mathbf{X}, \mathbf{y}|M_2)}. \tag{2.3}$$

The second factor in Equation (2.3) is called the Bayes Factor (Jeffreys, 1935). The Bayes factor is defined as the contribution of the data to the posterior

odds. It is given by

$$BF(M_1, M_2) = \frac{P(\mathbf{X}, \mathbf{y}|M_1)}{P(\mathbf{X}, \mathbf{y}|M_2)},$$

which is equal to the ratio of the posterior probabilities when both models have equal prior probabilities. Model selection using BIC and Bayes factor are equivalent assuming that the two models are regarded as equally probable a-priori, such that choosing the model with the smallest BIC is equivalent to choosing the model with the greatest posterior probability (Wasserman, 2000). The Bayes Factor for comparing two models can be interpreted using the evaluation ranges (Kass and Raftery, 1995) that are summarized in Table 2.5

Table 2.5: Bayes Factor comparison values.

$BF(M_1, M_2)$	Strength for evidence
1-3	Not worth more than the bare mention
3-10	Substantial
10-100	Strong
100-1000	Decisive

Finally, there are different versions of Mallows C_p for generalized linear models. Hurvich and Tsai (1995) proposed a version of Mallows C_p based on the Pearson χ^2 goodness-of-fit statistic via

$$C_p = \frac{(n - \bar{p})\chi_p}{\chi_{\bar{p}}} + 2p - n,$$

where $\bar{p}$ is the number of explanatory variables contained in the full model which includes all available candidate variables and χ_p, $\chi_{\bar{p}}$ are the Pearson χ^2 statistics evaluated under the model indicated by the subscripts respectively.

2.3 Dimensionality reduction methods for high-dimensional data

There are many situations, for example in very high-dimensional data, when the number of parameters is large, or when there is a collinearity between variables (columns of $\mathbf{X}$ are highly correlated), where using the likelihood alone does not produce estimators with good properties and so those estimators may not be suitable for model selection. Standard methods for variable selection are computationally too expensive to compare all possible models

with different variables. In addition, in high dimensions estimating many parameters increases the overall error of the estimations.

In high dimensions it is reasonable to assume that the relevant variables lie in low-dimensional space. The last can consists of only a few original important variables (this process is known as variable selection) or a few new variables that are transformation of the original ones (this process is known as variable extraction).

2.3.1 Variable selection in high-dimensional data

Variable selection (VS) is also called feature subset selection, feature selection or attributes selection in the literature. If it is possible to compute all the possible models then we can use a criterion to select the best subset of important variables, otherwise we need to apply one of many strategies for searching for important variables. In high-dimensional data we can not fit a full model and we can not compute all the possible models. In the last case we study different ways of searching the space in order to include or exclude a variable.

Stepwise procedures (search strategies)

The search for important variables can be done by either a deterministic stepwise procedure or a stochastic one. Stepwise procedures, also known as stepwise regression, are used to improve the fit at each step using a search strategy that identifies significant variables. Existing methods have been developed for evaluating a number p of variables (complexity $O(p^2)$) by adding or deleting one variable at each step.

Specifically, for the stepwise procedure three sequential strategies are given below which select to add or delete a variable in the model (Hastie et al., 2001).

The first is sequential forward selection (SFS). It starts with an empty set of variables. Then, adds variables, one at a time, in order to improve the classification performance. One way do it is to create a criterion that estimates the classification performance using cross validation methods (details in Subsection 2.1.1) and adds the variable if the classification performance is improved. The process stops when no other useful variables are identified based on the specific criterion that is selected. A full sequence though to the model with all variables would visit $p(p+1)/2$ models. Generalized sequential forward selection is used when not one-at-a-time but the best q-subset (in the sense that it improves the criterion) of the candidate variables is added

into the model. The disadvantage of both simple and generalized sequential forward selection is that once a variable is retained, it cannot be discarded. This is called the nesting problem.

The second is sequential backward selection, known as backward elimination (Miller, 2002). This process starts with every candidate variable in the model. Remove variables, one-at-a-time, until all remaining variables contribute significantly to the classification performance or until e.g. AIC is minimised. As with the forward stepwise algorithm, a total of $p(p+1)/2$ models would be visited in a full sequence. Sequential backward selection requires more computational time than the forward one and also suffers from the nesting problem. Similar to the first method, generalized sequential backward selection deletes the least significant q-subset of the candidate variables at each step.

The last one, stepwise regression, which overcomes the nesting problem, is a combination of forward and backward in various ways. For example, the process starts with forward selection until no further variables are added, then backward selection is run until no further variables are dropped, then forward selection again, and so on until successive forward and backward steps both produce no change in the model. Note that this method has two criteria one for adding and one for removing variables and the criterion for adding variables should be more stringent than the criterion for removing variables so as to avoid infinite loops. Stepwise regression is fast and simple to implement for selection, but often misses the best model by becoming trapped in local minima. Alternatives to the aforementioned deterministic method are stochastic algorithms, for example simulated annealing or genetic algorithms (Murphy, 2012), which are able to escape local minima.

The complexity of Bayesian variable selection methods is harder to quantify than that of stepwise methods ($O(p^2)$). It depends on the model, the complexity of Metropolis-Hastings procedures for high-dimensional data (choice of proposal, good mixing) and the prior specification. For example, for the Bayesian variable selection with a linear model under a sparsity constraint the mixing time is linear in the number of covariates up to a logarithmic factor (Yang et al., 2015).

In the case of Bayesian variable selection with probit model for healthy and cancer tissues of BE ($p = 447$) for example, the method proposed about 20000 models in order to find a good model with only two variables. In this case, the forward stepwise method would start with a null model and visit only 893 models (at the first step the method evaluates 447 models in order to add the first variable in the model and, at the second, from the remaining

variables, the method evaluates 446 more models in order to add the second variable in the model), before stopping on a model with two variables. The backward stepwise is not feasible here, but it would in principle start with a full model and evaluate 198915 models ($= 447 \times 445$), before finally coming up with a model with two variables. Bayesian variable selection visits many more models than a simple forward stepwise algorithm, but explores the model space more thoroughly and in an efficient way.

Penalized regression methods

In multiple linear regression the ordinary least squares estimate of $\boldsymbol{\beta}$ is given by $\hat{\boldsymbol{\beta}}^{ols} = \text{argmin}_{\boldsymbol{\beta}}\{-2\log[L(\boldsymbol{\beta};\mathbf{y})]\}$, where the likelihood of the model is normal. The analytical solution if $\mathbf{X}'\mathbf{X}$ has full rank, is $\hat{\boldsymbol{\beta}}^{ols} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, where $\mathbf{X}$ and $\mathbf{y}$ are assumed to be centered. However, ordinary least squares estimates do not always exist; if $\mathbf{X}$ is not of full rank, $\mathbf{X}'\mathbf{X}$ is not invertible and there is no unique solution for $\hat{\boldsymbol{\beta}}^{ols}$. In particular, in the case of high-dimensional data, $(\mathbf{X}'\mathbf{X})^{-1}$ does not exist and in this case a regularization term, also known as a penalty term or shrinkage term, may be included. The penalty term makes the penalized estimates biased, but can also substantially reduce the variance. Different methods add different penalties as described below.

Ridge regression

Ridge regression was one of the first penalized regression methods. It was introduced by Hoerl and Kennard (1970). It penalizes the size of the regression coefficients but does not shrink coefficients to zero, so it is not on its own a method of VS. Ridge regression was extended to the generalized linear model (Friedman et al., 2010) through the criterion

$$\hat{\boldsymbol{\beta}}^{ridge} = \text{argmin}_{\boldsymbol{\beta}}\{-2\log[L(\boldsymbol{\beta};\mathbf{y})] + \lambda\|\boldsymbol{\beta}\|_2^2\}, \tag{2.4}$$

where $\|\boldsymbol{\beta}\|_2 = \sqrt{\sum_{j=1}^{p}\beta_j^2}$ (L_2 norm, also known as the Euclidean norm) and λ is a positive regularization parameter. In the linear case the analytical solution of Equation (2.4) is $\hat{\boldsymbol{\beta}}^{ridge} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{y}$, where the matrix $\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_p$ is always invertible. There is a unique solution for $\hat{\boldsymbol{\beta}}^{ridge}$ via Equation (2.4). The solution is indexed by the shrinkage parameter λ which is chosen by minimizing the mean squared fitting error, typically using cross validation. This parameter controls the amount of regularization and thus the size of the coefficients. The higher λ is, the more all coefficients are

shrunk towards zero. Note that if $\lambda \to 0$, then the least squares solution is obtained, $\hat{\boldsymbol{\beta}}^{ridge} \to \hat{\boldsymbol{\beta}}^{ols}$. If $\lambda \to \infty$, then $\hat{\boldsymbol{\beta}}^{ridge} \to 0$, which means that the model consists only of the intercept. In the orthonormal design case, the relation between those coefficients is $\hat{\boldsymbol{\beta}}^{ridge} = \frac{1}{1+\lambda}\hat{\boldsymbol{\beta}}^{ols}$.

Least absolute shrinkage and selection operator regression

Tibshirani (1996) introduced for the linear model the least absolute shrinkage and selection operator (LASSO), where coefficients are solutions of an L_1 optimization problem. LASSO was extended to the generalized linear model (Park and Hastie, 2007) via

$$\hat{\boldsymbol{\beta}}^{lasso} = \text{argmin}_{\boldsymbol{\beta}}\{-2\log[L(\boldsymbol{\beta};\mathbf{y})] + \lambda\|\boldsymbol{\beta}\|_1\}, \tag{2.5}$$

where $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p}|\beta_j|$ (L_1 norm). As λ increases, more coefficients are set to zero (fewer variables are selected), and among the nonzero coefficients, more shrinkage is employed. So, as λ increases, the number of nonzero components of $\boldsymbol{\beta}$ decreases.

The L_1 norm penalty of the LASSO approach makes the solution nonlinear and requires a quadratic programming algorithm. For linear models, the analytical solution of Equation (2.5) is more complicated than for L_2 shrinkage. In fact, it becomes $\hat{\boldsymbol{\beta}}^{lasso} = (\mathbf{X}'\mathbf{X})^{-1}\left(\mathbf{X}'\mathbf{y} - \frac{\lambda}{2}\mathbf{w}\right)_+$ where $\mathbf{w} = (w_1, \ldots, w_p)', w_j \in \{-1, +1\}$, depending on the sign of the corresponding regression coefficient and $_+$ denotes the positive part of the portion that is in parentheses. Based on the last equation, the relationship with ordinary least squares in the orthonormal case is given by $\hat{\boldsymbol{\beta}}^{lasso} = \text{sign}(\hat{\boldsymbol{\beta}}^{ols})(|\hat{\boldsymbol{\beta}}^{ols}| - \lambda/2)_+$, where $(|\hat{\boldsymbol{\beta}}^{ols}| - \lambda/2)_+ = |\hat{\boldsymbol{\beta}}^{ols}| - \lambda/2$ if $|\hat{\boldsymbol{\beta}}^{ols}| - \lambda/2 > 0$ and 0 otherwise. If $\hat{\boldsymbol{\beta}}^{ols} > \lambda/2$, then $\hat{\boldsymbol{\beta}}^{lasso} > 0$. If $\hat{\boldsymbol{\beta}}^{ols} < -\lambda/2$, then $\hat{\boldsymbol{\beta}}^{lasso} < 0$. This is a least squares problem with 2^p inequality constraints (there are 2^p possible sign patterns for the coefficients). LASSO works as a VS method. If $p > n$, the number of selected variables is bounded by the number of samples (Park and Casella, 2008). LASSO has one important limitation. If many variables are correlated LASSO fails to do grouped selection. So, it tends to select one variable from a group and ignore the others.

When there are many correlated variables, since LASSO gives nonzero weight to only one of them, Fused LASSO (Tibshirani et al., 2005) may be used. This is defined as

$$\hat{\boldsymbol{\beta}}^{f-lasso} = \text{argmin}_{\boldsymbol{\beta}}\{-2\log[L(\boldsymbol{\beta};\mathbf{y})] + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\sum_{j=2}^{p}|\beta_j - \beta_{j-1}|\},$$

where λ_1 and λ_2 are a positive regularization parameters. Here, the variables (and consequently the coefficients) are assumed to have a meaningful ordering. For example, in spectroscopic data, the wavenumbers have a natural order.

Zou (2006) considers the LASSO penalty and proposes a new penalty called the adaptive LASSO, which minimizes a penalized marginal likelihood function with a weighted L_1 penalty. The adaptive LASSO imposes different penalties on different coefficients: unimportant covariates receive larger penalties than important ones and vice versa. In this way, important variables can be protectively preserved in models while unimportant ones are more likely to be shrunk to zero via adaptive LASSO

$$\hat{\boldsymbol{\beta}}^{ad-lasso} = \text{argmin}_{\boldsymbol{\beta}}\{-2\log[L(\boldsymbol{\beta};\mathbf{y})] + \lambda\sum_{j=1}^{p}|\beta_j|w_j\},$$

where the nonnegative weights are chosen adaptively by data. They can be estimated by $w_j = (|\hat{\beta}_j^{ini}|)^{-\xi}$, $j = 1, 2, \ldots, p$, where ξ is a positive constant and $\hat{\beta}_j^{ini}$ is an initial root-n-consistent estimator of β_j. In the presence of collinearity ridge regression can be used to estimate the $\hat{\beta}_j^{ini}$, otherwise ordinary least squares is a possible choice. The w_j are regarded as leverage factors, which adjust penalties on the coefficients by taking large values for unimportant covariates and small values for important ones. The choice of ξ in the adaptive LASSO is important to assure good solutions. This penalty can be considered as an approximation to the L_k penalties with $k = 1 - \xi$. Any root-n-consistent estimates of β_j can be used for the adaptive weights without changing the asymptotic properties of the adaptive LASSO solution.

Bridge regression

Bridge regression (Fu, 1998) is a more general formulation of ridge regression and LASSO since it uses the L_κ norm $\|\boldsymbol{\beta}\|_\kappa = \left(\sum_{j=1}^{p}|\beta_j|^\kappa\right)^{1/\kappa}, \kappa \geq 0$. Ridge regression and the LASSO are special cases of bridge regression and correspond to $\kappa = 2$ and $\kappa = 1$ respectively.

Figure 2.3 gives an illustration of the L_κ norm for two coefficients ($p = 2$). Different norms produce different penalty functions that are represented with different colors. Also shown in the figure is the ordinary least squares estimate $\hat{\boldsymbol{\beta}}^{ols}$ and likelihood ellipses around it. The penalty shrinks the estimate towards zero, with the constrained regression coefficient estimates being given by the first point at which an ellipse contacts the constraint region. For example, for $\kappa = 1$, (LASSO) the enclosed region is a diamond. The ellipsoidal

outer contour of the residual sum of squares can touch the 'corner' of the diamond which corresponds to setting the coefficient at zero (top 'corner' corresponds to $\beta_1 = 0$) and so LASSO can perform VS (the resulting model here will only contain β_2). On the other hand, for example, for $\kappa = 2$ (ridge) the constraint region is a disk and the ellipsoidal inner contour hits the disk away from the axis. Since there are no 'corners' where the elliptical contours hit the constraint region and the coefficients are not usually set to zero, ridge shrinkage cannot perform VS. For $\kappa < 1$ the constraint region is nonconvex. For $\kappa \geq 2$, a smaller ellipsoidal inner contour touches the constraint region at a different point and the penalty shrinks some coefficients towards zero, but does not perform VS. When $p > 2$, the constraint region may have many 'corners' and consequently it is very likely that some coefficients will be zero.

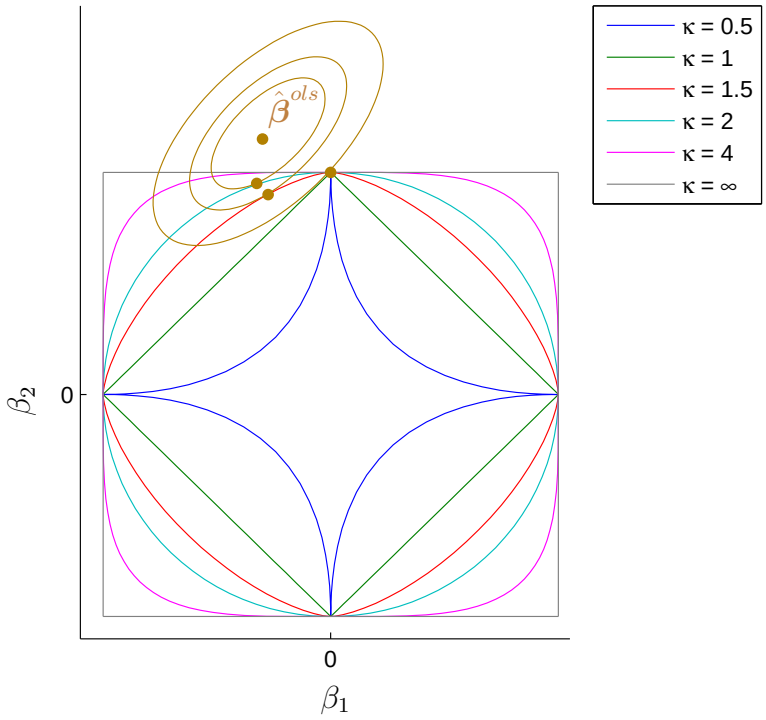


Figure 2.3: Illustrations of the L_κ norm for various values of κ.

Elastic net

For linear models Zou and Hastie (2005) proposed the elastic net, which is a convex combination of ridge and LASSO. The elastic net was extended to generalized linear models by Friedman et al. (2010)

$$\hat{\boldsymbol{\beta}}^{e-net} = \text{argmin}_{\boldsymbol{\beta}}\{-2\log[L(\boldsymbol{\beta};\mathbf{y})] + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_2^2\}, \qquad (2.6)$$

with λ_2 the ridge penalty parameter, penalizing the sum of the squared regression coefficients and λ_1, the LASSO penalty, penalizing the sum of the absolute values of the regression coefficients.

If $\lambda_1 = 0$, then ridge regression is obtained. If $\lambda_2 = 0$, then LASSO is obtained. The elastic net method seems to be particularly useful when dealing with highly correlated variables: the ridge term shrinks coefficients of correlated variables toward each other, whereas the LASSO term picks one among the correlated variables and puts all weight on it. Elastic net selects a group of highly correlated variables once one variable among them is selected, in contrast with LASSO which selects only one of them. Elastic net produces a sparse model with good prediction accuracy, while encouraging a grouping effect. For example, for linear models the analytical solution of Equation (2.6) is $\hat{\boldsymbol{\beta}}^{e-net} = (\mathbf{X}'\mathbf{X} + \lambda_2\mathbf{I}_p)^{-1}\left(\mathbf{X}'\mathbf{y} - \frac{\lambda_1}{2}\mathbf{w}\right)_+$. The adaptive version of elastic net (Zou and Zhang, 2009) is given by

$$\hat{\boldsymbol{\beta}}^{ad-e-net} = \text{argmin}_{\boldsymbol{\beta}}\{-2\log[L(\boldsymbol{\beta};\mathbf{y})] + \lambda_2\|\boldsymbol{\beta}\|_2^2 + \lambda_1\sum_{j=1}^{p}|\beta_j|w_j\},$$

where the nonnegative weights can estimated by $w_j = (|\hat{\beta}_j^{\,e-net}|)^{-\xi}$. To avoid dividing by zeros in the adaptive elastic net estimator, the nonnegative weights can estimated by $w_j = (|\hat{\beta}_j^{\,e-net}| + 1/n)^{-\xi}$.

Summary

All the aforementioned penalized methods can be formulated as

$$\hat{\boldsymbol{\beta}}^{penalized} = \text{argmin}_{\boldsymbol{\beta}}\{-2\log[L(\boldsymbol{\beta};\mathbf{y})] + \lambda_1 f_1(\boldsymbol{\beta}) + \lambda_2 f_2(\boldsymbol{\beta})\}, \tag{2.7}$$

where $\lambda_1, \lambda_2 > 0$ and f_1, f_2 represents specific choices of functions, namely different norms. A summary of the penalized methods is given in Table 2.6, according to notation in Equation (2.7).

Similarities and differences between Bayesian and penalized methods

All of these penalized methods can be given a Bayesian interpretation, with the log of the prior distribution replacing the penalty. These prior distributions (some of which are distinctly odd) are shown in Table 2.6. The regularization parameter (usually denoted by λ) of the penalized methods determines a scale parameter of the prior of Bayesian methods. Then, a good choice of scale parameter (Bayesian framework), or alternatively, a good

Table 2.6: Summary of a penalized regression methods, including also the closed form equation for the shrinkage for linear model (whenever is possible). The last column contains the corresponding Bayesian representation via a prior on the coefficients.

Model	λ_1	λ_2	$f_1(\boldsymbol{\beta})$	$f_2(\boldsymbol{\beta})$	$\hat{\boldsymbol{\beta}}^{\text{penalized}}$	Bayesian prior on the coefficients
Ordinary least squares	0	0	0	0	$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$	Independent uniform
Elastic net	λ_1	λ_2	$\sum_{j=1}^{p} \lvert\beta_j\rvert$	$\sum_{j=1}^{p} \beta_j^2$	$(\mathbf{X}'\mathbf{X} + \lambda_2\mathbf{I}_p)^{-1}\left(\mathbf{X}'\mathbf{y} - \frac{\lambda_1}{2}\mathbf{w}\right)_+$	Product of Laplace common scale and Gaussian prior distributions (Bornn et al., 2010)
Adaptive elastic net	λ_1	λ_2	$\sum_{j=1}^{p} \lvert\beta_j\rvert w_j$	$\sum_{j=1}^{p} \beta_j^2$		Product of Laplace adaptive scale and Gaussian prior distributions
Ridge (Bridge $\kappa = 2$)	0	λ_2	0	$\sum_{j=1}^{p} \beta_j^2$	$(\mathbf{X}'\mathbf{X} + \lambda_2\mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{y}$	β_j has independent Gaussian prior with mean zero and strictly positive scalar variance λ_2^{-1}
LASSO (Bridge $\kappa = 1$)	λ_1	0	$\sum_{j=1}^{p} \lvert\beta_j\rvert$	0	$\text{sign}(\hat{\boldsymbol{\beta}}^{ols})(\lvert\hat{\boldsymbol{\beta}}^{ols}\rvert - \lambda_1/2)_+$	β_j independent with Laplace common scale (or double exponential prior)
Adaptive LASSO	λ_1	0	$\sum_{j=1}^{p} \lvert\beta_j\rvert w_j$	0		β_j independent with Laplace adaptive scale prior
Fused LASSO	λ_1	λ_2	$\sum_{j=1}^{p} \lvert\beta_j\rvert$	$\sum_{j=2}^{p} \lvert\beta_j - \beta_{j-1}\rvert$		Product of independent Laplace and normal-exponential-gamma (Shimamura et al., 2016)

choice of regularization parameter (penalized framework) that can guide us to the 'best' model. Penalized methods just maximize the likelihood, but Bayesian methods explore more of the space and quantify the model uncertainty. For example, LASSO is a fast and efficient method for selecting a single model (only contains variables that do not touch to the 'corner' (see Figure 2.3)), but as a penalized method it does not allow us to estimate model uncertainty. The last is important within the Bayesian set-up, especially for BVS with the aim to do predictions. In the Bayesian perspective, the LASSO penalty arises from a double exponential, or Laplace prior (the prior has a 'corner' at zero and so the posterior mode may be identically zero). Bayesian methods are also more flexible since they can estimate the tuning parameters jointly with the other parameters, in contrast with penalized methods that require tuning of one or more parameters via, for example, cross-validation.

2.3.2 Variable extraction in high-dimensional data

Some variable extraction methods, for example LDA, do not work for high-dimensional data, since the covariance matrix is singular and the classification rule involves a linear combination of all the variables. In this case, we can apply, for example, penalized LDA (Witten and Tibshirani, 2011). Another way is to first apply principal component analysis (PCA) using all the variables and then estimate the LDA projection using only the extracted variables (Barber, 2012). PCA does not take responses into account, so it is an unsupervised learning method. Kernel PCA applies a non-linear transformation to a potentially very high-dimensional space, instead of a linear transformation.

Alternative techniques (instead of PCA) that are commonly used for dimensionality reduction in high-dimensional data are independent component analysis, canonical correlation analysis and partial least squares discriminant analysis (Hastie et al., 2001). These methods play an important role in overcoming the collinearity problem. Since they are variable extraction methods, the new variables that they produce are combinations of the original one and thus the new variables lose their natural meaning. This may be very important for example in medical applications.

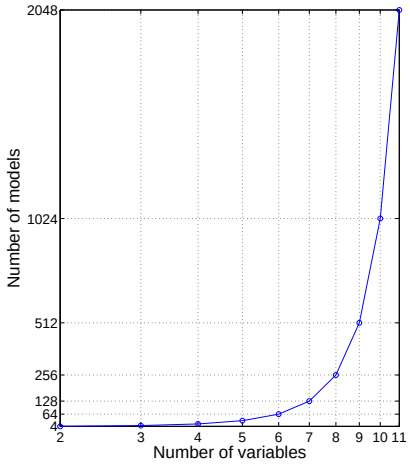
Chapter 3

Bayesian variable selection

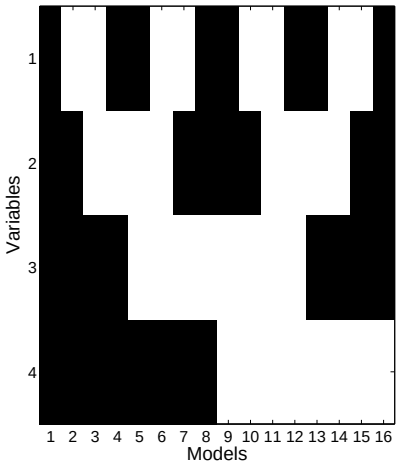
Another way to carry out variable selection (VS) is within the Bayesian framework. The model selection problem becomes part of the estimation: rather than searching for the single optimal model, we estimate the posterior probability of all models. Through Bayes theorem, we can calculate or approximate the posterior distribution which is proportional to likelihood times the prior. In Bayesian variable selection (BVS), the penalization of models with many variables is achieved via the prior distributions of the unknown parameters. In practice, we are interested in estimating both the probabilities of individual models and the marginal posterior probability that a variable should be included in the model. When the likelihood or marginal likelihood can not be computed in closed form we can approximate it using Markov chain Monte Carlo (MCMC) methods.

In practice, when the number of variables, p, is very large it can be impossible to calculate posterior probabilities of all 2^p possible models. Figure 3.1 shows how the number of models increases dramatically, as the number of variables in the model increases. In this case BVS not only provides intuitive probabilistic interpretation, but also efficiently explores the model space in a stochastic way to ensure that the models with high probabilities would show up earlier and more frequently during the MCMC simulation process.

BVS approaches are usually described in the context of linear models. This review will describe these methods and also discuss the extension to the probit model when the responses are categorical.



(a) Number of models versus number of variables.



(b) Gray code (0-1 coding, where 1/black means variable included in the model and 0/white variable not included) for a model with 4 variables.

Figure 3.1: (a) Number of models versus number of variables and (b) In a model with $p = 4$ variables there are $2^4 = 16$ possible models.

3.1 Probit model using latent variables

In this study we focus on BVS using a probit model. The representation of the probit model with binary responses ($M = 2$) is given by

$$P(y_i = 1|\mathbf{X}_{i,:}) = \Phi(\alpha + \mathbf{X}_{i,:}\boldsymbol{\beta}), \tag{3.1}$$

where y_i is the i-th response in the $n \times 1$ response vector $\mathbf{y}$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a $p \times 1$ vector of regression coefficients, α is an unknown intercept, $\mathbf{X}_{i,:}$ is the i-th row vector of $n \times p$ matrix $\mathbf{X}$ and Φ is the standard normal CDF, which is a link function between the linear predictor and the probability of the response taking value one. The probit model with binary responses may also be presented as a normal linear model for a latent variable vector $\mathbf{z} = (z_1, \ldots, z_n)'$ such that

$$z_i \sim N(\alpha + \mathbf{X}_{i,:}\boldsymbol{\beta}, \sigma^2), \tag{3.2}$$

for $i = 1, \ldots, n$ with $y_i = 1$ iff $z_i > 0$ (Albert and Chib, 1993). In the (univariate) probit model, if both $\boldsymbol{\beta}$, σ^2 are unknown parameters, then they are not identifiable from the model. $P(z_i > 0|\mathbf{X}_{i,:}, \sigma^2) = \Phi(\alpha/\sigma^2 + \mathbf{X}_{i,:}(\boldsymbol{\beta}/\sigma^2))$ depends only on the ratio between $\alpha, \boldsymbol{\beta}$ and σ^2. As such there are many combinations of $\boldsymbol{\beta}$ and σ^2 that can give the same response and likelihood.

The common way to overcome this identifiability problem for a probit model is to assume that the variance is fixed, i.e., $\sigma^2 = 1$. Even though throughout this chapter the variance will be taken as equal to one, the notation σ^2 is still used since later on we will assign a prior distribution to the variance instead of fixing it. The luck of identifiability means that this must be a proper prior and implies that the choice of prior, to be discussed later, will be important. For fixed i, there is one latent variable and two possible responses ($m = 0, 1$). The relationship between y_i and z_i is then simply given by

$$y_i = \begin{cases} 1, & \text{if } z_i > 0, \\ 0, & \text{if } z_i \leq 0, \end{cases} \tag{3.3}$$

and the corresponding illustration is presented in Figure 3.2.

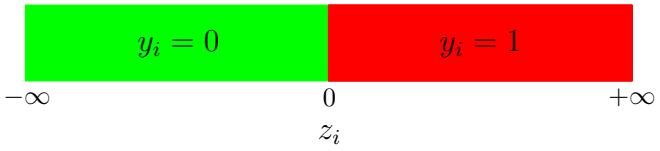


Figure 3.2: Graphical representation of the relationship between binary responses and the continuous latent variable (Equation (3.3)).

The two representations, without and with the latent variable z_i, are equivalent, since for $\sigma^2 = 1$

$$\begin{aligned} P(y_i = 1|\mathbf{X}_{i,:}) &= P(z_i > 0|\mathbf{X}_{i,:}) = P(z_i - \alpha - \mathbf{X}_{i,:}\boldsymbol{\beta} > -\alpha - \mathbf{X}_{i,:}\boldsymbol{\beta}|\mathbf{X}_{i,:}) \\ &= 1 - \Phi(-\alpha - \mathbf{X}_{i,:}\boldsymbol{\beta}) = \Phi(\alpha + \mathbf{X}_{i,:}\boldsymbol{\beta}). \end{aligned}$$

In order to perform the VS for the probit model a standard approach is to use a $p \times 1$ indicator vector $\boldsymbol{\gamma}$ Lee et al. (2003), where the j-th element γ_j is defined such that

$$\gamma_j = \begin{cases} 1, & \text{if } \beta_j \neq 0, \\ 0, & \text{if } \beta_j = 0, \end{cases} \tag{3.4}$$

for $j = 1, \ldots, p$.

In the case of using indicator variables, a model is built for VS where the goal is to not take into account those columns $\mathbf{X}$ for which $\beta_j = 0$. One such representation of the model of Equation (3.1) is to select a submodel of it, which corresponds to the probit model for VS such that

$$P(y_i = 1|\mathbf{X}_{i,\boldsymbol{\gamma}}) = \Phi(\alpha + \mathbf{X}_{i,\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}), \tag{3.5}$$

where $\mathbf{X}_{i,\boldsymbol{\gamma}}$ is i-th row vector of the $n \times p_{\boldsymbol{\gamma}}$ ($p_{\boldsymbol{\gamma}} = \sum_{j=1}^{p} \gamma_j$) matrix $\mathbf{X}_{\boldsymbol{\gamma}}$ which consists of the columns of $\mathbf{X}$ corresponding only to $\gamma_j = 1$, and $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is the corresponding $p_{\boldsymbol{\gamma}} \times 1$ vector of nonzero unknown regression coefficients. The latent variable of Equation (3.2) in the context of VS takes the form

$$z_i \sim N(\alpha + \mathbf{X}_{i,\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2), \tag{3.6}$$

where in this chapter $\sigma^2 = 1$.

Both probit representations, referred as the probit model (Equation (3.1)) and the probit model for VS (Equation (3.5)) assume that $\mathbf{X}$ has been centered, so that its columns sum to zero, which yields $\text{rank}(\mathbf{X}) \leq \min\{n-1, p\}$ (Sha et al., 2004). The last comment is important in the case of large p, small n.

3.2 Prior distributions

The choice of prior distributions for the unknown parameters is very important in BVS approaches. With σ^2 fixed, three priors, one for α, one for $\boldsymbol{\beta}$ (in the case of probit model) or for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ (in the case of probit model for VS) and one for $\boldsymbol{\gamma}$ need to be specified.

3.2.1 Priors for intercept

The intercept represents the overall mean of the model and it is a common parameter for all possible models. So, since there is usually no information for the intercept a priori, a non-informative, also known as vague, diffuse prior can be used, $p(\alpha) \propto 1, \alpha \in \mathbb{R}$ (Russu et al., 2012). Sha et al. (2004) and Brown et al. (1998a) use a normal prior for α

$$\alpha \sim N(\alpha_0, \sigma^2 h), \tag{3.7}$$

where α_0 corresponds to the mean (typical choice $\alpha_0 = 0$) and h is a hyperparameter scaling the variance of the univariate normal distribution. Usually, h is fixed at a large value, see for example Lamnisos et al. (2009). This corresponds to little prior information for α being available. At the other extreme, when $h = 0$, this yields $\alpha = 0$, which means that the study of BVS is without the intercept for the probit model (George and McCulloch, 1993; Lee et al., 2003).

3.2.2 Priors for coefficients

In contrast to the intercept α, there is extensive literature on the prior for $\boldsymbol{\beta}$, since this along with the prior for $\boldsymbol{\gamma}$ is the most important in BVS. The stochastic search VS approach (George and McCulloch, 1993) is widely used and assumes that $p(\beta_j, \gamma_j) = p(\beta_j|\gamma_j)p(\gamma_j)$ and β_j, $j = 1, \ldots, p$, are independent a priori conditional on $\boldsymbol{\gamma}$, where the coefficients come from mixture distributions that is often described as spike and slab:

$$p(\beta_j|\gamma_j) = (1 - \gamma_j)g_1 + \gamma_j g_2, \tag{3.8}$$

where g_1, g_2 are any continuous or discrete PDF/PMF that correspond to the spike and the slab respectively. Specifically, the g_1 is responsible for driving the coefficients to zero and g_2 allows for nonzero coefficients.

One widely used case of the last equation is a mixture of two normal distributions with different variances

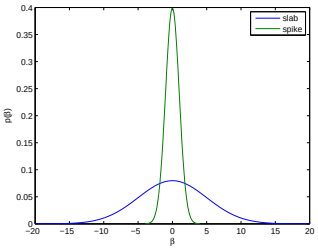
$$p(\beta_j|\gamma_j) = (1 - \gamma_j)N(\beta_0, c_1^2\sigma_\beta^2) + \gamma_j N(\beta_0, c_2^2\sigma_\beta^2), \tag{3.9}$$

where β_0 is the mean (typical choice $\beta_0 = 0$), and $c_2 \gg c_1$. In this case the first normal distribution is responsible for β_j to be relatively close to zero compared to β_j's that belong to the second normal distribution. This is one case of spike and slab priors with both distributions being normal. This prior applies the same shrinkage to all β_j and the relative size of the columns of $\mathbf{X}$ is important. If the columns of $\mathbf{X}$ have substantially different scales, standardization may be appropriate. In this thesis we are working mainly with spectroscopic data and they are all the same type of variables. As such, it is enough to center the spectroscopic data, since they are all in the same units (absorbance).

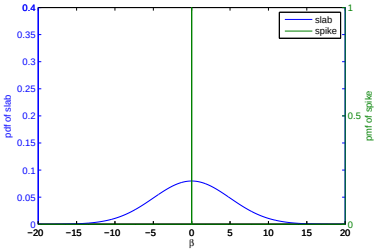
A special case of spike and slab (Equation (3.10)) is one where the spike is a discrete point mass at zero and the slab is a normal distribution, namely

$$p(\beta_j|\gamma_j) = (1 - \gamma_j)I_0 + \gamma_j N(\beta_0, c^2\sigma^2), \tag{3.10}$$

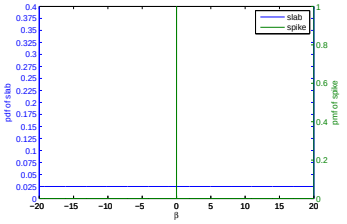
where I_0 is a point mass at zero and β_j are iid given $\boldsymbol{\gamma}$. $\gamma_j = 0$ is equivalent to $\beta_j = 0$ (β_j is exactly zero) and the prior on the nonzero coefficients is normal. A normal prior for the slab part is preferred against, for example, a uniform prior (Mitchell and Beauchamp, 1988) because a normal prior allows efficient sampling of the posterior. A graphical representation of different examples of spike and slab priors is given in Figure 3.3.



(a) Spike N(0,1) and slab N(0,5^2).



(b) Spike at 0 and slab N(0,5^2).



(c) Spike at 0 and slab U(-20,20).

Figure 3.3: Three cases of spike and slab priors for the coefficients.

In the BVS setup a commonly used prior form for nonzero coefficients is a multivariate normal (MVN) with a flexible choice for the covariance matrix (not necessarily assuming independence between nonzero coefficients) via

$$\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma} \sim MVN(\boldsymbol{\beta}_{0\boldsymbol{\gamma}}, \sigma^2\mathbf{H}_{\boldsymbol{\gamma}}), \mathbf{H}_{\boldsymbol{\gamma}} = c_1(\mathbf{X}_{\boldsymbol{\gamma}}{}'\mathbf{X}_{\boldsymbol{\gamma}})^+ + c_2\mathbf{D}_{\boldsymbol{\gamma}}, \tag{3.11}$$

where $\boldsymbol{\beta}_{0\boldsymbol{\gamma}}$ is the corresponding mean vector (typical choice $\boldsymbol{\beta}_{0\boldsymbol{\gamma}} = \mathbf{0}$), c_1 and c_2 are constants, $(\mathbf{X}_{\boldsymbol{\gamma}}{}'\mathbf{X}_{\boldsymbol{\gamma}})^+$ is the Moore-Penrose inverse matrix (or pseudoinverse matrix) and $\mathbf{D}_{\boldsymbol{\gamma}}$ is a $p_{\boldsymbol{\gamma}} \times p_{\boldsymbol{\gamma}}$ diagonal matrix. With σ^2 still fixed at 1, setting $c_1 = 0, c_2 = c^2, \mathbf{D}_{\boldsymbol{\gamma}} = \mathbf{I}_{p_{\boldsymbol{\gamma}}}$ in Equation (3.11) leads to Equation (3.10).

The simplest choice in Equation (3.11) is $\mathbf{H}_{\boldsymbol{\gamma}} = c\mathbf{I}_{p_{\boldsymbol{\gamma}}}$ ($c_1 = 0, \mathbf{D}_{\boldsymbol{\gamma}} = \mathbf{I}_{p_{\boldsymbol{\gamma}}}$), where $\mathbf{I}_{p_{\boldsymbol{\gamma}}}$ is the $p_{\boldsymbol{\gamma}} \times p_{\boldsymbol{\gamma}}$ identity matrix, which means that the nonzero coefficients are independent. Setting $c_2 = 0$ in Equation (3.11) yields the generalized g-prior (gsg-prior) $\mathbf{H}_{\boldsymbol{\gamma}} = c(\mathbf{X}_{\boldsymbol{\gamma}}{}'\mathbf{X}_{\boldsymbol{\gamma}})^+$. The Moore-Penrose inverse matrix is suggested for the coefficient prior in the probit model (Ai-Jun and Xin-Yuan, 2010) so as to avoid the problem of singular inverse matrix when $p_{\boldsymbol{\gamma}} > n$. If $\mathbf{X}_{\boldsymbol{\gamma}}$ is a full column rank matrix, then $(\mathbf{X}_{\boldsymbol{\gamma}}{}'\mathbf{X}_{\boldsymbol{\gamma}})^+ = (\mathbf{X}_{\boldsymbol{\gamma}}{}'\mathbf{X}_{\boldsymbol{\gamma}})^{-1}$ and thus a special widely used case of Equation (3.11), for $c_1 = g$ (g scalar) and $c_2 = 0$, is a Zellner's g-prior $\mathbf{H}_{\boldsymbol{\gamma}} = g(\mathbf{X}'_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}})^{-1}$ (Zellner, 1986). A diagonalized version of the g-prior is a special (not so usual) case of the g-prior, where $\mathbf{H}_{\boldsymbol{\gamma}} = c\,\mathrm{diag}((\mathbf{X}_{\boldsymbol{\gamma}}{}'\mathbf{X}_{\boldsymbol{\gamma}})^{-1})$, which has a computational advantage

since it is easy to invert a diagonal matrix. Zellner's g-prior cannot be used if $p_{\boldsymbol{\gamma}} > n$, or if some variables (out of p) are linear combinations of others, in which case the gsg-prior can used. The normal prior for nonzero coefficients can be replaced by a Cauchy prior (Zellner and Siow, 1980; Gelman et al., 2008), $\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma} \sim \text{Cauchy}(\boldsymbol{\beta}_{0\boldsymbol{\gamma}}, g(\mathbf{X}_{\boldsymbol{\gamma}}'\mathbf{X}_{\boldsymbol{\gamma}})^{-1})$, where g is a scalar. However, it does not lead to a closed form expression for the marginal likelihood, so it is not usually preferred.

Instead of using the Moore-Penrose inverse matrix in Equation (3.11), another idea to address singularity related issues is by limiting the number of selected covariates at each iteration (Baragatti and Pommeret, 2012), which is a computational advantage. Alternatively Gupta and Ibrahim (2007) proposed to use a ridge parameter, so that $\mathbf{H}_{\boldsymbol{\gamma}}$ can be written as

$$\mathbf{H}_{\boldsymbol{\gamma}} = c_1(\mathbf{X}_{\boldsymbol{\gamma}}'\mathbf{X}_{\boldsymbol{\gamma}} + \lambda\mathbf{I}_{p_{\boldsymbol{\gamma}}})^{-1} + c_2\mathbf{D}_{\boldsymbol{\gamma}}. \tag{3.12}$$

For $\lambda = 0$, Equation (3.12) is equivalent to Equation (3.11) if $\mathbf{X}_{\boldsymbol{\gamma}}$ is a full column rank matrix. Actually, Equation (3.12) for $c_2 = 0$ corresponds to the Bayesian formulation of the frequentist penalized ridge regression (Subsection 2.3.1).

In the aforementioned cases, the choice of c_1 and c_2 is crucial because they control the amount of shrinkage of the nonzero regression coefficients. Values of c_1 and c_2 that are too small lead to over-shrinkage and bad out-of-sample prediction but a value of c_1 that is too large leads to Lindley's paradox (the model with no regressors is favoured regardless of the data). This suggests that there are values of c_1 between those extreme values that will return a good out-of-sample prediction. For the probit case using $\mathbf{H}_{\boldsymbol{\gamma}} = c\mathbf{I}_{p_{\boldsymbol{\gamma}}}$, Sha et al. (2004) suggest that c should correspond to the ratio of prior to posterior precision (inverse of variance) being between 0.1 and 0.005, that is

$$c^\star(\bar{\lambda}, 0.1) < c < \max\{c^\star(\bar{\lambda}, 0.005), c^\star(\lambda_{0.1}, 0.5)\}, \tag{3.13}$$

where $c^\star(\lambda, p) = (1-p)/(p\lambda)$, $\bar{\lambda}$ is the mean of the r nonzero eigenvalues of the sample variance matrix ($r = \text{rank}(X)$) and $\lambda_{0.1}$ the eigenvalue such that 10% of eigenvalues are less than it. George and Foster (2000) estimate c by maximizing its marginal likelihood. Strimenopoulou and Brown (2008) describe an empirical Bayes method for maximum a-posteriori estimation for the logistic model. Lamnisos et al. (2012) select c so as to minimize the criterion log predictive score, which is used as a measure of performance. Cross validation densities are used to specify c and sampling from them is

implemented via importantce sampling. Alternatively, it may be useful to try to put a prior on c.

In addition, to allow incorporation of variable to variable interaction, a partial least squares g-prior was introduced by Peng et al. (2013). Here the prior $\mathbf{H}_{\boldsymbol{\gamma}}$ for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is based on the scores from a partial least squares analysis of the $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$.

In contrast with the above mentioned priors, Kuo and Mallick (1998) assume independent prior distributions for β_j and γ_j, i.e., $p(\beta_j, \gamma_j) = p(\beta_j)p(\gamma_j)$. In this case a conjugate normal prior is assigned to $\boldsymbol{\beta}$

$$\boldsymbol{\beta} \sim MVN(\boldsymbol{\beta}_0, \sigma^2\mathbf{H}), \mathbf{H} = c_1(\mathbf{X}'\mathbf{X})^+ + c_2\mathbf{D}, \tag{3.14}$$

which form is be similar to the prior of Equation (3.11) but without taking into account $\boldsymbol{\gamma}$ (because of independence), with $\mathbf{H}_{\boldsymbol{\gamma}}$ replaced by $\mathbf{H}$, $\mathbf{D}_{\boldsymbol{\gamma}}$ by $\mathbf{D}$ ($\mathbf{D} = \text{diag}(d_1, \ldots, d_p)$), with typical choices being $\boldsymbol{\beta}_0 = \mathbf{0}$ and $\sigma^2 = 1$. As it stands, this is not equivalent to spike and slab priors. However, the equivalence can seen if we consider $\gamma_j\beta_j$, $j = 1, \ldots, p$. Then the prior of $\gamma_j\beta_j$ is the spike point mass at zero and a normal slab.

Instead of using an indicator vector $\boldsymbol{\gamma}$ for VS Bae and Mallick (2004) modeled $\boldsymbol{\beta}$ via a prior on the covariance matrix $\mathbf{H}$, for example, a Laplace prior distribution can be used for the covariance matrix with mode zero and prior variance $\nu = 2/\lambda^2$, where λ is a penalty factor. The zero mode encodes a prior belief of no effect, the prior variance determines the strength of this belief and hence the sparseness of the fitted model. The Laplace prior does not produce an analytically tractable solution for the posterior, since it is not a conjugate prior. The Bayesian LASSO uses Laplace priors on the coefficients (Park and Casella, 2008). The Laplace prior has also been used for BVS for survival regression, where the method simultaneously performs regression parameter estimation (via a penalized maximum likelihood approach) and BVS (Tachmazidou et al., 2010).

Different priors on the coefficients induce different amounts of shrinkage in their estimates. This varies from a small amount of shrinkage, or even no shrinkage, to coefficients with a large amount of shrinkage, or even complete removal of coefficients. Prior densities with heavier tails will lead to less shrinkage and those that are more peaked at zero will lead to larger shrinkage. For example, comparing Equation (3.9) with Equation (3.10) (independent coefficients), in the last case the amount of shrinkage is much higher for the coefficients that are close to zero, corresponding to removing these coefficients altogether. Then, focusing on the prior of the nonzero coefficients

and allowing some dependence between them, see Equation (3.11), further shrinkage can be achieved via c_1. For the aforementioned equation, different types of shrinkage can be achieved via $c_1(\mathbf{X}_{\boldsymbol{\gamma}}{}'\mathbf{X}_{\boldsymbol{\gamma}})^+$ (for $c_2 = 0$) versus $c_2\mathbf{I}_{\boldsymbol{\gamma}}$ (for $c_1 = 0$). The first one simply shrinks all coefficients by the same scalar factor regardless of how well they are estimated from the data and may lead to unstable estimations of coefficients, whereas the second is a ridge type shrinkage that can stabilize collinearity. If we are interested in out-of-sample prediction, this latter type of shrinkage is preferable. Finally, Cauchy prior helps in not shrinking the large coefficients too much compared to the normal distribution.

3.2.3 Priors for indicator vector

A prior on the model space can be specified through $\boldsymbol{\gamma}$. BVS commonly uses a Bernoulli prior (George and McCulloch, 1997, 1993; Lee et al., 2003)

$$p(\boldsymbol{\gamma}) = \prod_{j=1}^{p} w_j^{\gamma_j}(1 - w_j)^{1-\gamma_j}, \tag{3.15}$$

where $w_j, 0 \leq w_j \leq 1$, is the probability that each column of $\mathbf{X}$ enters into the model (probability of success $P(\gamma_j = 1)$ for the Bernoulli prior) independently of the remaining columns of $\mathbf{X}$. In Equation (3.15) to distinguish groups of variables, different w_j can be chosen for different groups. In the special case where $w_j = w$ for $j = 1, \ldots, p$, Equation (3.15) yields

$$p(\boldsymbol{\gamma}) = w^{p_{\boldsymbol{\gamma}}}(1 - w)^{p-p_{\boldsymbol{\gamma}}}, \tag{3.16}$$

so that the distribution of $p_{\boldsymbol{\gamma}}$ is binomial, $Bin(p, w)$. The value of w can be set to control the number of selected variables a priori.

In the case that there is prior knowledge about the plausible values that w can take, w can be fixed. In the case of $p \gg n$, small values of w are chosen, so as to restrict the number of variables in the model. For example, for a dataset with 1000 variables and a smaller number of observations, setting $w = 0.01$ means that only 10 variables are expected to be selected before observing the data. So, the last prior penalizes larger models when w is given a small value. Alternatively, in the case of $p \gg n$, sets Dobra (2009) a maximum model size p_{max} so that we take into account only models where $p_{\boldsymbol{\gamma}} < p_{max}$, with a uniform prior across the models. The particular choice of $w = 1/2$ in Equation (3.16), yields the uniform prior where all models have equal probabilities. In this case, $p(\boldsymbol{\gamma}) = 2^{-p}$ which is easy to specify and

reduces the computational cost since it can be omitted from the marginal conditional distribution of γ_j. However, for our application we will not select $w = 1/2$ because giving all models have equal prior probabilities will not lead to the selection of a sparse model.

As an alternative to specifying w, a distribution can be assigned to it. Uncertainty on w can be modeled by imposing a Beta hyperprior for w, ($w \sim \text{Beta}(\delta_1, \delta_2)$, $p(w) = \frac{w^{\delta_1 - 1}(1-w)^{\delta_2 - 1}}{B(\delta_1, \delta_2)}$, where B is the Beta function) as defined in Kohn et al. (2001). The distribution of $p_{\boldsymbol{\gamma}}$ is a beta-binomial distribution. The distribution of $\boldsymbol{\gamma}$, without the binomial coefficient term, is

$$p(\boldsymbol{\gamma}) = \int p(\boldsymbol{\gamma}|w)p(w)dw = \frac{B(\delta_1 + p_{\boldsymbol{\gamma}}, p - p_{\boldsymbol{\gamma}} + \delta_2)}{B(\delta_1, \delta_2)},$$

where the parameters $\delta_1 > 0$ and $\delta_2 > 0$ can be specified based on the prior belief about the model size $p_{\boldsymbol{\gamma}}$. The choice of two parameters arises from solving the linear simultaneous equations of mean $E(p_{\boldsymbol{\gamma}}) = pE(w)$ and variance $\text{var}(p_{\boldsymbol{\gamma}}) = \text{var}(E(p_{\boldsymbol{\gamma}}|w)) + E(\text{var}(p_{\boldsymbol{\gamma}}|w))$ of $p_{\boldsymbol{\gamma}}$, in terms of δ_1 and δ_2. This hyperprior provides control of model sizes, in the sense that the unknown probability of success w has a prior, and the number of variables $p_{\boldsymbol{\gamma}}$ that enter in the model can be controlled.

Instead of putting a prior on w, a prior on model size $p_{\boldsymbol{\gamma}}$, $f(p_{\boldsymbol{\gamma}})$, can be used (Chipman et al., 2001). Then,

$$p(\boldsymbol{\gamma}) = \binom{p}{p_{\boldsymbol{\gamma}}}^{-1} f(p_{\boldsymbol{\gamma}}),$$

where the first term is the inverse binomial coefficient, which indicates that there are $\binom{p}{p_{\boldsymbol{\gamma}}}$ different models of size $p_{\boldsymbol{\gamma}}$, where $p_{\boldsymbol{\gamma}} = 0, \ldots, r-1$ ($r = \text{rank}(X)$). The case where all the different model sizes are equally likely, leads to the simple form $p(\boldsymbol{\gamma}) = \binom{p}{p_{\boldsymbol{\gamma}}}^{-1}$ (Abramovich and Grinshtein, 2010).

3.2.4 Bayesian variable selection via dependent indicator variables

One limitation of the priors for the indicator vectors described above is that the presence or absence of a variable is independent of the presence or absence of the other variables (Subsection 3.2.3). Here we consider the possibility that γ_j are dependently distributed with a two stage first-order Markov model used to represent the dependence, so as to build a more realistic model.

A probit model with binary responses is studied, where the model for latent variables includes an intercept α. We assume that the prior of α is

normal (Equation (3.7)) and the prior of the nonzero coefficients $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is normal (Equation (3.11) for $c_2 = 0$) with $\mathbf{H}_{\boldsymbol{\gamma}} = c(\mathbf{X}'_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}})^{+}$. Under the consideration that the inclusion of X_j is dependent on the inclusion of $X_{j'}$ for all $|j-j'| = 1$, the prior assumption for $\boldsymbol{\gamma}$ is considered to be a two stage first-order Markov model with transition probabilities

$$p(\gamma_j = 0|\gamma_{j-1} = 0), p(\gamma_j = 1|\gamma_{j-1} = 0),$$
$$p(\gamma_j = 0|\gamma_{j-1} = 1), p(\gamma_j = 1|\gamma_{j-1} = 1).$$

We specify the aforementioned probabilities via 2×2 matrix in three different ways:

- Case 1: An empirical approach

$$\begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} 1-\eta w & \eta w \\ 1-w & w \end{pmatrix} \end{array} \tag{3.17}$$

  where the probability to include the specific variable when the previous one is not included in the model is η times higher than the probability to include the specific variable when the previous one is included. In this case, the stationary distribution for the Markov chain is

$$\left(\frac{1-w}{1-w-\eta w}, \frac{\eta w}{1-w-\eta w} \right).$$

  This means that, leaving the chain to run for a long time the proportion of 0's is equal to the first element of the stationary distribution, and the proportion of 1's to the second element of the stationary distribution. The last term reflects the number of 1's that we expect to see and in practice we need this to be small in order to achieve sparsity in the solution. We select η and w together under the restriction that $\eta \ll 1/w$.

- Case 2: If $w = P(\gamma_j = 1)$ is the probability of success, then $q = 1 - w = P(\gamma_j = 0)$ is the probability of failure. In addition, we let $\rho = corr(\gamma_{j-1}, \gamma_j), \rho > 0$. In this case the t-step transition matrix of a Markov chain is

$$\begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} 1-w(1-\rho^t) & w(1-\rho^t) \\ q(1-\rho^t) & 1-q(1-\rho^t) \end{pmatrix}. \end{array}$$

So, the Markov chain represents a sequence of correlated Bernoulli random variables $\gamma_j, j = 1, \ldots, p$ (Minkova and Omey, 2014). The stationary distribution depends on the correlation coefficient and is

$$\left(\frac{q(1-\rho)}{q(1-\rho) + w(1-\rho)}, \frac{w(1-\rho)}{q(1-\rho) + w(1-\rho)} \right).$$

- Case 3: Muenz and Rubinstein (1985) employed logistic regression models to describe the transition probabilities from one stage to another

$$P(\gamma_j = 0|\gamma_{j-1} = 0) = \frac{e^{\alpha + \mathbf{X}_{i,:}\boldsymbol{\beta}_0}}{1 + e^{\alpha + \mathbf{X}_{i,:}\boldsymbol{\beta}_0}}, P(\gamma_j = 1|\gamma_{j-1} = 0) = \frac{e^{\alpha + \mathbf{X}_{i,:}\boldsymbol{\beta}_1}}{1 + e^{\alpha + \mathbf{X}_{i,:}\boldsymbol{\beta}_1}},$$

where $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ are the vector of coefficients in the two logistic regressions. We apply the last idea, instead of logistic, to a probit regression as follows

$$P(\gamma_j = 0|\gamma_{j-1} = 0) = \Phi(\alpha + \mathbf{X}_{i,:}\boldsymbol{\beta}_0), P(\gamma_j = 1|\gamma_{j-1} = 0) = \Phi(\alpha + \mathbf{X}_{i,:}\boldsymbol{\beta}_1)$$

where $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ coefficients can calculated via the MLE approach, Albert and Chib (1993). In this case, the stationary distribution for the Markov chain is more complicated and can be derived as

$$\left(\frac{\frac{e^{\alpha + \mathbf{X}_{i,:}\boldsymbol{\beta}_1}}{1+e^{\alpha + \mathbf{X}_{i,:}\boldsymbol{\beta}_1}}}{1 - \frac{e^{\alpha + \mathbf{X}_{i,:}\boldsymbol{\beta}_0}}{1+e^{\alpha + \mathbf{X}_{i,:}\boldsymbol{\beta}_0}} + \frac{e^{\alpha + \mathbf{X}_{i,:}\boldsymbol{\beta}_1}}{1+e^{\alpha + \mathbf{X}_{i,:}\boldsymbol{\beta}_1}}}, 1 - \frac{\frac{e^{\alpha + \mathbf{X}_{i,:}\boldsymbol{\beta}_1}}{1+e^{\alpha + \mathbf{X}_{i,:}\boldsymbol{\beta}_1}}}{1 - \frac{e^{\alpha + \mathbf{X}_{i,:}\boldsymbol{\beta}_0}}{1+e^{\alpha + \mathbf{X}_{i,:}\boldsymbol{\beta}_0}} + \frac{e^{\alpha + \mathbf{X}_{i,:}\boldsymbol{\beta}_1}}{1+e^{\alpha + \mathbf{X}_{i,:}\boldsymbol{\beta}_1}}} \right).$$

According to Ai-Jun and Xin-Yuan (2010), integrating out α and $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ from the joint posterior can reduce the strong posterior correlation between $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and $\boldsymbol{\gamma}$ and between $\mathbf{z}$ and $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$. The dependence between some of the indicator variables is taken under consideration when carrying out the posterior inference.

Rather than specifying η and w, in case 1, we could assign a prior on $\eta|w \sim \text{Bin}(p, w)$ and a hyperprior on w, where $w \sim \text{Beta}(\delta_1, \delta_2)$. This hyperprior could be integrated out similarly to the beta-binomial model. In case 2, it would be complicated to assign a prior on the hyperparameter ρ of a probit model. In the simple case where two random variables have a bivariate normal distribution, the prior on the correlation coefficients can chosen to be proportional to $(1+\rho)^r$, where r will determine the weight the prior will have in estimation (e.g. $r = -3/2$ multiple parameter Jeffreys' rule) (Schisterman et al., 2003). However, it is difficult to see how to use this here.

3.3 Bayesian inference for probit models

First we consider inference for a probit model using a latent variable (Section 3.1), but without variable selection there are two main approaches.

In the first approach (Albert and Chib, 1993), $\boldsymbol{\beta}$ and $\mathbf{z}$ are updated sequentially. The latent variable $\mathbf{z}$ conditioned on $\boldsymbol{\beta}$ follows a univariate normal distribution truncated at zero

$$z_i|\mathbf{X}_{i,:}, y_i, \boldsymbol{\beta} \propto \begin{cases} N(\mathbf{X}_{i,:}\boldsymbol{\beta}, \sigma^2) \text{ truncated on the left at 0}, & \text{if } y_i = 1, \\ N(\mathbf{X}_{i,:}\boldsymbol{\beta}, \sigma^2) \text{ truncated on the right at 0}, & \text{if } y_i = 0, \end{cases} \tag{3.18}$$

where these authors fixed $\sigma^2 = 1$. The PDF of an example of the univariate normal truncated at zero is given in Figure 3.4.

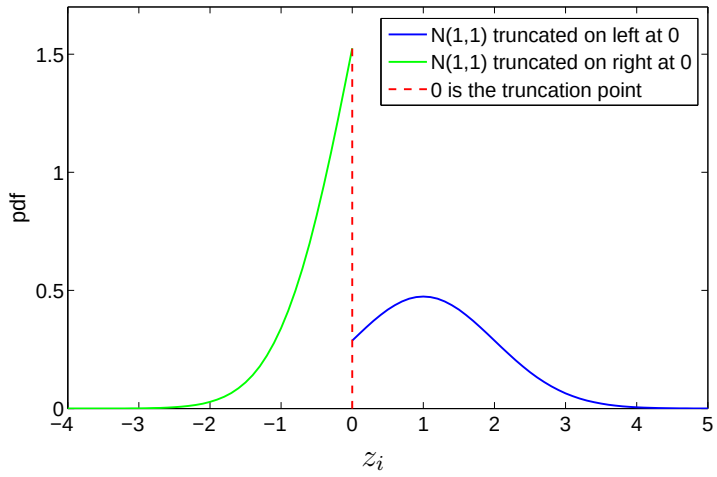


Figure 3.4: PDF of the normal distribution with mean and variance one truncated on the left (right) at zero.

The coefficient vector $\boldsymbol{\beta}$ conditioned on $\mathbf{z}$ follows, for the normal prior on $\boldsymbol{\beta}$ of Equation (3.14) with $\sigma^2 = 1$, a multivariate normal distribution

$$\boldsymbol{\beta}|\mathbf{z}, \mathbf{y} \sim MVN(\mathbf{W}, \mathbf{V}), \mathbf{W} = \mathbf{V}(\mathbf{H}^{-1}\boldsymbol{\beta}_0 + \mathbf{X}'\mathbf{z}), \mathbf{V} = (\mathbf{H}^{-1} + \mathbf{X}'\mathbf{X})^{-1}, \tag{3.19}$$

where a typical choice is $\boldsymbol{\beta}_0 = \mathbf{0}$. Sampling from Equation (3.19) is straightforward. A random sample from the truncated normal distribution (Equation (3.18)) can be drawn based on the exponential accept reject method algorithm of Robert (1995) or alternatively based on the fast method of Chopin (2011). Based on Equations (3.18) and (3.19) the Gibbs algorithm can be applied for the parameters $\mathbf{z}$ and $\boldsymbol{\beta}$.

The graphical representation of this model is given in Figure 3.5 via a directed acyclic graph (Madigan et al., 1995) which is a graph that shows

the conditional dependence structure between random variables. The latent variable $\mathbf{z}$ depends on coefficients $\boldsymbol{\beta}$, design matrix $\mathbf{X}$ and the variance and the response $\mathbf{y}$ depends on the latent variable $\mathbf{z}$.

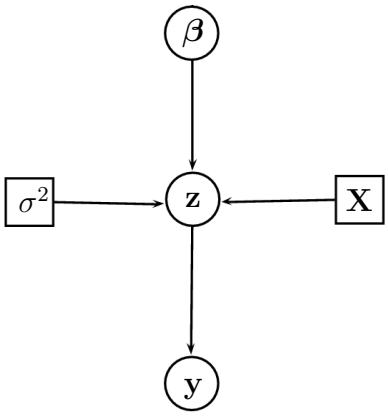


Figure 3.5: Graphical model for a probit model based on Albert and Chib (1993): circles denote random variables, squares constants and arrows indicate a dependency between variables or constants. Here $\sigma^2 = 1$.

There is strong correlation between $\boldsymbol{\beta}$ and $\mathbf{z}$ following from the definition of the latent variable (Equation (3.2)). Alternately updating these variables is likely to cause mixing problems in the Markov chain. The second approach (Holmes and Held, 2006) reduces the correlation and improves this mixing problem by updating $\boldsymbol{\beta}$ and $\mathbf{z}$ jointly using the factorization

$$p(\boldsymbol{\beta}, \mathbf{z}|\mathbf{y}) = p(\mathbf{z}|\mathbf{y})p(\boldsymbol{\beta}|\mathbf{z}),$$

where the first term on the right hand side follows a multivariate truncated normal distribution, the Z_i being independent given $\boldsymbol{\beta}$ and $\mathbf{y}$, but not given just $\mathbf{y}$, and the second term follows the multivariate normal distribution in Equation (3.19) with $\boldsymbol{\beta}_0 = \mathbf{0}$ and $\mathbf{W} = \mathbf{V}\mathbf{X}'\mathbf{z}$. Since it is difficult to sample directly from the truncated multivariate normal, Gibbs sampling is used in order to sample component-wise from the full conditional distribution of $p(\mathbf{z}|\mathbf{y})$ (Albert and Chib, 1993). Sampling component-wise from this conditional distribution means sampling from univariate truncated normals, where here the means and variances are updated at each step from the leave-one-out marginal predictive densities. By sampling from $p(\mathbf{z}|\mathbf{y})$, samples are less likely to be stuck far from the distribution of interest and so the mixing of the chain is improved.

The above methods can extended to perform inference for a probit model that is used for VS as we will describe in the next section.

3.4 Bayesian variable selection inference methods

In this section, BVS methods are described based on the probit model with binary responses as introduced by Albert and Chib (1993) (Section 3.1). Different MCMC schemes for sampling the unknown parameters come from different prior assumptions, different approaches to integrating out some parameters or/and joint updating some of the unknown parameters.

3.4.1 Literature review on sampling for MCMC in Bayesian variable selection

There are two main approaches to do BVS (with and without an indicator vector), and the review starts with methods that do not use the indicator vector.

No indicator vector

Since the indicator vector is not available, sparsity in the models can be introduced in different ways. For example, Bae and Mallick (2004) proposed, instead of a one-level, a two-level hierarchical Bayesian model assuming a prior that favours sparseness. In the two-level hierarchical Bayesian model, the prior distribution for $\boldsymbol{\beta}$ has mean zero and unknown variances $\mathbf{H} = \mathbf{D}$ (Equation (3.14) with $c_1 = 0$ and $c_2 = 1$). They put three different priors on the elements of $\mathbf{D}$, Inverse-Gamma, exponential and a non-informative Jeffreys prior. Different priors result in different degrees of sparseness. Similar to Albert and Chib, Equation (3.18) is used. With this form for $\mathbf{D}$, Equation (3.19) becomes

$$p(\boldsymbol{\beta}|\mathbf{z}, \mathbf{y}, \mathbf{D}) \propto MVN(\mathbf{V}\mathbf{X}^{'}\mathbf{z}, \mathbf{V}),$$

where $\mathbf{V} = (\mathbf{X}^{'}\mathbf{X} + \mathbf{D}^{-1})^{-1}$. For $p > n$, the covariance matrix can be calculated faster using the Woodbury-Sherman-Morrison matrix identity, $\mathbf{V} = \mathbf{D} - \mathbf{D}\mathbf{X}^{'}(\mathbf{X}\mathbf{D}\mathbf{X}^{'} + \mathbf{I})^{-1}\mathbf{X}\mathbf{D}$, since this reduces the dimension of the matrix to be inverted from p to n. The full conditional distribution of $\mathbf{z}$ is a truncated normal distribution (Equation (3.18)). To complete the sampling, the full conditional distribution for $\mathbf{D}$, $p(\mathbf{D}|\mathbf{z}, \mathbf{y}, \boldsymbol{\beta})$ is needed and this is proportional to Inverse-Gamma, Inverse-Gaussian and product of Gamma respectively for the three priors above. The full conditional distribution of $\mathbf{D}$ for the Inverse-

Gamma prior with parameters d_1, d_2 is

$$p(\mathbf{D}|\mathbf{z}, \mathbf{y}, \boldsymbol{\beta}) \propto \prod_{j=1}^{p} \text{Inverse-Gamma}\left(\frac{d_1+1}{2}, \frac{2}{d_2+\beta_j^2}\right),$$

for the exponential prior with parameter u is

$$p(\mathbf{D}^{-1}|\mathbf{z}, \mathbf{y}, \boldsymbol{\beta}) \propto \prod_{j=1}^{p} \text{Inverse-Gaussian}\left(\frac{\sqrt{u}}{\beta_j}, u\right),$$

and for the Jeffreys prior is

$$p(\mathbf{D}^{-1}|\mathbf{z}, \mathbf{y}, \boldsymbol{\beta}) \propto \prod_{j=1}^{p} \text{Gamma}\left(\frac{1}{2}, \frac{2}{\beta_j^2}\right).$$

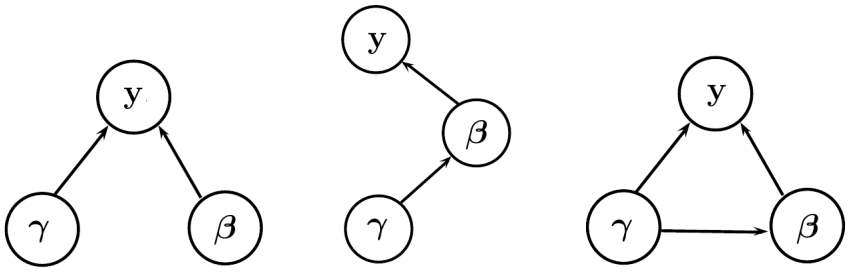
The selection is then based on the posterior mean of the elements of $\mathbf{D}$, which correspond to the variances of $\boldsymbol{\beta}$. The variables with significantly large variances are selected (i.e. we eliminate variables with small variances). However, there is not a clear way to select how many variables are important and should be included in the model. For example, coefficients that are bigger in absolute value than a selected threshold can be considered as selected ones, but the choice of threshold is sensitive. If the threshold is very large, then the model may consist of few variables or in the extreme case be the null model (the model with no variables in). On the other had, if the threshold is very low, then the model will include many variables or in the extreme case the model could be full (include all the variables in the model). There is more literature about exploiting the sparsity, however in this study we will focus on the BVS using an indicator vector.

Indicator vector

Instead of building sparse models via putting a suitable prior on the variance of coefficients and then have the difficulty of specifying the threshold for important variables, a more direct approach is to use an indicator vector. The main advantage of using an indicator vector is that we can calculate the most frequently visited indicator vector in the MCMC iterations, which contains the combination of important variables. Now methods that use an indicator vector in order to perform BVS are studied. The methods are studied in the linear context and in the probit context if the method has already been extended. This is because there is not much literature on BVS using the probit model, and in addition the probit model is an extension of

the linear model.

The simple version of BVS for linear models using an indicator vector has three variables, $\boldsymbol{\gamma}$, $\mathbf{y}$, and $\boldsymbol{\beta}$. Existing methods can be divided into three main cases according to the (in)dependence assumptions between the variables: (i) $\boldsymbol{\beta}$ is independent of $\boldsymbol{\gamma}$ (Kuo and Mallick, 1998), (ii) $\mathbf{y}$ and $\boldsymbol{\gamma}$ are independent given $\boldsymbol{\beta}$ (George and McCulloch, 1993), and (iii) $\mathbf{y}$ depends on $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ (Dellaportas et al., 2000). In order to visualize the relationships between the variables of the model and the independenence structure so as to do inference, the graphical representation of the three methods is shown (Figure 3.6). In the first subfigure, $\boldsymbol{\beta}$ is independent of $\boldsymbol{\gamma}$ and for this reason there is no arrow between them. However, $\mathbf{y}$ depends on both $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ and the dependence is shown using arrows. In the second subgfigure, since $\mathbf{y}$ and $\boldsymbol{\gamma}$ are independent given $\boldsymbol{\beta}$, there is no arrow directly from $\boldsymbol{\gamma}$ to $\mathbf{y}$. The third subfigure has all the arrows since $\mathbf{y}$ depends on both $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ and also $\boldsymbol{\beta}$ depends on $\boldsymbol{\gamma}$.



(a) Kuo and Mallick (b) George and McCulloch (c) Dellaportas et al.

Figure 3.6: Graphical representations for three different methods.

Firstly, Kuo and Mallick (1998) used the indicator vector in the model where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are independent a priori. In this case, the full conditional posterior distributions are given by

$$\begin{aligned} p(\beta_j|\mathbf{y},\boldsymbol{\gamma},\boldsymbol{\beta}_{\setminus j}) &\propto \begin{cases} p(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\gamma})p(\beta_j|\boldsymbol{\beta}_{\setminus j}), \text{ if } \gamma_j = 1, \\ p(\beta_j|\boldsymbol{\beta}_{\setminus j}), \text{ if } \gamma_j = 0, \end{cases} \\ p(\gamma_j = 1|\mathbf{y},\boldsymbol{\gamma}_{\setminus j},\boldsymbol{\beta}) &= p(\mathbf{y}|\boldsymbol{\beta},\gamma_j = 1,\boldsymbol{\gamma}_{\setminus j})p(\gamma_j = 1,\boldsymbol{\gamma}_{\setminus j}), \end{aligned} \tag{3.20}$$

where $\boldsymbol{\gamma}_{\setminus j} = (\gamma_1,\ldots,\gamma_{j-1},\gamma_{j+1},\ldots,\gamma_p)'$ and $\boldsymbol{\beta}_{\setminus j} = (\beta_1,\ldots,\beta_{j-1},\beta_{j+1},\ldots,\beta_p)'$ respectively.

Secondly, in the case where $\mathbf{y}$ and $\boldsymbol{\gamma}$ are independent given $\boldsymbol{\beta}$, the method is known as stochastic search VS (George and McCulloch, 1993) for linear models. One example for the prior of $\boldsymbol{\beta}$ is given by Equation (3.9). In the

case of stochastic search VS for a linear model, the full conditional posterior distributions of β_j and γ_j are given by

$$\begin{aligned} p(\beta_j|\mathbf{y},\boldsymbol{\gamma},\boldsymbol{\beta}_{\setminus j}) &\propto p(\mathbf{y}|\boldsymbol{\beta})p(\beta_j|\gamma_j), \\ p(\gamma_j=1|\mathbf{y},\boldsymbol{\gamma}_{\setminus j},\boldsymbol{\beta}) &= p(\boldsymbol{\beta}|\gamma_j=1,\boldsymbol{\gamma}_{\setminus j})p(\gamma_j=1,\boldsymbol{\gamma}_{\setminus j}). \end{aligned} \tag{3.21}$$

The stochastic search VS approach has been extended to the multivariate linear model (Brown et al., 1998b) and to the generalized linear model (George and McCulloch, 1997). The stochastic search VS approach has been extended to the probit model by Ai-Jun and Xin-Yuan (2010).

Finally, Dellaportas et al. (2000) consider Gibbs VS via a partition of $(\boldsymbol{\beta}_{\boldsymbol{\gamma}},\boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}})$ corresponding to the components of $\boldsymbol{\beta}$ that are included and not included in the model respectively. Different priors are assigned to different partitions: the prior $\boldsymbol{\beta}|\boldsymbol{\gamma}$ is partitioned into the model prior $\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}$ and the pseudoprior $\boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}}|\boldsymbol{\beta}_{\boldsymbol{\gamma}},\boldsymbol{\gamma}$. In this case, the full conditional posterior distributions for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, $\boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}}$ and $\boldsymbol{\gamma}$ are respectively given by

$$\begin{aligned} p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\mathbf{y},\boldsymbol{\gamma},\boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}}) &\propto p(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\gamma})p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma})p(\boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}}|\boldsymbol{\beta}_{\boldsymbol{\gamma}},\boldsymbol{\gamma}), \\ p(\boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}}|\mathbf{y},\boldsymbol{\gamma},\boldsymbol{\beta}_{\boldsymbol{\gamma}}) &\propto p(\boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}}|\boldsymbol{\beta}_{\boldsymbol{\gamma}},\boldsymbol{\gamma}), \\ p(\gamma_j=1|\mathbf{y},\boldsymbol{\gamma}_{\setminus j},\boldsymbol{\beta}) &= p(\mathbf{y}|\boldsymbol{\beta},\gamma_j=1,\boldsymbol{\gamma}_{\setminus j})p(\boldsymbol{\beta}|\gamma_j=1,\boldsymbol{\gamma}_{\setminus j})p(\gamma_j=1,\boldsymbol{\gamma}_{\setminus j}). \end{aligned} \tag{3.22}$$

Equation (3.22) contains both likelihood and prior. There are two special cases of it: Equation (3.21), where the first term of the general case is omitted since $\mathbf{y}$ is independent of $\boldsymbol{\gamma}$ given $\boldsymbol{\beta}$, and Equation (3.20), where the second term of the general case is omitted since $\boldsymbol{\gamma}$ is independent of $\boldsymbol{\beta}$ given $\mathbf{y}$.

The advantage of the approach in Kuo and Mallick (1998) is that it is simple, as it only requires one to specify the prior for the coefficients. However, there is no scope for improving the efficiency of sampling because of the independence between the coefficients and indicator vector. Dellaportas et al. (2000) and George and McCulloch (1993) methods can use similar priors for $\boldsymbol{\beta}$, for example a mixture of two normal distributions. In the latter method the pseudopriors (for β_j with $\gamma_j=0$) are kept close to zero by defining their mean to be zero and the variance to be small and all the prior parameters have impact on the posterior. However, the pseudoprior in Dellaportas et al. (2000) does not affect the posterior distribution of gamma, and the latter may not be distributed around zero for $\gamma_j=0$. For instance, the pseudoprior may be chosen in a way that helps to increase the efficiency of the sampling (e.g. via proposal densities estimated using a pilot run of the model with all the variables in). In the last case the specification of pseudopriors is difficult.

In summary, the three BVS methods for linear models have been studied for generalized linear models by simply doing variable selection for the linear predictor (Dellaportas et al., 2000). Here we will focus on BVS for a probit model with latent variables where the methods that we will study below are based on the stochastic search VS idea.

In the probit case, instead of modelling directly the discrete response vector $\mathbf{y}$, the continuous latent variables $\mathbf{z}$ are modelled and there is an one to one relationship between responses and latent variables. Lee et al. (2003) consider a model with three unknown parameters: $\mathbf{z}$, $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and $\boldsymbol{\gamma}$. Instead of drawing from the full conditionals, they draw $\boldsymbol{\gamma}$ from the marginal distribution after integrating out $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$. This can reduce the strong posterior correlation between $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and $\boldsymbol{\gamma}$ and improve the mixing. The steps of the Gibbs sampling, under the prior assumption of Zellner's g-prior for $\boldsymbol{\beta}$ with covariance matrix $\mathbf{H}_{\boldsymbol{\gamma}} = c(\mathbf{X}'_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}})^{-1}$ (special case of Equation (3.11)) are the following

$$\begin{aligned} p(\boldsymbol{\gamma}|\mathbf{z}) &\propto \exp\left[-1/2\left(\mathbf{z}'\mathbf{z} - \frac{c}{1+c}\mathbf{z}'\mathbf{X}_{\boldsymbol{\gamma}}(\mathbf{X}'_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}})^{-1}\mathbf{X}'_{\boldsymbol{\gamma}}\mathbf{z}\right)\right] p(\boldsymbol{\gamma}), \\ \boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}, \mathbf{z} &\sim MVN(\mathbf{V}_{\boldsymbol{\gamma}}\mathbf{X}'_{\boldsymbol{\gamma}}\mathbf{z}, \mathbf{V}_{\boldsymbol{\gamma}}), \mathbf{V}_{\boldsymbol{\gamma}} = \frac{c}{1+c}(\mathbf{X}'_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}})^{-1}, \\ z_i|\mathbf{X}_{i,\boldsymbol{\gamma}}, y_i, \boldsymbol{\beta}_{\boldsymbol{\gamma}} &\propto \text{truncated normal (see Equation (3.18), mean: } \mathbf{X}_{i,\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}). \end{aligned} \tag{3.23}$$

Ai-Jun and Xin-Yuan (2010) consider a model with four unknown parameters, α, $\mathbf{z}$, $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and $\boldsymbol{\gamma}$. From the joint posterior they integrate out α and $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ so as to avoid the convergence problem in MCMC due to $\mathbf{X}_{\boldsymbol{\gamma}}$ not being full column rank which leads to $\frac{c}{1+c}(\mathbf{X}_{\boldsymbol{\gamma}}{}'\mathbf{X}_{\boldsymbol{\gamma}})^{+}$ not being positive definite when using the generalized g-prior. This integration not only can reduce the strong posterior correlation between $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and $\boldsymbol{\gamma}$ but also between $\mathbf{z}$ and $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$. So, integrating out can speed up the computations and the Gibbs sampling is based on the following two steps

$$\begin{aligned} p(\boldsymbol{\gamma}|\mathbf{z}, \mathbf{X}, \mathbf{y}) &\propto |\mathbf{H}_{\boldsymbol{\gamma}}|^{-1/2} \exp\left[-1/2(\mathbf{z}'\mathbf{H}_{\boldsymbol{\gamma}}^{-1}\mathbf{z})\right] p(\boldsymbol{\gamma}), \\ \mathbf{H}_{\boldsymbol{\gamma}} &= \mathbf{I}_n + h\mathbf{1}_n\mathbf{1}'_n + c\mathbf{X}_{\boldsymbol{\gamma}}(\mathbf{X}'_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}})^{+}\mathbf{X}'_{\boldsymbol{\gamma}}, \\ \mathbf{z}|\boldsymbol{\gamma}, \mathbf{X}, \mathbf{y} &\propto MVN(\mathbf{0}, \mathbf{H}_{\boldsymbol{\gamma}}) \prod_{i=1}^{n} \mathbb{1}(z_i \in A_i), \end{aligned} \tag{3.24}$$

where in Equation (3.24) the last expression is the multivariate truncated normal, $\mathbb{1}(.)$ is the indicator function of the set A_i with

$$A_i = \begin{cases} \{z_i : z_i > 0\}, & \text{if } y_i = 1, \\ \{z_i : z_i \leq 0\}, & \text{if } y_i = 0. \end{cases}$$

Since it is difficult to sample directly from the multivariate truncated normal the univariate case of it (Equation (3.18)) is used via Gibbs sampling (Kotecha and Djuric, 1999).

Russu et al. (2012) also consider a model with four unknown parameters, α, $\mathbf{z}$, $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and $\boldsymbol{\gamma}$. Similar to the previous case, they integrate out α and $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, but in this case only from the marginal likelihood of $\mathbf{z}$, and use the special case prior of Equation (3.11) for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ with $\mathbf{H}_{\boldsymbol{\gamma}} = c\mathbf{I}_{p_{\boldsymbol{\gamma}}}$. In this case, the conditional distributions for Gibbs sampling are

$$\begin{aligned} p(\boldsymbol{\gamma}|\mathbf{z}) &\propto \exp\left[-1/2\left(\mathbf{z}'\mathbf{z} - \mathbf{z}'\mathbf{X}_{\boldsymbol{\gamma}}(c^{-1}\mathbf{I}_{p_{\boldsymbol{\gamma}}} + \mathbf{X}_{\boldsymbol{\gamma}}'\mathbf{X}_{\boldsymbol{\gamma}})^{-1}\mathbf{X}_{\boldsymbol{\gamma}}'\mathbf{z}\right)\right] p(\boldsymbol{\gamma}), \\ \boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}, \mathbf{z} &\sim MVN(\mathbf{V}_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}}'\mathbf{z}, \mathbf{V}_{\boldsymbol{\gamma}}), \mathbf{V}_{\boldsymbol{\gamma}} = (c^{-1}\mathbf{I}_{p_{\boldsymbol{\gamma}}} + \mathbf{X}_{\boldsymbol{\gamma}}'\mathbf{X}_{\boldsymbol{\gamma}})^{-1}, \\ z_i|\mathbf{X}_{i,\boldsymbol{\gamma}}, y_i, \boldsymbol{\beta}_{\boldsymbol{\gamma}} &\propto \text{truncated normal (see Equation (3.18), mean: } \mathbf{X}_{i,\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}). \end{aligned} \tag{3.25}$$

Russu et al. (2012) use the fast scan Metropolis-Hastings (Richardson et al., 2010) to sample from $p(\boldsymbol{\gamma}|\mathbf{z})$.

Finally, the Holmes and Held (2006) idea may be extended in the VS context using joint updates, where the full conditional distribution can be expressed as $p(\boldsymbol{\gamma}|\mathbf{z})p(\alpha, \boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}, \mathbf{z})$.

In summary, Ai-Jun and Xin-Yuan (2010), the MCMC approach after integrating out the intercept and the coefficients from the joint posterior is computationally more stable and efficient than Lee et al. (2003). In addition, the Russu et al. (2012) method seems more efficient than both.

3.4.2 Bayesian variable selection via alternative priors on the coefficients

Similar to Section 3.1, we use a probit model in VS setup, where the model is described by Equation (3.5) with one latent variable (Equation (3.6)).

Following the idea of George and McCulloch (1993) for linear models, a spike and slab prior is proposed, with the multivariate normal prior of Equation (3.11) with

$$\mathbf{H}_{\boldsymbol{\gamma}} = \mathbf{D}_{\boldsymbol{\gamma}}\mathbf{R}\mathbf{D}_{\boldsymbol{\gamma}},$$

where $\mathbf{R}$ is a prior correlation matrix, and $\mathbf{D}_{\boldsymbol{\gamma}} \equiv \text{diag}[\tau_1, \ldots, \tau_{p_{\boldsymbol{\gamma}}}]$. The aforementioned prior (referred to us as DgRDg) is flexible in the sense that $\mathbf{D}_{\boldsymbol{\gamma}}\mathbf{R}\mathbf{D}_{\boldsymbol{\gamma}}$ is always symmetric and positive definite, compared to the prior $\mathbf{H}_{\boldsymbol{\gamma}} = c(\mathbf{X}_{\boldsymbol{\gamma}}'\mathbf{X}_{\boldsymbol{\gamma}})^{-1}$. In this case, selecting a single crucial parameter c (Lamnisos et al., 2012) is avoided and we can select different values for the τ_i or for blocks of them, based on prior knowledge. Under the consideration

that the inclusion of $\mathbf{X}_j$ is independent of the inclusion of $\mathbf{X}_{j'}$ for all $j \neq j'$, the prior assumption for $\boldsymbol{\gamma}$ is Bernoulli (Equation (3.15)). This model has three unknown parameters $\mathbf{z}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and $\boldsymbol{\gamma}$. Figure 3.7 represents the relationship between the unknown parameters and observations using a graphical model.

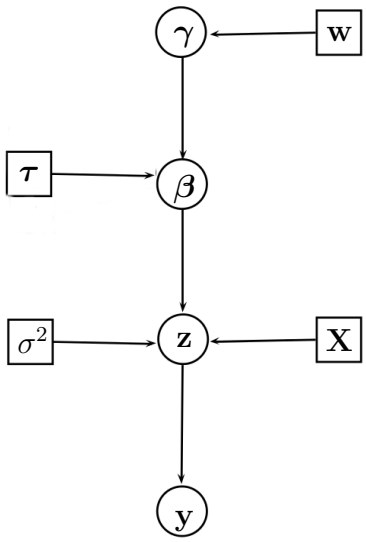


Figure 3.7: Graphical model of DgRDg.

We compute the full conditional distributions to implement Gibbs sampling. The following steps are a variation of the calculations of Lee et al. (2003). In the first step, we integrate out $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, since $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is conditionally independent of $\mathbf{y}$ given $\mathbf{z}$, and the sampling from $p(\mathbf{z}|\boldsymbol{\gamma})$ is then from the following distribution

$$p(\mathbf{z}|\boldsymbol{\gamma}) \propto MVN(\mathbf{0}, (\mathbf{I}_n - \mathbf{X}_{\boldsymbol{\gamma}}\mathbf{V}_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}}^{'})^{-1}),$$

where $\mathbf{V}_{\boldsymbol{\gamma}} = (\mathbf{D}_{\boldsymbol{\gamma}}^{-1}\mathbf{R}^{-1}\mathbf{D}_{\boldsymbol{\gamma}}^{-1} + \mathbf{X}_{\boldsymbol{\gamma}}'\mathbf{X}_{\boldsymbol{\gamma}})^{-1}$. To apply the data augmentation method (a method to construct iterative optimization via latent variables for computational convenience), it is necessary to be able to sample not only from $p(\mathbf{z}|\boldsymbol{\gamma})$ but also from $p(\boldsymbol{\gamma}|\mathbf{z})$. In this case, it is straightforward to sample from the last distribution, namely

$$p(\boldsymbol{\gamma}|\mathbf{z}) \propto MVN(\mathbf{0}, (\mathbf{I}_n - \mathbf{X}_{\boldsymbol{\gamma}}\mathbf{V}_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}}')^{-1}) \prod_{j=1}^{p} w_j^{\gamma_j}(1-w_j)^{1-\gamma_j}.$$

To speed up the stochastic search VS process, we can sample component-wise from $p(\gamma_j|\mathbf{z}, \boldsymbol{\gamma}_{\setminus j})$ (Ai-Jun and Xin-Yuan, 2010).

In the next step of the Gibbs sampler the full conditional distribution of the nonzero coefficients is calculated as

$$p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\mathbf{z}, \boldsymbol{\gamma}) \propto MVN(\mathbf{V}_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}}'\mathbf{z}, \mathbf{V}_{\boldsymbol{\gamma}}). \tag{3.26}$$

Finally, the posterior distribution of z_i given $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and y_i has a truncated normal distribution (see Equation (3.18), mean: $\mathbf{X}_{i,\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}$). So, the Gibbs algorithm in this case consists of three steps.

3.5 Sampling from the posterior of the indicator vector

This section starts by summarizing some existing VS methods that use Gibbs sampling for a probit model and have already been studied in the previous section. However, in all the cases it is not easy to sample from the conditional distribution of the indicator vector. For this reason, the second subsection describes different ways to specify an appropriate proposal for $\boldsymbol{\gamma}$ for VS.

3.5.1 Gibbs variable selection

The Gibbs sampler may be used to sample from the joint distribution if the full conditional distribution for each parameter is known (e.g. see Equations (3.23), (3.24)). Table 3.1 gives a summary of BVS in a probit model using Gibbs sampling.

Table 3.1: Summary of the BVS methods in a probit model using Gibbs. Notation '-' means that the corresponding method does not use this step.

	Parameters						
Authors	α	$\boldsymbol{\beta}$ ($\boldsymbol{\beta}_{\boldsymbol{\gamma}}$)	$\boldsymbol{\gamma}$	$\mathbf{z}$	$\mathbf{D}$	Integrate Out	Jointly
Ai-Jun and Xin-Yuan	α	$\boldsymbol{\beta}_{\boldsymbol{\gamma}}$	$\boldsymbol{\gamma}$	$\mathbf{z}$	-	α, $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ from $\boldsymbol{\gamma}, \mathbf{z}\|\mathbf{X}, \mathbf{y}$	-
Russu et al.	α	$\boldsymbol{\beta}_{\boldsymbol{\gamma}}$	$\boldsymbol{\gamma}$	$\mathbf{z}$	-	α, $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ from $\boldsymbol{\gamma}\|\mathbf{z}$	-
Lee et al.	-	$\boldsymbol{\beta}_{\boldsymbol{\gamma}}$	$\boldsymbol{\gamma}$	$\mathbf{z}$	-	$\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ from $\boldsymbol{\gamma}\|\mathbf{z}$	-
Holmes and Held	α	$\boldsymbol{\beta}_{\boldsymbol{\gamma}}$	$\boldsymbol{\gamma}$	$\mathbf{z}$	-	-	$\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \mathbf{z}$
Bae and Mallick	-	$\boldsymbol{\beta}$	-	$\mathbf{z}$	$\mathbf{D}$	-	-

Of the Gibbs steps, sampling from the indicator vector $\boldsymbol{\gamma}$ is the difficult task. At that step we will apply a Metropolis-Hastings algorithm as described in the next subsection.

3.5.2 Metropolis-Hastings algorithm

Within Gibbs, a Metropolis-Hastings (MH) algorithm can be used to take random samples from $p(\boldsymbol{\gamma}|\mathbf{Z})$ since there is no analytical solution for it and

within Gibbs sampling mixing is very slow. The efficient performance of the MH algorithm depends on a good choice of proposal (or pseudoprior) distribution. In the VS context, the proposal relates to $\boldsymbol{\gamma}$.

The model proposal $q(\boldsymbol{\gamma}^*|\boldsymbol{\gamma})$, where $\boldsymbol{\gamma}^*$ is the new candidate, influences the convergence of the MH algorithm (Algorithm 1). So, the selection of the model proposal is crucial. An asymmetric transition kernel can be considered for the MH algorithm:

$$q(\boldsymbol{\gamma}^*|\boldsymbol{\gamma}) = q_{d'} \text{ if } \sum_{j=1}^{p}(\gamma_j^* - \gamma_j) = d', \tag{3.27}$$

where $q_{d'}$ is the probability that candidate $\boldsymbol{\gamma}^*$ will have $d^{'}$ more variables than $\boldsymbol{\gamma}$ (George and McCulloch, 1997). If $d' < 0$, the candidate value represents a more parsimonious model than the one from the previous step.

Algorithm 1 Metropolis-Hastings (sampling from $p(\boldsymbol{\gamma}|\mathbf{z})$)

$t = 0$
$\boldsymbol{\gamma} = \boldsymbol{\gamma}^{(t)}$
while $t < N$ **do**
  $\boldsymbol{\gamma}^* \sim q(\boldsymbol{\gamma}^*|\boldsymbol{\gamma}^{(t)})$
  $\alpha(\boldsymbol{\gamma}^{(t)}, \boldsymbol{\gamma}^*) = min\left(\frac{q(\boldsymbol{\gamma}^{(t)}|\boldsymbol{\gamma}^*)}{q(\boldsymbol{\gamma}^*|\boldsymbol{\gamma}^{(t)})}\frac{p(\boldsymbol{\gamma}^*)}{p(\boldsymbol{\gamma}^{(t)})}\frac{p(\mathbf{z}^{(t)}|\boldsymbol{\gamma}^*)}{p(\mathbf{z}^{(t)}|\boldsymbol{\gamma}^{(t)})}, 1\right)$
  if $u \sim U[0,1] < \alpha$ **then**
    $\boldsymbol{\gamma}^{(t+1)} = \boldsymbol{\gamma}^\star$
  else
    $\boldsymbol{\gamma}^{(t+1)} = \boldsymbol{\gamma}^{(t)}$
  end if
  $t = t + 1$
end while

A widely applied and useful proposal comes from the special case of the MH algorithm, the Metropolis algorithm, which uses a symmetric proposal distribution. Then, the proposal ratio ($q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^*)/q(\boldsymbol{\gamma}^*|\boldsymbol{\gamma}) = 1$) can be omitted from the acceptance rate $\alpha(\boldsymbol{\gamma}, \boldsymbol{\gamma}^*)$ in Algorithm 1 and the symmetric transition kernel (special case of Equation (3.27)) has the form of

$$q(\boldsymbol{\gamma}^*|\boldsymbol{\gamma}) = q_d \text{ if } \sum_{j=1}^{p}|\gamma_j^* - \gamma_j| = d,$$

where q_d is the probability that candidate $\boldsymbol{\gamma}^*$ will have d new components. Note that for large values of d the corresponding algorithm makes big jumps to different $\boldsymbol{\gamma}$ values which requires high computational time per iteration.

The simplest symmetric transition kernel, $q_d = 1/p$ if $d = 1$, is

$$q(\boldsymbol{\gamma}^*|\boldsymbol{\gamma}) = \begin{cases} \frac{1}{p}, & \text{if } \sum_{j=1}^{p} |\gamma_j^* - \gamma_j| = 1, \\ 0, & \text{otherwise,} \end{cases}$$

and requires less computational cost compared to the algorithm that uses the symmetric proposal with $d > 1$.

A classical choice of the simplest symmetric proposal is a new candidate $\boldsymbol{\gamma}^*$ is selected by randomly choosing one of the following three transition moves with equal probability 1/3 (Fearn et al., 2002):

i. Add: Randomly choose a 0 in $\boldsymbol{\gamma}$ and change it to a 1,

ii. Delete: Randomly choose a 1 in $\boldsymbol{\gamma}$ and change it to a 0, and

iii. Swap: Choose independently and randomly a 0 and a 1 in $\boldsymbol{\gamma}$, and switch their values (in order to get $\boldsymbol{\gamma}^*$).

Alternatively, the new candidate can randomly be drawn by selecting one of the two transition moves with equal probability 1/2: add or delete a variable and swap two variables (Brown et al., 1998a). The simplest symmetric proposal corresponds to the case of local moves where the proposed value $\boldsymbol{\gamma}^*$ differs from the current value in a single component (with the add or delete moves) or in two components (with the swap move). Then the algorithm consists only of local moves and the model proposal is not efficient for VS problems with large numbers of variables. The algorithm can spend a large amount of time trying to add one variable before proposing to delete one, since the probability of adding one variable, $(p - p_{\boldsymbol{\gamma}})/p$, is very close to 1 if $p \gg p_{\boldsymbol{\gamma}}$. It may produce a low acceptance rate between models. Below we will present some methods that are more flexible in moves and may have higher acceptance rates.

On the other hand, high acceptance rates are often associated with poor mixing. To achieve a suitable acceptance rate, Lamnisos et al. (2009) introduced a new model proposal to combine local moves with more global ones by changing a block of variables simultaneously. In this case, the maximum number of variables N that can be changed from the current model $\boldsymbol{\gamma}$ at each iteration t is fixed, and $N^{(t)} \sim Bin(N - 1, \psi)$, where $N^{(t)} + 1$ variables are changed at the t-th step and ψ influences the proportion of local to global moves. Small values of ψ mean more local moves and large values mean more global moves. If $\psi = 0$, then the model proposal reduces to the local model proposal. When ψ increases, the number of variables proposed

to add or delete or swap increases on average. In order to find proposals that achieve good mixing, adaptive methods are used by Lamnisos et al. (2013), where a scale parameter ζ defined on $[0, 1]$ controls the differences between the current and the proposed model. In this case the adaptive version of MH algorithm (Algorithm 1) contains, instead of a fixed proposal, an adaptive one that changes at each step of the algorithm. It starts with $\zeta^{(t)}$, selects $\boldsymbol{\gamma}^*$ from $\boldsymbol{\gamma}^* \sim q_{\zeta^{(t)}}(\boldsymbol{\gamma}^*|\boldsymbol{\gamma}^{(t)})$, computes the acceptance rate α using $q_{\zeta^{(t)}}(\boldsymbol{\gamma}^{(t)}|\boldsymbol{\gamma}^*)/q_{\zeta^{(t)}}(\boldsymbol{\gamma}^*|\boldsymbol{\gamma}^{(t)})$ (instead of $q(\boldsymbol{\gamma}^{(t)}|\boldsymbol{\gamma}^*)/q(\boldsymbol{\gamma}^*|\boldsymbol{\gamma}^{(t)})$) and at the last step computes the $\zeta^{(t+1)} = \rho(\zeta^{(t)} + s^{(t)}(\alpha(\boldsymbol{\gamma}^{(t)}, \boldsymbol{\gamma}^*) - \bar{\tau}))$, where

$$\rho(\zeta) = \begin{cases} 0, & \text{if } \zeta < 0, \\ \zeta, & \text{if } \zeta \in [0, 1], \\ 1, & \text{if } \zeta > 1, \end{cases}$$

$s^{(t)} = \zeta_0/t$, where ζ_0 is a free parameter usually set to be 0.5, and $\bar{\tau}$ is a value chosen in the range 0.25 to 0.4. The scale parameter $\zeta^{(t+1)}$ decreases when the acceptance rate is small and increases when the acceptance rate is high. The sequence of scale parameters converges to a value that results in the target acceptance rate $\bar{\tau}$.

Zucknick and Richardson (2008) assume that there is a sparse dependence structure in $\mathbf{X}$, and then only those variables that are correlated need to be updated together, using a joint MH proposal. They proposed a dependence structure among the variables, in the case of logistic regression with a latent variable (Holmes and Held, 2006), in order to update them in block at each iteration. The blocks update idea defines neighbours for each variable based on correlation (or partial correlation). Block update improves the mixing compared to the simplest inclusion or exclusion of a variable.

In the case of a linear model with large p and small n an efficient sampling of $\boldsymbol{\gamma}$ could also be based on the parallel tempering strategy, which is used to weaken the dependence of the function from its parameters by adding a 'temperature' (Bottolo and Richardson, 2010). Different chains have different temperatures, which 'flatten' the likelihood. Evolutionary Monte Carlo extends the idea of parallel tempering and proposes efficient moves when updating the indicator vector. This method is known as evolutionary stochastic search. In this case the indicator vector is updated using local moves, based on Fast Scan MH (the accept/reject step depends also on the temperature of each chain), and global moves that allow the algorithm to escape from local models, which include four crossover operators and two exchange operators. Specifically, the algorithm either performs local moves or applies

one of the crossover operators; each choice has probability 1/2. After this is done, the algorithm picks uniformly between the two exchange operators and applies one of them (Bottolo and Richardson, 2010). An extension of the evolutionary stochastic search approach of linear models from univariate to multidimensional responses is also based on the evolutionary Monte Carlo approach and is called hierarchical evolutionary stochastic search (Richardson et al., 2010). This approach hierarchically relates sparse regressions to responses, associating each response with a small subset of the variables via a VS, and then linking the selection indicators in a hierarchical manner. The implementation of the hierarchical evolutionary stochastic search and evolutionary stochastic search is provided by Bottolo et al. (2011); Saadi et al. (2016).

3.5.3 MCMC convergence

Independent of which MCMC algorithm is used, a common way to check the convergence of the chains to the stationary distribution is to run multiple chains. We expect to see from the diagnostic plot posterior probability versus variable index (ID) that similar variables are being selected for all chains and from the diagnostic plot number of selected variables versus number of iterations that the chains have mixed well. Note that the chains would have to be run for a very long time to give reasonable samples and results.

In order to keep the variance of the MCMC estimator low, the autocorrelation between the samples is studied. A commonly used measure of the efficiency of any given sampler (measure of mixing) is the effective sample size

$$ESS = \frac{n}{1 + 2\sum_{k=1}^{n} \rho_k},$$

where ρ_k is the autocorrelation function with lag k.

If there is significant autocorrelation between the samples, then it can be reduced by systematically using every t^*-th sample ($t^* > 1$) and discarding the others, which is known as the thinning process. Instead of collecting T samples after a suitable burn-in period (without thinning), in this case we select the same number of samples after a suitable burn-in period and after thinning. So, actually we can generate a lot of samples but to save on storage (memory) T uncorrelated samples after the thinning process are saved and used to identify the best model.

3.6 Parameter estimation

Two different ways to carry out parameter estimation for VS are identified in order to find an appropriate model for accurate prediction. After the MCMC is finished, we are searching for replications of the proposed models, and we re-calculate the log-relative posterior probability of gamma for the distinct models using the sample mean of the latent variables. Then we calculate the normalized posterior probabilities of visited models and we can select the model with the maximum posterior probability. We refer to the latter as the best model ($\hat{\boldsymbol{\gamma}}$). Alternatively, variables that have largest posterior marginal probabilities can be selected via $p(\gamma_j = 1|\mathbf{X}, \mathbf{y})$ or more simply by empirical frequencies in MCMC

$$\hat{p}(\gamma_j = 1|\mathbf{X}, \mathbf{y}) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}(\gamma_j^{(t)} = 1). \tag{3.28}$$

In the last case the most frequently visited variables are discovered. Variables with the high posterior inclusion probabilities are relevant for classification.

The first method (that is based on the highest posterior probability) is direct, whereas the second one (that is based on the marginal posterior probabilities) is indirect since it requires a threshold on the top frequencies of the variables. However, if the aim of the study is to identify only important variables and not carry out predictions, then the important variables can be directly specified by Equation (3.28).

3.7 Prediction

When the MCMC algorithm is done, the Monte Carlo posterior predictive estimate of the new observation Y_{new} is given by

$$\hat{P}(y_{\text{new}} = 1|\mathbf{X}) = \frac{1}{T}\sum_{t=1}^{T} p(y_{\text{new}} = 1|\mathbf{X}, \mathbf{z}^{(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\gamma}^{(t)}, \alpha^{(t)}, \mathbf{H}^{(t)}) \tag{3.29}$$

where T is the total number of iterations after suitable burn-in period and $\mathbf{z}^{(t)}$, $\boldsymbol{\beta}^{(t)}, \boldsymbol{\gamma}^{(t)}, \alpha^{(t)}$ and $\mathbf{H}^{(t)}$ are the MCMC samples from the posterior distribution. This is a general form for the predictive distribution for all the aforementioned methods ($\boldsymbol{\beta}^{(t)}$ can be replaced by $\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{(t)}$ for some methods). In the special case where the sampling takes place without using indicator latent variables the term $\boldsymbol{\gamma}^{(t)}$ can be omitted from Equation (3.29). If the model does not contain the intercept or has intercept but it is known then the term $\alpha^{(t)}$ can

omitted. Note that $\mathbf{H}^{(t)}$ exists in the equation only in the case of the two-level hierarchical model (Bae and Mallick, 2004). So, predictions can be made with the aim to correctly classify new observations for all the aforementioned variations of models and priors.

Instead of averaging all the models of the MCMC steps to draw predictions, in practice we can calculate a single model prediction. Finding a best model either by selecting the model with the highest posterior probability among the visited models or by considering a threshold (for example median model) for the estimated marginal inclusion probabilities, the single model prediction can be obtained calculating the least squares predictions of the latent variables and then using the relationship between latent variables and responses. Alternatively, BVS allows to pick, instead of the best model, the top few best models and averaging among them. The last approach is known as Bayesian model averaging (Brown et al., 2002; Wasserman, 2000).

Chapter 4

Existing methods and variations applied to some datasets

This chapter is focused on the application of a probit model with binary outcomes. The results of BVS are given under different assumptions: a general form of prior distribution for the coefficients (Section 4.1), a DgRDg prior approach for the coefficients (Section 4.2) and a first order dependence between indicator variables (Section 4.3). The three variations are applied to a simulated dataset. Apart from the simulated results, the second method implemented is on two literature datasets. In the last section of the chapter some preliminary results are presented applying existing methods to the BE dataset that motivated this research.

4.1 Bayesian variable selection with a flexible prior for the coefficient vector

We simulate data with binary responses using latent variables and then perform VS using this data and assigning a flexible prior for the coefficient vector via Equation (3.11).

For the simulation study, we identify the number of variables (p), the number of samples (n), the indices j of the important elements of $\boldsymbol{\beta}$ denoted by β_j, the values that correspond to the important (nonzero) coefficients by v (most $v = 0$), and a diagonal covariance matrix. We simulate from $\mathbf{X} \sim MVN(\mathbf{0}, \mathbf{D})$, where $\mathbf{D}$ is a diagonal matrix, then we construct $\mathbf{z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, which is the latent variable that includes an error term ($\boldsymbol{\epsilon} \sim MVN(\mathbf{0}, \sigma_s^2\mathbf{I})$). At the next step we calculate the vector of responses by generating a random

draw from a Bernoulli distribution $\mathbf{y} \sim$Bernoulli$(\Phi(\mathbf{z}))$.

In order to implement BVS the flexible choice of prior of Equation (3.11) is used, via $\mathbf{H}_{\boldsymbol{\gamma}} = c_1(\mathbf{X}'_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}})^{+} + c_2\mathbf{D}_{\boldsymbol{\gamma}}$. We initialize the $\boldsymbol{\beta}^{(0)}$ setting four nonzero entries, $\boldsymbol{\gamma}^{(0)}$ selecting randomly four variables equal to one and we initialize $\mathbf{z}^{(0)}$ as follows

$$z_i^{(0)} = \begin{cases} 1, & \text{if } y_i = 1, \\ -1, & \text{if } y_i = 0. \end{cases} \tag{4.1}$$

The different settings of the simulated data and the results of VS using these data are given in Table 4.1, where in all cases $\mathbf{D} = \mathbf{I}$ and the error term σ_s^2 of the latent variable in the simulation is 2. Using a variety of values for the constants c_1 and c_2, all the simulations correctly identified the important variables individually and the best model.

Table 4.1: Simulation results for the flexible prior (Equation (3.11)).

n	p	$\beta_j = v$		c_1	c_2	Found	Found
		j	v			best γ_j	best model
100	15	1,4,5,10	10,12,10,12	100	0	✓	✓
100	15	1,4,5,10	10,12,10,12	50	100	✓	✓
100	15	1,4,5,10	10,12,10,12	100	50	✓	✓
100	15	1,4,5,10	10,12,10,12	0	100	✓	✓
100	15	1,4,5,10	10,12,10,12	100	0	✓	✓

4.2 Bayesian variable selection via DgRDg prior

Firstly, simulation of some data from the probit model is done in order to check if the idea described in Subsection 3.4.2 works well. This method uses DgRDg prior for the covariance matrix of nonzero coefficients $\mathbf{H}_{\boldsymbol{\gamma}} = \mathbf{D}_{\boldsymbol{\gamma}}\mathbf{R}\mathbf{D}_{\boldsymbol{\gamma}}$ (in Equation (3.11)). Afterwards, an illustration of the approach using two widely studied datasets, the Colon cancer and the Leukemia datasets is presented. These two datasets are selected because they also belong to the large p, small n problem, have binary outcomes, as when we study BE responses in pairs, and also they are commonly studied for VS. The key point is to do VS when $n < p$ using a Bayesian methods for a probit model with binary responses and the DgRDg covariance matrix on $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ prior. Once we pick the important variables, there are many different ways to make predictions. The most-natural, given that we have used BVS, would be to use the Bayes model for the predictions. However, because we will compare our results with analyses from the machine learning literature, we choose to do the prediction step

using two classifiers: support vector machine (SVM) and k-nearest neighbor (k-NN) using the selected variables. The comparison of the classification accuracy of this approach (combination of BVS for training and classifier for prediction) with the accuracy of existing VS approaches is presented. It would be interesting to compare other measures of performance, but those tend not to be available for the analyses in the literature.

Simulation studies

For the simulations of this approach the diagonal values of covariance matrix $\mathbf{D}$ are equal to one and the error term σ_s^2 of the latent variable in simulation is 1. The results are given in Table 4.2 (taking 10000 samples after 5000 burn-in period) for $\mathbf{D}_{\boldsymbol{\gamma}} = 2\mathbf{I}$ and $\mathbf{R} = \mathbf{I}$. Both best individually variables and best combination of variables (best model) work well. The number of times that each variable or combination of them appears, out of 10000 iterations, is given in parentheses. Note that the frequency of individual variables is much higher than the frequency of the models, as might be expected. In addition, the frequency of the 2nd best model is well below the frequency of the best model, which means that it is not necessary to increase the number of iterations.

Table 4.2: Simulation results for DgRDg method.

n	p	$\beta_j = v$		best γ_j	best model	2nd best model
		j	v	(frequency)	(frequency)	(frequency)
20	40	2	3	✓(9051)	✓(3904)	1676
20	40	1, 5	2,3	✓(7897,5946)	✓(3051)	1159
20	40	5, 30	3,2	✓(7522,6711)	✓(3843)	402
20	40	20, 30	4,5	✓(6667,4518)	✓(2163)	1278
27	476	20, 40	3,2	✓(3897,3818)	✓(1800)	300

Colon cancer study

The colon cancer data contains expression levels of 2000 variables for 62 different cases. Among them, 40 are tumor tissue and 22 are normal colon tissue. Similar to other research studies on this dataset, firstly, a base 10 logarithm is applied on the data and secondly each tissue is standardized to mean zero and unit variance. The data can be downloaded from `http://microarray.princeton.edu/oncology/affydata/index.html`.

In this analysis we specify the parameters of the Subsection 3.4.2 $w_j =$

$0.005, j = 1, \ldots, p$, $\mathbf{D}_{\boldsymbol{\gamma}} = 2\mathbf{I}$ and $\mathbf{R} = \mathbf{I}$. If any prior knowledge is available, the selection of the $\mathbf{R}$ matrix can be different than the identity matrix. The initialization is similar to one that described in Section 4.1. The check of convergence is done running three different chains with 5, 10 and 15 randomly selected variables (randomly selected indexes of the indicator vector) as starting vectors.

Since no test set for the Colon study is available, in order to evaluate the performance of the classification methods the LOOCV method is used. However if the number of best variables is selected to optimise LOOCV accuracy, then this accuracy will be biased. In order to avoid nesting cross validations, inside the LOOCV two random training sets (with 51 observations) and two random test sets (with 10 observations) are used to find the best number of variables. The VS approach is applied not for 2000 variables but for the top 50 variables based on t-statistics (Antoniadis et al., 2003). In the MCMC 10000 samples after a 5000 burn in period are taken to estimate the posterior variable inclusion probabilities. Finally, the external LOOCV procedure is the following: (i) omit one observation of the training set; (ii) based on the remaining $n-1$ observations, reduce the set of available variables to the top 50 variables as ranked in terms of the t-statistic; (iii) the p^* most significant variables were chosen from the 50 variables by BVS using the DgRDg approach; and (iv) these p^* variables were used to classify the left out sample and (v) repeat $n-$times the steps (i)-(iv).

We propose the DgRDg method to find the best model with binary responses and we make predictions using SVM and k-NN algorithms. The results are given in Table 4.3, which also contains accuracies for comparison with the existing methods. The DgRDg method achieves accuracy 88.71%, using only 8 variables. This percentage is close to the accuracy of the first method, which uses half of the variables (1000). In addition, the DgRDg method achieved at least the same classification accuracy as the other methods that are listed in this table, while using only 8 variables.

Leukemia study

The original Leukemia data contains 7129 variables and 72 patients and it can be downloaded from `http://www-genome.wi.mit.edu/mpr/data_set_ALL_AML.html`. The pre-processing described in Ai-Jun and Xin-Yuan (2010) is followed. After the pre-processing, thresholding, filtering and base-10 logarithmic transformation of the data, the Leukemia dataset is studied with 3571 variables. The cases are divided into two types of acute leukemias: 47

Table 4.3: Performance comparison using LOOCV for Colon cancer study.

Method	# of variables	Accuracy (%)
SVM Furey et al. (2000)	1000, 2000	90.32
LogitBoost, optimal Dettling (2004)	2000	87.10
Classification tree Dettling (2004)	200	85.48
MAVE-LD Antoniadis et al. (2003)	50	83.87
1-NN Dettling (2004)	25	85.48
LogitBoost, estimated Dettling (2004)	25	80.65
LogitBoost,100 iterations Dettling (2004)	10	85.48
AdaBoost,100 iterations Dettling (2004)	10	83.87
gsg-SVSS Ai-Jun and Xin-Yuan (2010)	14	88.71
gsg-SVSS Ai-Jun and Xin-Yuan (2010)	10	88.71
gsg-SVSS Ai-Jun and Xin-Yuan (2010)	6	87.10
DgRDg + SVM	8	88.71
DgRDg + k-NN	8	88.71

samples for acute lymphoblastic leukemia (ALL) and 25 for acute myeloid leukemia (AML). This dataset is analysed using 38 samples (27 are ALL and 11 are AML) as the training set and 34 samples (20 are ALL and 14 are AML) as the test set. The values of the hyperpriors are set the same as in the Colon study. The convergence is checked running three chains with different initial values. In this case 10000 iterations are used after a 5000 iteration burn-in period in order to find the most significant variables.

Since there is a test set available, we did not apply LOOCV. However, since there is no tuning set the classification accuracy in the training set is used in order to specify the number of best variables. More details for the accuracy of the classifiers using this method on the training set are given in Figure 4.1. The accuracy is perfect, 100%, on the training set using only the best 3 or 4 most important variables ranked by the posterior inclusion probabilities. The corresponding index numbers of these variables are: 979, 2481, 456 for both classifiers, and a 4th variable for SVM has index 3441. So these are the variables that are used in order to calculate the accuracy on the test set.

The test set is used in order to evaluate the performance of VS using two different classifiers. Table 4.4 shows that the DgRDg method achieves high accuracy, 97.06% using only 3 variables out of 3571 based on k-NN classifier. The same percentage of accuracy is reached using the top 4 variables based on SVM. Compared to other methods that are listed in Table 4.4, DgRDg method has at least the same accuracy using only 3 or 4 variables, depending on the classifier.

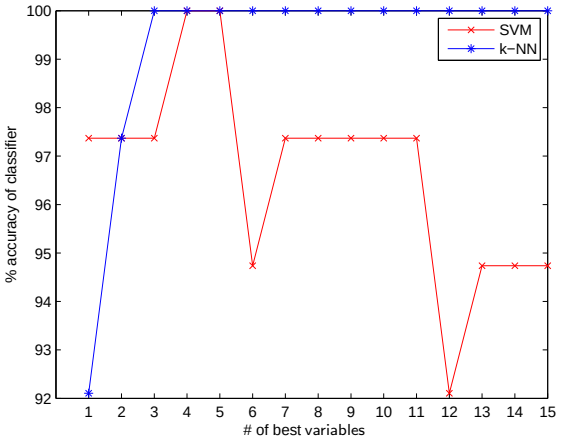


Figure 4.1: Accuracy on the training set of the Leukemia study versus the number of selected variables using DgRDg method.

Table 4.4: Performance comparison using the test set of Leukemia study.

Method	# of variables	Testing Accuracy (%)
SVM Furey et al. (2000)	25, 1000	94.12, 88.24
WVM Golub et al. (1999)	25	85.29
MAVE-LD Antoniadis et al. (2003)	50	97.06
MAVE-NPLD Antoniadis et al. (2003)	25	97.06
gsg-SVSS Ai-Jun and Xin-Yuan (2010)	14	97.06
gsg-SVSS Ai-Jun and Xin-Yuan (2010)	10	97.06
gsg-SVSS Ai-Jun and Xin-Yuan (2010)	6	97.06
DgRDg + k-NN	3	97.06
DgRDg + SVM	4	97.06

In conclusion, we consider the prior $\mathbf{H}_{\boldsymbol{\gamma}} = \mathbf{D}_{\boldsymbol{\gamma}}\mathbf{R}\mathbf{D}_{\boldsymbol{\gamma}}$ on the coefficients when the responses are binary. The BVS approach selects a small subset of variables that inform the response binary outcome. The proposed procedure is compared with other existing methods and achieves better or comparable accuracy in classification with variables using different classifiers. Finally, in the Colon cancer and Leukemia studies just 8 and 4 variables respectively give very high classification accuracies.

4.3 Bayesian variable selection with dependent indicator variables

In this section we consider that the γ_j are dependently distributed and a two-stage first-order Markov model is used to represent the dependence (Subsection 3.2.4). This idea is applied to simulated data.

For each simulation of this approach the number of samples, the number of variables and the important coefficients are varied as is noted in Tables 4.5 and 4.6. In addition, for all simulations we set the diagonal values of the covariance matrix equal to one, and the error term of the latent variable is 0.1.

For the analysis we set $c = 10$ in the g-prior of the coefficient vector (special case of Equation (3.11)) and $h = 10^4$ (in Equation (3.7)). In case 1 and 2 of the Subsection 3.2.4, we set $w = 1/40$ (for the first three simulations) and $w = 2/40$ (for the remaining simulations). Then, 20000 samples are selected after 5000 burn-in iterations. In the cases where more samples are taken, the number of samples are noted in the footnotes of those tables that contain the results of the VS.

In case 1 the probability to include the corresponding variable when the previous one is not included, can be for example 3 times higher than the probability to include the corresponding coefficient when the previous one is included ($\eta = 3$ in Equation (3.17)). Table 4.5 contains the results for the case 1 (empirical approach). Both the best individual variables and the best model are identified via this approach. In the last two simulations, the best model has much lower frequency than each γ_j, but the method still found the correct model.

Table 4.5: Simulation results for case 1 of the first-order Markov model.

n	p	$\beta_j = v$		best γ_j	best model	2nd best model
		j	v	(frequency)	(frequency)	(frequency)
20	40	1	2	✓(14267)	✓(7560)	20
20	40	10	-2	✓(13971)	✓(6382)	79
20	40	20	0.9	✓(13777)	✓(8276)	41
20	40	10, 11	2,2	✓(39565,13643) [a]	✓(310)	88
20	40	32, 33	2,3	✓(40135,14865) [b]	✓(421)	91

[a] Number of samples: 40000.
[b] Number of samples: 50000.

In case 2, instead of fixing the probability $p(\gamma_j = 1|\gamma_{j-1} = 0)$ regardless to the probability $p(\gamma_j = 1|\gamma_{j-1} = 1)$, we update the probabilities taking into account the autocorrelation. Those probabilities are updated at each step step of the Markov chain. Those results are presented in Table 4.6. The method gives the expected results.

Table 4.6: Simulation results for case 2 of the first-order Markov model.

n	p	$\beta_j = v$		best γ_j	best model	2nd best model
		j	v	(frequency)	(frequency)	(frequency)
20	40	1	2	✓(14213)	✓(6691)	26
20	40	10	-2	✓(13751)	✓(6001)	48
20	40	20	0.9	✓(13543)	✓(7598)	41
20	40	10, 11	2,2	✓(37121,16185) [a]	✓(311)	127
20	40	32, 33	2,3	✓(29393,10064) [b]	✓(401)	151

[a] Number of samples: 40000.
[b] Number of samples: 50000.

In conclusion, the idea of the first order Markov chain on $\boldsymbol{\gamma}$'s prior, which may be more realistic in real problems, gave the expected results in these simulations.

4.4 Bayesian variable selection for BE diagnosis

We have studied methods to apply BVS on binary data. After first providing some information about the collection process of the biopsy, then we focus on applying BVS methods on binary responses, healthy versus cancer samples.

Biopsy collection

Each patient had previously been diagnosed with BE prior to their assessment and had consented for additional biopsy collection under the clinical trial at UCL Hospital for the purpose of this project at the time of endoscopy. Biopsies were taken at various levels of the metaplastic segment. Four biopsies were taken from each level according to the position on the oesophagus: anterior, posterior, left and right. In the case of a visual abnormality, additional targeted biopsies were collected, referred to as multiple biopsies.

In this project two different data collection methods have been used: adjacent-paired dataset (APD) and intercepted-matched dataset (IMD). The differences of these methods are outlined in Table 4.7. Based on the first row of it, a weakness of the APD is that the two different samples could potentially be at different BE stages, which would cause a discrepancy between the IR and histological analyses. IMD collection overcomes this problem by using the same sample for both IR and histological analyses. In IMD, due to the slippery nature of the sample, the spectrum was first visualized in real time before being recorded, which allowed the position of the sample to be

optimized to gain an optimum amide II band ($\sim 1556\ cm^{-1}$, Figure 1.7). In the IMD process, the spectrum was taken, the sample was lifted, the prism and lens cleaned, and the sample replaced on the prism in the opposite orientation and another spectrum recorded. When a sufficient number of spectra had been collected, the samples were then placed in formalin and sent to histology (Foreman, 2016).

Table 4.7: Differences between adjacent-paired and intercepted-matched data collection.

	Adjacent-paired dataset	Intercepted-matched dataset
Analysis of the biopsy	Two adjacent biopsies: one biopsy was analysed by attenuated total reflection and FTIR spectroscopy and the other was clinically analysed by the UCLH histopathology lab.	Only one biopsy was taken. It was analysed by attenuated total reflection and FTIR spectroscopy first and then the same biopsy was analysed by the UCLH histopathology lab.
Collected by	Dr. Rehan Haidry, Gastrointestinal consultant at UCLH	Liberty Foreman, Institute of Structural & Molecular Biology, UCL

A Bruker Optics IFS 66/s FTIR spectrometer recorded the region of $2200 - 900\ cm^{-1}$ with 4 cm^{-1} resolution, with 3 reflections on a silicon prism with ZnSe optics. In this section the APD is analysed. The pre-processing methods that are applied are water vapor correction, water subtract correction, normalize by Amide II, and second derivatives by Savitzky Golay with 17 pt smoothing. Each stage of BE has a different number of spectra, as given in Table 4.8.

Table 4.8: Number of samples for each stage of APD.

SQ	NDBE	LGD	HGD	OAC	Total
19	18	7	17	8	69

Here we focus on the binary study of the SQ and OAC recorded biopsies of APD. Figure 4.2 gives an illustration of those records. This dataset has only 27 samples and 676 variables that correspond to 676 wavenumbers. Some MCMC methods are applied for VS. The ideas of Lamnisos et al. (2009) together with Holmes and Held (2006) are applied in the SQ versus OAC case with $c_2 = 5$, $c_1 = 0$ (special case of Equation (3.11)), $h = 10^4$ (in Equation (3.7)), $w = 0.0104$ (probability of success). In addition, we set the

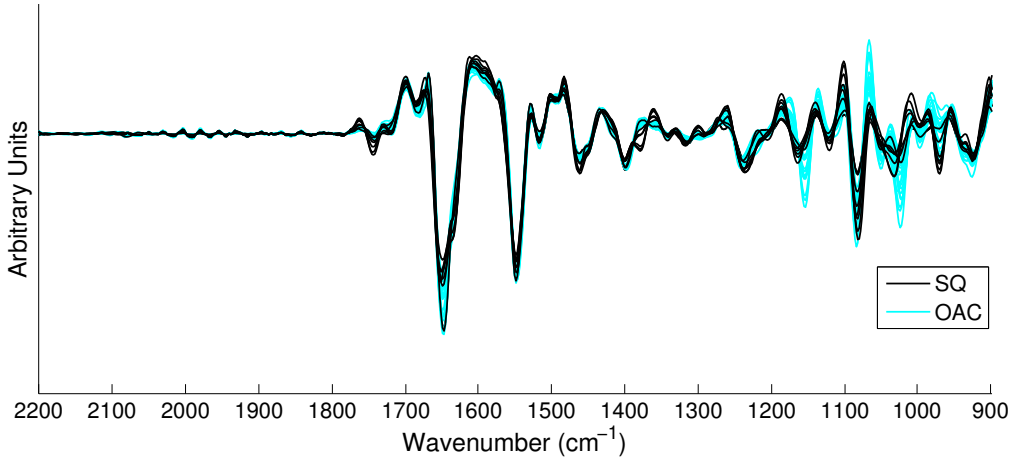


Figure 4.2: Second derivative spectra comparing all SQ biopsies with all OAC biopsies from the APD.

maximum number of proposed new variables to two. Due to the absence of a test set and to the small sample size of the training set, we apply LOOCV. The top 20 variables were the following 387, 388, 386, 389, 345, 414, 344, 343, 413, 385, 346, 412, 336, 335, 337, 347, 395, 415, 404 and 403. The adaptive approach of the last method (Lamnisos et al., 2013) was also implemented using the same values on the MCMC but with a Beta prior on variable inclusion probability instead of fixing w. The top 20 variables were 387, 388, 386, 389, 414, 345, 346, 344, 385, 413, 415, 335, 343, 336, 396, 412, 347, 334, 397 and 337. The non-adaptive and adaptive method identify the same 13 top variables but in a slightly different order.

In the adaptive and non adaptive case, the first four variables are considered to come from the same component since they belong to the same peak. Each one of them and the top 1, 2, 3 and 4 of them give a high classification accuracy (after applying LOOCV) compared to the accuracy without VS for both methods (Table 4.9). In this case, we predict the healthy or cancer stage using the selected variables and different classifiers than before, for example Naive Bayes, J48 which is a classification tree approach (Hall et al., 2009), and random forest. Variable selection improves by 4% the classification accuracy compared with not applying VS (Table 4.9). VS also increases the accuracy when we use other binary pairs of ATD.

Table 4.9: Accuracy for some classifiers using selected variables (by the adaptive approach) or using all variables (without VS).

	Naive Bayes	J48	Random forest
1st Top, 2nd Top, 3rd Top, 4th Top	96.29	96.29	96.29
Top 1-2, Top 1-3, Top 1-4	96.29	96.29	96.29
No VS	92.59	92.59	92.59

In this section we presented some preliminary results on two stages from the BE dataset. Applying VS on healthy and cancer samples and using only the selected variables we can improve the classification accuracy even though the sample size is small. More results about the real application on the proposed methodologies will be presented in Chapter 9.

Chapter 5

Variable selection methods for multi-class problems via MCMC

When studying a multi-class classification problem (multi-class responses), it is important to note whether the response is ordinal (consisting of ordered categories) or nominal (consisting of un-ordered categories). There are many methods to solve multi-class classification problems where the responses are purely nominal or purely ordinal (Section 5.1), and some of the them use variable selection in order to improve classification performance. However, the literature for BVS for multi-class responses is limited. Before extending the methodology to a mixture of ordinal and nominal categories in Chapter 5, we look here at the methods for just the nominal or just the ordinal multinomial probit model, as we will discuss in Sections 5.2 and 5.3.

5.1 Introduction

In statistics and machine learning, many methods have been proposed for classification, but they are usually relevant either to pure nominal or to pure ordinal responses. Classification methods, such as support vector machines and k-nearest neighbour have been introduced for use with nominal responses (Murphy, 2012). Some of them, for example support vector machines (Chu and Keerthi, 2007), have also been applied to ordinal responses. In classical statistics, the most common models are multinomial probit model for pure nominal/ordinal responses and multinomial logit model for pure nominal/ordinal responses (McCullagh, 1984). The Bayesian approach for the multinomial probit model involves latent variables Albert and Chib (1993).

In the context of a large number of regression predictors, it is also important to introduce sparsity in the model in order to improve prediction accuracy. For example, classification trees and random forests have been introduced for variable selection when the categorical responses are pure nominal or pure ordinal ((Piccarreta, 2008) and (Janitza et al., 2014) respectively). Dimensionality reduction techniques, for example LDA, create a few new variables using linear transformation of the original variables, for either pure nominal or pure ordinal responses (Witten and Tibshirani, 2011). A non-linear version of LDA called kernel discriminant analysis has been proposed for pure nominal (Scholkopft and Mullert, 1999) and pure ordinal responses (Sun et al., 2010). Details about frequentist approaches for variable selection have been presented in Subsections 2.2.2 and 2.3.1. A famous approach for variable selection is the LASSO, and the LASSO representation for ordinal responses is described by Park and Hastie (2007).

In Bayesian statistics, penalization takes place via prior distributions. As in the binary case, priors that offer penalisation in variable selection are usually a mixture of two distributions, known as a spike and slab prior (George and McCulloch, 1993). A classical choice is a spike at zero and a normal slab. Alternatively, the Laplacian (or double-exponential) prior (Zou and Hastie, 2005) can be used for penalisation, which corresponds to the Bayesian formulation of the LASSO. We focus in BVS on multi-class classification problems. In the Bayesian framework, there are extensions of the binary studies, Ai-Jun and Xin-Yuan (2010) and Lee et al. (2003), to the multi-class studies, Aijun et al. (2013) and Sha et al. (2004), respectively. References in multi-class BVS have studied either just the (nominal) multinomial probit model (Sha et al., 2004; Zhou et al., 2006) or just the ordinal multinomial probit model (Kwon et al., 2007). The last two methods will be thoroughly studied in the next two sections.

5.2 Bayesian variable selection in the probit model with nominal responses

In this section we will present a probit model with multi-class nominal responses using latent variables, which is an extension of the binary case, in order to use this model for variable selection. We will discuss the main similarities and differences between two methods (i.e. binary and multi-class nominal responses).

5.2.1 Model

Let us assume that the vector of categorical responses $\mathbf{y}$ can take M values $(m = 0, 1, \ldots, M-1)$, which is a general case of binary $(m = 0, 1)$. Assume that the zero response is the 'baseline' category of the multinomial probit model. Also assume $\mathbf{Z}$ is an $n \times (M-1)$ matrix of latent variables distributed as multivariate normal with common variance $\boldsymbol{\Sigma}$ across different possible responses

$$\mathbf{Z}_{i,:} = \boldsymbol{\alpha}^{'} + \mathbf{X}_{i,:}\mathbf{B} + \mathbf{E}_{i,:}, \mathbf{E}_{i,:} \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}), i = 1, 2, \ldots, n, \tag{5.1}$$

where $\mathbf{Z}_{i,:} = (Z_{i,1}, Z_{i,2}, \ldots, Z_{i,M-1})$ is the row vector of the $\mathbf{Z}$ matrix that refers to the i-th sample, $\mathbf{Z}_{i,:}$ are independent for $i = 1, \ldots, n$, $\boldsymbol{\alpha}$ is the $(M-1) \times 1$ vector of the intercept, $\mathbf{X}_{i,:}$ is the i-th row of the $n \times p$ matrix $\mathbf{X}$ and $\mathbf{B}$ is a $p \times (M-1)$ matrix of regression coefficients. Brown et al. (1998b) generalized the methods of George and McCulloch (1997) from univariate to multivariate regression with $(M-1)$-variate responses (one of these is the 'baseline' category). Following this we replace the $\boldsymbol{\beta}$ vector by $\mathbf{B}$ matrix and here we work with matrix distributions.

The equivalent matrix normal (MN) distribution (Appendix A) to Equation (5.1) can be expressed as

$$\mathbf{Z} - \mathbf{1}_n\boldsymbol{\alpha}^{'} - \mathbf{XB} \sim MN(\mathbf{I}_n, \boldsymbol{\Sigma}), \tag{5.2}$$

where $\mathbf{1}_n$ is a n dimensional column vector of ones. In analogy with the univariate probit model, the multivariate probit model the unknown parameters $\mathbf{B}$, $\boldsymbol{\Sigma}$ are not identifiable from the model. Since the assignments to groups depend only on the relative sizes of the elements of $\mathbf{Z}$, rescaling $\mathbf{Z}$ would give the same fit to the data. According to the original paper (Sha et al., 2004), one way to handle identifiability in a probit model is to fix $\boldsymbol{\Sigma}$ (even though in practice the authors assign a prior to it). This process of fixing $\boldsymbol{\Sigma}$, usually as an identity matrix, adds a strong restriction. It is a generalization of the restriction that was added to the univariate probit model (with binary responses) by setting the variance equal to one.

There are less drastic solutions to address the identifiability issue than setting $\boldsymbol{\Sigma} = \mathbf{I}$, since this imposes more constrains than are strictly necessary. Some of these methods are: setting the first diagonal element of the covariance matrix $\sigma_{11} = 1$ (McCulloch and Rossi, 1994) or drawing samples from an inverse Wishart conditional on $\sigma_{11} = 1$ (Linardakis and Dellaportas, 2003), or fixing the trace of the covariance matrix (Jiao and van Dyk, 2015).

Unfortunately these constraints make the MCMC more difficult. In the case of $\sigma_{11} = 1$ Imai and van Dyk (2005) proposed an efficient MCMC algorithm which expands the constrained covariance matrix ($\sigma_{11} = 1$) into an unconstrained covariance matrix (the strategy is known as parameter expansion for data augmentation).

Another way to handle non-identifiability is to restrict the covariance matrix to a correlation matrix, since this is not only positive semi-define but also has diagonal elements equal to one, while off-diagonal elements belong to the interval $[-1, 1]$ (Chib and Greenberg, 1998). However, this method is computationally expensive since there is no conjugate prior for a correlation matrix. Talhouk et al. (2012) adopted this restriction and used the parameter expansion framework to build an efficient algorithm for inference in the multivariate probit model case. Their idea is to expand the correlation matrix into a covariance matrix, update this covariance matrix using standard simulation steps and project it back to a correlation matrix. In addition, they extend the prior to accommodate sparse structure in the correlation matrix.

In the original paper (Sha et al., 2004) the authors decided to assign a proper prior to $\mathbf{\Sigma}$. Using a proper prior resolves the identifiability problem but replaces it with the problem of picking the appropriate proper prior. Too strong a prior and we may introduce bias; too weak and the near non-identifiability can cause problems for the MCMC. Even with a proper prior there will be some aspects of the model where the prior essentially carries through to determine the posterior so that interpreting the coefficients and extrapolating could be problematic.

To continue the study of the model, we denote by $Z_i^\star = \max_{1 \le r \le M-1} \{Z_{i,r}\}$. The relation between responses and latent variables is given by

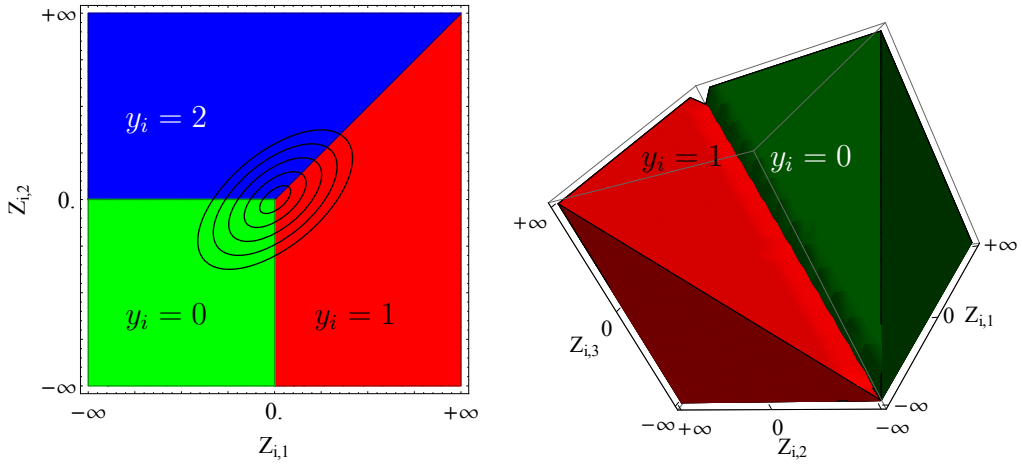
$$y_i = \begin{cases} 0, & \text{if } Z_i^\star \le 0 \\ r, & \text{if } Z_i^\star > 0 \text{ and } Z_{i,r} = Z_i^\star. \end{cases} \tag{5.3}$$

Denote the probability that the i-th object belongs to the m-th class ($m = 0, 1, \ldots, M-1$) by p_{im} . Then, the probabilities of the multinomial probit model with nominal responses based on latent variables can be expressed as

$$\begin{aligned} p_{i0} &= P(y_i = 0) = P(Z_i^\star \le 0), \\ p_{ir} &= P(y_i = r) = P(Z_i^\star > 0, Z_{i,r} = Z_i^\star), \text{ for } r = 1, \ldots, M-1. \end{aligned}$$

The simplest case of Equation (5.3) is the binary case ($M = 2$) where just one latent variable (Equation (3.2)) is used in the model and the relationship with responses is simply given by Equation (3.3). The graphical represen-

tation of Equation (5.3) for $M = 3$ ($y_i = \{0, 1, 2\}$) is given in Figure 5.1a. In this case, there are two latent variables, $Z_{i,1}$ and $Z_{i,2}$ (for each i). If the maximum of the two latent variables is nonpositive or equivalently $Z_{i,1} \leq 0$ and $Z_{i,2} \leq 0$ (zero is the 'baseline' category), then the response is $y_i = 0$ (green region). If $Z_{i,1} > Z_{i,2}$ and $Z_{i,1} > 0$, we focus on the red region, where $y_i = 1$. Similarly, if $Z_{i,2} > Z_{i,1}$ and $Z_{i,2} > 0$, then the response is $y_i = 2$ (blue region). The axes are from minus infinity to plus infinity, since both latent variables can take any real value. The contours indicate probability densities for $Z_{i,:}$. The contours in the figure are centered at zero but in general will be centered at $\alpha_r + \mathbf{X}_{i,:}\mathbf{B}_{:,r}$, where here $r = 1, 2$.



(a) Region using 2 (= $M - 1$) latent variables.

(b) Volumes using 3 (= M) latent variables.

Figure 5.1: Graphical representation of Equations (5.3) and (5.4) respectively, for $M = 3$ and fixed i.

In the case of nominal responses, an alternative representation for the regression is to use M latent variables, $\mathbf{Z}_i = (Z_{i,1}, Z_{i,2}, \ldots, Z_{i,M})$, which correspond to M regression equations (there is no 'baseline'). In this case the relation with the responses is simple and is the following

$$y_i = r - 1, \quad \text{if } Z_{i,r} = Z_i^{\star\star}, \tag{5.4}$$

where $Z_i^{\star\star} = \max_{1 \leq r \leq M} \{Z_{i,r}\}$. The graphical representation of Equation (5.4) for $M = 3$ is given in Figure 5.1b. In this case we need three latent variables, $Z_{i,1}, Z_{i,2}$ and $Z_{i,3}$. Based on their relationship we can define three volumes that correspond to three classes. For example, if $Z_{i,1} > Z_{i,2}$ and $Z_{i,1} > Z_{i,3}$, the point belongs to the volume that corresponds to $y_i = 0$ (green volume). Similarly, if $Z_{i,2}$ is the maximum latent variable, then the class is $y_i = 1$ (red volume). The missing space is cyan, which corresponds to the volume

where $y_i = 2$, but it is not shown to reveal how the cube looks inside. In statistics, it is common to select one category as the baseline category of the model because it is one way to avoid redundant comparisons between categories and to help with the interpretation of the model. In addition, in the MCMC approach, selecting a baseline category produces a model with one fewer $(M-1)$ latent variable, which is a computational advantage during the MCMC process compared to use of M latent variables.

In order to perform variable selection in the multi-class multivariate case, a common $p \times 1$ indicator vector $\boldsymbol{\xi}$ is used across different latent variables. The j-th element of ξ_j is defined such that

$$\xi_j = \begin{cases} 1, & \text{if } B_{j,r} \neq 0 \text{ for all } r, \\ 0, & \text{if } B_{j,r} = 0 \text{ for all } r, \end{cases} \tag{5.5}$$

where $B_{j,r}$ is the entry in the j-th row and r-th column of $\mathbf{B}$, for $j = 1, \ldots, p$ $r = 1, \ldots, M-1$. Selection of the j-th variable corresponds to $\xi_j = 1$, which requires all the coefficients of the j-th row to be nonzero. Later we will propose a model with two indicator vectors and we will use the two notations ($\boldsymbol{\gamma}$ and $\boldsymbol{\xi}$).

Note that incorporating $\boldsymbol{\xi}$ into Equation (5.2) yields an expression which will simplify algebra later,

$$\mathbf{Z} - \mathbf{1}_n \boldsymbol{\alpha}' - \mathbf{X}_{\boldsymbol{\xi}} \mathbf{B}_{\boldsymbol{\xi}} \sim MN(\mathbf{I}_n, \boldsymbol{\Sigma}), \tag{5.6}$$

where $\mathbf{X}_{\boldsymbol{\xi}}$ refers to those columns of $\mathbf{X}$ (out of p) that correspond to selected variables and $\mathbf{B}_{\boldsymbol{\xi}}$ refers to those rows of $\mathbf{B}$ (out of p) that correspond to selected variables.

5.2.2 Prior distributions

The prior distributions of the unknown parameters $\boldsymbol{\alpha}$, $\mathbf{B}_{\boldsymbol{\xi}}$, $\boldsymbol{\xi}$ and $\boldsymbol{\Sigma}$ are specified below.

The prior for the intercept of the standard binary probit model is typically set as a univariate normal (Equation (3.7)), and here it is extended to the matrix normal

$$\boldsymbol{\alpha}' - \boldsymbol{\alpha}_0{}' \sim MN(h, \boldsymbol{\Sigma}). \tag{5.7}$$

Similarly, the prior of the nonzero coefficients for the standard binary probit model is typically set as a multivariate normal (Equation (3.11)) and

here it is extended to a the matrix normal,

$$\mathbf{B}_{\boldsymbol{\xi}} - \mathbf{B}_{0\boldsymbol{\xi}} \sim MN(\mathbf{H}_{\boldsymbol{\xi}}, \boldsymbol{\Sigma}), \tag{5.8}$$

where $\mathbf{H}_{\boldsymbol{\xi}}$ refers to the entries of $\mathbf{H}$ that correspond to selected variables $\xi_j = 1$.

A proper conjugate prior is assigned to $\boldsymbol{\Sigma}$ as follows

$$\boldsymbol{\Sigma} \sim IW(\delta; \mathbf{Q}), \tag{5.9}$$

where $IW(\delta; \mathbf{Q})$ is the inverse Wishart distribution with shape parameter $\delta = n - (M-1) + 1 = n - M + 2$, n is the degrees of freedom and $M - 1$ is the dimension of the covariance matrix. The scale matrix hyperparameter $\mathbf{Q}$ usually takes the form $d\mathbf{I}_{M-1}$, where d is a constant. As already noted, the non-identifiability means that this choice may be an important one. With this conjugate form for the regression prior, the choice of $\mathbf{H}_{\boldsymbol{\xi}}$ will control the shrinkage in the estimation of $\mathbf{B}_{\boldsymbol{\xi}}$ whilst the choice of $\mathbf{Q}$ with control the scale of $\mathbf{Z}$. It seems clear from Equation (5.10) below that $\mathbf{Q} = \mathbf{I}$ would have a similar result so far as scaling of $\mathbf{Z}$ is concerned to fixing $\boldsymbol{\Sigma} = \mathbf{I}$, the main difference being that $\mathbf{Z}$ will have a matrix T-distribution instead of a matrix normal. Since there are different parametrizations of inverse Wishart distribution, the following inference is done based on the parametrization that is presented in Appendix A.

Finally, several priors on the indicator vector $\boldsymbol{\xi}$ can be assigned as discussed for $\boldsymbol{\gamma}$ in Subsection 3.2.3. Here a special case of Equation (3.15) where $w_j = w$ (same probability of success for Bernoulli trials across different latent variables) is selected as is commonly used in the literature.

5.2.3 Posterior inference

Here, similar to the binary case, $\boldsymbol{\alpha}$, $\mathbf{B}_{\boldsymbol{\xi}}$ and $\boldsymbol{\Sigma}$ can be integrated out from the joint posterior. The inference is done via two Gibbs steps. Setting $\boldsymbol{\alpha}_0 = \mathbf{0}$ and $\mathbf{B}_{0\boldsymbol{\xi}} = \mathbf{0}$, the full conditional probabilities can be explicitly derived as follows.

Sampling for latent matrix $\mathbf{Z}$ is given by

$$\mathbf{Z}|\boldsymbol{\xi}, \mathbf{X}, \mathbf{y} \sim MT(\delta; \mathbf{P}_{\boldsymbol{\xi}}, \mathbf{Q}) \prod_{i=1}^{n} \mathbb{1}(\mathbf{Z}_{i,:} \in F_i), \tag{5.10}$$

where MT is the matrix Student distribution (Appendix A), $\mathbf{P}_{\boldsymbol{\xi}} = \mathbf{I}_n +$

$h\mathbf{1}_n\mathbf{1}_n^{'} + \mathbf{X}_{\boldsymbol{\xi}}\mathbf{H}_{\boldsymbol{\xi}}\mathbf{X}_{\boldsymbol{\xi}}^{'}$ and $\mathbb{1}(.)$ is the indicator function of the set F_i with

$$F_i = \begin{cases} \{\mathbf{Z}_{i,:} : Z_i^{\star} \leq 0\}, & \text{if } y_i = 0, \\ \{\mathbf{Z}_{i,:} : Z_i^{\star} > 0 \text{ and } Z_{i,r} = Z_i^{\star}\}, & \text{if } y_i = r. \end{cases} \tag{5.11}$$

Since it is difficult to sample directly from Equation (5.10), a Gibbs sampler of full conditional distributions of the truncated Student distribution can be applied. An efficient way to sample from that distribution is given by Geweke (1991), where the exponential rejection sampling method is optimized.

Sampling from the posterior distribution of $\boldsymbol{\xi}$ given all other parameters is done via

$$p(\boldsymbol{\xi}|\mathbf{Z}, \mathbf{X}, \mathbf{Y}) \propto p(\boldsymbol{\xi}) \left|\mathbf{I}_n + \mathbf{X}_{\boldsymbol{\xi}}\mathbf{H}_{\boldsymbol{\xi}}\mathbf{X}_{\boldsymbol{\xi}}^{'}\right|^{-\frac{M-1}{2}} \left|\mathbf{Q}_{\boldsymbol{\xi}}\right|^{-\frac{\delta+n+M-2}{2}},$$

where $\mathbf{Q}_{\boldsymbol{\xi}} = \mathbf{Q} + \mathbf{Z}^{'}(\mathbf{I}_n - \mathbf{X}_{\boldsymbol{\xi}}\mathbf{V}_{\boldsymbol{\xi}}^{-1}\mathbf{X}_{\boldsymbol{\xi}}^{'})\mathbf{Z}$, $\mathbf{V}_{\boldsymbol{\xi}} = \mathbf{X}_{\boldsymbol{\xi}}^{'}\mathbf{X}_{\boldsymbol{\xi}} + \mathbf{H}_{\boldsymbol{\xi}}^{-1}$, $\mathbf{X}$ is centered around column means (subtract the column mean from each element) and h is large. It is not easy to sample from the last formula, so the Metropolis algorithm according to Brown et al. (1998a) is applied within the Gibbs step. The candidate indicator vector is updated by randomly picking between adding or deleting a variable, or swapping two variables. In order to speed up the computations, the QR deletion or addition algorithm can be applied (Brown et al., 1998b). The data augmentation for $\mathbf{H}_{\boldsymbol{\xi}} = c_3\mathbf{I}_{p_{\boldsymbol{\xi}}}$ ($p_{\boldsymbol{\xi}} = \sum_{j=1}^{p}\xi_j$) yields $\left|\mathbf{I}_n + \mathbf{X}_{\boldsymbol{\xi}}\mathbf{H}_{\boldsymbol{\xi}}\mathbf{X}_{\boldsymbol{\xi}}^{'}\right| = \left|\mathbf{H}_{\boldsymbol{\xi}}^{1/2}\mathbf{X}_{\boldsymbol{\xi}}^{'}\mathbf{X}_{\boldsymbol{\xi}}\mathbf{H}_{\boldsymbol{\xi}}^{1/2} + \mathbf{I}_n\right| = \left|\widetilde{\mathbf{X}}_{\boldsymbol{\xi}}^{'}\widetilde{\mathbf{X}}_{\boldsymbol{\xi}}\right|$, where

$$\widetilde{\mathbf{X}}_{\boldsymbol{\xi}} = \begin{pmatrix} \mathbf{X}_{\boldsymbol{\xi}}\mathbf{H}_{\boldsymbol{\xi}}^{1/2} \\ \mathbf{I}_{p_{\boldsymbol{\xi}}} \end{pmatrix} \tag{5.12}$$

is a $(n+p_{\boldsymbol{\xi}}) \times p_{\boldsymbol{\xi}}$ augmented matrix. Then, $\mathbf{Q}_{\boldsymbol{\xi}} = \widetilde{\mathbf{Z}}^{'}\widetilde{\mathbf{Z}} - \widetilde{\mathbf{Z}}^{'}\widetilde{\mathbf{X}}_{\boldsymbol{\xi}}(\widetilde{\mathbf{X}}_{\boldsymbol{\xi}}^{'}\widetilde{\mathbf{X}}_{\boldsymbol{\xi}})^{-1}\widetilde{\mathbf{X}}_{\boldsymbol{\xi}}^{'}\widetilde{\mathbf{Z}}$, where

$$\widetilde{\mathbf{Z}} = \begin{pmatrix} \mathbf{Z} \\ \mathbf{0} \end{pmatrix} \tag{5.13}$$

is a $(n + p_{\boldsymbol{\xi}}) \times (M-1)$ augmented matrix.

5.2.4 Classification and prediction

As we discussed in Section 3.6, we can derive the marginal posterior probabilities for each variable and for each model.

We focus on how to use the best model in order to do accurate predictions. Let us denote by $\mathbf{X}_f$ the $n_f \times p$ matrix of new (future) measurements. The

least squares model prediction, based on the best model ($\hat{\boldsymbol{\xi}}$), is given by

$$\hat{\mathbf{Z}}_f = \mathbf{1}_{n_f}\widetilde{\boldsymbol{\alpha}}' + \mathbf{X}_{f\hat{\boldsymbol{\xi}}}\widetilde{\mathbf{B}}_{\hat{\boldsymbol{\xi}}}, \tag{5.14}$$

where $\hat{\mathbf{Z}}_f$ is a $n_f \times (M-1)$ estimated matrix of latent variables for new measurements, $\widetilde{\boldsymbol{\alpha}} = \bar{\hat{\mathbf{Z}}}$ (mean of MCMC samples) and $\widetilde{\mathbf{B}}_{\hat{\boldsymbol{\xi}}} = \left(\mathbf{X}'_{\hat{\boldsymbol{\xi}}}\mathbf{X}_{\hat{\boldsymbol{\xi}}} + \mathbf{H}^{-1}_{\hat{\boldsymbol{\xi}}}\right)^{-1}\mathbf{X}'_{\hat{\boldsymbol{\xi}}}\hat{\mathbf{Z}}$. Alternatively, instead of using just the one best model, the average of the best models can be used, so as to improve predictions. This approach is known as Bayesian model averaging (Brown et al., 2002; Wasserman, 2000). In both cases, taking into account the estimated latent variables $\hat{\mathbf{Z}}_f$, the labels of future measurements, $\hat{\mathbf{Y}}_f$ ($n_f \times 1$), can be predicted according to Equation (5.3).

5.2.5 Hyperparameter settings

Usually, there is no information available about the intercept a priori. A non-informative prior is assigned to it by selecting a large value for h.

The hyperparameter of the indicator vector, w, controls the number of selected variables a priori. In the case of $p \gg n$, small values of w are chosen, so as to restrict the number of variables in the model.

For the hyperameter of the coefficient matrix Sha et al. (2004) set $\mathbf{H}_{\boldsymbol{\xi}} = c_3\mathbf{I}_{p_{\boldsymbol{\xi}}}$ (easy to calibrate). In these cases, the choice of c is crucial, because it controls the amount of shrinkage of the regression coefficients. Independent of whether the responses are binary or multi-class, a good choice of c is given via Equation (3.13).

5.3 Bayesian variable selection in the probit model with ordinal responses

In the current section we will discuss the probit model with multi-class ordinal responses, which can be considered as an extension of the binary case in the sense that it uses more than one boundary instead of just the zero cutoff that is used in the binary case. We will comment on important differences between the model with multi-class ordinal, multi-class nominal and binary responses.

5.3.1 Model

Similarly to the study of the probit model with nominal responses, let $\mathbf{z}$ be the $n \times 1$ vector of latent variables that is distributed as normal with common

variance σ^2 across different groups

$$z_i = \alpha + \mathbf{X}_{i,:}\boldsymbol{\beta} + \epsilon_i, \epsilon_i \sim N(0, \sigma^2), i = 1, 2, \ldots, n, \tag{5.15}$$

where z_i is the i-th latent variable, the scalar α is the intercept and $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients (just one column compared to the coefficient matrix $\mathbf{B}$ in the multi-class nominal case). The probit model with ordinal responses can seen as a general case of the binary probit. The identifiability problem concerning scales is the same, and the solution via fixing $\sigma^2 = 1$ is again the simplest one, though we may be instead use a proper prior for σ^2. Now, however, we have an additional problem with the (multiple) boundaries, and it is common to fix one of these to avoid get another identifiability issue.

The equivalent matrix form of the last equation can be expressed as a multivariate normal distribution

$$\mathbf{z} - \mathbf{1}_n\alpha - \mathbf{X}\boldsymbol{\beta} \sim MVN(\mathbf{0}, \sigma^2\mathbf{I}_n). \tag{5.16}$$

The form of the covariance between observations is $\sigma^2\mathbf{I}_n$, where throughout this subsection $\sigma^2 = 1$ for identifiability, similar to the case of a probit model with binary responses. Even though throughout this subsection the variance is fixed at one, σ^2 is still used since later on we will assign a proper prior to it as an alternative way to handle the non-identifiability. In the ordinal case (and also in the binary case) $\mathbf{z}$ is a latent vector in contrast with nominal case with more than two classes where $\mathbf{Z}$ is a latent matrix.

The relationship between the latent variables (Equation (5.15)) and the response, according to Albert and Chib (1993), is the following

$$y_i = m, \text{if } k_m < z_i \leq k_{m+1}, \tag{5.17}$$

where $\mathbf{k} = (k_0, \ldots, k_M)$, $k_0 = -\infty$ and $k_M = +\infty$ by definition, k_1 is fixed at 0 to avoid non-identifiability problem, and $k_2, k_3, \ldots, k_{M-1}$ are the unknown boundaries for the ordinal responses. The corresponding ordinal multinomial probit model that is based on the latent variable, can be expressed as

$$p_{im} = P(y_i = m) = P(k_m < z_i \leq k_{m+1}).$$

The graphical representation of Equation (5.17) for $M = 3$ ($y_i = \{0, 1, 2\}$) is given in Figure 5.2. The three ordinal responses can be specified by just one latent variable y_i (i is fixed) and four boundaries. Again, the green region corresponds to $y_i = 0$, the red to $y_i = 1$ and the blue to $y_i = 2$. The length

of the rectangles (distance between two boundaries) will vary for different classes.

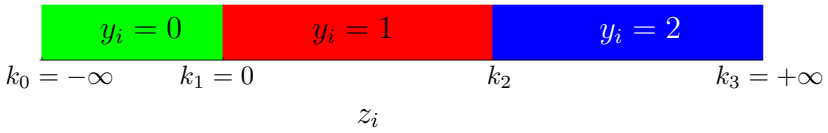


Figure 5.2: Graphical representation of Equation (5.17) for $M = 3$ ordinal responses.

In order to perform variable selection for the multi-class multivariate case, the $p \times 1$ indicator vector $\boldsymbol{\gamma}$ is used as defined in the binary case (Equation (3.4)). Incorporating $\boldsymbol{\gamma}$ in Equation (5.16) yields a convenient algebraic expression

$$\mathbf{z} - \mathbf{1}_n\alpha - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}} \sim MVN(\mathbf{0}, \sigma^2\mathbf{I}_n),$$

where again in this subsection $\sigma^2 = 1$ (Kwon et al., 2007).

5.3.2 Prior distributions

In this subsection the prior distributions of the unknown parameters α, $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, $\boldsymbol{\gamma}$ and $\mathbf{k}$ are studied. Now, an extra prior is needed for the unknown boundaries $\mathbf{k}$ of the ordinal responses. Kwon et al. (2007) place a flat prior on the components of the boundaries. $k_2, k_3, \ldots, k_{M-1}$ are uniformly distributed on the interval $(0, +\infty)$ subject to the constrain that $k_2 < k_3 < \ldots < k_{M-1}$ (since $k_0 = -\infty, k_M = +\infty$ by definition and $k_1 = 0$ by choice).

As in the binary case, the same univariate normal prior, $\alpha \sim N(\alpha_0, \sigma^2 h)$ (Equation (3.7), $\sigma^2 = 1$) can be assigned to the intercept. The prior for the nonzero coefficients is assumed multivariate normal $\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma} \sim MVN(\boldsymbol{\beta}_{0\boldsymbol{\gamma}}, \sigma^2\mathbf{H}_{\boldsymbol{\gamma}})$ (Equation (3.11), $\sigma^2 = 1$). In the absence of any additional information we set $\alpha_0 = 0$ and $\boldsymbol{\beta}_{0\boldsymbol{\gamma}} = \mathbf{0}$. If $\mathbf{H}_{\boldsymbol{\gamma}}$ is fixed, its value has to be considered carefully because the amount of shrinkage in the estimation of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is determined by $\mathbf{H}_{\boldsymbol{\gamma}}$. Thus we cannot learn from the data what is the appropriate level of shrinkage. If we choose an inappropriate value for $\mathbf{H}_{\boldsymbol{\gamma}}$ there is a risk of underfitting or overfitting on the training data. The former will be noticed in a poor fit, but the latter will only be apparent when we try. One way to select a sensible amount of shrinkage is to look at the size of the eigenvalues of $\mathbf{X}_{\boldsymbol{\gamma}}'\mathbf{X}_{\boldsymbol{\gamma}}$. Alternatively, a prior can be assigned to the shrinkage parameter so that we can learn from the data, but in practice this is computationally expensive (there is one extra step on the Gibbs algorithm, sampling from the shrinkage parameter which does not belong to any standard distribution and a MH algorithm is used). With respect to the prior of the indicator vector

$\boldsymbol{\gamma}$, we assume that each element of it is an independently and identically distributed Bernoulli random variable (Equation (3.15)) with $w_j = w$.

5.3.3 Posterior inference

Similar to the nominal case, inference is done via a Gibbs sampler. However, an extra Gibbs step is included in order to sample the unknown boundaries.

Integrating out α and $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ (using the fact that $\alpha_0 = 0$ and $\boldsymbol{\beta}_{0\boldsymbol{\gamma}} = \mathbf{0}$) sampling for the latent vector $\mathbf{z}$ is given by

$$\mathbf{z}|\boldsymbol{\gamma}, \mathbf{k}, \mathbf{X}, \mathbf{y} \sim MVN(\mathbf{0}, \mathbf{P}_{\boldsymbol{\gamma}}) \prod_{i=1}^{n} \mathbb{1}(z_i \in R_i), \tag{5.18}$$

where here $\mathbf{P}_{\boldsymbol{\gamma}} = \mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n^{'} + \mathbf{X}_{\boldsymbol{\gamma}}\mathbf{H}_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}}^{'}$ and

$$R_i = \{z_i : k_m < z_i \leq k_{m+1}\}, \text{ if } y_i = m. \tag{5.19}$$

Equation (5.18) is a multivariate truncated normal distribution (Appendix A), which cannot be sampled from directly. A Gibbs sampler for sampling from the multivariate truncated normal distribution, as proposed by Kotecha and Djuric (1999) can be applied. This technique uses the fact that the full conditional distributions are (univariate) truncated normal.

The full conditional posterior distribution of $\boldsymbol{\gamma}$ is calculated based on the following factorization

$$p(\boldsymbol{\gamma}|\mathbf{z}, \mathbf{k}, \mathbf{X}, \mathbf{y}) \propto p(\boldsymbol{\gamma})p(\mathbf{z}|\boldsymbol{\gamma}, \mathbf{k}, \mathbf{X}, \mathbf{y}). \tag{5.20}$$

Similarly to the nominal case study, since the last full conditional distribution does not have a closed form, the Metropolis algorithm (Brown et al., 1998a) is applied within that Gibbs step.

Finally, according to Albert and Chib (1993), the full conditional density of k_ν given the rest, for $\nu = 2, \ldots, M-1$, is given by

$$\begin{aligned} p(k_\nu|\boldsymbol{\gamma}, \mathbf{z}, \mathbf{X}, \mathbf{y}, \mathbf{k}_{\setminus\nu}) \propto \prod_{i=1}^{n} [&\mathbb{1}(y_i = \nu - 1)\mathbb{1}(k_{\nu-1} < z_i \leq k_\nu) \\ &+ \mathbb{1}(y_i = \nu)\mathbb{1}(k_\nu < z_i \leq k_{\nu+1})], \end{aligned}$$

where $\mathbf{k}_{\setminus\nu} = (k_0, \ldots, k_{\nu-1}, k_{\nu+1}, \ldots, k_M)$.

Actually, the update of each boundary parameter can be implemented via

$$k_\nu|\boldsymbol{\gamma}, \mathbf{z}, \mathbf{X}, \mathbf{y}, \mathbf{k}_{\setminus\nu} \sim U\big(\max[\{z_i : y_i = \nu - 1\}, k_{\nu-1}], \min[\{z_i : y_i = \nu\}, k_{\nu+1}]\big). \tag{5.21}$$

So, sampling from boundaries is the easiest step of the Gibbs sampling. For models with ordinal responses, if we alternately sample boundaries and latent variables neither of them can move very much between iterations because of the constraints. This can cause slow mixing. To improve that Cowles (1996) proposed the latent variables and boundaries to be updated simultaneously using a Hastings-with-Gibbs, instead of sampling individually from their full conditionals. In that case the proposal is univariate truncated normal (at left and right) with mean the corresponding boundary of the previous step and fixed variance. However, since it is difficult to fixed this variance Nandram and Chen (1996) proposed an improved method using Dirichlet proposal distribution based on a rescaling transformation of boundaries, coefficients and latent variables.

5.3.4 Classification and prediction

Variables that have the largest marginal posterior probabilities are found similarly to the nominal case (see Section 3.6). However, here we recalculate the posterior probabilities of gamma for distinct by using not only the sample mean of the latent variables but also the sample mean of the boundaries.

In addition we can identify the best model. The least squares model prediction of ordinal responses is based on the best model and is given by

$$\hat{\mathbf{z}}_f = \mathbf{1}_{n_f}\widetilde{\alpha} + \mathbf{X}_{f\hat{\gamma}}\widetilde{\boldsymbol{\beta}}_{\hat{\gamma}},$$

where here $\hat{\mathbf{z}}_f$ is the estimated $n_f \times 1$ vector of latent variables, $\widetilde{\alpha}$ is the mean of the estimated vector of latent variables and $\widetilde{\boldsymbol{\beta}}_{\hat{\gamma}} = \left(\mathbf{X}'_{\hat{\gamma}}\mathbf{X}_{\hat{\gamma}} + \mathbf{H}^{-1}_{\hat{\gamma}}\right)^{-1}\mathbf{X}'_{\hat{\gamma}}\hat{\mathbf{z}}$. Alternatively, as in the nominal case, here we can apply Bayesian model averaging. Using the estimated latent variable $\hat{\mathbf{z}}_f$ we are able to do predictions via Equation (5.17).

5.3.5 Hyperparameter settings

We assign a vague prior to the intercept, selecting h to be large. For the covariance matrix of selected coefficients we select $\mathbf{H}_{\boldsymbol{\gamma}} = c_2\mathbf{I}_{p_{\boldsymbol{\gamma}}}$ (Equation (3.11) for $\sigma^2 = 1, c_1 = 0, \mathbf{D}_{\boldsymbol{\gamma}} = \mathbf{I}_{p_{\boldsymbol{\gamma}}}$), to give a ridge type shrinkage. Without loss of generality we assume that $k_1 = 0$.

Chapter 6

Decomposed Bayesian variable selection in the probit model with a mixture of nominal and ordinal responses

In the previous chapter we reviewed BVS approaches for multi-class responses in the purely nominal and purely ordinal case. However, there are some cases where the multi-class response is a mixture of nominal and ordinal. In this chapter, we propose a decomposed BVS method in the multinomial probit model with a mixture of both types of responses using latent variables. Our approach consists of two distinct parts: one part treats the ordinal responses as a single nominal category and separates nominal responses, whereas the other part separates ordinal responses within this category. We present the decomposed approach for BVS under two different settings: the variance of the latent variables can be known or unknown i.e., fixed or given a prior distribution. We develop efficient posterior sampling and apply the decomposed methodology on simulated data. We compare the classification accuracy of our method to existing ones. This proposed decomposed BVS method has been published in a conference paper (Kotti et al., 2016c).

6.1 Introduction

References in multi-class BVS have studied either the purely nominal multinomial probit model with unknown covariance matrix $\mathbf{\Sigma}$ across different latent variables (Sha et al., 2004) or the purely ordinal multinomial probit model with known variance σ^2 of the single latent variable (Kwon et al.,

2007). Firstly, we are interested in studying the remaining two cases of BVS: the (nominal) multinomial probit model but with known covariance matrix $\boldsymbol{\Sigma}$ across different latent variables (proposed method 1 in Figure 6.1) and the ordinal multinomial probit model with unknown variance σ^2 of the latent variable (proposed method 2 in Figure 6.1). Secondly, having the four methodological approaches available, we will combine them as follows: the method 1 with the Kwon et al. (2007) approach (covariance matrix and variance known) and the approach of Sha et al. (2004) with method 2 (covariance matrix and variance unknown), with aim to propose a BVS approach with mixture of nominal and ordinal responses.

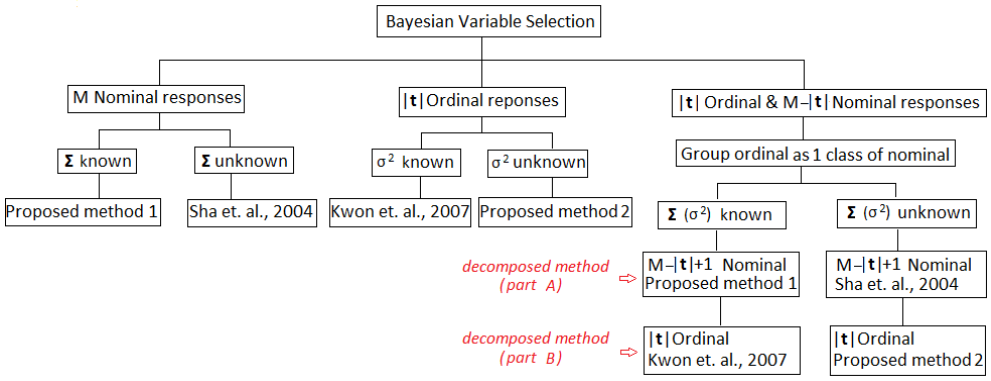


Figure 6.1: BVS for pure nominal, pure ordinal and mixture of nominal and ordinal responses, suggesting in the last case a decomposed approach.

6.2 Methods 1 and 2

Before developing the decomposed methodology, we study the methodological parts of the probit model with nominal responses and $\boldsymbol{\Sigma}$ known (method 1) and of the probit model with ordinal responses and σ^2 unknown (method 2), in detail in the following section.

6.2.1 Model and prior distributions

Probit model with nominal responses, $\boldsymbol{\Sigma}$ known

The multinomial probit model for nominal responses using $M-1$ latent variables has already been explained in Subsection 5.2.1, but here we assume that the distribution of the latent variables (Equation (5.2)) has a known covariance matrix $\boldsymbol{\Sigma}$. It is sufficient to fix $\boldsymbol{\Sigma}$, in practice to an identity matrix, for identifiability. Nominal responses have a common indicator vector $\boldsymbol{\xi}$ (Equation (5.5)).

Prior distributions for nominal responses case

The priors for the unknown parameters $\boldsymbol{\alpha}$, $\mathbf{B}_{\boldsymbol{\xi}}$ and $\boldsymbol{\xi}$ are presented in Subsection 5.2.2. Note that, since we assume that $\boldsymbol{\Sigma}$ is fixed, we do not assign any prior to it.

Probit model with ordinal responses, σ^2 unknown

We discussed the multinomial probit model for ordinal responses using one latent variable in Subsection 5.3.1. However, here we assume that the distribution of the each component of the latent variable (Equation (5.15)) has an unknown variance σ^2. We will assign a proper prior to σ^2, which in principle solves the non-identifiability problem, though, as discussed earlier, problems of MCMC convergence and interpretation of results may remain.

To distinguish the indicator vectors of method 1 and 2, we denote the indicator vector that is used for the ordinal responses by $\boldsymbol{\gamma}$. In addition, the vector of latent variables related to the ordinal responses is denoted by $\mathbf{z}$ (in contrast to the latent matrix $\mathbf{Z}$ of nominal responses).

Prior distributions for ordinal responses case

In this case, the unknown parameters are α, $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, $\boldsymbol{\gamma}$, $\mathbf{k}$ and σ^2. We have assigned prior distributions on the first four parameters (Subsection 5.3.2). In addition, we assign a conjugate prior for σ^2, $\sigma^2 \sim IG(d_1, d_2)$.

6.2.2 Posterior inference

Nominal responses case

Similar to the study of Sha et al. (2004), $\boldsymbol{\alpha}$ and $\mathbf{B}_{\boldsymbol{\xi}}$ are integrated out from the joint posterior. Inference is performed via two Gibbs steps. Setting $\boldsymbol{\alpha}_0 = \mathbf{0}$ and $\mathbf{B}_{0\boldsymbol{\xi}} = \mathbf{0}$, we can sample from the latent matrix $\mathbf{Z}$ according to

$$\mathbf{Z}|\boldsymbol{\xi}, \mathbf{X}, \mathbf{y} \sim MN(\mathbf{P}_{\boldsymbol{\xi}}, \boldsymbol{\Sigma}) \prod_{i=1}^{n} \mathbb{1}(\mathbf{Z}_{i,:} \in F_i), \tag{6.1}$$

where F_i is given by Equation (5.11).

Since it is difficult to sample directly from Equation (6.1), we use a a Gibbs sampler and we derive the full conditional distributions of the truncated matrix normal distribution. An efficient way to sample from that distribution, which will be multivariate truncated normal, is given by Geweke (1991), where an exponential rejection sampling method is optimized.

We are then interested in sampling from the posterior distribution of $\boldsymbol{\xi}$ given by

$$p(\boldsymbol{\xi}|\mathbf{Z},\mathbf{X},\mathbf{y}) \propto p(\boldsymbol{\xi}) \left|\mathbf{I}_n + \mathbf{X}_{\boldsymbol{\xi}}\mathbf{H}_{\boldsymbol{\xi}}\mathbf{X}_{\boldsymbol{\xi}}^{'}(\mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n^{'})^{-1}\right|^{-\frac{M-1}{2}} \cdot |\mathbf{P}_{\boldsymbol{\xi}}|^{-\frac{M-1}{2}} \exp\left\{-\frac{1}{2}\text{tr}\left[\boldsymbol{\Sigma}^{-1}\mathbf{Z}^{'}\mathbf{P}_{\boldsymbol{\xi}}^{-1}\mathbf{Z}\right]\right\}. \tag{6.2}$$

Note that, if $\mathbf{X}$ is centered and h is large, then it is true that $\left|\mathbf{I}_n + \mathbf{X}_{\boldsymbol{\xi}}\mathbf{H}_{\boldsymbol{\xi}}\mathbf{X}_{\boldsymbol{\xi}}^{'}(\mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n^{'})^{-1}\right| \approx \left|\mathbf{I}_n + \mathbf{X}_{\boldsymbol{\xi}}\mathbf{H}_{\boldsymbol{\xi}}\mathbf{X}_{\boldsymbol{\xi}}^{'}\right|$. Since it is not easy to sample from this full conditional distribution, the Metropolis algorithm according to Brown et al. (1998a) is applied within that Gibbs step. The candidate indicator vector is updated by randomly choosing between adding or deleting a variable or swapping two variables. In order to speed up the computations, the QR deletion or addition algorithm can be applied. More details about the algebraic calculations are presented in Appendix B.

Ordinal responses case

The unknown parameters α, $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and σ^2 are integrated out from the joint posterior. Similar to the nominal case inference is performed via a Gibbs sampler.

Setting $\alpha_0 = 0$ and $\boldsymbol{\beta}_{0\boldsymbol{\gamma}} = \mathbf{0}$, the full conditional distribution of the vector of latent variable becomes a multivariate Student distribution (MVT),

$$\mathbf{z}|\boldsymbol{\gamma},\mathbf{k},\mathbf{X},\mathbf{y} \sim MVT\left(2d_1;\mathbf{0},\frac{d_2}{d_1}\mathbf{P}_{\boldsymbol{\gamma}}\right)\prod_{i=1}^{n}\mathbb{1}(z_i \in R_i), \tag{6.3}$$

where R_i is given by Equation (5.19) and $\mathbf{P}_{\boldsymbol{\gamma}} = \mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n^{'} + \mathbf{X}_{\boldsymbol{\gamma}}\mathbf{H}_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}}^{'}$. Since we cannot easily sample from a truncated multivariate Student distribution, a Gibbs sampler is used to sample from the full conditional distributions using the method of Geweke (1991).

The conditional posterior distribution of $\boldsymbol{\gamma}$ then becomes

$$p(\boldsymbol{\gamma}|\mathbf{z},\mathbf{k},\mathbf{X},\mathbf{y}) \propto p(\boldsymbol{\gamma})\left(\left|\mathbf{I}_n + \mathbf{X}_{\boldsymbol{\gamma}}\mathbf{H}_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}}^{'}(\mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n^{'})^{-1}\right|\right)^{-1/2}|\mathbf{P}_{\boldsymbol{\gamma}}|^{-1/2}. \tag{6.4}$$

Similarly to the nominal case the Metropolis algorithm Brown et al. (1998a) is applied within that Gibbs step.

Finally, updating the boundaries is not affected by whether the variance is fixed or not, and their distribution is described completely by Equation (5.21). Appendix C contains more details about how to calculate those three full

conditional probabilities.

6.2.3 Classification and prediction

The predictive step for each approach is described in Subsection 5.2.4 for nominal responses and in Subsection 5.3.4 for ordinal responses.

6.2.4 Hyperparameter settings

Nominal responses case

We have to specify two fewer hyperparameters (δ, $\mathbf{Q}$), compared to the study of BVS with nominal responses and $\boldsymbol{\Sigma}$ unknown (Subsection 5.2.5), since here $\boldsymbol{\Sigma}$ is fixed.

Ordinal responses case

The hyperparameters here are the same as described in the ordinal case with σ^2 known (Subsection 5.3.5). In addition, the prior for σ^2 is assumed to be an Inverse-Gamma(d_1, d_2) distribution, which is a univariate version of the Inverse-Wishart$(\delta, \mathbf{Q})$ distribution, where $\mathbf{Q}$ is the scale matrix. Based on this fact, $d_1 = \delta/2$ and $d_2 = q_{11}/2$, where $\mathbf{Q} = d\mathbf{I}_{M-1}$. The value $\delta = 3$ is the smallest integer value so that the expectation $E(\boldsymbol{\Sigma}) = \mathbf{Q}/(\delta - 2) = \mathbf{Q}$ exists (Brown et al., 1998b), and so we take $d_1 = 1.5$.

6.3 Decomposed Bayesian variable selection for a mixture of response types

6.3.1 Method

Let us assume that from the total of M responses, $|\mathbf{t}|$ are ordinal, where $\mathbf{t} = (t_0, \ldots, t_{|\mathbf{t}|-1})$, and the remaining $M - |\mathbf{t}|$ are nominal responses. In this case we combine the existing and proposed methods as is noted in the last part of Figure 6.1. We assign the same coding to all ordinal responses and consider them as a single new class of nominal response. In this part of the decomposed method, we apply the BVS approach using $M - |\mathbf{t}|$ latent variables/regression equations that refer to the $M - |\mathbf{t}| + 1$ nominal responses (part A of Algorithm 2), assuming that zero is the 'baseline'. Note, that in this case the indicator vector is common (Equation (5.5)) across the different regression equations. From the posterior inference of the indicator vector

we identify important variables or most frequently visited models. In the next part of the decomposed method, we apply the BVS approach for $|\mathbf{t}|$ ordinal responses using just one latent variable (part B of Algorithm 2). In this case we use a different indicator vector than before, which is related to the inclusion or exclusion of the coefficients that refer only to the ordinal responses (Equation (3.4)). 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 important variables or most frequently visited models for the case of mixture of nominal and ordinal responses.

Algorithm 2 Decomposed Bayesian variable selection : 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 a single nominal category)

0: Initialize values $\boldsymbol{\xi}^{(0)}$ and $\mathbf{Z}^{(0)}$
1: Draw $\boldsymbol{\xi}^{(j_A)}$ from $p(\boldsymbol{\xi}|\mathbf{Z}^{(j_A-1)}, \mathbf{X}, \mathbf{y})$
2: Draw $\mathbf{Z}^{(j_A)}$ from $p(\mathbf{Z}|\boldsymbol{\xi}^{(j_A)}, \mathbf{X}, \mathbf{y})$
3: Repeat steps 1 and 2 until the number of iterations is achieved and stop (Results: $VS_A^\dagger$ and $MS_A^\dagger$)

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

0: Initialize values $\boldsymbol{\gamma}^{(0)}, \mathbf{z}^{(0)}$ and $\mathbf{k}^{(0)}$
1: Draw $\boldsymbol{\gamma}^{(j_B)}$ from $p(\boldsymbol{\gamma}|\mathbf{z}^{(j_B-1)}, \mathbf{k}^{(j_B-1)}, \mathbf{X}, \mathbf{y})$
2: Draw $\mathbf{k}^{(j_B)}$ from $p(\mathbf{k}|\mathbf{z}^{(j_B-1)}, \boldsymbol{\gamma}^{(j_B)}, \mathbf{X}, \mathbf{y})$
3: Draw $\mathbf{z}^{(j_B)}$ from $p(\mathbf{z}|\boldsymbol{\gamma}^{(j_B)}, \mathbf{k}^{(j_B)}, \mathbf{X}, \mathbf{y})$
4: Repeat steps 1, 2 and 3 until the number of iterations is achieved and stop (Results: $VS_B^\dagger$ and $MS_B^\dagger$)

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

$^\dagger VS_A$ (VS_B): the set of selected variables for nominal (ordinal) responses using marginal probabilities, MS_A (MS_B): the most frequently visited model for nominal (ordinal) responses using posterior probabilities, VS: the final set of selected variables (nominal and ordinal responses jointly), and MS: the corresponding set of variables in the most probable model.

We denote the sample of j-th iteration of part A with the upper index (j_A) and of part B with the upper index (j_B). We construct the Gibbs steps as summarized in Algorithm 2. The algorithm consists of two parts, A and B, and the final conclusion, which is the combination of both. Since the two parts are independent, the order in which they are computed does not

matter. Note that the indicator vector and latent variables are different for each part and refer to the corresponding approach (with nominal or ordinal responses). In this algorithm the union has been used to summarize the results of variable selection in parts A and B. However, for predictions we use either VS_A or VS_B, but not both (see Subsection 6.3.2). Using the union we can say (without going into the details of how the decomposed method works) that those are the important variables, and then specifically some of them are important for case A and some of them for case B.

6.3.2 Classification and prediction

The classification procedure for a new sample is done according to the following process: First, we use the best model (the model with the highest posterior probability) for nominal responses and we do predictions according to Equation (5.3). 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 use the best model for ordinal responses and we do predictions according to Equation (5.17).

6.4 Simulation results

6.4.1 Simulations

The experimental study was performed using simulated data from the probit model with multi-class nominal and ordinal responses.

For the simulation study first we fix the number of variables, the number of samples, the total number of possible responses and the ordinal responses. For each simulation part (nominal and ordinal), we identify the indices of the important variables via the indicator vectors. Based on those two indicator vectors we can extract the joint indicator vector. Then, the important variables determine the nonzero coefficients that have to be specified separately for each type of responses. The probability of success for the Bernoulli distribution may differ for the two parts of the simulation.

We consider that $X_{i,j}$ are i.i.d. and $X_{i,j} \sim N(0,1)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. In part A, we treat the ordinal responses as one nominal response, and without loss of generality we assign them the coding of the first ordinal response (t_0). Then, we construct $\mathbf{Z} = \mathbf{XB} + \boldsymbol{\epsilon}$, which is the matrix of latent variables where $\boldsymbol{\epsilon}$ is an $n \times (M - |\mathbf{t}|)$ matrix, $\epsilon_{i,g}$ are i.i.d. and $\epsilon_{i,g} \sim N(0, \sigma_A^2)$ for $i = 1, \ldots, n$ and $g = 1, \ldots, M - |\mathbf{t}|$ where by σ_A^2

we denote the error variance from part A. Here $\mathbf{B}$ has most of the entries zero as the indicator vector is sparse. Based on the simulated $\mathbf{Z}$ we assign values to the response vector $\mathbf{y}$ with coding $0, \ldots, M - |\mathbf{t}|$. In the next part, we select those rows of $\mathbf{X}$ and $\mathbf{y}$ that correspond to $y_i = t_0$ and denote them by $\mathbf{X}_{t_0,:}$ and $\mathbf{y}_{t_0}$ respectively. Then, in order to separate the ordinal responses, we construct $\mathbf{z} = \mathbf{X}_{t_0,:}\boldsymbol{\beta} + \boldsymbol{\epsilon}^{t_0}$, where $\mathbf{z}$ refers to the vector of latent variables of just ordinal responses and $\boldsymbol{\epsilon}^{t_0}$ to the error term of the ordinal responses, $\epsilon_i^{t_0}$ are i.i.d. and $\epsilon_i^{t_0} \sim N(0, \sigma_B^2)$ for $i = 1, \ldots, n_0$ (n_0: number of samples with ordinal responses). $\boldsymbol{\beta}$ has most of its entries zero. We use $|\mathbf{t}| - 1$ quantiles of the $\mathbf{z}$ to choose the boundary vector $\mathbf{k}$. The reason that we are using quantiles is because we are interested in ensuring that in each of $|\mathbf{t}|$ intervals $[k_{\nu'-1}, k_{\nu'})$, $\nu' = 1, \ldots, |\mathbf{t}|$, there fall a large enough number of latent variables or in the ideal case approximately the same number of latent variables (balanced ordinal responses). Based on the simulated $\mathbf{z}$ and $\mathbf{k}$ we assign ordinal values to the response vector $\mathbf{y}_{t_0}$ with coding $t_0, \ldots, t_{|\mathbf{t}|-1}$. Finally, we plug in the values of $\mathbf{y}_{t_0}$ into $\mathbf{y}$ in the positions that have $y_i = t_0$ and we construct the response vector that has a mixture of nominal and ordinal responses.

We run two different simulations to cover the following scenarios: (i) $\boldsymbol{\Sigma}$ and σ^2 are known with $n \gg p$, (ii) $\boldsymbol{\Sigma}$ and σ^2 are unknown with $n \gg p$ and (iii) $\boldsymbol{\Sigma}$ and σ^2 are unknown with $n \ll p$. In the first two cases, for generating simulated data we set $n = 100$, $p = 10$, $M = 6$, $\mathbf{t} = [3, 4, 5]$ and $\sigma_A = \sigma_B = 1$. Thus, $\sigma^2 = 1$ and $\boldsymbol{\Sigma} = \mathbf{I}_3$. The majority of $\mathbf{B}$'s entries (related to the nominal responses) are zero except for $B_{[3,8],1} = [-1, 0.4]$, $B_{[3,8],2} = [0.5, -0.9]$ and $B_{[3,8],3} = [0.6, -0.4]$, where $B_{[j,f],r} = [B_{j,r}, B_{f,r}]$, for $j, f = 1, \ldots, p$ and $r = 1, \ldots, M - |\mathbf{t}|$ (latent variables). In addition, the majority of $\boldsymbol{\beta}$'s entries (related to the ordinal responses) are zero except for $\beta_1 = -0.5$.

For scenario (i) in the BVS analysis we set $\boldsymbol{\Sigma}$ equal to $\mathbf{I}_3$, $\sigma^2 = 1$, $w_{(nom)} = 2/10$ and $w_{(ord)} = 1/10$. In addition, we select the hyperparameters as follows: $h = 10^6$ and $c_2 = c_3 = 5$. In part A, we initialize $\boldsymbol{\xi}^{(0)}$ randomly selecting two (out of p) variables to be one and the remaining are zero. Then, we initialize $\mathbf{Z}^{(0)}$ that is related to the nominal responses (included the one extra group of ordinal) according to

$$Z_{j,r}^{(0)} = \begin{cases} -1, & \text{if } y_j = 0, \\ 1, & \text{if } y_j = r, \\ 0, & \text{otherwise}, \end{cases} \tag{6.5}$$

where $r = 1, \ldots, M - |\mathbf{t}|$ (here $r = 1, 2, 3$). We run four different chains with 3000 iterations after 1000 burn-in iterations. In part B, we initialize the $\boldsymbol{\gamma}^{(0)}$ randomly selecting one variable to be one. Then, we initialize $\mathbf{z}^{(0)}$ that is related only to the ordinal responses as follows

$$z_i^{(0)} = \begin{cases} -1, & \text{if } y_i = t_0, \\ 1, & \text{otherwise}, \end{cases} \tag{6.6}$$

and $\mathbf{k}^{(0)}$ drawing the non-fixed boundaries uniformly in (0,1000). We run four different chains with 3000 iterations after 1000 burn-in iterations. Figure 6.2 contains the results of variable and model selection of the two parts, averaging over the chains. Our proposed algorithm correctly identifies the individually important variables 3 and 8 (part A) and 1 (part B). The remaining variables have marginal posterior probabilities close to zero. In addition, the combination of the variables 3 and 8 is the best model for part A (bottom of Figure 6.2: first line of $x-$axis) and the model with just one variable (the first) is the best model for part B (bottom of Figure 6.2: second line of $x-$axis), with posterior probability much higher than the second, third, etc. best models.

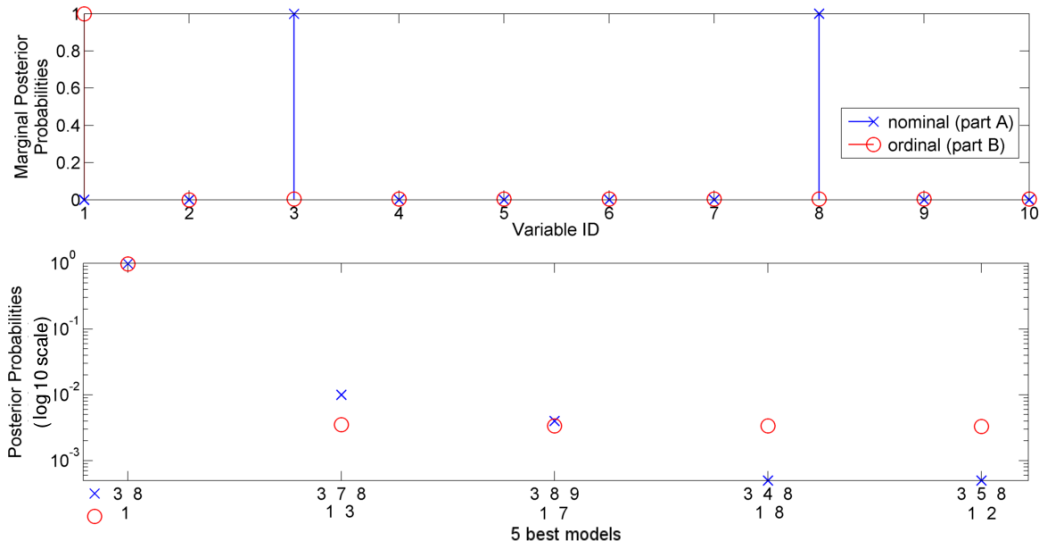


Figure 6.2: Marginal posterior probabilities (top) and posterior probabilities on a log scale (bottom) of average of chains for scenario (i).

For scenario (ii) in the variable selection approach, we set the values for the hyperparameters of the unknown covariance matrix $\boldsymbol{\Sigma}$, $\delta = 3$, $d = 1$, $\mathbf{Q} = \mathbf{I}_3$ and the hyperparameters of the variance σ^2, $d_1 = \delta/2 = 1.5$ and $d_2 = 0.5$ (inverse Gamma is the univariate case of inverse Wishart). The remaining parameters and hyperparameters are the same as in scenario (i). The figure of the scenario (ii) would be similar to the Figure 6.2 and the

results are similar to those reported in scenario (i).

Finally, we focus on the more interesting scenario (iii), where $n \ll p$ and $\boldsymbol{\Sigma}$ is unknown. To generate data from the model we set $n = 100$, $p = 200$, $M = 5$, $\mathbf{t} = [1, 2, 3]$, $w_{(nom)} = 2/200$, $w_{(ord)} = 3/200$ and $\sigma_A = \sigma_B = 1$. 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 entries (related to the ordinal responses) are zero except that $\beta_5 = -1.4$, $\beta_{100} = 1.2$ and $\beta_{150} = 1.3$. The hyperparameters are fixed as in scenario (ii). We run four different chains with 5000 iterations after 2000 burn-in iterations. Figure 6.3 contains the results of variable and model selection for the two parts, averaging across chains. Our proposed algorithm correctly identifies the individually important variables 3, 8 (part A) and 5, 100, 150 (part B) and the corresponding best models.

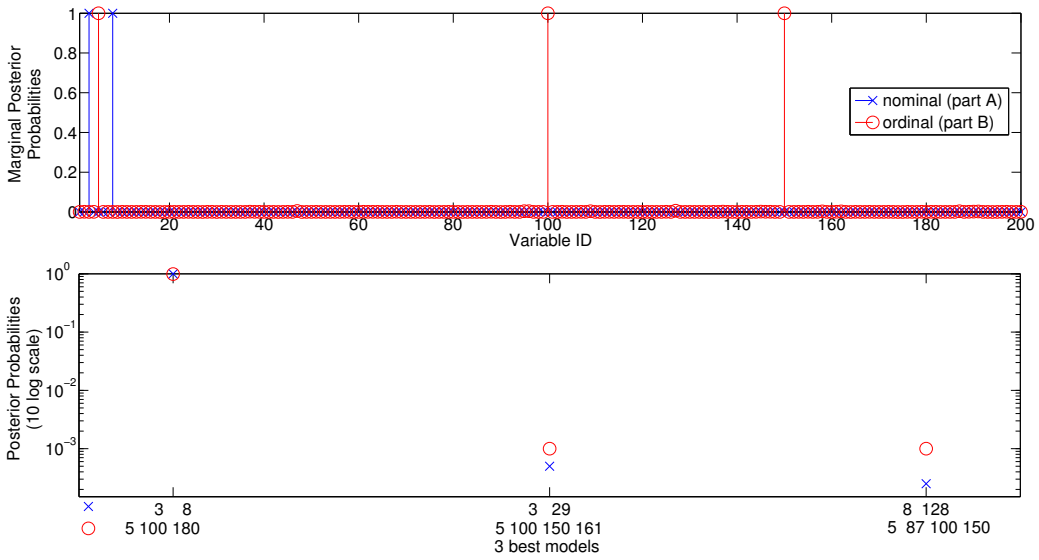


Figure 6.3: Marginal posterior probabilities (top) and posterior probabilities on a log scale (bottom) of average of chains for scenario (iii).

6.4.2 Predictions

In order to perform prediction of a new (future) sample for the last scenario, we generate new data $\mathbf{X}_f$ (a hundred samples, referred to as the test set) according to the parameters of scenario (iii). The process of generating the test set is the same as the process of the training set, except that here we use the boundary vector $\mathbf{k}$ from the training set rather than deriving it from the simulated $\mathbf{z}$. We pick from the test design matrix only the variables that had been selected after applying Algorithm 2 and we make predictions. We repeat the process of generating test sets one hundred times. Based on

Table 6.1: Comparison of classification accuracy for one test set after applying variable selection approaches for scenario (iii).

	Accuracy (%)	
	Nominal	Ordinal
'Highest' accuracy	66	
Our proposed method	60	
LASSO	50	47
Classification trees	41	29
Random forests	48	31
SVM	36	35

the inherent amount of error that the simulated data has, in our case the highest classification accuracy that could be achieved 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.

In order to compare the proposed method with existing methods, we select one test set (out of a hundred). Table 6.1 contains the results of the comparison with some other methods that have been proposed for pure nominal or pure ordinal responses. The highest classification accuracy for this test set is 66% and our method achieves classification accuracy 60%, beating existing methods. The methods presented in this table have already been cited in the previous chapters. There is a wide range of toolboxes to apply those methods for nominal responses. On the other hand, focusing on ordinal responses there are just a few toolboxes, recently developed, that implement variable selection for the following methods: LDA, kernel discriminant analysis and SVM (Gutiérrez et al., 2016), LASSO (Archer and Williams, 2012), classification trees (Galimberti et al., 2012) and random forest (Hothorn et al., 2015).

6.5 Discussion and conclusion

We have presented a decomposed Bayesian probit model for variable and model selection with mixtures of nominal and ordinal responses. For this purpose, we use latent variables. The decomposed approach consists of two parts: treat the ordinal responses as one nominal category and apply BVS for nominal responses, and then apply BVS just to the ordinal responses. We used two indicator vectors to represent the presence, or absence, of a predictor in the regression (one for each part of the decomposed method): the model with nominal responses (increased by one) has a common indicator vector across different regression equations and the model with ordinal responses has a different indicator vector from the model with nominal responses.

The Bayesian methodology is applied twice, for the nominal and for the ordinal responses. This means that the we had to select two different burn-in periods and sampling from conditional distributions of indicator vector twice. In this chapter we passed this hurdle by running in parallel the BVS approach using nominal and ordinal responses, which allowed the proposed algorithm to be computationally efficient and simple.

However, when the number of variables is very large, this procedure becomes very time intensive. For this reason in the next chapter we will propose a new methodology, where the same latent variable has a double role. This requires us to sample from the conditional distribution of the indicator vector one time (instead of two).

Chapter 7

Bayesian variable selection in the probit model with a mixture of nominal and ordinal responses using a common indicator vector

In this chapter we first build a new (one step) model for multi-class classification problems with a mixture of nominal and ordinal responses. The model makes use of latent variables. In this case, the latent variable that corresponds to the sequence of ordinal responses has a double role: to discriminate the ordinal responses from nominal ones and to order the responses within the sequence. This double role allows us to use a one step model, instead of the decomposed model that was studied in the previous chapter. We propose a BVS approach based on that model. Variable selection is achieved by using a common indicator vector across different responses. This proposed BVS method with a common indicator vector has been presented in a workshop (Kotti et al., 2015).

7.1 Method

7.1.1 Model

We develop an extension of the usual probit model to the multi-class nominal/ordinal case, assuming that zero response is the 'baseline' category and it is also a part of the ordinal sequence. As in the previous chapter, we study a

classification problem with M classes (coding $m = 0, 1, \ldots, M-1$), considering that $|\mathbf{t}|$ of the responses are ordinal ($\mathbf{t} = (t_0, \ldots, t_{|\mathbf{t}|-1})$). The remaining s responses are nominal, where $s = M - |\mathbf{t}| + 1$, because in this case the zero response is nominal as well as part of the ordinal sequence.

We use a latent variable representation of the response classes, where subsets of s continuous latent variables correspond to the different response classes. Assume $\mathbf{Z}$ is the $n \times s$ matrix of latent variables that is distributed as multivariate normal with common variance $\boldsymbol{\Sigma}$ ($s \times s$) across different latent variables

$$\mathbf{Z}_{i,:} = \boldsymbol{\alpha}' + \mathbf{X}_{i,:}\mathbf{B} + \mathbf{E}_{i,:}, \mathbf{E}_{i,:} \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}), i = 1, 2, \ldots, n,$$

where $\mathbf{Z}_{i,:} = (Z_{i,1}, Z_{i,2}, \ldots, Z_{i,s})$ is the row vector of $\mathbf{Z}$ that refers to the i-th sample, similarly $\mathbf{X}_{i,:}$ is the row vector of $\mathbf{X}$, $\mathbf{B}$ is a $p \times s$ matrix of regression coefficients and $\boldsymbol{\alpha}$ is the $s \times 1$ vector of intercepts. Note that in the proposed model the matrix of latent variables contains the latent variables of both nominal and ordinal responses. Specifically, the first column of the matrix of latent variables and the first column of the matrix of coefficients refer to the ordinal responses.

This model has the same identifiability issues as our earlier ones, slightly complicated by the fact that the first latent variable has a different role to the others so that the question arises whether we want to impose the same constraint on it. We will solve the problem in the same ways as before, either by fixing $\boldsymbol{\Sigma}$ or by assuming a proper prior distribution to $\boldsymbol{\Sigma}$. In either case we can see no obvious reason for treating the first latent variable differently (or to be more precise, it is far from clear exactly how one should treat it differently) and so we will typically fix $\boldsymbol{\Sigma} = \mathbf{I}$ or give $\boldsymbol{\Sigma}$ an inverse Wishart prior with $\mathbf{Q} \propto \mathbf{I}$.

The equivalent matrix normal distribution of the last equation can be expressed as

$$\mathbf{Z} - \mathbf{1}_n\boldsymbol{\alpha}' - \mathbf{XB} \sim MN(\mathbf{I}_n, \boldsymbol{\Sigma}). \tag{7.1}$$

We focus on the simple case where $M = 5$ and $|\mathbf{t}| = 4$, which requires two latent variables. The graphical representation for this is given in Figure 7.1, where responses 0 and 4 are nominal, and the sequence of 0, 1, 2 and 3 are ordinal in the sense that zero is also a part of the ordinal sequence and it is the 'baseline' category. If $Z_{i,2} > Z_{i,1}$ and $Z_{i,2} > 0$, then the response is $y_i = 4$ (blue region). Afterwards, if we know that $Z_{i,1} > Z_{i,2}$ and $Z_{i,1} > 0$, we are interested to discriminate the ordinal responses. For this purpose $Z_{i,1}$ axis of positive values split vertically in three trapezoids, since the splitting

should verify the last inequality. The boundary components k_2 and k_3 are responsible for the separation between the three ordinal responses which are the following: $y_i = 1$ (gray region), $y_i = 2$ (red region) and $y_i = 3$ (yellow region). Since zero is the 'baseline' category, if $Z_{i,2} > Z_{i,1}$ and $Z_{i,2} \leq 0$ or $Z_{i,1} > Z_{i,2}$ and $-\infty \leq Z_{i,1} \leq k_1$ ($k_1 = 0$), then the response is $y_i = 0$ (green region), which corresponds to zero as nominal response and also as a part of the ordinal sequence respectively. The contours indicate probability densities of the latent variables. To simplify the illustration we draw the contours centered at zero, though in general they will not be.

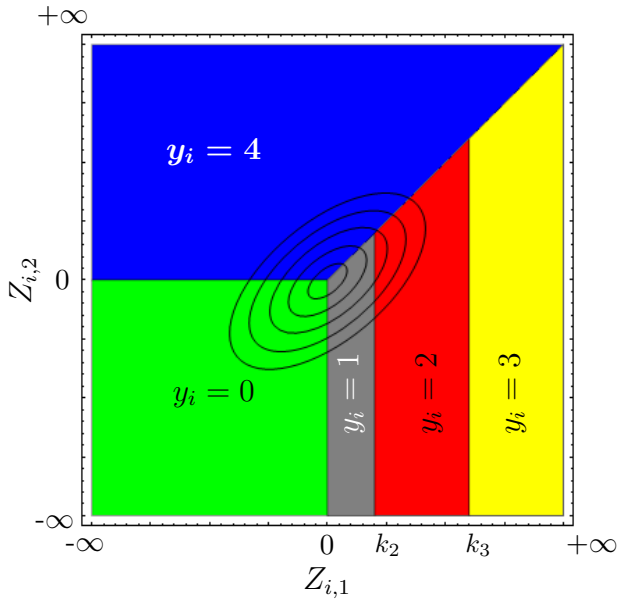


Figure 7.1: Graphical representation of the relationship between responses ($M = 5$, $|\mathbf{t}| = 4$) and two latent variables. Responses 0, 4 are nominal and the sequence of 0, 1, 2, and 3 are ordinal (zero is part of the ordinal sequence is also the 'baseline').

We denote by $Z_i^\star = \max_{1 \leq r \leq s}\{Z_{i,r}\}$. The relation between responses and the s latent variables is described in two steps. Firstly,

$$y_i = \begin{cases} 0, & \text{if } Z_i^\star \leq 0, \\ r + |\mathbf{t}| - 2, & \text{if } Z_i^\star > 0 \text{ and } Z_i^\star = Z_{i,r}. \end{cases} \tag{7.2}$$

Actually, Equation (7.2) distinguishes response zero from nominal responses and from the group of ordinal responses. Afterwards, knowing that the second part of this equation is true for $r = 1$, the relationship between ordinal responses and the corresponding latent variables $Z_{i,1}$ is given by

$$y_i = t_d, \text{ if } k_d < Z_{i,1} \leq k_{d+1}, \tag{7.3}$$

for $d = 0, \ldots, |\mathbf{t}| - 1$ where $\mathbf{k} = (k_0, k_1, \ldots, k_{|\mathbf{t}|})$, $k_0 < k_1 < \ldots < k_{|\mathbf{t}|}$, is the boundary vector for the ordinal responses with $k_0 = -\infty$ and $k_{|\mathbf{t}|} = +\infty$, and where we fix $k_1 = 0$ (zero is the 'baseline') for reasons of identifiability. It is important to note that the ordinal responses have a common latent variable and are specified via the boundary vector. Nominal responses each have a different latent variable. In summary, the relationship between all possible responses and the latent variables is given via Equations (7.2) and (7.3).

The multinomial probit model with a mixture of nominal and ordinal responses, based on the latent variables and the boundaries, can be expressed as

$$\begin{aligned} p_{i0} &= P(y_i = 0) = P(Z_i^\star \leq 0), \\ p_{it_d} &= P(y_i = t_d) = P(k_d < Z_{i,1} \leq k_{d+1} | Z_{i,1} = Z_i^\star), \\ p_{iv} &= P(y_i = v) = P(Z_i^\star > 0, Z_i^\star = Z_{i,v-|\mathbf{t}|+2}), \end{aligned}$$

where $v \in \{0 : (M-1)\} \backslash \{t_0, \ldots, t_{|\mathbf{t}|-1}\}$ and $\backslash$ denotes the relative complement.

In order to perform variable selection for the multi-class multivariate case with a mixture of response types, a common $p \times 1$ indicator vector $\boldsymbol{\xi}$ is used across different latent variables. In this chapter we will use a common indicator vector and in the next chapter we will use a different indicator vector for different regression equations. The j-th element of ξ_j is defined such that

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

where $B_{j,r}$ is the entry in the j-th row and r-th column of $\mathbf{B}$, for $r = 1, \ldots, s$. Selection of the j-th variable corresponds to $\xi_j = 1$, which requires all the coefficients of the j-th row of $\mathbf{B}$ to be nonzero.

Note that incorporating $\boldsymbol{\xi}$ into Equation (7.1) simplifies the algebra later on

$$\mathbf{Z} - \mathbf{1}_n \boldsymbol{\alpha}' - \mathbf{X}_{\boldsymbol{\xi}} \mathbf{B}_{\boldsymbol{\xi}} \sim MN(\mathbf{I}_n, \boldsymbol{\Sigma}). \tag{7.4}$$

In summary, the model and the dependence between the variables are presented in a directed graphical model (Figure 7.2). Circles denote random variables and squares constants. For example, w is a constant that comes from the prior knowledge of the indicator vector.

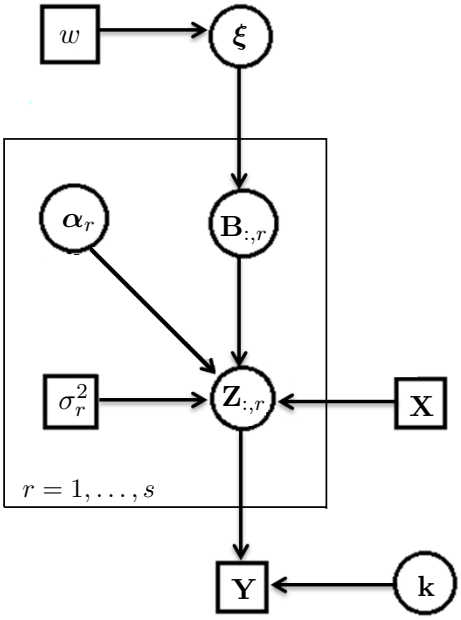


Figure 7.2: Directed graphical model for the probit model with latent variables and a common indicator vector. Circles denote random variables, squares constants. $\mathbf{Z}_{:,r}$ is the r-th column vector of latent variables, $\mathbf{B}_{:,r}$ is the r-th column vector of coefficients vector and $\sigma_r^2 = 1$.

7.1.2 Prior distributions

In this section the priors for the unknown parameters $\boldsymbol{\alpha}$, $\mathbf{B}_{\boldsymbol{\xi}}$, $\boldsymbol{\xi}$, $\mathbf{k}$ and $\boldsymbol{\Sigma}$ when it is not fixed are specified. The priors of the first three parameters are similar to those that are presented in Subsection 5.2.2. $k_2, k_3, \ldots, k_{|\mathbf{t}|-1}$ are uniformly distributed on the interval $(0, +\infty)$ subject to the constraint that $k_2 < k_3 < \ldots < k_{|\mathbf{t}|-1}$ ($k_0 = -\infty, k_{|\mathbf{t}|-1} = +\infty$ and $k_1 = 0$). When $\boldsymbol{\Sigma}$ is not fixed we assume $\boldsymbol{\Sigma} \sim IW(\delta; \mathbf{Q})$, where here $\delta = n - s + 1$ according to the parametrization of Dawid (1981).

7.1.3 Posterior inference

The posterior inference is carried out under two different settings for $\boldsymbol{\Sigma}$, either assigning a prior distribution to it (case A) or fixing to a specific value (case B), but inference follows very similar steps. First, the unknown parameters $\boldsymbol{\alpha}$, $\mathbf{B}_{\boldsymbol{\xi}}$ (and $\boldsymbol{\Sigma}$ if it is not fixed) are integrated out from the joint posterior. In the absence of prior information setting $\boldsymbol{\alpha}_0 = \mathbf{0}$ and $\mathbf{B}_{0\boldsymbol{\xi}} = \mathbf{0}$, inference is done via three Gibbs steps. Let us denote the sample of the j-th iteration with the upper index (j), and construct the Gibbs steps as summarized in Algorithm 3. Appendices D and E contain details about the algebra calculations under the two different settings for $\boldsymbol{\Sigma}$.

Algorithm 3 Gibbs sampling for mixture of responses using an indicator vector

0: Initialize values $\boldsymbol{\xi}^{(0)}, \mathbf{Z}^{(0)}$ and $\mathbf{k}^{(0)}$
1: Draw $\boldsymbol{\xi}^{(j)}$ from $p(\boldsymbol{\xi}|\mathbf{Z}^{(j-1)}, \mathbf{k}^{(j-1)}, \mathbf{X}, \mathbf{y})$
2: Draw $\mathbf{k}^{(j)}$ from $p(\mathbf{k}|\mathbf{Z}^{(j-1)}, \boldsymbol{\xi}^{(j)}, \mathbf{X}, \mathbf{y})$
3: Draw $\mathbf{Z}^{(j)}$ from $p(\mathbf{Z}|\boldsymbol{\xi}^{(j)}, \mathbf{k}^{(j)}, \mathbf{X}, \mathbf{y})$
4: Repeat steps 1, 2 and 3 until the maximum number of iterations is achieved and stop.

Case A: $\boldsymbol{\Sigma}$ has a distribution

Derivations for calculating the full conditional distributions are carried out after integrating out $\boldsymbol{\Sigma}$. The derivation of the first step of Gibbs sampling (Algorithm 3) is easy if the algebra calculations are first done for the third step of this algorithm.

The third step of Algorithm 3, sampling for the latent matrix $\mathbf{Z}$, is performed via

$$\mathbf{Z}|\boldsymbol{\xi}, \mathbf{k}, \mathbf{X}, \mathbf{y} \sim MT(\delta; \mathbf{P}_{\boldsymbol{\xi}}, \mathbf{Q}) \prod_{i=1}^{n} \mathbb{1}(\mathbf{Z}_{i,:} \in G_i), \tag{7.5}$$

where $\mathbf{P}_{\boldsymbol{\xi}} = \mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n^{'} + \mathbf{X}_{\boldsymbol{\xi}}\mathbf{H}_{\boldsymbol{\xi}}\mathbf{X}_{\boldsymbol{\xi}}^{'}$. The indicator function contains the truncations of the matrix Student distribution based on both boundaries and latent variables and it is given by

$$G_i = \begin{cases} \{\mathbf{Z}_{i,:} : Z_i^{\star} \leq 0\}, & \text{if } y_i = 0 \\ \{\mathbf{Z}_{i,:} : Z_i^{\star} = Z_{i,1} \text{ and } k_d < Z_{i,1} \leq k_{d+1}\}, & \text{if } y_i = t_d \\ \{\mathbf{Z}_{i,:} : Z_i^{\star} > 0 \text{ and } Z_i^{\star} = Z_{i,v-|\mathbf{t}|+2}\}, & \text{if } y_i = v. \end{cases} \tag{7.6}$$

Since it is difficult to sample directly from Equation (7.5), a Gibbs sampler can be applied (Geweke, 1991).

Now, sampling from the posterior distribution of $\boldsymbol{\xi}$ (first step of Algorithm 3) is implemented according to the factorization $p(\boldsymbol{\xi}|\mathbf{Z}, \mathbf{k}, \mathbf{X}, \mathbf{y}) \propto p(\boldsymbol{\xi})p(\mathbf{Z}|\boldsymbol{\xi})$. This factorization notes that $\boldsymbol{\xi}$ depends on $\mathbf{Z}$, $\mathbf{k}$, and $\mathbf{y}$, but once we condition on $\mathbf{Z}$, then $\mathbf{k}$ and $\mathbf{y}$ are independent of $\boldsymbol{\xi}$ (Figure 7.2). Then,

$$p(\boldsymbol{\xi}|\mathbf{Z}, \mathbf{k}, \mathbf{X}, \mathbf{y}) \propto p(\boldsymbol{\xi}) \left|\mathbf{I}_n + \mathbf{X}_{\boldsymbol{\xi}}\mathbf{H}_{\boldsymbol{\xi}}\mathbf{X}_{\boldsymbol{\xi}}^{'}(\mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n^{'})^{-1}\right|^{-\frac{s}{2}} \left|\mathbf{Q}_{\boldsymbol{\xi}}\right|^{-\frac{\delta+n+s-1}{2}} \tag{7.7}$$

where $\mathbf{Q}_{\boldsymbol{\xi}} = \mathbf{Q} + \mathbf{Z}^{'}\mathbf{G}^{-1}\mathbf{Z} - \mathbf{Z}^{'}\mathbf{G}^{-1}\mathbf{X}_{\boldsymbol{\xi}} \left[\mathbf{X}_{\boldsymbol{\xi}}^{'}\mathbf{G}^{-1}\mathbf{X}_{\boldsymbol{\xi}} + \mathbf{H}_{\boldsymbol{\xi}}^{-1}\right] \mathbf{X}_{\boldsymbol{\xi}}^{'}\mathbf{G}^{-1}\mathbf{Z}$, with $\mathbf{G} = \mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n^{'}$. Note that, if $\mathbf{X}$ is centered and h is large, then $\left|\mathbf{I}_n + \mathbf{X}_{\boldsymbol{\xi}}\mathbf{H}_{\boldsymbol{\xi}}\mathbf{X}_{\boldsymbol{\xi}}^{'}\right| \approx \left|\mathbf{I}_n + \mathbf{X}_{\boldsymbol{\xi}}\mathbf{H}_{\boldsymbol{\xi}}\mathbf{X}_{\boldsymbol{\xi}}^{'}(\mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n^{'})^{-1}\right|$. The full conditional distribution defined via

Equation (7.7) it is not easy to sample from, so we apply Metropolis within Gibbs to update $\boldsymbol{\xi}$ Brown et al. (1998a). This is very computationally intensive if p is large. For this reason, a QR-decomposition (Seber, 2000; Brown et al., 2002) is applied. The data augmentation for $\mathbf{H}_{\boldsymbol{\xi}} = c\mathbf{I}_{p_{\boldsymbol{\xi}}}$ ($p_{\boldsymbol{\xi}}$ is the number of selected variables) yields $\left|\mathbf{I}_n + \mathbf{X}_{\boldsymbol{\xi}}\mathbf{H}_{\boldsymbol{\xi}}\mathbf{X}_{\boldsymbol{\xi}}^{'}\right| = \left|\mathbf{H}_{\boldsymbol{\xi}}^{1/2}\mathbf{X}_{\boldsymbol{\xi}}^{'}\mathbf{X}_{\boldsymbol{\xi}}\mathbf{H}_{\boldsymbol{\xi}}^{1/2} + \mathbf{I}_n\right| = \left|\widetilde{\mathbf{X}}_{\boldsymbol{\xi}}^{'}\widetilde{\mathbf{X}}_{\boldsymbol{\xi}}\right|$, where $\widetilde{\mathbf{X}}$ is given via Equation (5.12). Then, $\mathbf{Q}_{\boldsymbol{\xi}} = \widetilde{\mathbf{Z}}^{'}\widetilde{\mathbf{Z}} - \widetilde{\mathbf{Z}}^{'}\widetilde{\mathbf{X}}_{\boldsymbol{\xi}}(\widetilde{\mathbf{X}}_{\boldsymbol{\xi}}^{'}\widetilde{\mathbf{X}}_{\boldsymbol{\xi}})^{-1}\widetilde{\mathbf{X}}_{\boldsymbol{\xi}}^{'}\widetilde{\mathbf{Z}}$, where

$$\widetilde{\mathbf{Z}} = \begin{pmatrix} \mathbf{Z} \\ \mathbf{0} \end{pmatrix}$$

is a $(n + p_{\boldsymbol{\xi}}) \times s$ augmented matrix.

Finally, the boundaries are related only to the ordinal responses. Following the basic idea of Albert and Chib (1993), the full conditional density of k_ν given the rest is given by

$$\begin{aligned} & p(k_\nu|\boldsymbol{\xi}, \mathbf{Z}, \mathbf{X}, \mathbf{y}, \mathbf{k}_{\setminus\nu}) \\ & \propto \prod_{i=1}^{n}[\mathbb{1}(y_i = t_{\nu-1})\mathbb{1}(k_{\nu-1} < Z_{i,1} \leq k_\nu) \\ & \quad + \mathbb{1}(y_i = t_\nu)\mathbb{1}(k_\nu < Z_{i,1} \leq k_{\nu+1})], \end{aligned} \tag{7.8}$$

where $\mathbf{k}_{\setminus\nu} = (k_0, \ldots, k_{\nu-1}, k_{\nu+1}, \ldots, k_{|\mathbf{t}|})$, for $\nu = 2, \ldots, |\mathbf{t}|-1$. In fact, the update of each boundary parameter can be implemented (using Equation (7.8)) via

$$k_\nu|\boldsymbol{\xi}, \mathbf{Z}, \mathbf{X}, \mathbf{y}, \mathbf{k}_{\setminus\nu} \sim U(\max[\{Z_{i,1} : y_i = t_{\nu-1}\}, k_{\nu-1}], \min[\{Z_{i,1} : y_i = t_\nu\}, k_{\nu+1}]). \tag{7.9}$$

Case B: $\boldsymbol{\Sigma}$ is fixed

In the case where $\boldsymbol{\Sigma}$ is taken to be fixed, the three Gibbs steps are derived for convenience in the same order as in the previous section. Initially, to sample the latent matrix $\mathbf{Z}$ we can use the relation

$$\mathbf{Z}|\boldsymbol{\xi}, \mathbf{k}, \mathbf{X}, \mathbf{y} \sim MN(\mathbf{P}_{\boldsymbol{\xi}}, \boldsymbol{\Sigma}) \prod_{i=1}^{n} \mathbb{1}(\mathbf{Z}_{i,:} \in G_i),$$

where G_i is given via Equation (7.6) and the mean of the latent matrix is the zero vector (it is not denoted here). A Gibbs sampler is used in order to take samples from the full conditional distributions of the truncated normal distribution (Geweke, 1991).

Sampling from the posterior distribution of $\boldsymbol{\xi}$, this time without integrat-

ing out $\boldsymbol{\Sigma}$ (which is fixed), is performed via

$$p(\boldsymbol{\xi}|\mathbf{Z},\mathbf{k},\mathbf{X},\mathbf{y}) \propto p(\boldsymbol{\xi}) \left|\mathbf{I}_n + \mathbf{X}_{\boldsymbol{\xi}}\mathbf{H}_{\boldsymbol{\xi}}\mathbf{X}_{\boldsymbol{\xi}}^{'}(\mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n^{'})^{-1}\right|^{-\frac{s}{2}} \cdot |\mathbf{P}_{\boldsymbol{\xi}}|^{-\frac{s}{2}} \exp\left\{-\tfrac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\mathbf{Z}'\mathbf{P}_{\boldsymbol{\xi}}^{-1}\mathbf{Z}\right]\right\},$$

where tr(.) denotes the trace of a square matrix. Again, this formula is not easy to sample from and a Metropolis algorithm (Brown et al., 1998a) is applied, within the Gibbs step, in order to generate posterior samples from $\boldsymbol{\xi}$. The Metropolis algorithm via QR fast updating is implemented in order to speed up the computations. Updating the boundaries of the ordinal responses follows exactly the same form as before (Equation (7.9)).

7.1.4 Classification and prediction

We focus on how to use the best model in order to make predictions. The least squares model prediction is given by Equation (5.14), where here $\hat{\mathbf{Z}}_f$ is the $n_f \times s$ estimated matrix of latent variables that it is related to a new (future) measurement ($\widetilde{\mathbf{B}}_{\hat{\boldsymbol{\xi}}}$ is $p \times s$ and $\hat{\mathbf{Z}}$ is $n \times s$). Then, the labels of future measurements, $\hat{\mathbf{y}}_f$ ($n_f \times 1$), can be predicted according to Equations (7.2) and (7.3).

7.1.5 Hyperparameter settings

The hyperparameters are specified as explained in Subsection 5.2.5. We assign a vague prior to the intercept selecting h to be large and fix $k_1 = 0$.

7.2 Simulation results

7.2.1 Simulations

An experimental study was performed using simulated data from the probit model with multi-class nominal and ordinal responses.

For the simulation study, we identify the indices of the important variables via the common indicator vector. The important variables then determine the nonzero coefficients of $\mathbf{B}$. In addition, we fix the number of variables (p), the number of samples (n), the sequence of ordinal responses $\mathbf{t}$ and the total number of responses (M).

We consider that $X_{i,j}$ are i.i.d. and $X_{i,j} \sim N(0,1)$ for $i = 1,\ldots,n$ and $j = 1,\ldots,p$. Then, for fixed $\mathbf{B}$, we construct $\mathbf{Z} = \mathbf{XB} + \boldsymbol{\epsilon}$, which is the matrix of latent variables where $\boldsymbol{\epsilon}$ is an $n \times s$ matrix, $\epsilon_{i,r}$ are i.i.d. and $\epsilon_{i,r} \sim N(0,\sigma^2)$

for $i = 1, \ldots, n$ and $r = 1, \ldots, s$. Focusing on the ordinal responses, we define $\mathbf{k}$ via $|\mathbf{t}| - 1$ quantiles of the corresponding latent variable. Based on the simulated $\mathbf{Z}$ and $\mathbf{k}$ we assign values to the response vector $\mathbf{y}$ with coding $0, \ldots, M-1$. The simulation process produces $\mathbf{X}$ and $\mathbf{y}$ that are required as inputs of the variable selection process.

We run one simulation to cover the following analysis scenarios: (i) $\boldsymbol{\Sigma}$ is fixed with $n \ll p$ and (ii) $\boldsymbol{\Sigma}$ has a distribution with $n \ll p$. In both cases, we set $n = 100$, $p = 200$, $M = 5$, $\mathbf{t} = [1, 2, 3]$ and $\sigma = 0.8$ for generating data. The majority of $\mathbf{B}$'s entries are zero except that $B_{[8,100,180],1} = [4, -6.5, 4]$ and $B_{[8,100,180],2} = [-5.5, 5, -2]$. This means that the two latent variables have the same important variables (coefficients 8, 100 and 180 are nonzero), because we use a common indicator vector for variable selection. To apply BVS, we assign values to parameters and hyperparameters separately in each scenario.

For scenario (i) variable selection is done assuming that $\boldsymbol{\Sigma}$ is equal to $\mathbf{I}_2$, which is not equal to $0.8\mathbf{I}_2$ as used in the simulation. In the variable selection approach we set $h = 10^6$, $\mathbf{H}_{\boldsymbol{\xi}} = c\mathbf{I}_{p_{\boldsymbol{\xi}}}$ with $c = 10$ and $w = 3/200$. We initialize the $\boldsymbol{\xi}^{(0)}$ randomly selecting three of the variables to be one and the rest are zero and the $\mathbf{k}^{(0)}$ by drawing the non-fixed boundaries uniformly in $(0, 1000)$. In addition, we initialize $\mathbf{Z}^{(0)}$ as described in Equation (6.5) for $r = 1, 2$.

We run four different chains with 30000 iterations after 20000 burn-in iterations. Averaging the four chains, our proposed algorithm correctly identifies the individually important variables 8, 100 and 180 (top of Figure 7.3) and the best model that consists of these variables (bottom of Figure 7.3). Note that the second, third, etc. best models have very small posterior probabilities compared to the posterior probability of the best model.

For the variable selection under scenario (ii) we set the values for the hyperparameters of the unknown covariance matrix $\boldsymbol{\Sigma}$, $\delta = 3$, $\mathbf{Q} = \mathbf{I}_2$ and the remaining parameters and hyperparameters are the same as in scenario (i). The results are similar to those reported in Figure 7.3 and in the previous paragraph.

7.2.2 Predictions

We generate some new data (test sets) according to the procedure that is described in Subsection 6.4.2 for scenario (ii). We pick from the test design matrix only variables 8, 100 and 180 that had been selected after applying Algorithm 3 and we make predictions for the one hundred new samples. The highest possible classification accuracy for this test set is 69% and our method achieves classification accuracy 66%. Table 7.1 contains the results

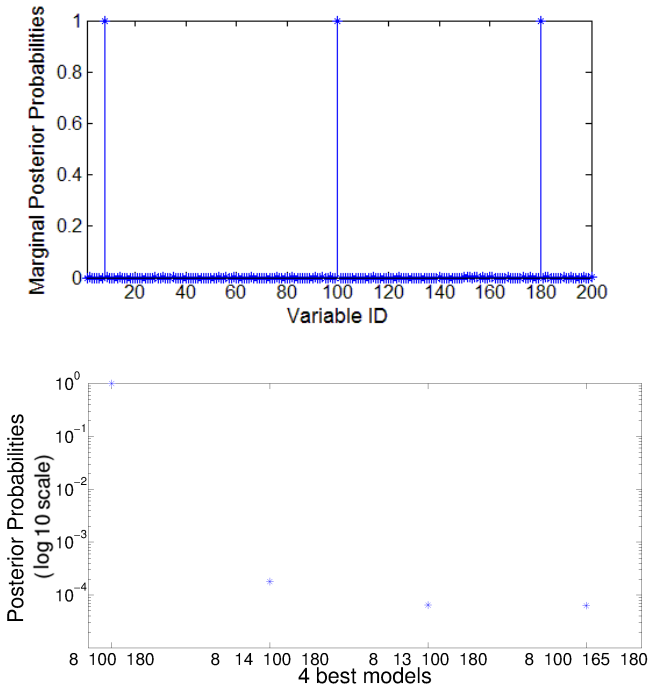


Figure 7.3: Marginal posterior probabilities (top) and posterior probabilities on a log scale (bottom) of average of chains for scenario (i).

of the comparison with some other methods (Section 5.1) that have been proposed for pure nominal or pure ordinal responses. Our approach beats the existing methods.

Table 7.1: Comparison of classification accuracy for the test set after applying different variable selection approaches.

	Accuracy (%)	
	Nominal	Ordinal
'Highest' accuracy	69	
Our proposed method	66	
LDA	60	41
LASSO	50	47
Classification trees	55	55
Random forests	59	42

7.3 Discussion and conclusion

We have described a Bayesian probit model for variable and model selection with mixture of ordinal and nominal responses and with a common indicator vector across latent variables. In this case latent vectors of nominal and ordinal responses are combined in one matrix in order to build a complete probit model. It is important to note that the latent variables of the ordinal responses have double utility: to discriminate ordinal from nominal responses

and also from each other. We derived efficient MCMC sampling for posterior inference. The computational advantage here is that we sample from the full conditional density of a single indicator vector (instead of two as we discussed in the previous chapter). The proposed algorithm is simple in the sense that the variable selection and prediction are each a one step process.

In this modelling process we consider a common indicator vector in variable selection, which discovers the same important variables for the rows of the coefficient matrix. However, in real problems sometimes different variables may be important for each of the nominal responses and for the sequence of ordinal responses. This requires different indicator vectors for each one of the nominal responses and one more indicator vector for the sequence of ordinal responses (or, equivalently, an indicator matrix). In the next chapter we will extend the current idea of variable selection proposing an indicator matrix. In practice this means that the important variables for the rows of coefficient matrix can differ.

In this study we consider that the shrinkage parameter and probability of success of the Bernoulli distribution are known. However, we could relax this assumption and assign prior distributions on them. Conjugate prior selection would use inverse Gamma and Beta distributions respectively.

Chapter 8

Bayesian variable selection in the probit model with a mixture of nominal and ordinal responses using an indicator matrix

In the previous chapter we proposed a BVS approach for a mixture of nominal and ordinal responses using a common indicator vector (assuming the same set of important variables across different responses).

However, in some applications we expect that different variables may be important for different responses. For this reason, in the current chapter, we propose a variable selection method using an indicator matrix (instead of the indicator vector), where each column corresponds to one latent variable (indicating the presence of the covariate). We apply our approach on simulated data and we compare the classification accuracy of this method with existing ones. This proposed BVS method using an indicator matrix has been published in a workshop (Kotti et al., 2016b).

8.1 Method

8.1.1 Model

The basic model is a probit model with a mixture of nominal and ordinal responses using latent variables (Equation (7.4)) and the relationship between responses and latent variables remains the same (Equations (7.2) and (7.3)).

As before, we use s latent variables. Let us assume a $\mathbf{Z}$ $(n \times s)$ matrix of latent variables, where each column is distributed as multivariate normal with mean zero and covariance matrix the identity. That is,

$$\mathbf{Z}_{:,r} = \mathbf{1}_n \alpha_r + \mathbf{X}\mathbf{B}_{:,r} + \mathbf{E}_{:,r}, \mathbf{E}_{:,r} \sim MVN(\mathbf{0}, \sigma_r^2 \mathbf{I}_n), \tag{8.1}$$

where $\mathbf{Z}_{:,r}$ is the r-th column vector of $\mathbf{Z}$, $\mathbf{B}_{:,r}$ is the r-th column vector of $\mathbf{B}$ $(p \times s)$ and σ_r^2 is the noise variance of the r-th latent variable, for $r = 1, \ldots, s$. The equivalent matrix normal distribution of the last equation can be expressed as

$$\mathbf{Z} - \mathbf{A} - \mathbf{X}\mathbf{B} \sim MN(\mathbf{I}_n, \mathbf{\Sigma}), \tag{8.2}$$

where $\mathbf{A} = (\mathbf{1}_n \alpha_1, \ldots, \mathbf{1}_n \alpha_s)$ and $\mathbf{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_s^2)$. The dimensions of those matrices are $n \times s$ and $s \times s$ respectively. We tackle the non-identifiability problem by fixing $\sigma_1^2, \ldots, \sigma_s^2$, typically with all $\sigma_r^2 = 1$.

In order to perform variable selection for the multi-class multivariate case with a mixture of response types, a $p \times s$ (s is the number of latent variables) indicator matrix $\mathbf{\Xi}$ is used, where each of its columns ($\mathbf{\Xi}_{:,r}$) represents the important variables for the r-th regression equation. The j-th element of the r-th latent variable of this matrix is denoted by $\Xi_{j,r}$ and is defined as

$$\Xi_{j,r} = \begin{cases} 1, & \text{if } B_{j,r} \neq 0, \\ 0, & \text{if } B_{j,r} = 0, \end{cases} \tag{8.3}$$

for $j = 1, \ldots, p$ and $r = 1, \ldots, s$. Selection of the j-th variable in the r-th regression equation corresponds to $\Xi_{j,r} = 1$.

The relationship between the aforementioned random variables is represented in Figure 8.1, which also includes the data and the fixed parameters of the proposed model.

Finally, note that incorporating $\mathbf{\Xi}_{:,r}$ into Equation (8.1) simplifies the algebra later on

$$\mathbf{Z}_{:,r} - \mathbf{1}_n \alpha_r - \mathbf{X}_{\mathbf{\Xi}_{:,r}} \mathbf{B}_{\mathbf{\Xi}_{:,r},r} \sim MVN(\mathbf{0}, \sigma_r^2 \mathbf{I}_n),$$

where $\mathbf{X}_{\mathbf{\Xi}_{:,r}}$ refers to those columns of $\mathbf{X}$ (out of p) that correspond to selected variables of the $r-$th regression equation and $\mathbf{B}_{\mathbf{\Xi}_{:,r},r}$ refers to those rows of $\mathbf{B}$ (out of p) that correspond to selected variables of the $r-$th regression equation and to the column of the $r-$th regression equation.

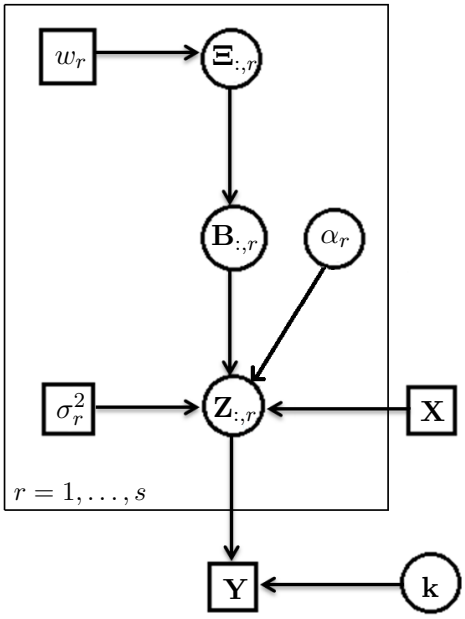


Figure 8.1: Graphical representation that indicates the relationship between random variables (circles) and data/constants (squares).

8.1.2 Prior distributions

In this section, the priors for the unknown parameters α_r, $\mathbf{B}_{\boldsymbol{\Xi}_{:,r},r}$, $\boldsymbol{\Xi}_r$ and $\mathbf{k}$ are presented for each one of the r regression equations.

The prior for the intercept is $\alpha_r - \alpha_{0r} \sim N(0, \sigma_r^2 h)$ and the prior for the nonzero coefficients, for different regression equations, is given by

$$\mathbf{B}_{\boldsymbol{\Xi}_{:,r},r} - \mathbf{B}_{0\boldsymbol{\Xi}_{:,r},r} \sim MVN(\mathbf{0}, \sigma_r^2 \mathbf{H}_{\boldsymbol{\Xi}_{:,r}}). \tag{8.4}$$

We generalize the form of the covariance matrix in Equation (8.4) from the case of common $\boldsymbol{\xi}$ to the case of different $\boldsymbol{\xi}$'s across different regression equations. We extend the specification of the covariance matrix (Brown et al., 2002) to the full g-prior $\mathbf{H}_{\boldsymbol{\Xi}_{:,r}} = c_r(\mathbf{X}'_{\boldsymbol{\Xi}_{:,r}}\mathbf{X}_{\boldsymbol{\Xi}_{:,r}})^+$, or to the diagonalized version of it $\mathbf{H}_{\boldsymbol{\Xi}_{:,r}} = c_r \operatorname{diag}\{(\mathbf{X}'_{\boldsymbol{\Xi}_{:,r}}\mathbf{X}_{\boldsymbol{\Xi}_{:,r}})^+\}$, where the 'plus' denotes the pseudoinverse matrix, or, for even greater simplicity, $\mathbf{H}_{\boldsymbol{\Xi}_{:,r}} = c_r \mathbf{I}_{p_{\boldsymbol{\Xi}_{:,r}}}$ ($p_{\boldsymbol{\Xi}_{:,r}}$ is the number of important variables that correspond to the r-th regression equation). We assume that the columns of the indicator matrix are independent and $\Xi_{j,r} \sim \text{Bernoulli}(w_r)$, with different probability of success across different nominal responses and the same probability of success across ordinal responses. As in the previous studies, we assume the the boundaries $k_2, k_3, \ldots, k_{|\mathbf{t}|-1}$ are uniformly distributed on the interval $(0, +\infty)$ subject to the constraint that $k_2 < k_3 < \ldots < k_{|\mathbf{t}|-1}$, and with k_1 fixed at 0.

8.1.3 Posterior inference

In our model, the unknown parameters α_r and $\mathbf{B}_{\boldsymbol{\Xi}_{:,r},r}$ can be integrated out from the joint posterior, which significantly simplifies computations. In the absence of prior information we set $\alpha_{0r} = 0$ and $\mathbf{B}_{0\boldsymbol{\Xi}_{:,r},r} = \mathbf{0}$. Inference is done via three Gibbs steps (details in Appendix F). We denote the sample of the j-th iteration with the upper index (j), and construct the Gibbs steps as summarized in Algorithm 4.

Algorithm 4 Gibbs sampling for mixture of responses using an indicator matrix

0: Initialize values $\boldsymbol{\Xi}^{(0)}, \mathbf{Z}^{(0)}$ and $\mathbf{k}^{(0)}$
1: Draw $\boldsymbol{\Xi}^{(j)}$ from $p(\boldsymbol{\Xi}|\mathbf{Z}^{(j-1)}, \mathbf{k}^{(j-1)}\mathbf{X}, \mathbf{Y})$
2: Draw $\mathbf{k}^{(j)}$ from $p(\mathbf{k}|\mathbf{Z}^{(j-1)}, \boldsymbol{\Xi}^{(j)}, \mathbf{X}, \mathbf{Y})$
3: Draw $\mathbf{Z}^{(j)}$ from $p(\mathbf{Z}|\boldsymbol{\Xi}^{(j)}, \mathbf{k}^{(j)}, \mathbf{X}, \mathbf{Y})$
4: Repeat steps 1, 2 and 3 until the desired number of iterations is achieved and stop.

Sampling the latent matrix $\mathbf{Z}$ is performed via

$$\mathbf{Z}|\boldsymbol{\Xi}, \mathbf{k}, \mathbf{X}, \mathbf{Y} \sim \prod_{r=1}^{s} \Big[MVN(\mathbf{0}, \mathbf{P}_{\boldsymbol{\Xi}_{:,r}})\Big] \prod_{i=1}^{n} \mathbb{1}(\mathbf{Z}_{i,:} \in G_i), \tag{8.5}$$

where $\mathbf{P}_{\boldsymbol{\Xi}_{:,r}} = \mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n^{'} + \mathbf{X}_{\boldsymbol{\Xi}_{:,r}}\mathbf{H}_{\boldsymbol{\Xi}_{:,r}}\mathbf{X}_{\boldsymbol{\Xi}_{:,r}}^{'}$ and the set G_i is given via Equation (7.6). Since it is difficult to sample directly from Equation (8.5), a Gibbs sampler can be applied for each latent variable in turn (Geweke, 1991).

Sampling from the posterior distribution of $\boldsymbol{\Xi}$ is performed via

$$\begin{aligned} p(\boldsymbol{\Xi}|\mathbf{Z}, \mathbf{k}, \mathbf{X}, \mathbf{Y}) \propto \prod_{r=1}^{s} \Big[& p(\boldsymbol{\Xi}_{:,r}) \left|\mathbf{I}_n + \mathbf{X}_{\boldsymbol{\Xi}_{:,r}}\mathbf{H}_{\boldsymbol{\Xi}_{:,r}}\mathbf{X}_{\boldsymbol{\Xi}_{:,r}}^{'}(nh + \sigma_r^2)^{-1}\right|^{-\frac{1}{2}} \cdot \\ & \left|\mathbf{P}_{\boldsymbol{\Xi}_{:,r}}\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left[\mathbf{Z}_{:,r}^{'}\mathbf{P}_{\boldsymbol{\Xi}_{:,r}}^{-1}\mathbf{Z}_{:,r}\right]\right\}\Big]. \end{aligned} \tag{8.6}$$

The Metropolis algorithm of Brown et al. (1998a) is applied within the Gibbs step, since the last expression is not easy to sample from directly. Similarly to the case of common $\boldsymbol{\xi}$, here we apply r times a QR decomposition to avoid squaring (Seber, 2000; Brown et al., 2002). The data augmentation for the $\mathbf{H}_{\boldsymbol{\Xi}_{:,r}} = c_r\mathbf{I}_{p_{\boldsymbol{\Xi}_{:,r}}}$, $\mathbf{X}$ centered and large h yields $\left|\mathbf{I}_n + \mathbf{X}_{\boldsymbol{\Xi}_{:,r}}\mathbf{H}_{\boldsymbol{\Xi}_{:,r}}\mathbf{X}_{\boldsymbol{\Xi}_{:,r}}^{'}\right| = \left|\mathbf{H}_{\boldsymbol{\Xi}_{:,r}}^{1/2}\mathbf{X}_{\boldsymbol{\Xi}_{:,r}}^{'}\mathbf{X}_{\boldsymbol{\Xi}_{:,r}}\mathbf{H}_{\boldsymbol{\Xi}_{:,r}}^{1/2} + \mathbf{I}_n\right| = \left|\widetilde{\mathbf{X}}_{\boldsymbol{\Xi}_{:,r}}^{'}\widetilde{\mathbf{X}}_{\boldsymbol{\Xi}_{:,r}}\right|$, where

$$\widetilde{\mathbf{X}}_{\boldsymbol{\Xi}_{:,r}} = \begin{pmatrix} \mathbf{X}_{\boldsymbol{\Xi}_{:,r}}\mathbf{H}_{\boldsymbol{\Xi}_{:,r}}^{1/2} \\ \mathbf{I}_{p_{\boldsymbol{\Xi}_{:,r}}} \end{pmatrix} \tag{8.7}$$

is a $(n + p_{\boldsymbol{\Xi}_{:,r}}) \times p_{\boldsymbol{\Xi}_{:,r}}$ matrix. Then,

$$\mathbf{Q}_{\boldsymbol{\Xi}_{:,r}} = \widetilde{\mathbf{Z}}_{:,r}^{'}\widetilde{\mathbf{Z}}_{:,r} - \widetilde{\mathbf{Z}}_{:,r}^{'}\widetilde{\mathbf{X}}_{\boldsymbol{\Xi}_{:,r}}(\widetilde{\mathbf{X}}_{\boldsymbol{\Xi}_{:,r}}^{'}\widetilde{\mathbf{X}}_{\boldsymbol{\Xi}_{:,r}})^{-1}\widetilde{\mathbf{X}}_{\boldsymbol{\Xi}_{:,r}}^{'}\widetilde{\mathbf{Z}}_{:,r},$$

where $\widetilde{\mathbf{Z}}_{:,r}$ is given by

$$\widetilde{\mathbf{Z}}_{:,r} = \begin{pmatrix} \mathbf{Z}_{:,r} \\ \mathbf{0} \end{pmatrix}$$

which is a $(n + p_{\boldsymbol{\Xi}_{:,r}}) \times 1$ augmented vector.

Finally, the full conditional distribution for each component of the boundaries is uniform (Equation (7.9)), since it is not affected by whether the variable selection is carried out via a common indicator vector or an indicator matrix.

8.1.4 Classification and prediction

The model prediction for the r-th regression equation, based on the best model ($\hat{\boldsymbol{\Xi}}_{:,r}$), is given by

$$\hat{\mathbf{Z}}_{f:,r} = \mathbf{1}_{n_f}\widetilde{\alpha}_r + \mathbf{X}_{f\hat{\boldsymbol{\Xi}}_{:,r}}\widetilde{\mathbf{B}}_{\boldsymbol{\Xi}_{:,r},r}, \tag{8.8}$$

where $\hat{\mathbf{Z}}_{f:,r}$ is a $n_f \times 1$ estimated vector of latent variables that is related to the r-th regression equation, $\widetilde{\mathbf{B}}_{\boldsymbol{\Xi}_{:,r},r} = \left(\mathbf{X}_{\hat{\boldsymbol{\Xi}}_{:,r}}^{'}\mathbf{X}_{\hat{\boldsymbol{\Xi}}_{:,r}} + \mathbf{H}_{\hat{\boldsymbol{\Xi}}_{:,r}}^{-1}\right)^{-1}\mathbf{X}_{\hat{\boldsymbol{\Xi}}_{:,r}}^{'}\hat{\mathbf{Z}}_{:,r}$ and $\widetilde{\alpha}_r = \bar{\hat{\mathbf{Z}}}_{:,r}$. Alternatively, instead of using just the one best model for each regression equation, the average of the best models can be used for each regression equation, which may improve the classification accuracy. In both cases, combining all estimated column vectors of latent variables, we can determine the estimated matrix of latent variables and we can make predictions according to Equations (7.2) and (7.3).

8.2 Simulation Results

8.2.1 Simulations

The experimental study was performed using simulated data from the probit model with multi-class nominal and ordinal responses using an indicator matrix for the purpose of variable selection.

The multi-class simulation study is similar to the one that we have already described in Subsection 7.2.1, except that here we can select different important variables for each regression equation. We set $n = 100$, $p = 476$,

$M = 6$, the sequence of responses $0, 1, 2, 3$ is ordinal and $\sigma_r = 1$, so that $\boldsymbol{\Sigma} = \mathbf{I}_3$, for generating simulated data with $n \ll p$. The majority of $\mathbf{B}$'s entries are zero except that $B_{[5,15],1} = [4, -5]$ (related to the ordinal responses), $B_{[10,155],2} = [5.5, 4]$ and $B_{[50,300],3} = [4.5, 4]$ (related to the nominal responses).

To apply variable selection we set in the analysis $\boldsymbol{\Sigma} = \mathbf{I}_3$, $h = 10^6$, $w_r = 2/476$, $\mathbf{H}_{\boldsymbol{\Xi}_{:,r}} = c_r \mathbf{I}_{p_{\boldsymbol{\Xi}_{:,r}}}$ and $c_r = 10$ for $r = 1, 2, 3$. We initialize $\boldsymbol{\Xi}^{(0)}$ selecting randomly two variables for each regression equation. Then, we initialize $\mathbf{Z}^{(0)}$ and $\mathbf{k}^{(0)}$ (details in Subsection 7.2.1). We run four different chains with 5000 iterations after 2000 burn-in iterations. Figure 8.2 contains the results of variable and model selection for the average of the chains. Our proposed algorithm correctly identifies the individually important variables for each regression equation. The rest of the variables have marginal posterior probabilities close to zero. In addition, the posterior probabilities of the best model are much higher than the second, third, etc., best models for each regression equation (bottom of Figure 8.2).

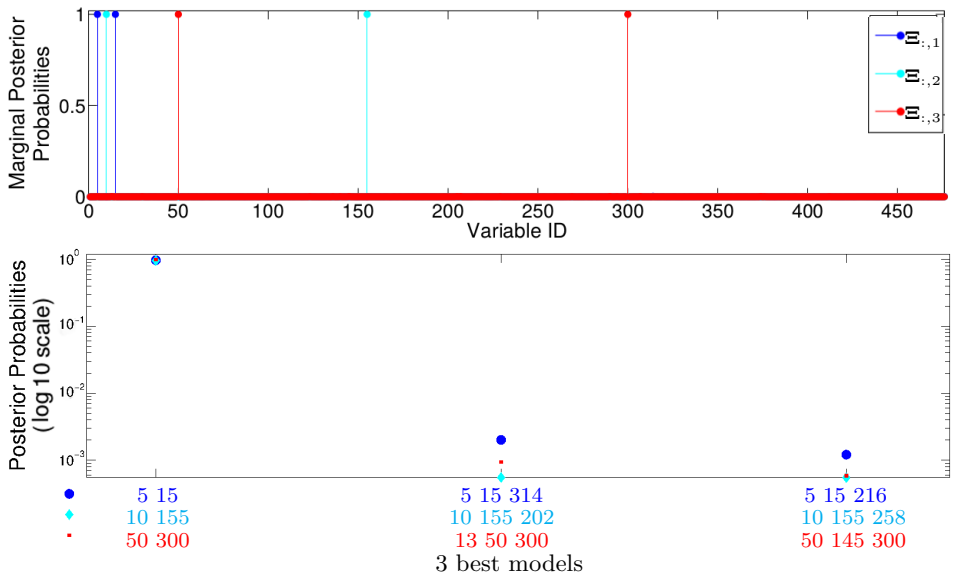


Figure 8.2: Marginal posterior probabilities (top) and posterior probabilities on a log scale (bottom) for BVS approach using an indicator matrix.

8.2.2 Predictions

In order to evaluate the best model, we generate a test set (as for the training set) with $n = 200$. For the test set, in order to generate the ordinal responses, we use the same boundaries as on the training set. Then, from the test design matrix, we pick only the variables that our approach has selected. Afterwards, we repeat the process of generating test sets 100 times and the classification accuracy of the test sets is on average 84.04%. Based on the amount of error

Table 8.1: Comparison of classification accuracy for one test set after applying different variable selection approaches.

	Accuracy (%)	
	Nominal	Ordinal
'Highest' accuracy	90	
The proposed method	83	
Penalized LDA	51	26
LASSO	65	26
Classification trees	45	21
Random forests	53	24

that the simulated data have, the highest classification accuracy that we can achieve is 90.50%. So, the proposed method produces a high classification accuracy on the test set, in the sense that it is very close to the highest possible. To compare our method with existing ones, we randomly select one test set. Table 8.1 contains the results of the comparison with some other methods that have been proposed for pure nominal or pure ordinal responses. In all cases the proposed method beats existing methods.

8.3 Discussion and conclusion

We have presented a Bayesian probit model for variable and model selection with a mixture of ordinal and nominal responses and with different indicator vectors across different latent variables. To build the appropriate model, we use different latent variables for the nominal responses and a common latent variable for the ordinal responses, where the last one has a double role as we discussed in the previous chapter. With respect to variable selection, using different indicator vectors allows us to identify different important variables for different regression equations, which is an advantage. This is a more realistic scenario for real applications, compared to the case of using common indicator vector. In the extreme case that the important variables are the same for all the different latent variables, then we could drop to the simple case of using a common indicator vector instead of the indicator matrix.

The new algorithm remains simple and efficient, because the indicator vectors are independent and we can sample from them in parallel at each step. This is an advantage, since sampling from indicator vectors is the most time-consuming step of our algorithm. In addition, we can sample from the latent variables in parallel too.

In this study we consider that c_r and w_r are known for each regression equation and indicator vector respectively. An advantage here is that the

shrinkage and the prior probability of success can be different across different regression equations, which is a realistic scenario. Alternatively, we could assign conjugate priors to them (inverse Gamma and Beta distribution respectively), at the expense of more computation.

Chapter 9

Application to Barrett's oesophagus (BE) for clinical diagnosis

In Chapters 6, 7 and 8 we proposed three BVS methods, all based on latent variables, with the aim of improving classification accuracy. In this chapter we explore the BE disease (intercepted-matched dataset) as a classification problem in which we are interested in finding a good classification rule as well as identifying the variables most important in diagnosing the stage of the disease. Our results of BVS by using an indicator matrix with an application to BE have been published in an international conference (Kotti et al., 2016a).

9.1 Data description and pre-processing

In the BE application FTIR spectra, which measure IR-absorbance over the range of wavenumbers as described in Section 1.2, are recorded on tissue samples. Standard data pre-processing was carried out on the spectra to improve the signal quality of the BE dataset. This involved liquid water subtraction, water vapour subtraction, normalization by Amide II band, spectral smoothing by a 13pt point Savitzky-Golay second derivative filter (Savitzky and Golay, 1964) and cropping the noisy part above wavenumber 1760 cm^{-1}.

Each patient has a unique ID number, the patient ID. This enables the identification of each spectrum and biopsy with an individual patient. A unique ID that identifies spectra and biopsies is also available, as there are multiple biopsies taken from different positions of the oesophagus in each patient and multiple spectra per biopsy. The multiple spectra of the same biopsy were averaged to produce a single spectrum per biopsy. Histopathol-

ogists labelled each biopsy as belonging to one of the five possible classes: healthy (H), Barrett's oesophagus type 1 (BE1), Barrett's oesophagus type 2 (BE2), Barrett's oesophagus type 3 (BE3) and cancer (C). We use this notation here because it is concise and highlights the mixed nominal/ordinal structure. The correspondence with the UK classification (Chapter 1) is: H is SQ, BE1 is NDBE, BE2 is LGD, BE3 is HGD, and C is OAC. Our dataset contains 309 spectra (samples) from 103 patients with one (possibly averaged spectrum) per biopsy. Where there are patients with multiple biopsies with different diagnoses, we refer to this as multi-labelled diagnosis. We split the dataset, with two thirds being the training set and one third being the test set such that all the spectra of one patient are in the same set. We would like to test our model on single-labelled diagnoses. Towards this goal, we randomly select single-labelled diagnoses, i.e. patients all of whose biopsies have the same diagnosis, to construct the test set. Then, the training set consists of whatever remains of the single-labelled diagnoses as well as all of the multi-labelled ones. Both sets now contain different patients and this is a realistic scenario, since actually the test set contains only unseen (for training) patients.

Table 9.1 presents a summary of BE dataset for the training, test, and the entire dataset. Healthy and early stage patients are more frequent than patients that are in late stages of the disease, which leads to an unbalanced dataset. This is typical in clinical diagnosis of diseases. In addition, it is very difficult to histologically distinguish BE2 from the remaining middle stages. This is one of the reasons that our dataset includes so few BE2 spectra (Figure 9.1).

Table 9.1: Summary of BE training set, test set and entire dataset.

	Training set		Test set		Entire dataset	
	Patients	Spectra	Patients	Spectra	Patients	Spectra
H	27	47	14	24	41	71
BE1	27	92	8	45	35	137
BE2	5	8	2	4	7	12
BE3	9	31	1	16	10	47
C	7	29	3	13	10	42
Total	75	207	28	102	103	309

Figure 9.2 contains barplots showing the number of spectra per patient for the training and test sets. All the biopsies in the test set come from patients with a single diagnosis. The training set has 46 such patients and the test set has 28 such patients. The total number of patients in the training set is greater than 46 (Table 9.1) because it includes patients that have multi-

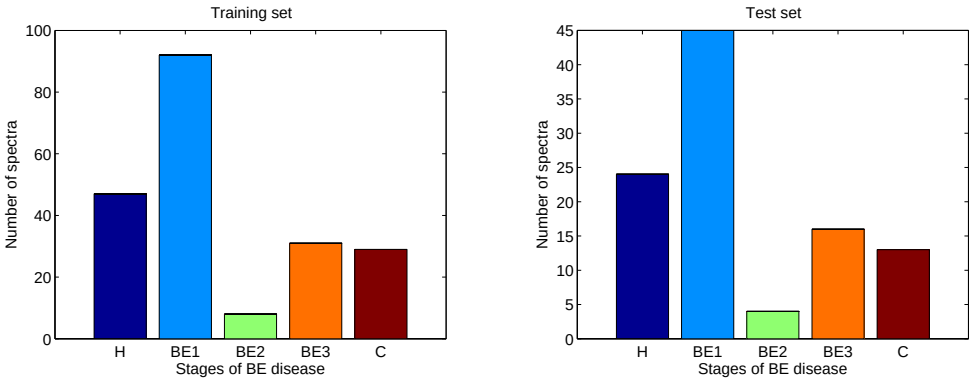


Figure 9.1: Barplots of the number of spectra versus the five possible labels on the training and test set.

labelled diagnosis. Counting all the spectra and assigning to each patient in a specific stage a number, we also visualize the distribution of the spectra for each patient in the training set in each stage of the BE disease (Figure 9.3).

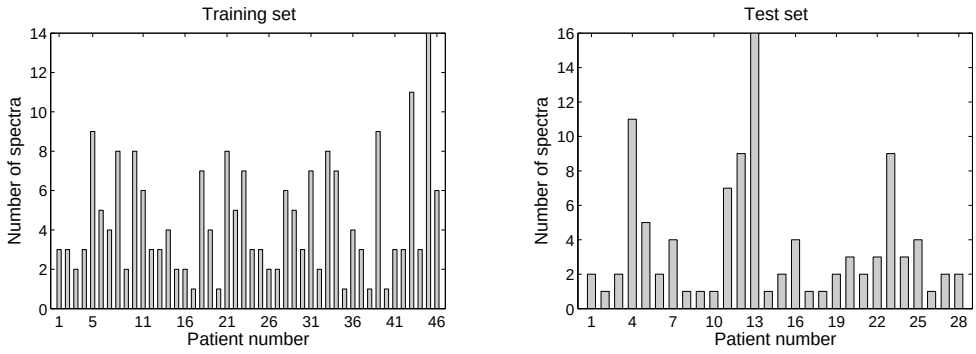


Figure 9.2: Barplots of the number of spectra versus the patient ID for the training and test set.

Further details about the spectra distribution of each patient is given at the top of Figure 9.4. The training set contains 46 patients with either unique or multi-labeled spectra. Different colors correspond to different labels/stages of the disease and the height of the bar indicates how many spectra each patient has. For example, patient one has three spectra, two of them are healthy and one is BE2. However, patient two has also three spectra, and all of them are diagnosed as BE1. The case where the same patient has spectra that correspond to three different labels (e.g. patient forty three) is rare. The test set has 28 patients with all of their spectra having only one label (bottom of Figure 9.4).

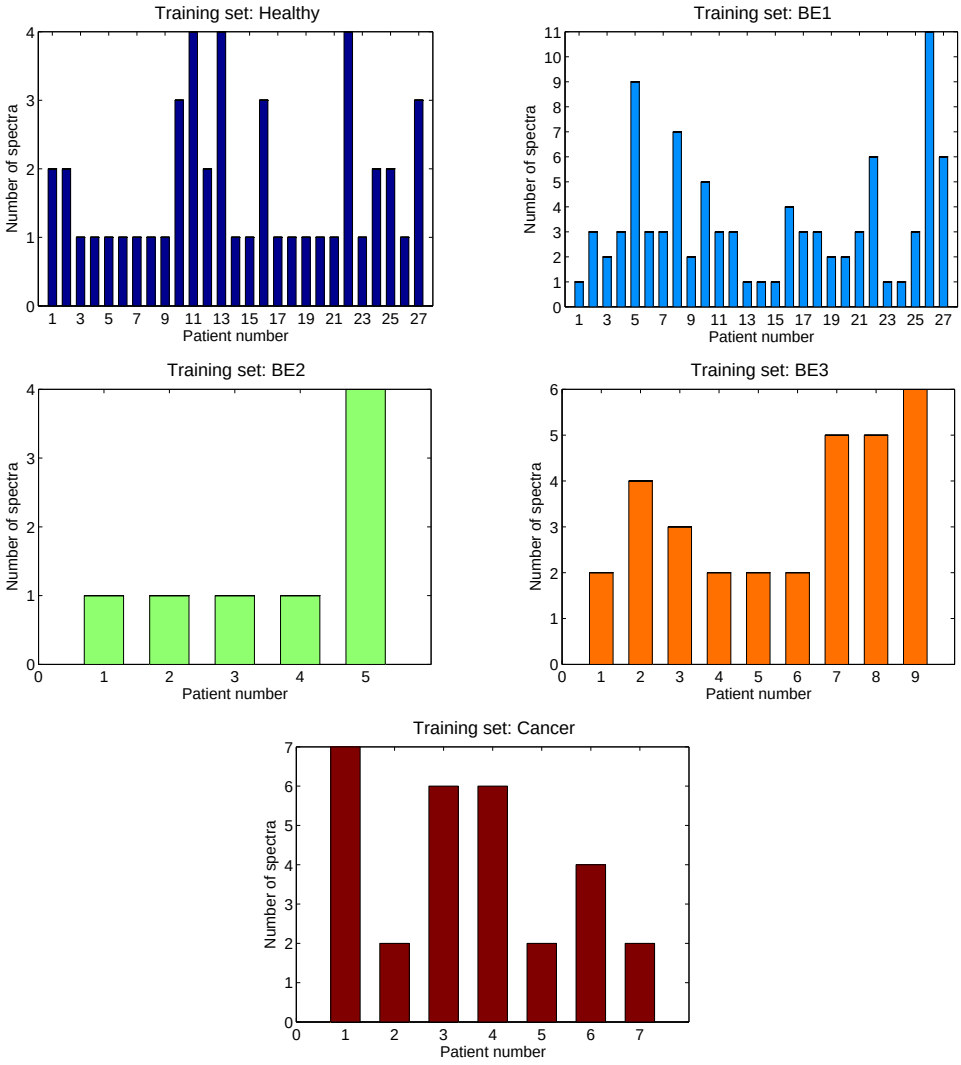


Figure 9.3: Barplots of the number of spectra per patient for each stage of the BE disease (H, BE1, BE2, BE3, C) for the training set.

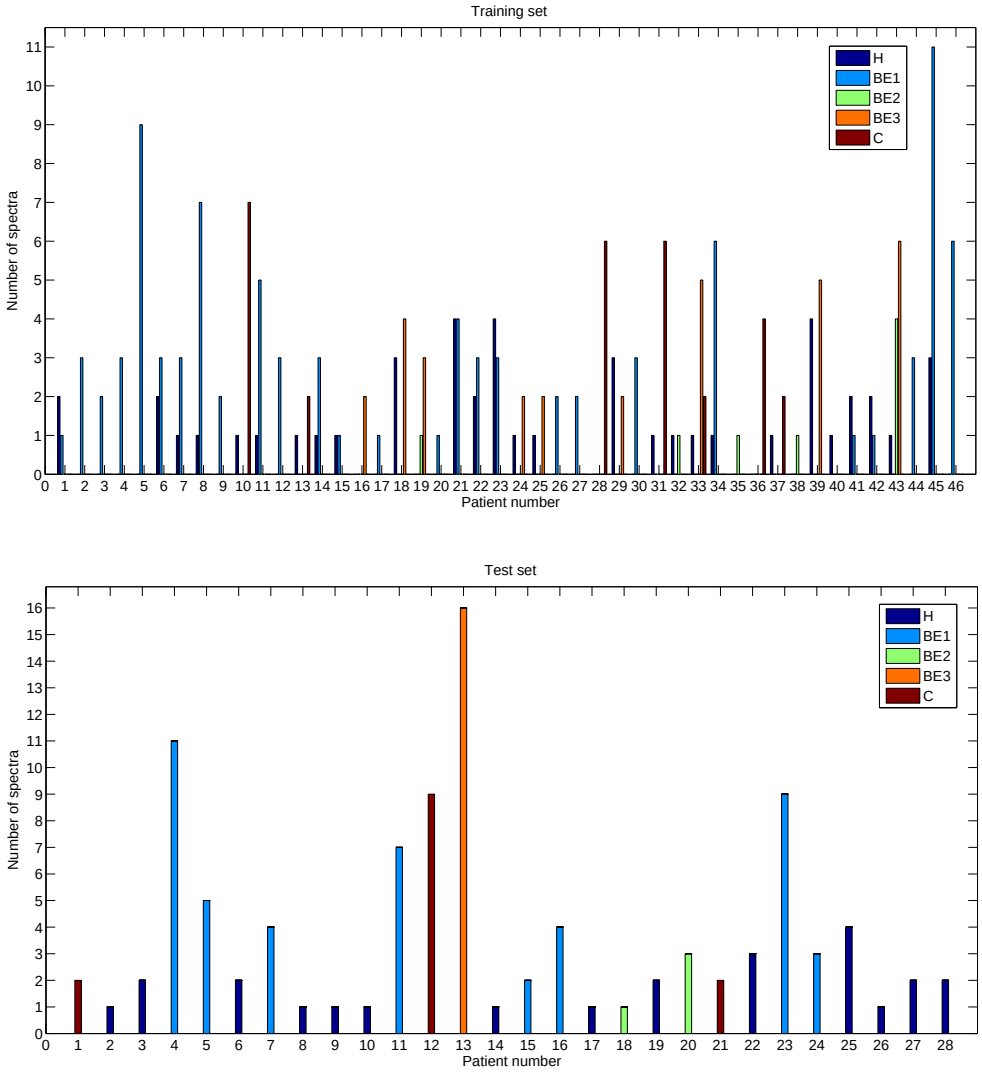


Figure 9.4: Number of spectra for each patient with the diagnoses shown for the training (top) and test (bottom) set.

9.2 Visualizing important variables for BE diagnosis

Initially, we plot the spectra with the aim of visually spotting the differences between spectra from different stages of the BE disease. These differences may suggest diagnostic variables and associated wavenumbers that could be used as potential biomarkers, or to improve the classification accuracy compared with using all variables. We will compare spectra from the simple case of the first versus the last stage and then between the middle stages, and finally we present more complex comparisons between those stages. The more complex comparisons of spectral differences between multiple labels correspond to different approaches to find potential important variables in each of

our methods.

Figure 9.5 contrasts the mean of the H versus the mean of the C spectra. Some potential diagnostic wavenumbers that may be able to distinguish H from C spectra are $1358, 1263, 1226,\ 1207, 1172, 1155, 1024$ and $995\ \text{cm}^{-1}$ (corresponding variable ID: $209, 258,\ 277, 287,\ 305, 314, 382$ and 397).

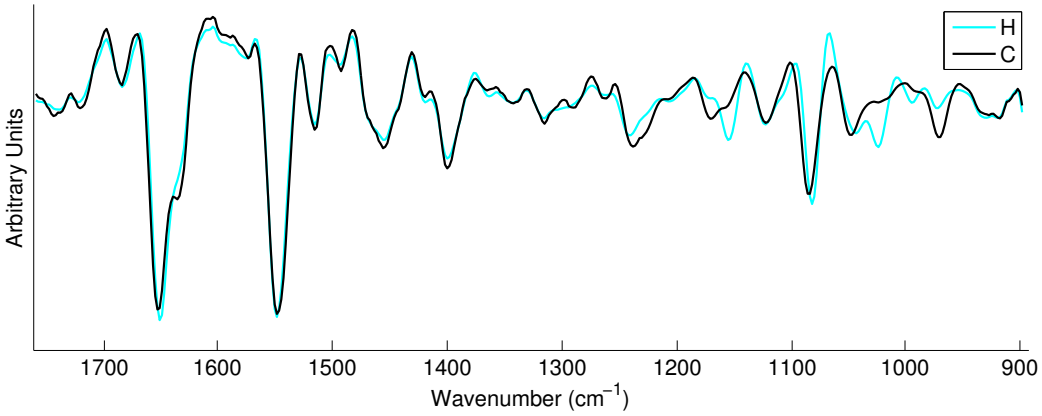


Figure 9.5: Mean second derivative spectra comparing H versus C on the training set.

Figure 9.6 compares the mean spectra for BE1 versus the means of BE2 and BE3. Some potential diagnostic variables that may be able to distinguish the spectra of the three stages are at wavenumbers 1315, 1273, 1261, 1213, 1161, 1082, 1061, 1043, 1030, 1007 and $972\ \text{cm}^{-1}$ (ID: 231, 253, 259, 284, 311, 352, 363, 372, 379, 391 and 409). The peaks and troughs are more than one variable wide, and may be represented by different selected variables.

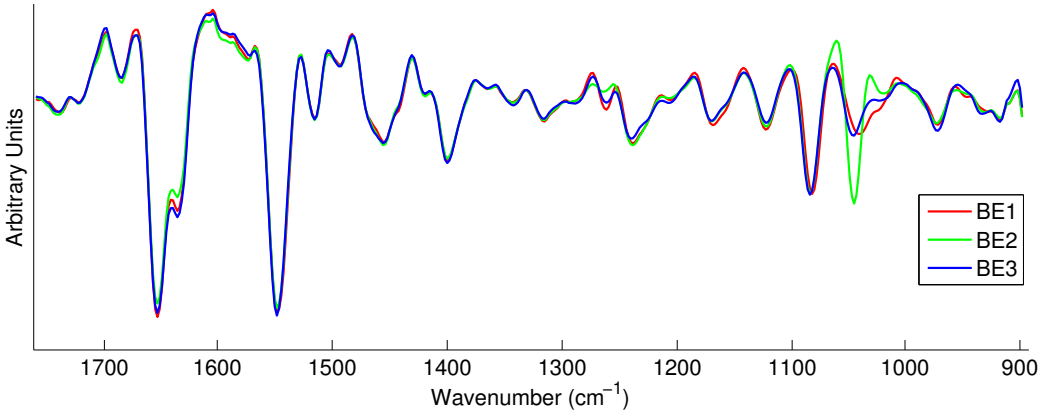


Figure 9.6: Mean second derivative spectra comparing the three BE progression stages on the training set.

Then we can combine the above mentioned results in three different ways:

(i) BE1 versus BE2 versus BE3 (Figure 9.6) and H versus BE1+BE2+BE3 versus C (Figure 9.7), where by BE1+BE2+BE3 we denote the merged

group of the three stages. Some common potential important variables of H versus BE1+BE2+BE3 versus C may be at wavenumbers 1635, 1357, 1292, 1263, 1172, 1155, 1035 and 977 cm^{-1} (ID: 65, 209, 243, 258, 305, 314, 376 and 406) and in general those variables differ from the important variables of BE1 versus BE2 versus BE3 noted above apart from the variable with ID 258,

(ii) H versus C (Figure 9.5) and H versus BE1 versus BE2 versus BE3 (Figure 9.8). In the latter case, differences between the spectra are noted at wavenumbers 1267, 1171, 1155, 1024, 995 and 947 cm^{-1} (ID: 256, 306, 314, 382, 397 and 422). Potentially common important variables are at wavenumbers 1267, 1171, 1155 and 1024 cm^{-1} (ID: 256, 306, 314 and 382), in the sense that they are so close that they belong to the same peak/trough, and

(iii) the same comparison as in (ii) but here we are interested in finding the potentially different important variables for each stage of the disease. According to Figure 9.9 those can be H: wavenumbers 1635, 1263, 1172, 1155, 1024 and 995 cm^{-1} (ID: 65, 258, 305, 314, 382 and 397), BE1: wavenumbers 1263, 1165 and 1122 cm^{-1} (ID: 285, 313 and 382), BE2: wavenumbers 1157, 1020 and 1032 cm^{-1} (ID: 313, 384 and 378), BE3: wavenumbers 1155, 1032 and 974 cm^{-1} (ID: 314, 384, and 408), and C: wavenumbers 1288, 1157 and 970 cm^{-1} (ID: 245, 313 and 410). We note that H, BE1, BE2, and BE3 have common variables with ID 314 and 384.

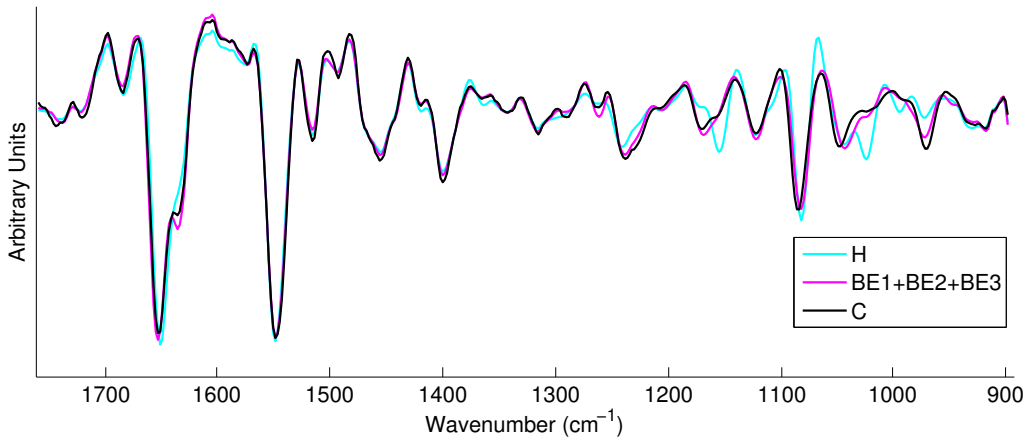


Figure 9.7: Mean second derivative spectra comparing the H versus BE1+BE2+BE3 versus C on the training set.

These variables are just indications and they will not necessarily appear among the important variables that belong to the best model though it will

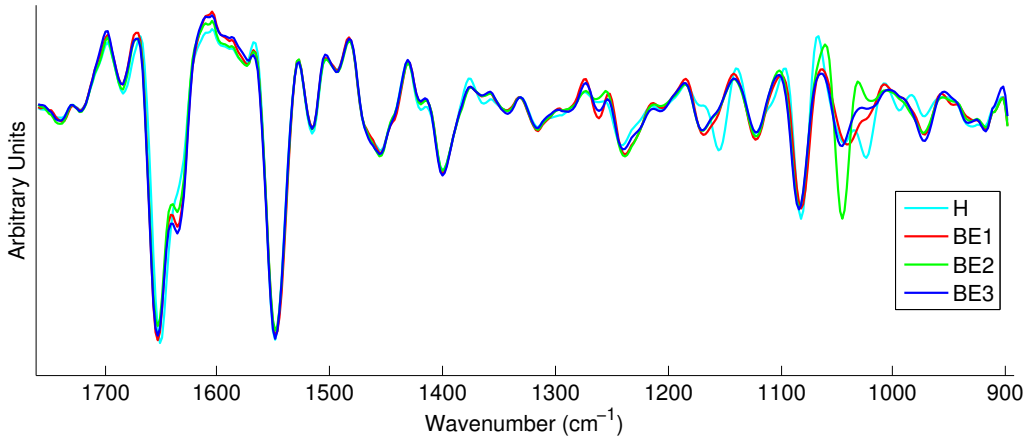


Figure 9.8: Mean second derivative spectra comparing H versus BE1 versus BE2 versus BE3 on the training set.

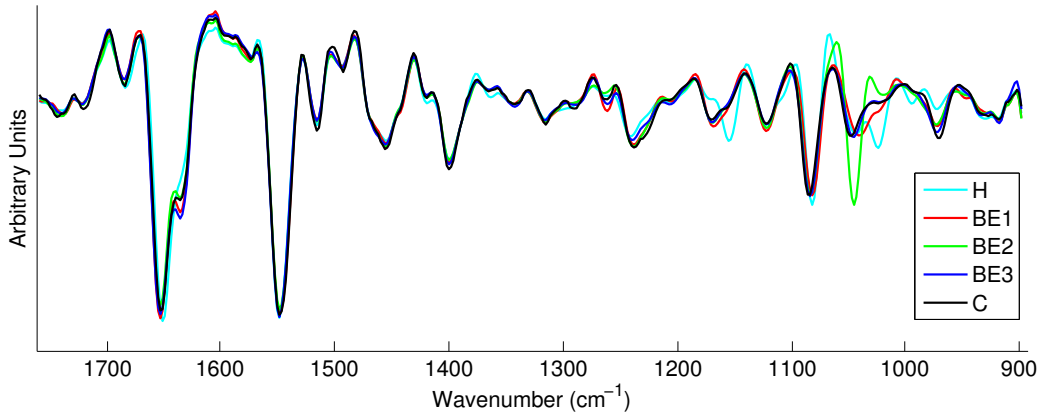


Figure 9.9: Mean second derivative spectra comparing the five stages of the BE disease on the training set.

be interesting to see if they do. Additionally, averaging the spectra can compromise details such as variation between spectra for the same diagnosis, and the comparisons above ignore the partial order of the response vector. All of these limitations are addressed by applying one of our three proposed methods. Actually, scenario (i) corresponds to the proposed idea of decomposed BVS (Chapter 6), scenario (ii) to the BVS using a common indicator vector (Chapter 7), and scenario (iii) to the BVS using an indicator matrix (Chapter 8).

9.3 Bayesian variable selection (BVS) on BE

We will now apply the three proposed methods on BE dataset, which consists of $M = 5$ stages. We assign the following coding: H = 0, BE1 = 1, BE2 = 2, BE3 = 3 and C = 4. From the five stages of BE disease, H and C are nominal and BE1, BE2, and BE3 is the sequence of ordinal responses ($\mathbf{t}$ =(BE1, BE2,

BE3)) with H added in this sequence in the non-decomposed cases. In the training set $n = 207$ spectra are recorded at 447 different wavenumbers, which correspond to $p = 447$ variables. The design matrix $\mathbf{X}$ of the training dataset is centered by column mean.

9.3.1 Decomposed BVS

Firstly, our decomposed approach is implemented in two parts (Algorithm 2): In part A, we merge BE1, BE2, and BE3 and treat them as one extra group of nominal responses (we refer to it as BE1+BE2+BE3). We apply BVS for the nominal responses H, BE1+BE2+BE3, and C (with unknown covariance matrix) using two latent variables. In part B, we apply BVS for the ordinal responses BE1, BE2, and BE3 by using one latent variable and a boundary vector that distinguishes the classes.

For part A, we set the values for the hyperparameters of the unknown covariance matrix $\mathbf{\Sigma}$, $\delta = 3$, $d = 1$, $\mathbf{Q} = \mathbf{I}_2$, and in addition $w_{(\text{nom})} = 3/447$ (the earlier graphical analysis suggests a small number of important variables may suffice), $h = 10^6$ (vague prior for the intercept) and $\mathbf{H}_{\boldsymbol{\xi}} = c_2\mathbf{I}_{p_{\boldsymbol{\xi}}}$ (special case of Equation (3.11), give a ridge type shrinkage). A simple way to find a range of possible values that the regularization parameter c_2 can take is to apply PCA on the covariance matrix, calculate the eigenvalues, sort and standardize them, and finally calculate the cumulative sum of the standardized eigenvalues. From the cumulative sum we can say that the top 5 to 10 eigenvalues explain 93.6% to 99.7% of the variability of the data. These correspond to 1.72 and 0.06, and thus reasonable values for the regularizing parameter lie between 0.6 and 17 (the reciprocals of these eigenvalues). Trying regularizing parameter values in this range, we select $c_2 = 10$, which corresponds to the highest classification accuracy on the training set. We run four different chains with 40000 iterations from which the 20000 iterations are the burn-in period. As we expect only few variables to inform the responses, we initialize the indicator vector of each one of the four chains randomly, selecting 1, 2, 3, and 5 variables respectively to be important. The components of the two latent variables are initialized according to Equation (6.5) for $r = 1, 2$.

For part B, we set the hyperparameters of the variance σ^2, $d_1 = \delta/2 = 1.5$ and $d_2 = 0.5$ (inverse Gamma is the univariate case of inverse Wishart), similarly $w_{(\text{ord})} = 3/447$, $h = 10^6$ and $\mathbf{H}_{\boldsymbol{\gamma}} = c_3\mathbf{I}_{p_{\boldsymbol{\gamma}}}$, with $c_3 = 10$. We run four different chains with 150000 iterations out of which 100000 iterations is the burn-in period. We initialize the indicator vector of each chain as in part A, the latent variable according to Equation (6.6), and the non-fixed boundaries

distributed uniformly in $(0, 1000)$.

Figure 9.10 shows the results for variable and model selection for the average of chains for both parts of the decomposed approach. Variable 314 is the most important for part A with marginal inclusion probability very close to 1 and 284 is the most important variable for part B with inclusion probability 0.93 (top of Figure 9.10). These variables were noted in the visualization of the mean spectra (Figures 9.7 and 9.6). In addition, these variables are included in the best models of the corresponding parts (bottom of Figure 9.10). The best model for each part has very high posterior probability compared to the second and third best models. Comparing the two best models we note that they are distinct. Overall, the variables with high (marginal) posterior probabilities in parts A and B are quite different, which is not surprising as we noticed at scenario (i) in Section 9.2.

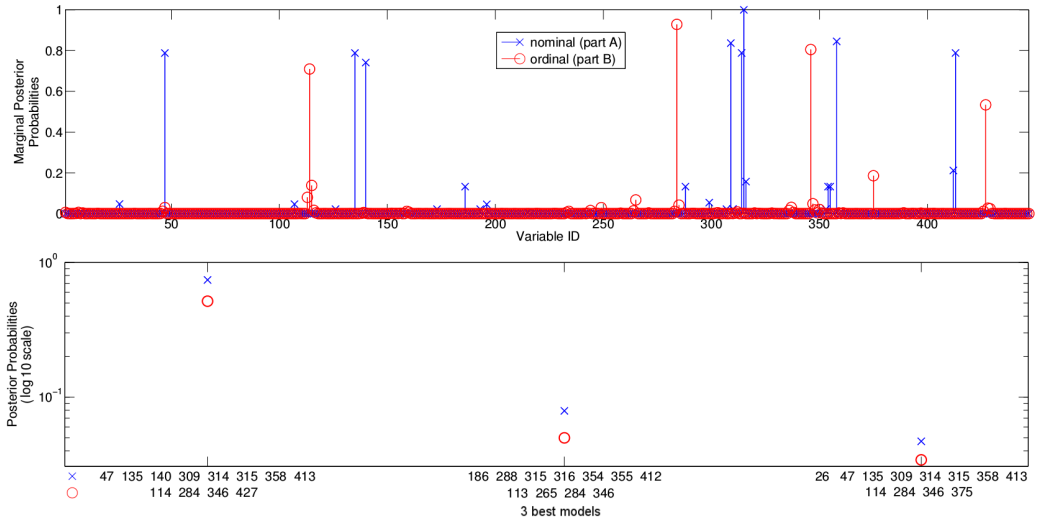


Figure 9.10: Marginal posterior probabilities of variables (top) and posterior probabilities of models on a log scale (bottom) based on the average of chains for decomposed BVS approach.

9.3.2 BVS with an indicator vector

In order to apply BVS with common indicator vector (BVS $\boldsymbol{\xi}$), we use two latent variables ($s = 2$), where the first one has a double role. The first latent variable distinguishes H from the sequence of ordinal responses as well as separating the ordinal classes, and the second latent variable distinguishes H from C spectra. In the BVS $\boldsymbol{\xi}$ approach, we are interested in finding the best model with common important variables across the two latent variables. The earlier graphical analysis suggests a few of the variables may be important jointly.

We set the values for the hyperparameters of the unknown covariance matrix $\mathbf{\Sigma}$, $\delta = 3$, $d = 1$, $\mathbf{Q} = \mathbf{I}_2$. Additionally, we set $w = 3/447$, $h = 10^6$, and $c_2 = 15$. We run four different chains with 100000 from which 30000 iterations are the burn-in period. We initialize the indicator vector of each one of the four chains by randomly selecting 1, 2, 3 and 5 variables respectively to be important. Latent variables are initialized according to Equation (6.5) for $r = 1, 2$ and boundaries as we described in Subsection 9.3.1, part A. Figure 9.11 contains the results of the BVS $\boldsymbol{\xi}$. Only two variables (ID: 314 and 384) are the same as or similar to those identified in scenario (ii) in Section 9.2 and those are combined with another five variables (ID: 127, 313, 316, 327, and 386) in order to get the best model (bottom of Figure 9.11). Note that those five variables are hard to discern by inspection of the mean spectra.

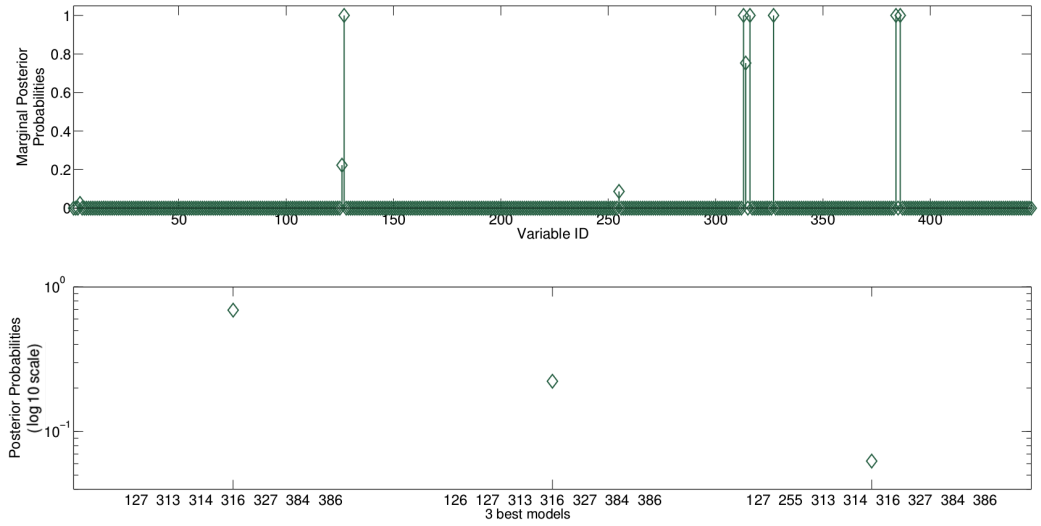


Figure 9.11: Marginal posterior probabilities of variables (top) and posterior probabilities of models on a log scale (bottom) based on the average of chains for BVS using a common indicator vector $\boldsymbol{\xi}$.

9.3.3 BVS with an indicator matrix

In order to apply BVS with an indicator matrix (BVS $\mathbf{\Xi}$), we use two latent variables ($s = 2$), where the first one has a double role as before. However, in this method, we wish to pick two best models, one for the sequence of ordinal responses and one for the nominal responses, which allows us to have flexibility.

We set the hyperparameters and handle the initialization as for the common $\boldsymbol{\xi}$. The only difference is that here we initialize two indicator vectors instead of a common one and we set $w_1 = w_2 = 2/447$. Here we set $\mathbf{H}_{\mathbf{\Xi}_{:,1}} = c_2\mathbf{I}_{p_{\mathbf{\Xi}_{:,1}}}$ and $\mathbf{H}_{\mathbf{\Xi}_{:,2}} = c_3\mathbf{I}_{p_{\mathbf{\Xi}_{:,2}}}$ with $c_2 = c_3 = 3$. Figure 9.12 con-

tains the results for both best variables and best models. It is clear that the individual important variables are different for $\boldsymbol{\Xi}_{,:1}$ (which corresponds to ordinal responses) and for $\boldsymbol{\Xi}_{,:2}$ (which corresponds to nominal responses) (top of Figure 9.12). Just one variable of each (314 for $\boldsymbol{\Xi}_{,:1}$ and 347 for $\boldsymbol{\Xi}_{:,2}$) is the same as in scenario (iii) in Section 9.2. Finally, variable 314 combined with 51 gives the best model for $\boldsymbol{\Xi}_{:,1}$.

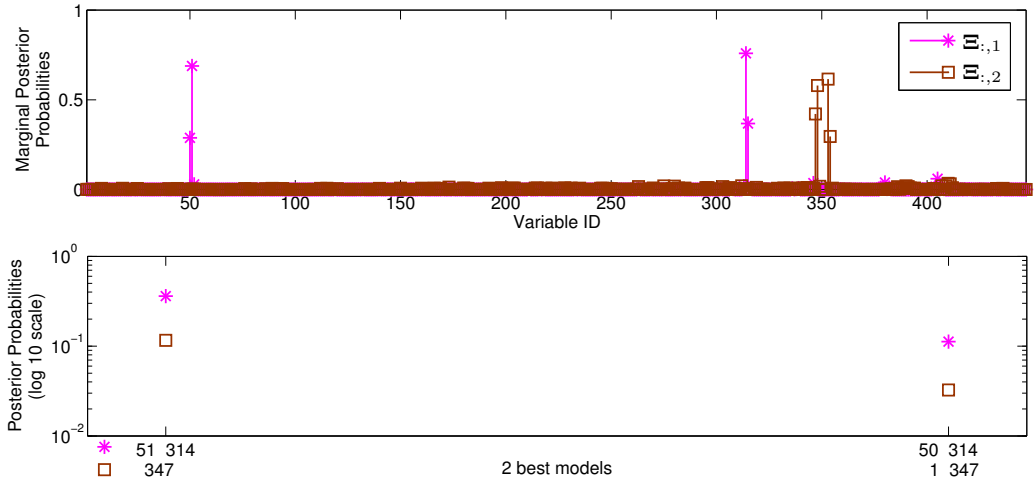


Figure 9.12: Marginal posterior probabilities of variables (top) and posterior probabilities of models on a log scale (bottom) based on the average of chains for BVS using an indicator matrix $\boldsymbol{\Xi}$.

In summary, the individual variables that our methods identify would be difficult to discern with just visual information about their spectral differences. In fact, only a small number of them would be selected if we only looked at the peaks or troughs where the five stages show significant variations in the spectra. Those peaks/troughs may vary slightly but the interpretations remain the same. For example, wavenumber 1155 cm^{-1} (ID: 314) is associated with DNA, similarly to wavenumbers 1157, 1152 and 1151 cm^{-1} (ID: 313, 315, 316). However, the best models (combination of important variables) of decomposed BVS and BVS $\boldsymbol{\xi}$ include ID 314 and 315 or 314 and 316 respectively in order to get accurate predictions.

9.3.4 Comparing the best models with different methods

In this section, we first compare the best models provided by the three proposed methods: decomposed BVS, BVS $\boldsymbol{\xi}$, and BVS $\boldsymbol{\Xi}$. Then, we contrast them with the best models of existing BVS methods and afterwards with non Bayesian methods that treat responses as pure nominal or pure ordinal. Table 9.2 summarizes the best models that different methods identify.

Comparing the three new methods, we notice that the variable with ID 314 is included in the best models for each approach, but the best models are different for each of the three cases. This means that there is a key variable (ID 314), although the aim of the three proposed methods is different (find two best models considering ordinal responses as one nominal, find one common best model and find two best models, one for nominal and one for ordinal responses). However, ID 314 is combined with different variables according to the method used in order to build a good model for predictions. In addition, comparing the BVS $\boldsymbol{\xi}$ approach with other new and existing BVS approaches, we note that variables with ID 313, 314, 316 seem to mainly come from the part A (decomposed approach) or pure nominal part, in contrast with 384, 386 that come from the BVS pure ordinal part.

A more specific comparison is between the nominal (ordinal) decomposed BVS and pure nominal (ordinal) BVS respectively. The BVS approach for pure nominal responses has the same most important variables as the nominal (part A) decomposed BVS, ID 314, 316 and 358, together with some extra variables, but does not have any common variable with the ordinal decomposed step. Then, comparing the variable ID's of the decomposed BVS approach for ordinal responses (part B) with the approach that uses pure ordinal responses, we note only that the variable with ID 346 may belong to the same peak as the 348.

Before continuing the discussion on comparisons with existing non Bayesian methods, we remind the reader of details and notation in Table 9.2. The methods are: Penalized LDA, LASSO, classification trees, random forests, and SVMs. Penalized LDA uses an L_1 penalty and finds the combinations of the important variables that are noted in the table, see Witten and Witten (2015). Both LASSO methods include the intercept in the best model. The LASSO method for ordinal responses minimizes the regularization parameter using BIC as criterion. Random forests for pure nominal responses find a best model with 91 variables that is used for predictions, but in this table we only write down the ID of the top 15 variables. Similarly, random forests for pure ordinal responses finds that the best model contains 177 variables that are used for predictions but in this table we include the ID of only the top 15 variables.

The existing BVS (pure) nominal method identifies that variable 316 (roughly, it can be 314, 315) belongs in the best model. On the other hand, the existing BVS (pure) ordinal method identifies that variable 346 (roughly, it can be 347, 348) belongs in the best model. This explains why, for example, classification trees for pure nominal responses has at the root of the tree

variable 316, whereas for the pure ordinal case there is a different one (346). In addition, those two ID's (314 for entirely nominal and 346 for entirely ordinal responses) are part of the best model of the non Bayesian methods that treat responses as pure nominal or ordinal. However, each existing Bayesian and non Bayesian method combines the two variables with different ones in order to create the best model that can be used for predictions.

Table 9.2: Best models (note the variable ID) for each method.

Method	Best Model
Decomposed BVS	Part A: 47, 135, 140, 309, 314, 315, 358, 413 Part B: 114, 284, 346, 427
BVS $\boldsymbol{\xi}$	127, 313, 314, 316, 327, 384, 386
BVS $\boldsymbol{\Xi}$	$\boldsymbol{\Xi}_{:,1}$: 51, 314 $\boldsymbol{\Xi}_{:,2}$: 347
BVS pure nominal	41, 52, 154, 273, 275, 314, 316, 346, 358, 367, 376, 402, 419
BVS pure ordinal	348, 358, 383, 387
Penalized LDA pure nominal	{1:447}$\setminus${8, 17, 18, 24, 72, 106, 119, 137, 142, 153, 154, 171, 184, 191, 193, 214, 218:220, 223, 245, 247, 248, 301, 335, 343, 363, 369:376, 388, 415, 443, 444, 447}
Penalized LDA pure ordinal	{1:447} $\setminus$ {8, 24, 72, 106:110, 119, 153, 154, 171, 184, 214:223, 256:261, 301, 335, 343, 363, 415:423, 436:447}
LASSO pure nominal	H: a_0, 162, 264, 286, 314, 315, 358, 427 BE1: a_1, 5, 14, 15, 45, 46, 110, 111, 151, 164, 212, 234, 235, 284, 290, 310, 368, 377, 401, 429, 430 BE2: a_2, 5, 57, 110, 111, 151, 152, 201, 262, 287, 334, 335, 340, 377, 378, 402, 414, 437 BE3: a_3, 54, 55, 171, 178, 210, 230, 262, 284, 285, 335, 344, 365, 378, 397, 398, 405, 421, 427, 431, 437, 447 C: a_4, 4, 97, 188, 195, 213, 215, 226, 236, 242, 243, 264, 291, 414, 432, 443, 447
LASSO pure ordinal	226, 238, 413, 347, 1, 359, 237, 381, 348, a_0, 358, 3, 186, a_1, a_3, a_2
Classification trees pure nominal	H: 316 BE1: 316, 346, 405, 14, 394 or 316, 346, 405, 57 or 316, 346, 265, 234, 414, 1, 314 BE2: 316, 346, 405, 14, 345, 4 or 316, 346, 265, 234, 414, 1 or 316, 346, 265, 234, 414, 1, 314 BE3: 316, 346, 405, 14, 345, 4 or 316, 346, 265 C: 316, 346, 405, 14, 394 or 316, 346, 405, 14, 345 or 316, 346, 405, 57 or 316, 346, 265, 234 or 316, 346, 265, 234, 414
Classification trees pure ordinal	H: 346, 315 BE1: 346, 243, 376 or 346, 315, 234 BE2: 346, 315, 243, 60 BE3: 346, 243, 405, 159 or 346, 243, 376 C: 346, 243, 405, 159 or 346, 243, 405 or 346, 315, 234, 60
Random forests pure nominal	24, 42, 272, 285, 287, 308 , 310, 312, 315, 316, 360, 367, 410, 416, 423
Random forests pure ordinal	306, 315, 380, 314, 337, 51, 348, 352, 384, 78, 265, 22, 178, 311, 406
SVM pure nominal	{1:447} $\setminus$ {8, 17, 18, 24, 72, 106, 119, 137, 142, 153, 154, 171, 184, 191, 193, 214, 218, 220, 223, 245, 248, 301, 335, 343, 363, 369, 370:376, 388, 415, 443, 444, 447}
SVM pure ordinal	{1:447} $\setminus$ {8, 24, 72, 106:110, 119, 153, 154, 171, 184, 214:223, 256:261, 301, 335, 343, 363, 415:423, 436:447}

9.3.5 Classification and prediction

We use the variables of the best models that our approaches have found in order to quantify the predictive ability of the best models on the test set. Table 9.3 contains these results, employing the overall classification accuracy on the test set as performance measure. The three proposed methods achieve at least 4% better performance than existing methods that treat all responses as nominal or ordinal. The decomposed BVS comes out with the highest classification accuracy, at least 8% higher than the best of the existing methods. The improvement of the accuracy is important because, on the one hand, it confirms that the proposed methodologies are properly modelling a classification problem with mixture type of responses and, on the other, they are beneficial for the study of BE disease.

Table 9.3: Comparison of overall classification accuracy for the three proposed BVS approaches with mixture of responses with existing methods, where the last one treats responses as pure nominal or pure ordinal.

	Accuracy (%)	
	Nominal	Ordinal
Decomposed BVS	74.5	
BVS $\boldsymbol{\xi}$	71.6	
BVS $\boldsymbol{\Xi}$	70.6	
BVS	64.7	66.7
Penalized LDA	47.1	44.1
LASSO	64.7	61.8
Classification trees	61.8	66.7
Random forests	63.7	66.7
SVM	63.7	44.1

A more fair comparison for the accuracy on the BE disease would be to apply decomposed BVS versus using the decomposed versions of five existing methods. The main idea of the decomposed method remains the same (Chapter 6, Algorithm 2): in part A apply a method for nominal responses H, BE1+BE2+BE3, and C, in part B apply a method only for ordinal responses BE1, BE2, and BE3 and finally combine the two results. Interestingly, our decomposed BVS approach has the highest classification accuracy, at least 5% better performance, compared to the remaining methods in Table 9.4. Comparing those decomposed methods with existing methods that treat responses as pure nominal or ordinal (Table 9.3) we conclude that the application of the decomposed methods can increase the classification accuracy in the case that the responses are a mixture of nominal and ordinal, but again the accuracy is not as good as our proposed decomposed BVS.

Finally, in Table 9.5 we compare the class-specific classification accuracy of our three proposed methods with existing ones that treat the responses as

Table 9.4: Comparison of overall classification accuracy of the proposed decomposed BVS with existing methods that are applied in a decomposed manner for nominal and ordinal responses.

	Accuracy (%)
Decomposed BVS	74.5
Decomposed penalized LDA	55.9
Decomposed LASSO	67.7
Decomposed classification trees	67.7
Decomposed random forests	69.7
Decomposed SVM	67.7

Table 9.5: Overall and for each stage of BE disease classification accuracy on the test set.

	Accuracy (%)					
Method	H	BE1	BE2	BE3	C	Overall
Decomposed BVS	100.0	86.7	0	37.5	53.9	74.5
BVS $\boldsymbol{\xi}$	100.0	82.2	0	0	92.3	71.6
BVS $\boldsymbol{\Xi}$	95.8	91.1	0	0	61.5	70.6
BVS pure nominal	95.8	80.0	0	31.3	15.4	64.7
BVS pure ordinal	83.3	82.2	0	31.3	46.2	66.7
Penalized LDA pure nominal	95.8	44.4	0	0	38.5	47.1
Penalized LDA pure ordinal	95.8	35.6	0	12.5	30.8	44.1
LASSO pure nominal	100.0	68.9	0	56.3	15.4	64.7
LASSO pure ordinal	75.0	82.2	0	6.3	53.9	61.8
Classification trees pure nominal	91.7	71.1	0	50.0	7.7	61.8
Classification trees pure ordinal	95.8	75.6	0	31.3	46.2	66.7
Random forests pure nominal	100.0	73.3	0	25.0	30.8	63.7
Random forests pure ordinal	91.8	91.1	0	0	38.5	66.7
SVM pure nominal	100.0	62.2	25.0	37.5	46.2	63.7
SVM pure ordinal	87.5	40.0	0	18.8	23.1	44.1

entirely nominal or ordinal. We may draw further conclusions based on this table. We can see that the easiest task is to correctly classify healthy spectra, but it is very difficult to distinguish class BE2 from the remaining stages. This difficulty arises partly from the fact that the BE2 spectra in the training set are very few, but may also reflect unreliability on the histologist's diagnosis. The classification accuracy of BE1 is higher than BE3, which verifies that it is easier to identify an early stage disease. This is important because it can be treated. Class BE3 is difficult to classify histopathologically and for this reason all methods have not so high accuracy for this class. On the other hand, the last stage of the disease is full cancer and in general we can diagnose with higher accuracy a cancer spectrum than a spectrum from an intermediate stage. Focusing on our methods, they achieve almost perfect accuracy for healthy spectra, the highest for BE1 and cancer spectrum and the overall accuracy is much higher when compared to other methods.

The three proposed methods are applied on the BE disease and illustrated to have prediction accuracy superior to that of some well-established machine

learning methods. This could be because BVS methods can potentially explore the space of the possible models more efficiently. Although there will be no method/model that works best for every problem, decomposed BVS works well for the particular problem of the BE diagnosis.

9.4 Discussion of the BE results

It is important that the proposed models and variable selection approaches accomplish finding the best model(s). Such models can give good predictions for an unseen patient of the test set, and not just discriminate well within the training set. We achieve this discrimination by including variable 314 in the best model but in each approach combining it with different variables. The model that gives the best classification results is the one that the decomposed BVS found. It is interesting that 1155 cm^{-1} (ID: 314) together with 1153 cm^{-1} (ID: 315) and others achieve very high overall predictive accuracy.

We try to provide some insight into the biological problem of predicting BE for each one of the three proposed models. In the decomposed model, to discriminate the H from BE1+BE2+BE and C, the combination of wavenumbers 1080, 1124 and 1171 cm^{-1} played an important role. Those wavenumbers were attributed to glycogen. Focusing then on the process of discovering the best combination of biomarkers for BE1, BE2 and BE3, we note that the important ones are at wavenumbers 1541, 1213, 1093 and 937 cm^{-1}, which were attributed to DNA/RNA combined with amide II and metabolites.

In the BVS $\boldsymbol{\xi}$, where all the latent variables use the same wavenumbers, important wavenumbers for the five possible BE stages were 1516, 1157, 1155, 1551, 1130, 1020 and 1016 cm^{-1}, which correspond mainly to DNA/RNA combining also with Amide II and glycogen.

Finally, in the BVS $\boldsymbol{\Xi}$, important wavenumbers included: for H versus C samples at wavenumber 1091 cm^{-1} which are assigned to DNA/RNA, and for H, BE1, BE2, BE3 the combination of wavenumbers 1662 and 1155 cm^{-1}, which are assigned to the combination of Amide I with DNA/RNA.

In summary, DNA/RNA is the main compound that is combined with Amide I, II, and/or glycogen in order to increase the classification accuracy of the five stages of BE (according to the different proposed methods).

Among the three proposed models, the decomposed BVS was the one preferred by the biologists. This is because it is important to end up with a model that distinguishes satisfactorily the middle stages of the BE disease and the model delivers some important variables for the middle stages that would be difficult to identify by eye. This step is crucial as it is difficult to

find biomarkers as BE progresses from BE2 to BE4. Combining this task with the easier step of finding potential biomarkers for healthy, BE2+BE3+BE4, and cancer, can give an approach to identifying potential biomarkers for the five stages of the BE disease according to the UK classification.

The outcomes of the BVS approaches are presented as individually important variables and best model (combination of important variables), where the first are used mainly for biomarker identification (based on marginal associations) and the second for predictions (based on the best model). Both ways are useful, even though there is no preference between the two from the biological point of view.

When using the marginal posterior probabilities we refer to the top variables, where a threshold of a probability is required such that variables above that threshold correspond to important variables and below it to unimportant variables. Here in order to present the top (individual) variables to biologist and doctors, we pick 0.5 as the threshold. This is known as median probability model (Barbieri and Berger, 2004). In high dimensional settings, a more advanced technique is to control the FDR as discussed in Chapter 1. We note that, even the simple threshold of 0.5 leads to identifying the same variables as those in the best model (the model with the highest probability) in all cases of the current study. There are also other applications, for example on the Ozone dataset (Berger and Molina, 2005), which exhibit the same interesting behavior.

As we mentioned above, it is challenging to correctly classify BE2 spectra (Table 9.5). We could try to merge BE2 with BE3. Then, H and C would be nominal and H, BE1, and BE2+BE3 the sequence of ordinal responses. Again, both the BVS $\boldsymbol{\xi}$ and the BVS $\boldsymbol{\Xi}$ approach would use two latent variables, but we would have to sample for just one component of the boundary vector.

Chapter 10

Conclusions

In this chapter, a summary of the achieved objectives in this work is given, focusing on both statistical methodology and the Barretts oesophagus (BE) application. Finally, some comments on the possible directions for future methodological research are presented.

10.1 Summary of thesis

In this thesis statistical models were developed to deal with classification problems with a mixture of nominal and ordinal responses, focusing on how to eliminate irrelevant variables in high-dimensional data, for example data from spectroscopy. This was achieved by extracting the most useful information from the data, which was then included in the best model that allows us in the BE case to correctly classify tissue as healthy, BE1, BE2, BE3, and cancer.

We built three models that are appropriate for the mixture of ordinal and nominal responses. Since the aim was to find a simple model that contains the important variables, we proposed Bayesian variable selection (BVS) approaches. The three methods were decomposed BVS, BVS $\boldsymbol{\xi}$, and BVS $\boldsymbol{\Xi}$. All of the methods are based on a probit model with latent variables and are able to incorporate prior knowledge, where it exists. The three methods were applied successfully in the BE disease class prediction. We demonstrated that those methods work well for high-dimensional data ($n \ll p$). We can select any of the three methods depending on the aim of the specific study. The guideline is the following: If the aim of the study is to identify one best model for nominal responses and one for ordinal, then the appropriate method is the decomposed BVS. On the other hand, if the goal is to identify different best models for each one of the nominal responses and one (common) model for

ordinal, then the method to use is the BVS $\boldsymbol{\Xi}$. However, in some cases, for example in the initial stage of a study, it may be interesting to identify just one (common) best model that encapsulates both nominal and ordinal responses. Then, the preferred method is a BVS $\boldsymbol{\xi}$. Finally, if we are interested only in predictions and we do not have knowledge about the interpretation of the application, then we can apply all of the proposed methods and keep the one that satisfies some criterion of our choice, e.g. classification accuracy.

We conclude that the three proposed methods can give higher classification accuracy than existing methods that treat responses as pure nominal or pure ordinal. This means that the models and the variable selection methods are successfully taking into account both types of responses. Additionally, applying the decomposition idea to any existing methods ha the potential to increase the classification accuracy compared to applying the same methods onto pure nominal or ordinal responses. If we contrast the existing decomposed methods with the decomposed BVS we can say that the latter performs better than frequentist methods, because the Bayesian method incorporates prior knowledge, e.g. the number of variables included in the model may be controlled by the prior. Finally, from the three methods, the highest classification accuracy is achieved via the decomposed BVS in the case of the BE application. This means that we found two best models: the first discriminates nominal responses H from C and BE1+BE2+BE3 (group of ordinal responses that treated as one nominal response) and the second discriminates between the ordinal responses (BE1, BE2, BE3). Then, the classification process follows the same pattern. This idea works nicely, since it splits a complex problem into two simpler problems (through the decomposition).

The advantages that arise from the BE application are beneficial for the patients, medicine, and hospital management as well. Speed of the diagnosis (without waiting for histological results) and low cost of it (without the need of expert histopathologists) are the main advantages for diagnose of the BE disease via the best model. Therefore, the management options for the BE tend to become easier. The option to have more frequent follow ups may helps the patient to have better life quality. For the medical science, finding the combination of important variables is of prime importance and opens new possibilities for understanding the cell changes involved in the disease. Up to now, medical doctors know that variable 314 is a biomarker, but it is new knowledge that this variable, combined with the variables 47, 135, 140, 309, 315, 358 and 413, contains a lot of information from the tissues in order to discriminate H from C and BE1+BE2+BE3, even though those variables may or may not be biomarkers themselves. However, we notice that

the important variable 314 (and the others that were just mentioned) is not a potential biomarker for BE1, BE2 and BE3, and does not belong to the best model of the three middle stages.

Finally, correctly identifying the stage of the disease that the patient has, can assist the doctors in providing the appropriate treatment. Additionally, this reduces the cost of repeatedly taking extra biopsies to identify the stage of the disease as well as the risk of mistreating the patient, which could lead to health issues. Avoiding giving inappropriate treatments to the patient at a specific stage of the disease can provide a cost-effective strategy for managing BE. This strategy promises an economical plan for hospitals and governments.

10.2 Future work

This work could be extended or improved in several directions. In most published examples and in this study the prior information for the coefficients uses some standard choice for the covariance matrix, in order, for instance, to be invertible and not computationally expensive. However, we could try to extract some prior knowledge that comes from the specific BE study in order to use the benefits of the Bayesian approach of variable selection. One choice may be an AR(1) small bands covariance matrix for $\boldsymbol{\beta}$ (if the the lag between two variables is small, then they are dependent and if it is large are independent). In addition, we could add low order dependences between the indicator variables, similar to the binary case where we used a first order Markov model that captures the dependency. In the models under those more complex assumptions, and also in the proposed methods it will be useful to reduce the computational cost. However, the first aim remains the same: to set up an appropriate model for mixture of nominal and ordinal responses and to identify the best model(s) in order to achieve accurate predictions.

One natural extension of BVS $\boldsymbol{\Xi}$ would be to study this method under different assumptions, where the covariance matrix is unknown. In this case we would work with the block matrices, for example block matrix of coefficients, where each block will be related to different regression equations. However, in this case variable selection can be challenging, as we have to take into account correlations between the latent variables when computing the conditional distributions.

In the current study, we focus on the case of a mixture of nominal and ordinal responses where there is one sequence of ordinal responses. However, it might be interesting to construct a model for the mixture of responses using

more than one group of ordinal responses. One example is the emoticons that are mainly used in social media. Those emoticons can be very sad, sad, happy, and very happy (one group of ordinal responses), upset, angry, and furious (another group of ordinal responses), and shocked and scared (nominal responses). In this case, we can extend the three methodologies for model building and variable selection. In the modelling part we will introduce two boundary vectors, one for each group of ordinal responses, and consequently two of the latent variables will be related to those boundary vectors. In the variable selection process we will use one extra indicator vector for the second group of ordinal responses.

Another possible extension is to multi-label classification, for example the topics in the document/newspapers/media could be sports, politics (nominal) primary, secondary and high education (ordinal). The extension will contain two parts: modelling and variable selection approach. The variable selection with many labels is more challenging, especially if there is dependence between those labels.

Apart from the methodological extensions of the proposed methods, a practical direction would be to develop a novel and user-friendly software that speeds up the diagnosis of BE disease with a high accuracy using the proposed decomposed BVS but without the need for a histologist's diagnosis. Although, the diagnosis of the BE will become automatic, we will need to verify the clinical relevance via testing many samples.

Finally, the proposed approaches for variable selection can be applied in many scientific areas such as Bioinformatics (microarrays, gene expression), pattern recognition (image processing, face recognition, handwritten digit recognition signal, document classification), Neuroscience (functional MRI studies of the brain), Medicine (cancer types), Finance (loan applications, income), Social sciences (nursery application, students grade), Genetics and Genomics.

Appendix A

Multi and matrix variate distributions

i. **Truncated multivariate normal**

The definition of a truncated distribution at $\mathbf{a}$, $\mathbf{b}$ of a random variable X, is given by

$$f(X) = \begin{cases} \frac{P(\mathbf{a} \leq X \leq \mathbf{b})}{\int_{\mathbf{a}}^{\mathbf{b}} P(\mathbf{a} \leq X \leq \mathbf{b}) dX}, & \text{if } \mathbf{a} \leq X \leq \mathbf{b} \\ 0, & \text{otherwise.} \end{cases} \tag{A.1}$$

Based on this definition, the PDF for the truncated multivariate normal $(TMVN)$ $\mathbf{X} \sim TMVN(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{a}, \mathbf{b})$ can be expressed as

$$f(\mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{a}, \mathbf{b}) = \frac{\exp\left\{-\frac{1}{2}(\mathbf{X} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu})\right\}}{\int_{\mathbf{a}}^{\mathbf{b}} \exp\left\{-\frac{1}{2}(\mathbf{X} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu})\right\} d\mathbf{X}}, \tag{A.2}$$

for $\mathbf{a} \leq \mathbf{X} \leq \mathbf{b}$ and 0 otherwise. The constant term, $c(p) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}}$ of the PDF of a p-variate normal distribution does not appear in the last equation because it cancels out in ratio.

The first idea to generate variates from a truncated multivariate normal distribution is to draw from the untruncated distribution and accept only those samples inside the support region. However, this is not an efficient algorithm. The second approach to generating random samples is to use the Gibbs sampler and take samples from conditional univariate distributions are actually truncated univariate normal distributions Kotecha and Djuric (1999).

ii. **Matrix normal**

Let $\mathbf{Z}$ be a $p \times q$ random matrix. $\mathbf{Z}$ follows a matrix normal distribution,

$\mathbf{Z} - \mathbf{M} \sim MN(\mathbf{P}, \mathbf{Q})$, where $\mathbf{M}$ is the mean of $\mathbf{Z}$ and $p_{ii}\mathbf{Q}$ and $q_{jj}\mathbf{P}$ are the covariance matrixes of i-th row and j-th column respectively. If $\mathbf{P}$ and $\mathbf{Q}$ are positive definite, the PDF of the matrix normal is

$$f(\mathbf{Z}) = c(p,q)|\mathbf{P}|^{-\frac{q}{2}}|\mathbf{Q}|^{-\frac{p}{2}} \exp\left\{-\frac{1}{2}\text{tr}[\mathbf{P}^{-1}(\mathbf{Z}-\mathbf{M})\mathbf{Q}^{-1}(\mathbf{Z}-\mathbf{M})']\right\}, \tag{A.3}$$

with $c(p,q) = (2\pi)^{-\frac{pq}{2}}$, where $|.|$ represents the determinant of square matrix and tr(.) is the trace of the square matrix (i.e., the sum of the diagonal elements of it).

iii. **Inverse Wishart**

Let $\mathbf{V} \sim IW(\delta, \mathbf{Q})$, where $\mathbf{Q}$ is the positive definite scale matrix and $\delta = \nu - q + 1 > 0$ ($\nu \geq q$) is the parameter that denotes the degrees of freedom (ν degrees of freedom of Wishart distribution). Then $\mathbf{V}$ is positive definite and the PDF, according to Dawid (1981), is given by

$$f(\mathbf{V}) = c(q,\delta)|\mathbf{Q}|^{\frac{\delta+q-1}{2}}|\mathbf{V}|^{-\frac{\delta+2q}{2}} \exp\left\{-\frac{1}{2}\text{tr}(\mathbf{V}^{-1}\mathbf{Q})\right\}, \mathbf{V} > 0, \tag{A.4}$$

with $c(q,\delta) = 2^{\frac{-(\delta+q-1)p}{2}}/\mathbf{\Gamma}_q[(\delta+q-1)/2])$, where q is the dimension of $\mathbf{V}$ ($\mathbf{V} \in \mathbb{R}^{q\times q}$) and $\mathbf{\Gamma}_q$ is the multivariate generalization of the gamma function and is equal to

$$\mathbf{\Gamma}_q(x) = \pi^{\frac{q(q-1)}{4}} \prod_{j=1}^{q} \mathbf{\Gamma}[x + (1-j)/2].$$

This formula of inverse Wishart distribution it does not require constant adjustment of the shape parameter $\mathbf{V}$, since it is true that $E(\mathbf{V}) = \mathbf{Q}/(\delta - 2)$ for $\delta > 2$. The inverse Wishart distribution as defined here is also consistent under marginalization, a property that is not true by other parametrizations (it comes from different specification of degrees of the freedom).

iv. **Matrix Student distribution**

Let $\mathbf{\Sigma} \sim IW(\delta, \mathbf{Q})$ and given $\mathbf{\Sigma}$, $\mathbf{T} \sim MN(\mathbf{P}, \mathbf{\Sigma})$. Then the marginal distribution for $\mathbf{T}$ is a $p \times q$ matrix Student distribution and is denoted by $MT(\delta; \mathbf{P}, \mathbf{Q})$. The PDF of the matrix Student distribution exists for $\delta > 0$, $\mathbf{P} > 0$ and $\mathbf{Q} > 0$ and is

$$f(\mathbf{T}) = c(p,q,\delta)|\mathbf{P}|^{\frac{\delta+p-1}{2}}|\mathbf{Q}|^{-\frac{p}{2}}|\mathbf{P} + \mathbf{T}\mathbf{Q}^{-1}\mathbf{T}'|^{-\frac{\delta+p+q-1}{2}}, \tag{A.5}$$

with $c(p,q,\delta) = \pi^{-\frac{pq}{2}}\mathbf{\Gamma}_q[(\delta+p+q-1)/2])/\mathbf{\Gamma}_q[(\delta+p-1)/2])$.

Marginal and conditional distributions of the matrix $\mathbf{T}$ also have a matrix Student distribution. If $\mathbf{T}$ is partitioned as $\mathbf{T}' = (\mathbf{T}_1', \mathbf{T}_2')$ (T_i is $p_i \times q$ matrix, $i = 1, 2$, $p_1 + p_2 = p$), then the marginal distribution is $\mathbf{T}_2 \sim MT(\delta; \mathbf{P}_{22}, \mathbf{Q})$ and the conditional distribution of $\mathbf{T}_1$ given $\mathbf{T}_2$ is $\mathbf{T}_1 - \mathbf{P}_{12}\mathbf{P}_{22}^{-1}\mathbf{T}_2 \sim MT(\delta + p_2; \mathbf{P}_{11.2}, \mathbf{Q} + \mathbf{T}_2\mathbf{P}_{11}^{-1}\mathbf{T}_2')$, where $\mathbf{P}_{11.2} = \mathbf{P}_{11} - \mathbf{P}_{12}\mathbf{P}_{22}^{-1}\mathbf{P}_{21}$.

v. **Truncated multivariate Student**

Let $\mathbf{X}$ follow a truncated multivariate Student distribution with degrees of freedom ν. The PDF of $\mathbf{X} \sim TMNT(\nu; \boldsymbol{\mu}, \mathbf{\Sigma}, \mathbf{a}, \mathbf{b},)$ can be expressed as

$$f(\mathbf{X}, \boldsymbol{\mu}, \mathbf{\Sigma}, \mathbf{a}, \mathbf{b}, \nu) = \frac{\left[1 + \frac{1}{\nu}(\mathbf{X}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})\right]^{-\frac{\nu+p}{2}}}{\int_{\mathbf{a}}^{\mathbf{b}} \left[1 + \frac{1}{\nu}(\mathbf{X}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})\right]^{-\frac{\nu+p}{2}} d\mathbf{X}}, \quad \text{(A.6)}$$

for $\mathbf{a} \leq \mathbf{X} \leq \mathbf{b}$ and 0 otherwise. The constant term, $c(p,\nu) = (\nu\pi)^{-\frac{p}{2}} \cdot |\mathbf{\Sigma}|^{-\frac{1}{2}} \frac{\mathbf{\Gamma}_p[(\nu+p)/2]}{\mathbf{\Gamma}_p[\nu/2]}$ of the PDF of a p-variate Student T distribution does not appear in the last equation because in cancels out in the ratio.

The sampling method for the truncated multivariate normal distribution can be easily generalized to a method for sampling from the truncated multivariate Student T distribution (Geweke, 1991), since truncated multivariate normal can be obtained as the ratio of a truncated multivariate Student T to the square root of an independent chi-squared random variable divided by its degrees of freedom.

vi. **Inverse Gamma**

Let $\sigma^2 \sim IG(\nu_1, \nu_2)$, ν_1 is the shape ($\nu_1 > 0$) and ν_2 is the scale ($\nu_2 > 0$). The PDF of σ^2, according this parametrization, is given by

$$f(\sigma^2) = c(\nu_1, \nu_2)(\sigma^2)^{-\nu_1-1} \exp\left\{-\frac{\nu_2}{\sigma^2}\right\} \quad \text{(A.7)}$$

with $c(\nu_1, \nu_2) = \nu_2^{\nu_1}/\Gamma(\nu_1)$, where $\Gamma(.)$ is the gamma function and is equal to

$$\Gamma(\nu_1) = \int_0^\infty (\sigma^2)^{\nu_1-1} \exp(-\sigma^2) d\sigma^2.$$

The inverse Gamma distribution can be parameterized differently using as the second positive parameter the rate (instead of scale). In this study we select to work with the parametrization of shape and scale.

vii. **Normal inverse Gamma**

Let $\mathbf{X}$ be a random variable that distributed as multivariate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\sigma^2\boldsymbol{\Sigma}$, where $\sigma^2 \sim IG(\nu_1, \nu_2)$. The common distribution of them is the normal inverse gamma, $\mathbf{X}, \sigma^2 \sim NIG(\boldsymbol{\mu}, \sigma^2\boldsymbol{\Sigma}, \nu_1, \nu_2)$, and the PDF is given by

$$f(\mathbf{X}, \sigma^2) = c(\nu_1, \nu_2)\left(\frac{1}{\sigma^2}\right)^{n/2+\nu_1+1} \exp\left\{-\frac{(\mathbf{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu}) + 2\nu_2}{2\sigma^2}\right\}$$

where $c(\nu_1, \nu_2) = (2\pi)^{-n/2}\nu_2^{\nu_1}/\Gamma(\nu_2)|\boldsymbol{\Sigma}|^{-1/2}$.

Appendix B

Algebra calculations for the probit model with nominal responses, $\boldsymbol{\Sigma}$ known, common $\boldsymbol{\xi}$

This appendix contains the algebra calculations of the multi-class BVS with mixture of nominal and ordinal responses under the assumption that $\boldsymbol{\Sigma}$ is known and $\boldsymbol{\xi}$ is common across different classes, as described in Subsection 6.2.1. The prior distributions are described in the same subsection. This appendix actually contains the algebra calculations of part A of the decomposed approach for $\boldsymbol{\Sigma}$ known (Figure 6.1, Algorithm 2 at Chapter 6). Setting $\boldsymbol{\alpha}_0 = \mathbf{0}$ and $\mathbf{B}_{0\boldsymbol{\xi}} = \mathbf{0}$ the full conditional distributions are calculated below.

In this case, the unknown parameters are $\mathbf{Z}, \boldsymbol{\alpha}, \mathbf{B}_{\boldsymbol{\xi}}$ and $\boldsymbol{\xi}$. We assume that $\boldsymbol{\alpha}$, $\mathbf{B}_{\boldsymbol{\xi}}$ are independent, namely that $p(\boldsymbol{\alpha}, \mathbf{B}_{\boldsymbol{\xi}}) = p(\boldsymbol{\alpha})p(\mathbf{B}_{\boldsymbol{\xi}})$. The joint posterior is proportional to likelihood times the prior distribution and it is given by

$$
\begin{aligned}
& p(\mathbf{Z}, \boldsymbol{\alpha}, \mathbf{B}_{\boldsymbol{\xi}}, \boldsymbol{\xi}, \mathbf{k} | \mathbf{X}, \mathbf{y}) \\
\propto & \, |\boldsymbol{\Sigma}|^{-n/2} \exp\left\{ -\frac{1}{2} \mathrm{tr}\Big[\Big(\mathbf{Z} - \mathbf{1}_n \boldsymbol{\alpha}^{'} - \mathbf{X}_{\boldsymbol{\xi}} \mathbf{B}_{\boldsymbol{\xi}} \Big) \boldsymbol{\Sigma}^{-1} (\mathbf{Z} - \mathbf{1}_n \boldsymbol{\alpha}^{'} - \mathbf{X}_{\boldsymbol{\xi}} \mathbf{B}_{\boldsymbol{\xi}})^{'} \Big] \right\} \\
& \cdot \prod_{i=1}^{n} \mathbb{1}(\mathbf{Z}_{i,:} \in F_i) \, |\mathbf{H}_{\boldsymbol{\xi}}|^{-(M-1)/2} \, |\boldsymbol{\Sigma}|^{-p_{\boldsymbol{\xi}}/2} \exp\left\{ -\frac{1}{2} \mathrm{tr}\Big[\mathbf{H}_{\boldsymbol{\xi}}^{-1} \mathbf{B}_{\boldsymbol{\xi}} \boldsymbol{\Sigma}^{-1} \mathbf{B}_{\boldsymbol{\xi}}^{'} \Big] \right\} \qquad \text{(B.1)} \\
& \cdot |\boldsymbol{\Sigma}|^{-1/2} \exp\left\{ -\frac{h^{-1} \boldsymbol{\alpha}^{'} \boldsymbol{\Sigma}^{-1} \boldsymbol{\alpha}}{2} \right\} \\
& \cdot \prod_{j=1}^{p} w^{\xi_j} (1-w)^{1-\xi_j},
\end{aligned}
$$

where F_i is given by Equation (5.11).

First, in order to integrate out $\boldsymbol{\alpha}$ given $\mathbf{Z}, \mathbf{B}_{\boldsymbol{\xi}}$ and $\boldsymbol{\xi}$, the exponential

terms that are associated with $\boldsymbol{\alpha}$ (excluding $\frac{1}{2}\boldsymbol{\Sigma}^{-1}$) in Equation (B.1) can be rewritten as

$$
\begin{aligned}
&- \left(\mathbf{Z} - \mathbf{1}_n\boldsymbol{\alpha}' - \mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}}\right)^{'}\left(\mathbf{Z} - \mathbf{1}_n\boldsymbol{\alpha}' - \mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}}\right) - h^{-1}\boldsymbol{\alpha}\,\boldsymbol{\alpha}^{'} \\
&= - \left[(\mathbf{Z} - \mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}}) - \mathbf{1}_n\boldsymbol{\alpha}^{'}\right]^{'}\left[(\mathbf{Z} - \mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}}) - \mathbf{1}_n\boldsymbol{\alpha}^{'}\right] - h^{-1}\boldsymbol{\alpha}\,\boldsymbol{\alpha}^{'} \\
&= - (n + h^{-1})\boldsymbol{\alpha}\,\boldsymbol{\alpha}^{'} + \mathbf{1}_n^{'}(\mathbf{Z} - \mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}})\boldsymbol{\alpha} + \boldsymbol{\alpha}^{'}(\mathbf{Z} - \mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}})^{'}\mathbf{1}_n \\
&-(\mathbf{Z} - \mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}})^{'}(\mathbf{Z} - \mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}}) \\
&= - (n + h^{-1})\Big[\boldsymbol{\alpha}\,\boldsymbol{\alpha}^{'} - (n + h^{-1})^{-1}\mathbf{1}_n^{'}(\mathbf{Z} - \mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}})\boldsymbol{\alpha} \\
&\quad - (n + h^{-1})^{-1}\boldsymbol{\alpha}^{'}(\mathbf{Z} - \mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}})^{'}\mathbf{1}_n + (n + h^{-1})^{-1}(\mathbf{Z} - \mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}})^{'}(\mathbf{Z} - \mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}}) \\
&\quad \pm (n + h^{-1})^{-1}h^{-1}(\mathbf{Z} - \mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}})^{'}(\mathbf{Z} - \mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}})\Big] \\
&= - (n + h^{-1})\mathbf{V}\mathbf{V}^{'} - (nh + 1)^{-1}(\mathbf{Z} - \mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}})^{'}(\mathbf{Z} - \mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}}), \qquad \text{(B.2)}
\end{aligned}
$$

where $\mathbf{V} = \{\boldsymbol{\alpha}^{'} - (n+h^{-1})^{-1}\mathbf{1}_n^{'}(\mathbf{Z} - \mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}})\}$ and $(n+h^{-1})^{-1}h^{-1} = (nh+1)^{-1}$. Combining the factors $\frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1})$, $|\boldsymbol{\Sigma}|^{-1/2}$ and the exponential terms of the first and penultimate line in Equation (B.1) together with Equation (B.2) yield

$$
\begin{aligned}
&\int |\boldsymbol{\Sigma}|^{1/2}\exp\left\{-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\left[(n + h^{-1})\mathbf{V}\mathbf{V}^{'}\right.\right.\right. \\
&\left.\left.\left.+(nh + 1)^{-1}(\mathbf{Z} - \mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}})^{'}(\mathbf{Z} - \mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}})\right]\right)\right\} d\boldsymbol{\alpha} \\
&\propto \exp\left\{-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\left[(nh + 1)^{-1}(\mathbf{Z} - \mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}})^{'}(\mathbf{Z} - \mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}})\right]\right)\right\}, \qquad \text{(B.3)}
\end{aligned}
$$

where in Equation (B.3) there is a constant that comes from the matrix normal density $MN(\mathbf{I}_n, \boldsymbol{\Sigma})$.

Using Binomial inverse theorem, also known as Sherman-Morrison-Woodbury formula (Woodbury (1950); Plackett (1950)),

$$
(nh + 1)^{-1}(\mathbf{Z} - \mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}})^{'}(\mathbf{Z} - \mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}}) = (\mathbf{Z} - \mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}})^{'}(\mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n^{'})^{-1}(\mathbf{Z} - \mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}}). \qquad \text{(B.4)}
$$

Secondly, in order to integrate out $\mathbf{B}_{\boldsymbol{\xi}}$, the exponential terms that are associated with $\mathbf{B}_{\boldsymbol{\xi}}$ in Equation (B.1) together with the term that is in the squared brackets of Equation (B.3), using also Equation (B.4), can be rewrit-

ten as

$$
\begin{aligned}
&\mathbf{B}_{\boldsymbol{\xi}}^{'}\mathbf{H}_{\boldsymbol{\xi}}^{-1}\mathbf{B}_{\boldsymbol{\xi}} + (\mathbf{Z}-\mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}})^{'}(\mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n^{'})^{-1}(\mathbf{Z}-\mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}}) \\
=&\mathbf{B}_{\boldsymbol{\xi}}^{'}\mathbf{H}_{\boldsymbol{\xi}}^{-1}\mathbf{B}_{\boldsymbol{\xi}} + \mathbf{Z}^{'}(\mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n')^{-1}\mathbf{Z} - \mathbf{Z}^{'}(\mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n')^{-1}\mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}} \\
-&\mathbf{B}_{\boldsymbol{\xi}}^{'}\mathbf{X}_{\boldsymbol{\xi}}^{'}(\mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n')^{-1}\mathbf{Z} + \mathbf{B}_{\boldsymbol{\xi}}^{'}\mathbf{X}_{\boldsymbol{\xi}}^{'}(\mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n')^{-1}\mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}} \qquad \text{(B.5)} \\
=&\mathbf{B}_{\boldsymbol{\xi}}^{'}\mathbf{W}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}} - \mathbf{B}_{\boldsymbol{\xi}}^{'}\mathbf{N} - \mathbf{N}^{'}\mathbf{B}_{\boldsymbol{\xi}} + \mathbf{J} \pm \mathbf{N}^{'}\mathbf{W}_{\boldsymbol{\xi}}^{-1}\mathbf{N} \\
=&(\mathbf{B}_{\boldsymbol{\xi}} - \mathbf{W}_{\boldsymbol{\xi}}^{-1}\mathbf{N})^{'}\mathbf{W}_{\boldsymbol{\xi}}(\mathbf{B}_{\boldsymbol{\xi}} - \mathbf{W}_{\boldsymbol{\xi}}^{-1}\mathbf{N}) + \mathbf{J} - \mathbf{N}^{'}\mathbf{W}_{\boldsymbol{\xi}}^{-1}\mathbf{N},
\end{aligned}
$$

where $\mathbf{W}_{\boldsymbol{\xi}} = \mathbf{X}_{\boldsymbol{\xi}}^{'}(\mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n')^{-1}\mathbf{X}_{\boldsymbol{\xi}} + \mathbf{H}_{\boldsymbol{\xi}}^{-1}$, $\mathbf{N} = \mathbf{X}_{\boldsymbol{\xi}}^{'}(\mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n')^{-1}\mathbf{Z}$ and $\mathbf{J} = \mathbf{Z}^{'}(\mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n')^{-1}\mathbf{Z}$. Combining the factors $-\frac{1}{2}\text{tr}(\boldsymbol{\Sigma}^{-1})$, $|\boldsymbol{\Sigma}|^{-n/2}$, $|\mathbf{H}_{\boldsymbol{\xi}}|^{-(M-1)/2}$, $|\boldsymbol{\Sigma}|^{-p_{\boldsymbol{\xi}}/2}$ and the corresponding exponential term of Equation (B.1) together with the first term of Equation (B.5) yield the completed quadratic form of the matrix normal density $MN(\mathbf{W}_{\boldsymbol{\xi}}^{-1}, \boldsymbol{\Sigma})$. Continuing the integration of $\mathbf{B}_{\boldsymbol{\xi}}$,

$$
\begin{aligned}
&|\boldsymbol{\Sigma}|^{-n/2}\,|\mathbf{H}_{\boldsymbol{\xi}}|^{-(M-1)/2}\,|\boldsymbol{\Sigma}|^{-p_{\boldsymbol{\xi}}/2} \\
&\cdot \int \exp\left\{-\frac{1}{2}\text{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\mathbf{B}_{\boldsymbol{\xi}} - \mathbf{W}_{\boldsymbol{\xi}}^{-1}\mathbf{N}\right)^{'}\mathbf{W}_{\boldsymbol{\xi}}\left(\mathbf{B}_{\boldsymbol{\xi}} - \mathbf{W}_{\boldsymbol{\xi}}^{-1}\mathbf{N}\right)\right]\right\} d\mathbf{B}_{\boldsymbol{\xi}} \\
&\cdot \exp\left\{-\frac{1}{2}\text{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\mathbf{J} - \mathbf{N}^{'}\mathbf{W}_{\boldsymbol{\xi}}^{-1}\mathbf{N}\right)\right]\right\} \\
&\propto |\boldsymbol{\Sigma}|^{-n/2}\,|\mathbf{H}_{\boldsymbol{\xi}}|^{-(M-1)/2}\,|\boldsymbol{\Sigma}|^{-p_{\boldsymbol{\xi}}/2}\,|\boldsymbol{\Sigma}|^{p_{\boldsymbol{\xi}}/2}\left|\mathbf{W}_{\boldsymbol{\xi}}^{-1}\right|^{(M-1)/2} \\
&\cdot \exp\left\{-\frac{1}{2}\text{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\mathbf{J} - \mathbf{N}^{'}\mathbf{W}_{\boldsymbol{\xi}}^{-1}\mathbf{N}\right)\right]\right\} \\
&\propto |\boldsymbol{\Sigma}|^{-n/2}\left(|\mathbf{H}_{\boldsymbol{\xi}}|\,|\mathbf{W}_{\boldsymbol{\xi}}|\right)^{-(M-1)/2}\exp\left\{-\frac{1}{2}\text{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\mathbf{J} - \mathbf{N}^{'}\mathbf{W}_{\boldsymbol{\xi}}^{-1}\mathbf{N}\right)\right]\right\}, \qquad \text{(B.6)}
\end{aligned}
$$

where $\mathbf{J} - \mathbf{N}^{'}\mathbf{W}_{\boldsymbol{\xi}}^{-1}\mathbf{N} = \mathbf{Z}^{'}\mathbf{P}_{\boldsymbol{\xi}}^{*-1}\mathbf{Z}$ and $\mathbf{P}_{\boldsymbol{\xi}}^{*-1} = (\mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n')^{-1} - (\mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n')^{-1}\mathbf{X}_{\boldsymbol{\xi}} \cdot [\mathbf{X}_{\boldsymbol{\xi}}^{'}(\mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n')^{-1}\mathbf{X}_{\boldsymbol{\xi}} + \mathbf{H}_{\boldsymbol{\xi}}^{-1}]^{-1}\mathbf{X}_{\boldsymbol{\xi}}^{'}(\mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n')^{-1}$. Since $\mathbf{P}_{\boldsymbol{\xi}}^{*-1}\mathbf{P}_{\boldsymbol{\xi}} = \mathbf{I}_n$, $\mathbf{P}_{\boldsymbol{\xi}}^{*}$ takes the simple form $\mathbf{P}_{\boldsymbol{\xi}} = \mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n^{'} + \mathbf{X}_{\boldsymbol{\xi}}\mathbf{H}_{\boldsymbol{\xi}}^{-1}\mathbf{X}_{\boldsymbol{\xi}}^{'}$ and $\mathbf{Z}^{'}\mathbf{P}_{\boldsymbol{\xi}}^{*-1}\mathbf{Z} = \mathbf{Z}^{'}\mathbf{P}_{\boldsymbol{\xi}}^{-1}\mathbf{Z}$. From Equation (B.6) it is true that $\mathbf{Z}|\boldsymbol{\xi}, \mathbf{X} \sim MN(\mathbf{P}_{\boldsymbol{\xi}}, \boldsymbol{\Sigma})$. From this we can easily calculate the full conditional distribution of $\mathbf{Z}$ (we also condition on the $\mathbf{y}$), so that

$$
\begin{aligned}
&\mathbf{Z}|\boldsymbol{\xi}, \mathbf{X}, \mathbf{y} \sim MN(\mathbf{P}_{\boldsymbol{\xi}}, \boldsymbol{\Sigma})\prod_{i=1}^{n}\mathbb{1}(\mathbf{Z}_{i,:} \in F_i) \qquad \text{(B.7)} \\
&p(\mathbf{Z}|\boldsymbol{\xi}, \mathbf{X}, \mathbf{y}) \propto |\boldsymbol{\Sigma}|^{-n/2}\,|\mathbf{P}_{\boldsymbol{\xi}}|^{-(M-1)/2}\exp\left\{-\frac{1}{2}\text{tr}\left[\boldsymbol{\Sigma}^{-1}\mathbf{Z}^{'}\mathbf{P}_{\boldsymbol{\xi}}^{-1}\mathbf{Z}\right]\right\}\prod_{i=1}^{n}\mathbb{1}(\mathbf{Z}_{i,:} \in F_i).
\end{aligned}
$$

Then, we calculate the conditional probability of $\boldsymbol{\xi}$ given the remaining unknown parameters using Equations (B.1), (B.6) and (B.7), which is the

following

$$
\begin{aligned}
& p(\boldsymbol{\xi}|\mathbf{Z},\mathbf{X},\mathbf{y}) = p(\boldsymbol{\xi})p(\mathbf{X},\mathbf{y},\mathbf{Z}|\boldsymbol{\xi}) \\
=& p(\boldsymbol{\xi})p(\mathbf{X},\mathbf{y}|\mathbf{Z},\boldsymbol{\xi})p(\mathbf{Z}|\boldsymbol{\xi}) = p(\boldsymbol{\xi})p(\mathbf{X},\mathbf{y}|\mathbf{Z})p(\mathbf{Z}|\boldsymbol{\xi}) \\
\propto& p(\boldsymbol{\xi})p(\mathbf{Z}|\boldsymbol{\xi}) = p(\boldsymbol{\xi}) \int p(\mathbf{Z}|\boldsymbol{\alpha},\mathbf{B}_{\boldsymbol{\xi}},\boldsymbol{\xi})p(\boldsymbol{\alpha})p(\mathbf{B}_{\boldsymbol{\xi}}|\boldsymbol{\xi})d\boldsymbol{\alpha}d\mathbf{B}_{\boldsymbol{\xi}} \\
\propto& p(\boldsymbol{\xi})(|\mathbf{H}_{\boldsymbol{\xi}}|\,|\mathbf{W}_{\boldsymbol{\xi}}|)^{-(M-1)/2}\,|\mathbf{P}_{\boldsymbol{\xi}}|^{-(M-1)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\mathbf{Z}'\mathbf{P}_{\boldsymbol{\xi}}^{-1}\mathbf{Z}\right]\right\},
\end{aligned}
$$

where in general $|\mathbf{H}_{\boldsymbol{\xi}}|\,|\mathbf{W}_{\boldsymbol{\xi}}| = |\mathbf{H}_{\boldsymbol{\xi}}\mathbf{W}_{\boldsymbol{\xi}}| = \left|\mathbf{I}_n + \mathbf{X}_{\boldsymbol{\xi}}\mathbf{H}_{\boldsymbol{\xi}}\mathbf{X}_{\boldsymbol{\xi}}^{'}(\mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n^{'})^{-1}\right|$. In the special case that $\mathbf{X}$ is centered by column and h large it is true that $|\mathbf{H}_{\boldsymbol{\xi}}|\,|\mathbf{W}_{\boldsymbol{\xi}}| \approx \left|\mathbf{I}_n + \mathbf{X}_{\boldsymbol{\xi}}\mathbf{H}_{\boldsymbol{\xi}}\mathbf{X}_{\boldsymbol{\xi}}^{'}\right|$.

Appendix C

Algebra calculations for the probit model with ordinal responses, σ^2 unknown

This appendix contains the algebra calculations of the multi-class BVS with ordinal responses under the assumption that σ^2 is unknown as described in Subsection 6.2.1. The prior distributions are described in the same subsection. This is actually part B of the decomposed approach for σ^2 unknown (Figure 6.1, Algorithm 2 at Chapter 6). Setting $\alpha_0 = 0$ and $\boldsymbol{\beta}_{0\boldsymbol{\gamma}} = \mathbf{0}$ the full conditional distributions are calculated below.

In this case the unknown parameters are $\mathbf{z}, \alpha, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \sigma^2$ and $\mathbf{k}$. We assume that α, $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ are independent conditionally on σ^2, namely that $p(\alpha, \boldsymbol{\beta}_{\boldsymbol{\gamma}}|\sigma^2) = p(\alpha|\sigma^2)p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\sigma^2)$. The joint posterior is proportional to likelihood times the prior distribution and it is given by

$$
\begin{aligned}
& p(\mathbf{z}, \alpha, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \mathbf{k}, \sigma^2|\mathbf{X}, \mathbf{y}) \\
& \propto \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(z_i - \alpha - \mathbf{X}_{i,\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}\right)^2\right\} \\
& \cdot \prod_{i=1}^{n}\mathbb{1}(z_i \in R_i) \cdot \prod_{i=1}^{n}\left[\sum_{m=0}^{M-1}\mathbb{1}(y_i = m)\mathbb{1}(k_m < z_i < k_{m+1})\right] \\
& \cdot \left(\frac{1}{\sigma^2}\right)^{1/2} |\mathbf{H}_{\boldsymbol{\gamma}}|^{-1/2} \exp\left\{-\frac{\boldsymbol{\beta}_{\boldsymbol{\gamma}}\mathbf{H}_{\boldsymbol{\gamma}}^{-1}\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{'}}{2\sigma^2}\right\} \\
& \cdot \frac{1}{h^{1/2}\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{n\alpha^2}{2h\sigma^2}\right\} \\
& \cdot \prod_{j=1}^{p} w^{\gamma_j}(1-w)^{1-\gamma_j} \\
& \cdot \frac{d_2^{d_1}}{\Gamma(d_1)}(\sigma^2)^{-d_1-1}\exp\left\{-\frac{d_2}{\sigma^2}\right\},
\end{aligned}
\tag{C.1}
$$

where R_i is given via Equation (5.19).

First, in order to integrate out α given $\mathbf{z}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}$, $\mathbf{k}$ and σ^2, the exponential terms that are associated with α in Equation (C.1) can be rewritten as

$$
\begin{aligned}
&- \frac{\sum_{i=1}^{n} \left(z_i - \alpha - \mathbf{X}_{i,\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}\right)^2}{2} - \frac{n\alpha^2}{2h} \\
=& - \frac{\left(\mathbf{z} - \mathbf{1}_n\alpha - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}\right)^{'}\left(\mathbf{z} - \mathbf{1}_n\alpha - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}\right)}{2} - \frac{\alpha^2}{2h} \\
=& - \frac{\left(\mathbf{z} - \mathbf{1}_n\alpha - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}\right)^{'}\left(\mathbf{z} - \mathbf{1}_n\alpha - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}\right) + h^{-1}\alpha^2}{2} \\
=& - \frac{\left[\left(\mathbf{z} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}\right) - \mathbf{1}_n\alpha\right]^{'}\left[\left(\mathbf{z} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}\right) - \mathbf{1}_n\alpha\right] + h^{-1}\alpha^2}{2} \\
=& - \frac{(n+h^{-1})\alpha^2 - \mathbf{1}_n^{'}(\mathbf{z} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}})\alpha - \alpha(\mathbf{z} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}})^{'}\mathbf{1}_n - (\mathbf{z} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}})^{'}(\mathbf{z} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}})}{2} \\
=& - \frac{(n+h^{-1})\Big[\alpha^2 - (n+h^{-1})^{-1}\mathbf{1}_n^{'}(\mathbf{z} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}})\alpha - (n+h^{-1})^{-1}\alpha(\mathbf{z} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}})^{'}\mathbf{1}_n}{2} \\
&- \frac{(n+h^{-1})^{-1}(\mathbf{z} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}})^{'}(\mathbf{z} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}})}{2} \\
&\pm \frac{(n+h^{-1})^{-1}h^{-1}(\mathbf{z} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}})^{'}(\mathbf{z} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}})\Big]}{2} \qquad \text{(C.2)} \\
=& - (n+h^{-1})\mathbf{V}\mathbf{V}^{'} - (nh+1)^{-1}(\mathbf{z} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}})^{'}(\mathbf{z} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}),
\end{aligned}
$$

where $V = \{\alpha - (n+h^{-1})^{-1}\mathbf{1}_n^{'}(\mathbf{z} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}})\}$ and $(n+h^{-1})^{-1}h^{-1} = (nh+1)^{-1}$.

Combining the factor $1/\sigma^2$ and the exponential terms of Equation (C.1) together with the Equation (C.2) yield

$$
\begin{aligned}
&\int \frac{1}{h^{1/2}\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{\sigma^2}\left[\frac{(n+h^{-1})V^2}{2}\right.\right. \\
&\left.\left. + \frac{(nh+1)^{-1}(\mathbf{z} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}})^{'}(\mathbf{z} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}})}{2}\right]\right\} d\alpha \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\left[(nh+1)^{-1}(\mathbf{z} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}})^{'}(\mathbf{z} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}})\right]\right\}, \qquad \text{(C.3)}
\end{aligned}
$$

where there is a proportionality constant that comes from the normal distribution $N(V, \sigma^2(n+h^{-1})^{-1})$.

Using Binomial inverse theorem, also known as Sherman-Morrison-Woodbury formula (Woodbury (1950); Plackett (1950)),

$$
\frac{(nh+1)^{-1}(\mathbf{z} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}})^{'}(\mathbf{z} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}})}{2} = \frac{(\mathbf{z} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}})^{'}(\mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n^{'})^{-1}(\mathbf{z} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}})}{2}. \qquad \text{(C.4)}
$$

Secondly, in order to integrate out $\boldsymbol{\beta_\gamma}$, the exponential terms that are associated with $\boldsymbol{\beta_\gamma}$ in Equation (C.1) together with the term that is in the squared brackets of Equation (C.3), using also Equation (C.4), can be rewritten as

$$\begin{aligned}
&\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{'}\mathbf{H}_{\boldsymbol{\gamma}}^{-1}\boldsymbol{\beta}_{\boldsymbol{\gamma}} + (\mathbf{z} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}})^{'}(\mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n^{'})^{-1}(\mathbf{z} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}) \\
=&\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{'}\mathbf{H}_{\boldsymbol{\gamma}}^{-1}\boldsymbol{\beta}_{\boldsymbol{\gamma}} + \mathbf{z}^{'}(\mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n^{\prime})^{-1} - \mathbf{z}^{'}(\mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n^{\prime})^{-1}\mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}} \\
-&\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{'}\mathbf{X}_{\boldsymbol{\gamma}}^{'}(\mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n^{\prime})^{-1}\mathbf{z} + \boldsymbol{\beta}_{\boldsymbol{\gamma}}^{'}\mathbf{X}_{\boldsymbol{\gamma}}^{'}(\mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n^{\prime})^{-1}\mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}} \qquad \text{(C.5)} \\
=&\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{'}\mathbf{W}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}} - \boldsymbol{\beta}_{\boldsymbol{\gamma}}^{'}\mathbf{N} - \mathbf{N}^{'}\boldsymbol{\beta}_{\boldsymbol{\gamma}} + \mathbf{J} \pm \mathbf{N}^{'}\mathbf{W}_{\boldsymbol{\gamma}}^{-1}\mathbf{N} \\
=&(\boldsymbol{\beta}_{\boldsymbol{\gamma}} - \mathbf{W}_{\boldsymbol{\gamma}}^{-1}\mathbf{N})^{'}\mathbf{W}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}} - \mathbf{W}_{\boldsymbol{\gamma}}^{-1}\mathbf{N}) + \mathbf{J} - \mathbf{N}^{'}\mathbf{W}_{\boldsymbol{\gamma}}^{-1}\mathbf{N},
\end{aligned}$$

where $\mathbf{W}_{\boldsymbol{\gamma}} = \mathbf{X}_{\boldsymbol{\gamma}}^{'}(\mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n^{\prime})^{-1}\mathbf{X}_{\boldsymbol{\gamma}} + \mathbf{H}_{\boldsymbol{\gamma}}^{-1}$, $\mathbf{N} = \mathbf{X}_{\boldsymbol{\gamma}}^{'}(\mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n^{\prime})^{-1}\mathbf{z}$ and $\mathbf{J} = \mathbf{Z}^{'}(\mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n^{\prime})^{-1}\mathbf{z}$. Combining the factors $-\frac{1}{2\sigma^2}$, $|\mathbf{H}_{\boldsymbol{\gamma}}|^{-1/2}$ and the corresponding exponential term of Equation (C.1) together with the first term of Equation (C.3) yield the completed quadratic form of the multivariate normal density $MVN(\mathbf{0}, \sigma^2\mathbf{W}_{\boldsymbol{\gamma}}^{-1})$. Continuing the integration of $\boldsymbol{\beta_\gamma}$,

$$\begin{aligned}
&|\mathbf{H}_{\boldsymbol{\gamma}}|^{-1/2}\int \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}} - \mathbf{W}_{\boldsymbol{\gamma}}^{-1}\mathbf{N}\right)^{'}\mathbf{W}_{\boldsymbol{\gamma}}\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}} - \mathbf{W}_{\boldsymbol{\gamma}}^{-1}\mathbf{N}\right)\right\} d\boldsymbol{\beta}_{\boldsymbol{\gamma}} \\
&\cdot \exp\left\{-\frac{1}{2\sigma^2}\left(\mathbf{J} - \mathbf{N}^{'}\mathbf{W}_{\boldsymbol{\gamma}}^{-1}\mathbf{N}\right)\right\} \\
&\propto |\mathbf{H}_{\boldsymbol{\gamma}}|^{-1/2}\left|\sigma^2\mathbf{W}_{\boldsymbol{\gamma}}^{-1}\right|^{1/2}\exp\left\{-\frac{1}{2\sigma^2}\left(\mathbf{J} - \mathbf{N}^{'}\mathbf{W}_{\boldsymbol{\gamma}}^{-1}\mathbf{N}\right)\right\} \qquad \text{(C.6)} \\
&= \left(|\mathbf{H}_{\boldsymbol{\gamma}}|\,|\mathbf{W}_{\boldsymbol{\gamma}}|\right)^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}\left(\mathbf{J} - \mathbf{N}^{'}\mathbf{W}_{\boldsymbol{\gamma}}^{-1}\mathbf{N}\right)\right\}(\sigma^2)^{1/2},
\end{aligned}$$

where $\mathbf{J} - \mathbf{N}^{'}\mathbf{W}_{\boldsymbol{\gamma}}^{-1}\mathbf{N} = \mathbf{z}^{'}\mathbf{P}_{\boldsymbol{\gamma}}^{*-1}\mathbf{z} = \mathbf{z}^{'}\mathbf{P}_{\boldsymbol{\gamma}}^{-1}\mathbf{z}$ and $\mathbf{P}_{\boldsymbol{\gamma}}^{*-1} = (\mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n^{\prime})^{-1} - (\mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n^{\prime})^{-1}\mathbf{X}_{\boldsymbol{\gamma}}\cdot\,[\mathbf{X}_{\boldsymbol{\gamma}}^{'}(\mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n^{\prime})^{-1}\mathbf{X}_{\boldsymbol{\gamma}} + \mathbf{H}_{\boldsymbol{\gamma}}^{-1}]^{-1}\mathbf{X}_{\boldsymbol{\gamma}}^{'}(\mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n^{\prime})^{-1}$. Since $\mathbf{P}_{\boldsymbol{\gamma}}^{*-1}\mathbf{P}_{\boldsymbol{\gamma}} = \mathbf{I}_n$, $\mathbf{P}_{\boldsymbol{\gamma}}^{*}$ takes the simple form $\mathbf{P}_{\boldsymbol{\gamma}} = \mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n^{'} + \mathbf{X}_{\boldsymbol{\gamma}}\mathbf{H}_{\boldsymbol{\gamma}}^{-1}\mathbf{X}_{\boldsymbol{\gamma}}^{'}$. From Equation (C.6) $p(\mathbf{z}|\boldsymbol{\gamma}, \mathbf{k}, \mathbf{X}) \sim MVN(\mathbf{0}, \sigma^2\mathbf{P}_{\boldsymbol{\gamma}})$. From it we can easily calculate the full conditional distribution of $\mathbf{z}$ (also conditioning on the $\mathbf{y}$) so

that

$$
\begin{aligned}
&p(\mathbf{z}|\boldsymbol{\gamma},\mathbf{k},\mathbf{X},\mathbf{y},\sigma^2)\\
&\propto (2\pi)^{n/2}(\sigma^2)^{-n/2}\left|\mathbf{P}_{\boldsymbol{\gamma}}\right|^{-1/2}\exp\left\{-\frac{1}{2}\left[\frac{\mathbf{z}'\mathbf{P}_{\boldsymbol{\gamma}}^{-1}\mathbf{z}}{\sigma^2}\right]\right\}\frac{d_2^{d_1}}{\Gamma(d_1)}(\sigma^2)^{-d_1-1}\\
&\exp\left\{-\frac{d_2}{\sigma^2}\right\}\prod_{i=1}^{n}\mathbb{1}(z_i\in R_i)\\
&\mathbf{z}|\boldsymbol{\gamma},\mathbf{k},\mathbf{X},\mathbf{y},\sigma^2\sim MVN(\mathbf{0},\sigma^2\mathbf{P}_{\boldsymbol{\gamma}})\prod_{i=1}^{n}\mathbb{1}(z_i\in R_i)p(\sigma^2). \qquad \text{(C.7)}
\end{aligned}
$$

Integrating out σ^2 from the last equation, via the density $IG\left(d_1+\frac{n}{2},\frac{1}{2}\mathbf{z}'\mathbf{P}_{\boldsymbol{\gamma}}^{-1}\mathbf{z}+d_2\right)$

$$
\begin{aligned}
&p(\mathbf{z}|\boldsymbol{\gamma},\mathbf{k},\mathbf{X},\mathbf{y})\\
&\propto (2\pi)^{-n/2}\left|\mathbf{P}_{\boldsymbol{\gamma}}\right|^{-1/2}\frac{d_2^{d_1}}{\Gamma(d_1)}\prod_{i=1}^{n}\mathbb{1}(Z_i\in R_i)\int\exp\left\{-\frac{\frac{1}{2}\mathbf{z}'\mathbf{P}_{\boldsymbol{\gamma}}^{-1}\mathbf{z}+d_2}{\sigma^2}\right\}d\sigma^2\\
&\propto \frac{d_2^{d_1}}{2^{n/2}\pi^{n/2}\Gamma(d_1)\left|\mathbf{P}_{\boldsymbol{\gamma}}\right|^{1/2}}\cdot\frac{\Gamma\left(\frac{2d_1+n}{2}\right)}{\left[\frac{1}{2}\mathbf{z}'\mathbf{P}_{\boldsymbol{\gamma}}^{-1}\mathbf{z}+d_2\right]^{d_1+\frac{n}{2}}}\prod_{i=1}^{n}\mathbb{1}(z_i\in R_i)\\
&\propto \frac{d_2^{d_1}\Gamma\left(\frac{2d_1+n}{2}\right)}{2^{n/2}\pi^{n/2}\Gamma(d_1)\left|\mathbf{P}_{\boldsymbol{\gamma}}\right|^{1/2}\left[\frac{1}{2d_2}\mathbf{z}'\mathbf{P}_{\boldsymbol{\gamma}}^{-1}\mathbf{z}+1\right]^{d_1+\frac{n}{2}}}\prod_{i=1}^{n}\mathbb{1}(z_i\in R_i) \qquad \text{(C.8)}\\
&\propto \frac{\frac{d_2^{d_1}}{\Gamma}\left(\frac{2d_1+n}{2}\right)}{\pi^{n/2}\Gamma\left(\frac{2d_1}{2}\right)2^{n/2}\left|\mathbf{P}_{\boldsymbol{\gamma}}\right|^{1/2}\left[\frac{d_1}{2d_2d_1}\mathbf{z}'\mathbf{P}_{\boldsymbol{\gamma}}^{-1}\mathbf{z}+1\right]^{d_1+\frac{n}{2}}}\prod_{i=1}^{n}\mathbb{1}(z_i\in R_i)\\
&\propto \frac{\Gamma\left(\frac{2d_1+n}{2}\right)}{\pi^{n/2}\Gamma\left(\frac{2d_1}{2}\right)\left|2d_1\frac{d_2}{d_1}\mathbf{P}_{\boldsymbol{\gamma}}\right|^{1/2}\left[\frac{1}{2d_1}\mathbf{z}'\left(\frac{d_2}{d_1}\mathbf{P}_{\boldsymbol{\gamma}}\right)^{-1}\mathbf{z}+1\right]^{\frac{2d_1+n}{2}}}\prod_{i=1}^{n}\mathbb{1}(z_i\in R_i).
\end{aligned}
$$

where $\Gamma(\cdot)$ denotes the gamma function.
So, $\mathbf{z}|\boldsymbol{\gamma},\mathbf{k},\mathbf{X},\mathbf{y}\sim MVT\left(2d_1;\mathbf{0},\frac{d_2}{d_1}\mathbf{P}_{\boldsymbol{\gamma}}\right)\prod_{i=1}^{n}\mathbb{1}(z_i\in R_i)$.

Using Equations (C.1), (C.6) and (C.7) it is true that

$$
\begin{aligned}
p(\mathbf{z}|\boldsymbol{\gamma}) \propto & \int (|\mathbf{H}_{\boldsymbol{\gamma}}|\,|\mathbf{W}_{\boldsymbol{\gamma}}|)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}\left(\mathbf{J}-\mathbf{N}^{'}\mathbf{W}_{\boldsymbol{\gamma}}^{-1}\mathbf{N}\right)\right\}(\sigma^2)^{n/2} \\
& \cdot \frac{d_2^{d_1}}{\Gamma(d_1)}(\sigma^2)^{-d_1-1}\exp\left\{-\frac{d_2}{\sigma^2}\right\}d\sigma^2 \qquad \text{(C.9)} \\
& \propto (|\mathbf{H}_{\boldsymbol{\gamma}}|\,|\mathbf{W}_{\boldsymbol{\gamma}}|)^{-1/2}\frac{d_2^{d_1}}{\Gamma(d_1)}\int \exp\left\{-\frac{\mathbf{z}^{'}\mathbf{P}_{\boldsymbol{\gamma}}^{-1}\mathbf{z}+2d_2}{2\sigma^2}\right\}\left(\frac{1}{\sigma^2}\right)^{d_1+\frac{n}{2}+1}d\sigma^2 \\
& \propto (|\mathbf{H}_{\boldsymbol{\gamma}}|\,|\mathbf{W}_{\boldsymbol{\gamma}}|)^{-1/2}|P_{\boldsymbol{\gamma}}|^{-1}
\end{aligned}
$$

where the proportionality constants do not depend on $\boldsymbol{\gamma}$ and $|\mathbf{H}_{\boldsymbol{\gamma}}|\,|\mathbf{W}_{\boldsymbol{\gamma}}| = |\mathbf{H}_{\boldsymbol{\gamma}}\mathbf{W}_{\boldsymbol{\gamma}}| = \left|\mathbf{I}_n + \mathbf{X}_{\boldsymbol{\gamma}}\mathbf{H}_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}}^{'}(\mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n^{'})^{-1}\right|$. In the special case that $\mathbf{X}$ is centered by column and h large it is true that $|\mathbf{H}_{\boldsymbol{\gamma}}|\,|\mathbf{W}_{\boldsymbol{\gamma}}| \approx \left|\mathbf{I}_n + \mathbf{X}_{\boldsymbol{\gamma}}\mathbf{H}_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}}^{'}\right|$.

Then, after integrating out the three unknown parameters,

$$
\begin{aligned}
& p(\boldsymbol{\gamma}|\mathbf{z},\mathbf{k},\mathbf{X},\mathbf{y}) = p(\boldsymbol{\gamma})p(\mathbf{X},\mathbf{y},\mathbf{Z},\mathbf{k}|\boldsymbol{\gamma}) \\
=& p(\boldsymbol{\gamma})p(\mathbf{X},\mathbf{y},\mathbf{k}|\mathbf{Z},\boldsymbol{\gamma})p(\mathbf{Z}|\boldsymbol{\gamma}) = p(\boldsymbol{\gamma})p(\mathbf{X},\mathbf{y},\mathbf{k}|\mathbf{Z})p(\mathbf{Z}|\boldsymbol{\gamma}) \\
\propto & p(\boldsymbol{\gamma})p(\mathbf{Z}|\boldsymbol{\gamma}) \\
\propto & p(\boldsymbol{\gamma})(|\mathbf{H}_{\boldsymbol{\gamma}}|\,|\mathbf{W}_{\boldsymbol{\gamma}}|)^{-1/2}\,|\mathbf{P}_{\boldsymbol{\gamma}}|^{-1/2},
\end{aligned}
$$

The last factorization based on the fact that $\boldsymbol{\gamma}$ depends on $\mathbf{Z}$, $\mathbf{y}$ and $\mathbf{k}$ but once we condition on $\mathbf{Z}$, the $\mathbf{k}$ and $\mathbf{y}$ is independent of $\boldsymbol{\gamma}$. Then, conditional probability of $\boldsymbol{\xi}$ using Equations (B.6) and (B.7) is the following

This fully conditional density of k_ν given the rest is calculated easily from Equation (C.1), selecting only the terms that associated with the boundary vector. Then the full conditional density of each component is given by

$$
\begin{aligned}
p(k_\nu|\boldsymbol{\gamma},\mathbf{z},\mathbf{X},\mathbf{y},\mathbf{k}_{\setminus\nu}) \propto \prod_{i=1}^{n}[&\mathbb{1}(y_i=\nu-1)\mathbb{1}(k_{\nu-1}<z_i\le k_\nu) \\
&+\mathbb{1}(y_i=\nu)\mathbb{1}(k_\nu<z_i\le k_{\nu+1})],
\end{aligned} \qquad \text{(C.10)}
$$

where $\nu = 2,\ldots,M-1$ and $\mathbf{k}_{\setminus\nu} = (k_0,\ldots,k_{\nu-1},k_{\nu+1},\ldots,k_M)$.

Actually, sampling via Equations (C.10) is done via continuous uniform distribution, namely via

$$
k_\nu|\boldsymbol{\gamma},\mathbf{z},\mathbf{X},\mathbf{y},\mathbf{k}_{\setminus\nu} \sim U\big(\max[\{z_i : y_i=\nu-1\},k_{\nu-1}],\min[\{z_i : y_i=\nu\},k_{\nu+1}]\big).
$$

Appendix D

Algebra calculations for the probit model with mixture of nominal and ordinal responses - $\boldsymbol{\Sigma}$ is fixed, common $\boldsymbol{\xi}$

This appendix contains the algebra calculations of the multi-class Bayesian variable selection with mixture of nominal and ordinal responses, under the assumption that $\boldsymbol{\Sigma}$ is fixed and $\boldsymbol{\xi}$ is common across, as described in Subsection 7.1.1. The prior distributions are described in Subsection 7.1.2. Setting $\boldsymbol{\alpha}_0 = \mathbf{0}$ and $\mathbf{B}_{0\boldsymbol{\xi}} = \mathbf{0}$ the full conditional distributions are derived below.

In this case the unknown parameters are $\mathbf{Z}, \boldsymbol{\alpha}, \mathbf{B}_{\boldsymbol{\xi}}, \boldsymbol{\xi}$ and $\mathbf{k}$. We assume that $\boldsymbol{\alpha}$, $\mathbf{B}_{\boldsymbol{\xi}}$ are independent, namely that $p(\boldsymbol{\alpha}, \mathbf{B}_{\boldsymbol{\xi}}) = p(\boldsymbol{\alpha})p(\mathbf{B}_{\boldsymbol{\xi}})$. The joint posterior is proportional to likelihood times the prior distribution and it is given by

$$
\begin{aligned}
& p(\mathbf{Z}, \boldsymbol{\alpha}, \mathbf{B}_{\boldsymbol{\xi}}, \boldsymbol{\xi}, \mathbf{k} | \mathbf{X}, \mathbf{y}) \\
\propto & |\boldsymbol{\Sigma}|^{-n/2} \\
& \cdot \exp\left\{ -\frac{1}{2} \mathrm{tr}\Big[\Big(\mathbf{Z} - \mathbf{1}_n \boldsymbol{\alpha}^{'} - \mathbf{X}_{\boldsymbol{\xi}} \mathbf{B}_{\boldsymbol{\xi}} \Big) \boldsymbol{\Sigma}^{-1} (\mathbf{Z} - \mathbf{1}_n \boldsymbol{\alpha}^{'} - \mathbf{X}_{\boldsymbol{\xi}} \mathbf{B}_{\boldsymbol{\xi}})^{'} \Big] \right\} \prod_{i=1}^{n} \mathbb{1}(\mathbf{Z}_{i,:} \in G_i) \\
& \cdot \prod_{i=1}^{n} [\mathbb{1}(y_i = t_{\nu-1}) \mathbb{1}(k_{\nu-1} < Z_{i,t_0+1} \leq k_{\nu}) + \mathbb{1}(y_i = t_{\nu}) \mathbb{1}(k_{\nu} < Z_{i,t_0+1} \leq k_{\nu+1})] \\
& \cdot |\mathbf{H}_{\boldsymbol{\xi}}|^{-s/2} |\boldsymbol{\Sigma}|^{-p_{\boldsymbol{\xi}}/2} \exp\left\{ -\frac{1}{2} \mathrm{tr}\Big[\mathbf{H}_{\boldsymbol{\xi}}^{-1} \mathbf{B}_{\boldsymbol{\xi}} \boldsymbol{\Sigma}^{-1} \mathbf{B}_{\boldsymbol{\xi}}^{'} \Big] \right\} \\
& \cdot h^{-s/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left\{ -\frac{h^{-1} \boldsymbol{\alpha}^{'} \boldsymbol{\Sigma}^{-1} \boldsymbol{\alpha}}{2} \right\} \\
& \cdot \prod_{j=1}^{p} w^{\gamma_j} (1-w)^{1-\gamma_j} \prod_{i=1}^{n} \mathbb{1}(y_i = t_{\nu}) \frac{1}{k_{\nu+1} - k_{\nu-1}}
\end{aligned}
\tag{D.1}
$$

where G_i is given via Equation (7.6).

Firstly, in order to integrate out $\boldsymbol{\alpha}$ given $\mathbf{Z}, \mathbf{B}_{\boldsymbol{\xi}}, \boldsymbol{\xi}$ and $\mathbf{k}$, the exponential terms that are associated with $\boldsymbol{\alpha}$ in Equation (D.1) can be rewritten as

$$
\begin{aligned}
& -\left(\mathbf{Z}-\mathbf{1}_n\boldsymbol{\alpha}'-\mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}}\right)'\left(\mathbf{Z}-\mathbf{1}_n\boldsymbol{\alpha}'-\mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}}\right)+h^{-1}\boldsymbol{\alpha}\boldsymbol{\alpha}' \\
=& -(n+h^{-1})\boldsymbol{\alpha}\boldsymbol{\alpha}'-\mathbf{1}_n'(\mathbf{Z}-\mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}})\boldsymbol{\alpha}-\boldsymbol{\alpha}' \\
& +(\mathbf{Z}-\mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}})'\mathbf{1}_n-(\mathbf{Z}-\mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}})'(\mathbf{Z}-\mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}}) \\
=& -(n+h^{-1})\Big[\boldsymbol{\alpha}\boldsymbol{\alpha}'-(n+h^{-1})^{-1}\mathbf{1}_n'(\mathbf{Z}-\mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}})\boldsymbol{\alpha} \\
& +(n+h^{-1})^{-1}\boldsymbol{\alpha}'(\mathbf{Z}-\mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}})'\mathbf{1}_n \\
& +(n+h^{-1})^{-1}(\mathbf{Z}-\mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}})'(\mathbf{Z}-\mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}}) \\
& \pm(n+h^{-1})^{-1}h^{-1}(\mathbf{Z}-\mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}})'(\mathbf{Z}-\mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}})\Big] \\
=& -(n+h^{-1})\mathbf{V}\mathbf{V}'-(nh+1)^{-1}(\mathbf{Z}-\mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}})'(\mathbf{Z}-\mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}}),
\end{aligned}
\tag{D.2}
$$

where $\mathbf{V}=\{\boldsymbol{\alpha}'-(n+h^{-1})^{-1}\mathbf{1}_n'(\mathbf{Z}-\mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}})\}$ and $(n+h^{-1})^{-1}h^{-1}=(nh+1)^{-1}$. Combining the factors $\frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1})$, $|\boldsymbol{\Sigma}|^{-1/2}$ and the exponential terms of the first and penultimate line in Equation (D.1) together with the Equation (D.2) yield

$$
\begin{aligned}
& \int|\boldsymbol{\Sigma}|^{1/2}\exp\left\{-\tfrac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\left[(n+h^{-1})\mathbf{V}\mathbf{V}'\right]\right)\right\} \\
\cdot&\exp\left\{-\tfrac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\left[(nh+1)^{-1}(\mathbf{Z}-\mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}})'(\mathbf{Z}-\mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}})\right]\right)\right\}d\boldsymbol{\alpha} \\
\propto&\exp\left\{-\tfrac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\left[(nh+1)^{-1}(\mathbf{Z}-\mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}})'(\mathbf{Z}-\mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}})\right]\right)\right\},
\end{aligned}
\tag{D.3}
$$

where the proportionality hides the constant that comes from the matrix normal density $MN(\boldsymbol{\Sigma}, \mathbf{I}_n)$.

Using Binomial inverse theorem, also known as Sherman-Morrison-Woodbury formula (Woodbury, 1950; Plackett, 1950),

$$
\begin{aligned}
& (nh+1)^{-1}(\mathbf{Z}-\mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}})'(\mathbf{Z}-\mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}}) \\
&=(\mathbf{Z}-\mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}})'(\mathbf{I}_n+h\mathbf{1}_n\mathbf{1}_n')^{-1}(\mathbf{Z}-\mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}}).
\end{aligned}
\tag{D.4}
$$

Secondly, in order to integrate out $\mathbf{B}_{\boldsymbol{\xi}}$, the exponential term that are associated with $\mathbf{B}_{\boldsymbol{\xi}}$ in Equation (D.1) together with the term that is in the squared brackets of Equation (D.3) using also Equation (D.4) can be rewritten

as

$$
\begin{aligned}
&\mathbf{B}_{\boldsymbol{\xi}}^{'}\mathbf{H}_{\boldsymbol{\xi}}^{-1}\mathbf{B}_{\boldsymbol{\xi}}+(\mathbf{Z}-\mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}})^{'}(\mathbf{I}_n+h\mathbf{1}_n\mathbf{1}_n^{'})^{-1}(\mathbf{Z}-\mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}})\\
=&\mathbf{B}_{\boldsymbol{\xi}}^{'}\mathbf{H}_{\boldsymbol{\xi}}^{-1}\mathbf{B}_{\boldsymbol{\xi}}+\mathbf{Z}^{'}(\mathbf{I}_n+h\mathbf{1}_n\mathbf{1}_n^{\prime})^{-1}\mathbf{Z}-\mathbf{Z}^{'}(\mathbf{I}_n+h\mathbf{1}_n\mathbf{1}_n^{\prime})^{-1}\mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}}\\
-&\mathbf{B}_{\boldsymbol{\xi}}^{'}\mathbf{X}_{\boldsymbol{\xi}}^{'}(\mathbf{I}_n+h\mathbf{1}_n\mathbf{1}_n^{\prime})^{-1}\mathbf{Z}+\mathbf{B}_{\boldsymbol{\xi}}^{'}\mathbf{X}_{\boldsymbol{\xi}}^{'}(\mathbf{I}_n+h\mathbf{1}_n\mathbf{1}_n^{\prime})^{-1}\mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}}\\
=&\mathbf{B}_{\boldsymbol{\xi}}^{'}\mathbf{W}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}}-\mathbf{B}_{\boldsymbol{\xi}}^{'}\mathbf{N}-\mathbf{N}^{'}\mathbf{B}_{\boldsymbol{\xi}}+\mathbf{J}\pm\mathbf{N}^{'}\mathbf{W}_{\boldsymbol{\xi}}^{-1}\mathbf{N} \qquad \text{(D.5)}\\
=&(\mathbf{B}_{\boldsymbol{\xi}}-\mathbf{W}_{\boldsymbol{\xi}}^{-1}\mathbf{N})^{'}\mathbf{W}_{\boldsymbol{\xi}}(\mathbf{B}_{\boldsymbol{\xi}}-\mathbf{W}_{\boldsymbol{\xi}}^{-1}\mathbf{N})+\mathbf{J}-\mathbf{N}^{'}\mathbf{W}_{\boldsymbol{\xi}}^{-1}\mathbf{N},
\end{aligned}
$$

where $\mathbf{W}_{\boldsymbol{\xi}}=\mathbf{X}_{\boldsymbol{\xi}}^{'}(\mathbf{I}_n+h\mathbf{1}_n\mathbf{1}_n^{\prime})^{-1}\mathbf{X}_{\boldsymbol{\xi}}+\mathbf{H}_{\boldsymbol{\xi}}^{-1}$, $\mathbf{N}=\mathbf{X}_{\boldsymbol{\xi}}^{'}(\mathbf{I}_n+h\mathbf{1}_n\mathbf{1}_n^{\prime})^{-1}\mathbf{Z}$ and $\mathbf{J}=\mathbf{Z}^{'}(\mathbf{I}_n+h\mathbf{1}_n\mathbf{1}_n^{\prime})^{-1}\mathbf{Z}$. Combining the factors $\frac{1}{2}\text{tr}(\boldsymbol{\Sigma}^{-1})$, $|\boldsymbol{\Sigma}|^{-n/2}$, $|\mathbf{H}_{\boldsymbol{\xi}}|^{-s/2}$, $|\boldsymbol{\Sigma}|^{-p_{\boldsymbol{\xi}}/2}$ and the corresponding exponential term of Equation (D.1) together with the first term of Equation (D.5) yield the completed quadratic form of the matrix normal density $MN(\boldsymbol{\Sigma},\mathbf{W}_{\boldsymbol{\xi}}^{-1})$. Continuing the integration of $\mathbf{B}_{\boldsymbol{\xi}}$,

$$
\begin{aligned}
&|\boldsymbol{\Sigma}|^{-n/2}|\mathbf{H}_{\boldsymbol{\xi}}|^{-s/2}|\boldsymbol{\Sigma}|^{-p_{\boldsymbol{\xi}}/2}\\
&\int\exp\left\{-\frac{1}{2}\text{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\mathbf{B}_{\boldsymbol{\xi}}-\mathbf{W}_{\boldsymbol{\xi}}^{-1}\mathbf{N}\right)^{'}\mathbf{W}_{\boldsymbol{\xi}}\left(\mathbf{B}_{\boldsymbol{\xi}}-\mathbf{W}_{\boldsymbol{\xi}}^{-1}\mathbf{N}\right)\right]\right\}d\mathbf{B}_{\boldsymbol{\xi}}\\
&\exp\left\{-\frac{1}{2}\text{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\mathbf{J}-\mathbf{N}^{'}\mathbf{W}_{\boldsymbol{\xi}}^{-1}\mathbf{N}\right)\right]\right\}\\
&=|\boldsymbol{\Sigma}|^{-n/2}|\mathbf{H}_{\boldsymbol{\xi}}|^{-s/2}|\boldsymbol{\Sigma}|^{-p_{\boldsymbol{\xi}}/2}|\boldsymbol{\Sigma}|^{p_{\boldsymbol{\xi}}/2}\left|\mathbf{W}_{\boldsymbol{\xi}}^{-1}\right|^{s/2} \qquad \text{(D.6)}\\
&\cdot\exp\left\{-\frac{1}{2}\text{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\mathbf{J}-\mathbf{N}^{'}\mathbf{W}_{\boldsymbol{\xi}}^{-1}\mathbf{N}\right)\right]\right\}\\
&\propto|\boldsymbol{\Sigma}|^{-n/2}\left(|\mathbf{H}_{\boldsymbol{\xi}}|\,|\mathbf{W}_{\boldsymbol{\xi}}|\right)^{-s/2}\\
&\cdot\exp\left\{-\frac{1}{2}\text{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\mathbf{J}-\mathbf{N}^{'}\mathbf{W}_{\boldsymbol{\xi}}^{-1}\mathbf{N}\right)\right]\right\},
\end{aligned}
$$

where $\mathbf{J}-\mathbf{N}^{'}\mathbf{W}_{\boldsymbol{\xi}}^{-1}\mathbf{N}=\mathbf{Z}^{'}\mathbf{P}_{\boldsymbol{\xi}}^{*-1}\mathbf{Z}=\mathbf{Z}^{'}\mathbf{P}_{\boldsymbol{\xi}}^{-1}\mathbf{Z}$ and $\mathbf{P}_{\boldsymbol{\xi}}^{*-1}=(\mathbf{I}_n+h\mathbf{1}_n\mathbf{1}_n^{\prime})^{-1}-(\mathbf{I}_n+h\mathbf{1}_n\mathbf{1}_n^{\prime})^{-1}\mathbf{X}_{\boldsymbol{\xi}}\cdot\,[\mathbf{X}_{\boldsymbol{\xi}}^{'}(\mathbf{I}_n+h\mathbf{1}_n\mathbf{1}_n^{\prime})^{-1}\mathbf{X}_{\boldsymbol{\xi}}+\mathbf{H}_{\boldsymbol{\xi}}^{-1}]^{-1}\mathbf{X}_{\boldsymbol{\xi}}^{'}(\mathbf{I}_n+h\mathbf{1}_n\mathbf{1}_n^{\prime})^{-1}$. Since $\mathbf{P}_{\boldsymbol{\xi}}^{*-1}\mathbf{P}_{\boldsymbol{\xi}}=\mathbf{I}_n$, $\mathbf{P}_{\boldsymbol{\xi}}^{*}$ takes the simple form $\mathbf{P}_{\boldsymbol{\xi}}=\mathbf{I}_n+h\mathbf{1}_n\mathbf{1}_n^{'}+\mathbf{X}_{\boldsymbol{\xi}}\mathbf{H}_{\boldsymbol{\xi}}^{-1}\mathbf{X}_{\boldsymbol{\xi}}^{'}$. From Equation (D.6) it is true that $p(\mathbf{Z}|\boldsymbol{\xi},\mathbf{k},\mathbf{X})\sim MN(\mathbf{P}_{\boldsymbol{\xi}},\boldsymbol{\Sigma})$, which is a matrix normal distribution. From the last expression we can easily derive the full conditional distribution of $\mathbf{Z}$ given the rest, so that

$$
\begin{aligned}
\mathbf{Z}|\boldsymbol{\xi},\mathbf{k},\mathbf{X},\mathbf{y}&\sim MN(\mathbf{P}_{\boldsymbol{\xi}},\boldsymbol{\Sigma})\prod_{i=1}^{n}\mathbb{1}(\mathbf{Z}_{i,:}\in G_i) \qquad \text{(D.7)}\\
p(\mathbf{Z}|\boldsymbol{\xi},\mathbf{k},\mathbf{X},\mathbf{y}&\propto|\boldsymbol{\Sigma}|^{-n/2}|\mathbf{P}_{\boldsymbol{\xi}}|^{-s/2}\exp\left\{-\frac{1}{2}\text{tr}\left[\boldsymbol{\Sigma}^{-1}\mathbf{Z}^{'}\mathbf{P}_{\boldsymbol{\xi}}^{-1}\mathbf{Z}\right]\right\}\prod_{i=1}^{n}\mathbb{1}(\mathbf{Z}_{i,:}\in G_i).
\end{aligned}
$$

After integrating out $\boldsymbol{\alpha},\mathbf{B}_{\boldsymbol{\xi}}$, we derive the full conditional distribution of

$\boldsymbol{\xi}$. Using Equations (D.1), (D.6) and (D.7), and taking into account the relationship between the parameters of the model,

$$
\begin{aligned}
p(\boldsymbol{\xi}|\mathbf{Z},\mathbf{k},\mathbf{X},\mathbf{y}) &\propto p(\boldsymbol{\xi})p(\mathbf{Z}|\boldsymbol{\xi}) \\
&\propto p(\boldsymbol{\xi})(|\mathbf{H}_{\boldsymbol{\xi}}|\,|\mathbf{W}_{\boldsymbol{\xi}}|)^{-s/2}\,|\mathbf{P}_{\boldsymbol{\xi}}|^{-s/2}\exp\left\{-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\mathbf{Z}'\mathbf{P}_{\boldsymbol{\xi}}^{-1}\mathbf{Z}\right]\right\},
\end{aligned}
$$

where in general $|\mathbf{H}_{\boldsymbol{\xi}}|\,|\mathbf{W}_{\boldsymbol{\xi}}| = |\mathbf{H}_{\boldsymbol{\xi}}\mathbf{W}_{\boldsymbol{\xi}}| = \left|\mathbf{I}_n + \mathbf{X}_{\boldsymbol{\xi}}\mathbf{H}_{\boldsymbol{\xi}}\mathbf{X}_{\boldsymbol{\xi}}'\ (\mathbf{I}_n + h\mathbf{1}_n\mathbf{1}_n')^{-1}\right|$. In the special case that $\mathbf{X}$ is centered by column and h large it is true that $|\mathbf{H}_{\boldsymbol{\xi}}|\,|\mathbf{W}_{\boldsymbol{\xi}}| \approx \left|\mathbf{I}_n + \mathbf{X}_{\boldsymbol{\xi}}\mathbf{H}_{\boldsymbol{\xi}}\mathbf{X}_{\boldsymbol{\xi}}'\right|$.

This fully conditional density of k_j given the rest is derived only for ordinal responses. From Equation (D.1), easily select only the terms that associated with the boundary vector and then the full conditional density of its component is given by Equation (5.21).

Appendix E

Algebra calculations for the probit model with mixture of nominal and ordinal responses - $\boldsymbol{\Sigma}$ has a prior, common $\boldsymbol{\xi}$

This appendix contains the algebra calculations of the multi-class Bayesian variable selection with mixture of nominal and ordinal responses, under the assumption that $\boldsymbol{\xi}$ is common, but $\boldsymbol{\Sigma}$ has a distribution in this case, as described in Subsection 7.1.3. Setting $\boldsymbol{\alpha}_0 = \mathbf{0}$ and $\mathbf{B}_{0\boldsymbol{\xi}} = \mathbf{0}$ the full conditional distributions are derived below.

In this case the unknown parameters are $\mathbf{Z}, \boldsymbol{\alpha}, \mathbf{B}_{\boldsymbol{\xi}}, \boldsymbol{\xi}, \mathbf{k}$ and in addition $\boldsymbol{\Sigma}$. We assume that $\boldsymbol{\alpha}$, $\mathbf{B}_{\boldsymbol{\xi}}$ are independent conditionally on $\boldsymbol{\Sigma}$, $p(\boldsymbol{\alpha}, \mathbf{B}_{\boldsymbol{\xi}}|\boldsymbol{\Sigma}) = p(\boldsymbol{\alpha}|\boldsymbol{\Sigma})p(\mathbf{B}_{\boldsymbol{\xi}}|\boldsymbol{\Sigma})$. The joint posterior is proportional to likelihood times the

prior distribution and it is given by

$$
\begin{aligned}
& p(\mathbf{Z}, \boldsymbol{\alpha}, \mathbf{B}_{\boldsymbol{\xi}}, \boldsymbol{\xi}, \mathbf{k}, \boldsymbol{\Sigma} | \mathbf{X}, \mathbf{y}) \\
\propto & |\boldsymbol{\Sigma}|^{-n/2} \\
& \cdot \exp\left\{-\frac{1}{2}\mathrm{tr}\left[\left(\mathbf{Z}-\mathbf{1}_n\boldsymbol{\alpha}^{'}-\mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}}\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{Z}-\mathbf{1}_n\boldsymbol{\alpha}^{'}-\mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}}\right)^{'}\right]\right\}\prod_{i=1}^{n}\mathbb{1}(\mathbf{Z}_{i,:}\in G_i) \\
& \cdot \prod_{i=1}^{n}[\mathbb{1}(y_i=t_{\nu-1})\mathbb{1}(k_{\nu-1}<Z_{i,t_0+1}\leq k_\nu)+\mathbb{1}(y_i=t_\nu)\mathbb{1}(k_\nu<Z_{i,t_0+1}\leq k_{\nu+1})] \\
& \cdot |\mathbf{H}_{\boldsymbol{\xi}}|^{-s/2}|\boldsymbol{\Sigma}|^{-p_{\boldsymbol{\xi}}/2}\exp\left\{-\frac{1}{2}\mathrm{tr}\left[\mathbf{H}_{\boldsymbol{\xi}}^{-1}\mathbf{B}_{\boldsymbol{\xi}}\boldsymbol{\Sigma}^{-1}\mathbf{B}_{\boldsymbol{\xi}}^{'}\right]\right\} \qquad \text{(E.1)} \\
& \cdot h^{-s/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\left\{-\frac{h^{-1}\boldsymbol{\alpha}^{'}\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}{2}\right\} \\
& \cdot \prod_{j=1}^{p}w^{\gamma_j}(1-w)^{1-\gamma_j}\prod_{i=1}^{n}\mathbb{1}(y_i=t_\nu)\frac{1}{k_{\nu+1}-k_{\nu-1}} \\
& \cdot |\mathbf{Q}|^{\frac{\delta+s-1}{2}}|\boldsymbol{\Sigma}|^{-\frac{\delta+2s}{2}}\exp\left\{-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{Q}\right)\right\}
\end{aligned}
$$

where G_i is given via Equation (7.6).

Firstly, in order to integrate out $\boldsymbol{\alpha}$ given $\mathbf{Z}, \mathbf{B}_{\boldsymbol{\xi}}, \boldsymbol{\xi}, \mathbf{k}$ and $\boldsymbol{\Sigma}$, the exponential terms that are associated with $\boldsymbol{\alpha}$ (excluding $\frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1})$) in Equation (E.1) can be rewritten as Equation (D.3), where Equations (D.2) and (D.4) are also true.

Secondly, integrating out $\mathbf{B}_{\boldsymbol{\xi}}$ yields Equation (D.6). Similar to Equation (D.6) $\mathbf{Z}|\boldsymbol{\xi}, \mathbf{k}, \boldsymbol{\Sigma}, \mathbf{X} \sim MN(\mathbf{P}_{\boldsymbol{\xi}}, \boldsymbol{\Sigma})p(\boldsymbol{\Sigma})$. From the last expression, where the $\boldsymbol{\Sigma}$ density is given via the last line of Equation (E.1), we can easily derive the full conditional distribution of $\mathbf{Z}$ given the rest:

$$
\begin{aligned}
& \mathbf{Z}|\boldsymbol{\xi}, \mathbf{k}, \boldsymbol{\Sigma}, \mathbf{X}, \mathbf{y} \sim MN(\mathbf{P}_{\boldsymbol{\xi}}, \boldsymbol{\Sigma})\prod_{i=1}^{n}\mathbb{1}(\mathbf{Z}_{i,:}\in G_i)p(\boldsymbol{\Sigma}) \\
& p(\mathbf{Z}|\boldsymbol{\xi}, \mathbf{k}, \boldsymbol{\Sigma}, \mathbf{X}, \mathbf{y}) \\
\propto & |\boldsymbol{\Sigma}|^{-n/2}|\mathbf{P}_{\boldsymbol{\xi}}|^{-s/2}\exp\left\{-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\mathbf{Z}^{'}\mathbf{P}_{\boldsymbol{\xi}}^{-1}\mathbf{Z}\right]\right\}\prod_{i=1}^{n}\mathbb{1}(\mathbf{Z}_{i,:}\in G_i) \\
& \cdot |\mathbf{Q}|^{\frac{\delta+s-1}{2}}|\boldsymbol{\Sigma}|^{-\frac{\delta+2s}{2}}\exp\left\{-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{Q}\right)\right\} \qquad \text{(E.2)} \\
\propto & |\mathbf{P}_{\boldsymbol{\xi}}|^{-s/2}|\boldsymbol{\Sigma}|^{-\frac{n+\delta+2s}{2}}|\mathbf{Q}|^{\frac{\delta+s-1}{2}} \\
& \cdot \exp\left\{-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\mathbf{Z}^{'}\mathbf{P}_{\boldsymbol{\xi}}^{-1}\mathbf{Z}+\mathbf{Q}\right)\right]\right\}\prod_{i=1}^{n}\mathbb{1}(\mathbf{Z}_{i,:}\in G_i).
\end{aligned}
$$

In addition, integrating out $\boldsymbol{\Sigma}$ from (E.2) and using the determinant prop-

erty $\left|\mathbf{A}+\mathbf{USR}^{'}\right| = \left|\mathbf{S}^{-1}+\mathbf{R}^{'}\mathbf{A}^{-1}\mathbf{U}\right| |\mathbf{S}|\,|\mathbf{A}|$, we obtain

$$
\begin{aligned}
& p(\mathbf{Z}|\boldsymbol{\xi},\mathbf{k},\boldsymbol{\Sigma},\mathbf{X},\mathbf{y}) \\
\propto & \left|\mathbf{P}_{\boldsymbol{\xi}}\right|^{-s/2} |\mathbf{Q}|^{\frac{\delta+s-1}{2}} \left|\mathbf{Z}^{'}\mathbf{P}_{\boldsymbol{\xi}}^{-1}\mathbf{Z}+\mathbf{Q}\right|^{-\frac{n+\delta+s-1}{2}} \prod_{i=1}^{n} \mathbb{1}(\mathbf{Z}_{i,:}\in G_i) \\
\propto & \left|\mathbf{P}_{\boldsymbol{\xi}}\right|^{-s/2} |\mathbf{Q}|^{\frac{\delta+s-1}{2}} \\
\cdot & \left(\left|\mathbf{ZQ}^{-1}\mathbf{Z}^{'}+\mathbf{P}_{\boldsymbol{\xi}}\right| \left|\mathbf{P}_{\boldsymbol{\xi}}^{-1}\right| |\mathbf{Q}|\right)^{-\frac{n+\delta+s-1}{2}} \prod_{i=1}^{n} \mathbb{1}(\mathbf{Z}_{i,:}\in G_i) \\
\propto & \left|\mathbf{P}_{\boldsymbol{\xi}}\right|^{-s/2} |\mathbf{Q}|^{\frac{\delta+s-1}{2}} \\
\cdot & \left|\mathbf{ZQ}^{-1}\mathbf{Z}^{'}+\mathbf{P}_{\boldsymbol{\xi}}\right|^{-\frac{n+\delta+s-1}{2}} |\mathbf{P}_{\boldsymbol{\xi}}|^{\frac{n+\delta+s-1}{2}} |\mathbf{Q}|^{-\frac{n+\delta+s-1}{2}} \prod_{i=1}^{n} \mathbb{1}(\mathbf{Z}_{i,:}\in G_i) \\
\propto & \left|\mathbf{P}_{\boldsymbol{\xi}}\right|^{\frac{n+\delta-1}{2}} |\mathbf{Q}|^{-\frac{n}{2}} \left|\mathbf{ZQ}^{-1}\mathbf{Z}^{'}+\mathbf{P}_{\boldsymbol{\xi}}\right|^{-\frac{n+\delta+s-1}{2}} \prod_{i=1}^{n} \mathbb{1}(\mathbf{Z}_{i,:}\in G_i)
\end{aligned}
\tag{E.3}
$$

So,

$$
\mathbf{Z}|\boldsymbol{\xi},\mathbf{k},\boldsymbol{\Sigma},\mathbf{X},\mathbf{y} \sim MT(\delta;\mathbf{P}_{\boldsymbol{\xi}},\mathbf{Q}) \prod_{i=1}^{n} \mathbb{1}(\mathbf{Z}_{i,:}\in G_i)
$$

which is a truncated matrix Student distribution.

Using Equations (D.6) and the last line of Equation (E.1), we calculate the

$$
\begin{aligned}
& p(\mathbf{Z}|\boldsymbol{\xi}) \\
\propto & \int (|\mathbf{H}_{\boldsymbol{\xi}}|\,|\mathbf{W}_{\boldsymbol{\xi}}|)^{-s/2} |\mathbf{H}_{\boldsymbol{\xi}}|\,|\boldsymbol{\Sigma}|^{-n/2} \\
& \cdot \exp\left\{-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\mathbf{J}-\mathbf{N}^{'}\mathbf{W}_{\boldsymbol{\xi}}^{-1}\mathbf{N}\right)\right]\right\} \\
& \cdot |\boldsymbol{\Sigma}|^{-\frac{\delta+2s}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{Q}\right)\right\} d\boldsymbol{\Sigma} \\
= & \int (|\mathbf{H}_{\boldsymbol{\xi}}|\,|\mathbf{W}_{\boldsymbol{\xi}}|)^{-s/2} |\boldsymbol{\Sigma}|^{-\frac{n+\delta+2s}{2}} \\
& \cdot \exp\left\{-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\mathbf{Q}+\mathbf{J}-\mathbf{N}^{'}\mathbf{W}_{\boldsymbol{\xi}}^{-1}\mathbf{N}\right)\right]\right\} d\boldsymbol{\Sigma} \\
\propto & (|\mathbf{H}_{\boldsymbol{\xi}}|\,|\mathbf{W}_{\boldsymbol{\xi}}|)^{-s/2} \left|\mathbf{Q}_{\boldsymbol{\xi}}\right|^{-\frac{n+\delta+s-1}{2}},
\end{aligned}
\tag{E.4}
$$

where the constants do not depend of $\boldsymbol{\xi}$, and $\mathbf{Q}_{\boldsymbol{\xi}} = \mathbf{Q}+\mathbf{J}-\mathbf{N}^{'}\mathbf{W}_{\boldsymbol{\xi}}^{-1}\mathbf{N}$. The last line of Equation (E.4) is true, since the portion that is inside the integral is the density of $IW(n+\delta,\mathbf{Q}_{\boldsymbol{\xi}})$ without the term $\left|\mathbf{Q}_{\boldsymbol{\xi}}\right|^{(n+\delta+s-1)/2}$.

$\boldsymbol{\xi}$ depends on $\mathbf{Z}$, $\mathbf{k}$ and $\mathbf{y}$ (after integrated out $\mathbf{A}$, $\mathbf{B}_{\boldsymbol{\xi}}$ and $\boldsymbol{\Sigma}$), but once

we condition on $\mathbf{Z}$, the $\mathbf{y}$ is independent of $\boldsymbol{\xi}$. Then, conditional probability of $\boldsymbol{\xi}$ using Equation (E.4) is the following

$$\begin{aligned} & p(\boldsymbol{\xi}|\mathbf{Z},\mathbf{X},\mathbf{y},\mathbf{k}) = p(\boldsymbol{\xi})p(\mathbf{X},\mathbf{y},\mathbf{Z},\mathbf{k}|\boldsymbol{\xi}) \\ =& p(\boldsymbol{\xi})p(\mathbf{X},\mathbf{y},\mathbf{k}|\mathbf{Z},\boldsymbol{\xi})p(\mathbf{Z}|\boldsymbol{\xi}) = p(\boldsymbol{\xi})p(\mathbf{X},\mathbf{y},\mathbf{k}|\mathbf{Z})p(\mathbf{Z}|\boldsymbol{\xi}) \\ \propto& p(\boldsymbol{\xi})p(\mathbf{Z}|\boldsymbol{\xi}) = p(\boldsymbol{\xi})(|\mathbf{H}_{\boldsymbol{\xi}}|\,|\mathbf{W}_{\boldsymbol{\xi}}|)^{-s/2}\left|\mathbf{Q}_{\boldsymbol{\xi}}\right|^{-\frac{n+\delta+s-1}{2}}. \end{aligned} \tag{E.5}$$

The full conditional density of k_ν given the rest is derived only for ordinal responses as in the case where $\boldsymbol{\Sigma}$ is known. From Equation (E.1) the full conditional density is given by Equation (5.21).

Appendix F

Algebra calculations for the probit model with mixture of nominal and ordinal responses - $\boldsymbol{\Sigma}$ known, $\boldsymbol{\Xi}$

This appendix contains the algebraic calculations of the multi-class Bayesian variable selection with mixture of nominal and ordinal responses, under the assumption that the indicator vectors are different across different responses (which means that we work using an indicator matrix instead of indicator vector) and the latent variables are independent. The prior distributions are described in Subsection 8.1.2. Setting $\alpha_{0r} = 0$ and $\mathbf{B}_{0\boldsymbol{\Xi}_{:,r},r} = \mathbf{0}$ for the r-th regression equation, the full conditional distributions are calculated below.

In this case the unknown parameters (in matrix form) are $\mathbf{Z}, \mathbf{A}, \mathbf{B}, \boldsymbol{\Xi}$ and $\mathbf{k}$. We assume that α_r, $\mathbf{B}_{\boldsymbol{\Xi}_{:,r},r}$ are independent, namely $p(\alpha_r, \mathbf{B}_{\boldsymbol{\Xi}_{:,r},r}) = p(\alpha_r)p(\mathbf{B}_{\boldsymbol{\Xi}_{:,r},r})$. The joint posterior is proportional to likelihood times the

prior distribution and it is given by

$$
\begin{aligned}
& p(\mathbf{Z}, \mathbf{A}, \mathbf{B}, \boldsymbol{\Xi}, \mathbf{k} | \mathbf{X}, \mathbf{Y}) \\
= & \prod_{r=1}^{s} p(\mathbf{Y}|\mathbf{Z}_{:,r}, \mathbf{k}) p(\mathbf{k}) \prod_{r=1}^{s} \left[p(\mathbf{Z}_{:,r}|\alpha_r, \mathbf{B}_{\boldsymbol{\Xi}_{:,r},r}, \mathbf{X}) p(\alpha_r) p(\mathbf{B}_{\boldsymbol{\Xi}_{:,r},r}|\boldsymbol{\Xi}_{:,r}) p(\boldsymbol{\Xi}_{:,r}) \right] \\
= & \prod_{r=1}^{s} p(\mathbf{Y}|\mathbf{Z}_{:,r}, \mathbf{k}) p(\mathbf{k}) \\
& \cdot \prod_{r=1}^{s} \left[\frac{1}{(2\pi)^{-n/2} \sigma_r^n} \exp \left\{ -\frac{1}{2\sigma_r^2} \sum_{i=1}^{n} \left(Z_{i,r} - \alpha_r - \mathbf{X}_{i,\boldsymbol{\Xi}_{:,r}} \mathbf{B}_{\boldsymbol{\Xi}_{:,r},r} \right)^2 \right\} \right] \\
& \cdot \prod_{r=1}^{s} \left[(2\pi)^{-p_{\boldsymbol{\Xi}_{:,r}}/2} \left| \mathbf{H}_{\boldsymbol{\Xi}_{:,r}} \right|^{-1/2} \exp \left\{ -\frac{1}{2\sigma_r^2} \left(\mathbf{B}_{\boldsymbol{\Xi}_{:,r},r} \mathbf{H}_{\boldsymbol{\Xi}_{:,r}}^{-1} \mathbf{B}'_{\boldsymbol{\Xi}_{:,r},r} \right) \right\} \right] \qquad \text{(F.1)} \\
& \cdot \prod_{r=1}^{s} \left[(2\pi)^{-1/2} h^{-1/2} \exp \left\{ -\frac{n\alpha_r^2}{2\sigma_r^2 h} \right\} \right] \\
& \cdot \prod_{r=1}^{s} \left[\prod_{j=1}^{p} w_r^{\Xi_{j,r}} (1 - w_r)^{1 - \Xi_{j,r}} \right].
\end{aligned}
$$

Firstly, in order to integrate out $\mathbf{A}$ given $\mathbf{Z}, \mathbf{B}, \boldsymbol{\Xi}$ and $\mathbf{k}$, the exponentiated terms that are associated with α_r in Equation (F.1) can be rewritten as

$$
\begin{aligned}
& - \frac{\sum_{i=1}^{n} \left(Z_{i,r} - \alpha_r - \mathbf{X}_{i,\boldsymbol{\Xi}_{:,r}} \mathbf{B}_{\boldsymbol{\Xi}_{:,r},r} \right)^2}{2\sigma_r^2} - \frac{n\alpha_r^2}{2h\sigma_r^2} \\
= & - \frac{\left[\left(\mathbf{Z}_{:,r} - \mathbf{X}_{\boldsymbol{\Xi}_{:,r}} \mathbf{B}_{\boldsymbol{\Xi}_{:,r},r} \right) - \mathbf{1}_n \alpha_r \right]^2}{2\sigma_r^2} - \frac{\alpha_r^2 h^{-1}}{2\sigma_r^2} \\
= & - \frac{\left[\left(\mathbf{Z}_{:,r} - \mathbf{X}_{\boldsymbol{\Xi}_{:,r}} \mathbf{B}_{\boldsymbol{\Xi}_{:,r},r} \right) - \mathbf{1}_n \alpha_r \right]' \left[\left(\mathbf{Z}_{:,r} - \mathbf{X}_{\boldsymbol{\Xi}_{:,r}} \mathbf{B}_{\boldsymbol{\Xi}_{:,r},r} \right) - \mathbf{1}_n \alpha_r \right] + \alpha_r^2 h^{-1}}{2\sigma_r^2} \\
= & - \frac{(n + h^{-1}) \alpha_r^2 - \mathbf{1}'_n (\mathbf{Z}_{:,r} - \mathbf{X}_{\boldsymbol{\Xi}_{:,r}} \mathbf{B}_{\boldsymbol{\Xi}_{:,r},r}) \alpha_r}{2\sigma_r^2} \\
& - \frac{-\alpha_r (\mathbf{Z}_{:,r} - \mathbf{X}_{\boldsymbol{\Xi}_{:,r}} \mathbf{B}_{\boldsymbol{\Xi}_{:,r},r})' \mathbf{1}_n + (\mathbf{Z}_{:,r} - \mathbf{X}_{\boldsymbol{\Xi}_{:,r}} \mathbf{B}_{\boldsymbol{\Xi}_{:,r},r})' (\mathbf{Z}_{:,r} - \mathbf{X}_{\boldsymbol{\Xi}_{:,r}} \mathbf{B}_{\boldsymbol{\Xi}_{:,r},r})}{2\sigma_r^2} \\
= & - \frac{(n + h^{-1}) \Big[\alpha_r^2 - (n + h^{-1})^{-1} \mathbf{1}'_n (\mathbf{Z}_{:,r} - \mathbf{X}_{\boldsymbol{\Xi}_{:,r}} \mathbf{B}_{\boldsymbol{\Xi}_{:,r},r}) \alpha_r}{2\sigma_r^2} \\
& - \frac{(n + h^{-1})^{-1} \alpha_r (\mathbf{Z}_{:,r} - \mathbf{X}_{\boldsymbol{\Xi}_{:,r}} \mathbf{B}_{\boldsymbol{\Xi}_{:,r},r})' \mathbf{1}_n}{2\sigma_r^2} \\
& - \frac{(n + h^{-1})^{-1} (\mathbf{Z}_{:,r} - \mathbf{X}_{\boldsymbol{\Xi}_{:,r}} \mathbf{B}_{\boldsymbol{\Xi}_{:,r},r})' (\mathbf{Z}_{:,r} - \mathbf{X}_{\boldsymbol{\Xi}_{:,r}} \mathbf{B}_{\boldsymbol{\Xi}_{:,r},r})}{2\sigma_r^2} \\
& - \frac{\pm (n + h^{-1})^{-1} h^{-1} (\mathbf{Z}_{:,r} - \mathbf{X}_{\boldsymbol{\Xi}_{:,r}} \mathbf{B}_{\boldsymbol{\Xi}_{:,r},r})' (\mathbf{Z}_{:,r} - \mathbf{X}_{\boldsymbol{\Xi}_{:,r}} \mathbf{B}_{\boldsymbol{\Xi}_{:,r},r}) \Big]}{2\sigma_r^2} \\
= & - \frac{(n + h^{-1}) v_r v'_r}{2\sigma_r^2} - \frac{(nh + 1)^{-1} (\mathbf{Z}_{:,r} - \mathbf{X}_{\boldsymbol{\Xi}_{:,r}} \mathbf{B}_{\boldsymbol{\Xi}_{:,r},r})' (\mathbf{Z}_{:,r} - \mathbf{X}_{\boldsymbol{\Xi}_{:,r}} \mathbf{B}_{\boldsymbol{\Xi}_{:,r},r})}{2\sigma_r^2}, \qquad \text{(F.2)}
\end{aligned}
$$

where $v_r = \{ \alpha_r - (n + h^{-1})^{-1} \mathbf{1}'_n (\mathbf{Z}_{:,r} - \mathbf{X}_{\boldsymbol{\Xi}_{i,r}} \mathbf{B}_{\boldsymbol{\Xi}_{:,r},r}) \}$ and $(n + h^{-1})^{-1} h^{-1} =$

$(nh+1)^{-1}$. The second term of Equation (F.2) can be written as

$$\frac{(nh+1)^{-1}(\mathbf{Z}_{:,r}-\mathbf{X}_{\boldsymbol{\Xi}_{:,r}}\mathbf{B}_{\boldsymbol{\Xi}_{:,r},r})'(\mathbf{Z}_{:,r}-\mathbf{X}_{\boldsymbol{\Xi}_{:,r}}\mathbf{B}_{\boldsymbol{\Xi}_{:,r},r})}{2\sigma_r^2} = \frac{(\mathbf{Z}_{:,r}-\mathbf{X}_{\boldsymbol{\Xi}_{:,r}}\mathbf{B}_{\boldsymbol{\Xi}_{:,r},r})'(\mathbf{I}_n+h\mathbf{1}_n\mathbf{1}_n')^{-1}(\mathbf{Z}_{:,r}-\mathbf{X}_{\boldsymbol{\Xi}_{:,r}}\mathbf{B}_{\boldsymbol{\Xi}_{:,r},r})}{2\sigma_r^2}.$$

In summary, integrating out $\mathbf{A}$ yields

$$\begin{aligned}
&p(\mathbf{Z},\mathbf{B},\boldsymbol{\Xi},\mathbf{k}|\mathbf{X},\mathbf{Y})\\
=&\int p(\mathbf{Z},\mathbf{A},\mathbf{B},\boldsymbol{\Xi},\mathbf{k}|\mathbf{X},\mathbf{Y})d\mathbf{A}\\
=&\prod_{r=1}^{s}\int p(\mathbf{Z}_{:,r},\alpha_r,\mathbf{B}_{\boldsymbol{\Xi}_{:,r}},\boldsymbol{\Xi}_{:,r},\mathbf{k}|\mathbf{X},\mathbf{Y})d\alpha_r \qquad \text{(F.3)}\\
\propto&\prod_{r=1}^{s}\int \exp\left\{-\frac{(n+h^{-1})v_r^2}{2\sigma_r^2}-\frac{(nh+1)^{-1}(\mathbf{Z}_{:,r}-\mathbf{X}_{\boldsymbol{\Xi}_{:,r}}\mathbf{B}_{\boldsymbol{\Xi}_{:,r},r})'(\mathbf{Z}_{:,r}-\mathbf{X}_{\boldsymbol{\Xi}_{:,r}}\mathbf{B}_{\boldsymbol{\Xi}_{:,r},r})}{2\sigma_r^2}\right\}d\alpha_r\\
\propto&\prod_{r=1}^{s}\exp\left\{-\frac{(\mathbf{Z}_{:,r}-\mathbf{X}_{\boldsymbol{\Xi}_{:,r}}\mathbf{B}_{\boldsymbol{\Xi}_{:,r},r})'(\mathbf{I}_n+h\mathbf{1}_n\mathbf{1}_n')^{-1}(\mathbf{Z}_{:,r}-\mathbf{X}_{\boldsymbol{\Xi}_{:,r}}\mathbf{B}_{\boldsymbol{\Xi}_{:,r},r})}{2\sigma_r^2}\right\}
\end{aligned}$$

Secondly, in order to integrate out $\mathbf{B}$, the exponentiated terms that are associated with $\mathbf{B}_{\boldsymbol{\Xi}_{:,r},r}$ in Equation (F.1) together with the term that is in the squared brackets of Equation (F.3) can be rewritten as

$$\begin{aligned}
&\mathbf{B}_{:,r}'\mathbf{H}_{\boldsymbol{\Xi}_{:,r}}^{-1}\mathbf{B}_{\boldsymbol{\Xi}_{:,r},r}+(\mathbf{Z}_{:,r}-\mathbf{X}_{\boldsymbol{\Xi}_{:,r}}\mathbf{B}_{\boldsymbol{\Xi}_{:,r},r})'(\mathbf{I}_n+h\mathbf{1}_n\mathbf{1}_n')^{-1}(\mathbf{Z}_{:,r}-\mathbf{X}_{\boldsymbol{\Xi}_{:,r}}\mathbf{B}_{\boldsymbol{\Xi}_{:,r},r})\\
=&\mathbf{B}_{:,r}'\mathbf{H}_{\boldsymbol{\Xi}_r}^{-1}\mathbf{B}_{\boldsymbol{\Xi}_{:,r},r}+\mathbf{Z}_{:,r}'(\mathbf{I}_n+h\mathbf{1}_n\mathbf{1}_n')^{-1}\mathbf{Z}_{:,r}-\mathbf{Z}_{:,r}'(\mathbf{I}_n+h\mathbf{1}_n\mathbf{1}_n')^{-1}\mathbf{X}_{\boldsymbol{\Xi}_{:,r}}\mathbf{B}_{\boldsymbol{\Xi}_{:,r},r}\\
&-\mathbf{B}_{:,r}'\mathbf{X}_{\boldsymbol{\Xi}_{:,r}}'(\mathbf{I}_n+h\mathbf{1}_n\mathbf{1}_n')^{-1}\mathbf{Z}_{:,r}+\mathbf{B}_{:,r}'\mathbf{X}_{\boldsymbol{\Xi}_{:,r}}'(\mathbf{I}_n+h\mathbf{1}_n\mathbf{1}_n')^{-1}\mathbf{X}_{\boldsymbol{\Xi}_{:,r}}\mathbf{B}_{\boldsymbol{\Xi}_{:,r},r} \qquad \text{(F.4)}\\
=&\mathbf{B}_{:,r}'\mathbf{W}_{\boldsymbol{\Xi}_{:,r}}\mathbf{B}_{\boldsymbol{\Xi}_{:,r},r}-\mathbf{B}_{:,r}'\mathbf{N}-\mathbf{N}'\mathbf{B}_{r_{\boldsymbol{\Xi}_{:,r}}}+\mathbf{J}\pm\mathbf{N}'\mathbf{W}_{\boldsymbol{\Xi}_{:,r}}\mathbf{N}\\
=&(\mathbf{B}_{\boldsymbol{\Xi}_{:,r},r}-\mathbf{W}_{\boldsymbol{\Xi}_{:,r}}^{-1}\mathbf{N})'\mathbf{W}_{\boldsymbol{\Xi}_{:,r}}(\mathbf{B}_{\boldsymbol{\Xi}_{:,r},r}-\mathbf{W}_{\boldsymbol{\Xi}_{:,r}}^{-1}\mathbf{N})+\mathbf{J}-\mathbf{N}'\mathbf{W}_{\boldsymbol{\Xi}_{:,r}}\mathbf{N},
\end{aligned}$$

where $\mathbf{W}_{\boldsymbol{\Xi}_{:,r}}=\mathbf{X}_{\boldsymbol{\Xi}_{:,r}}'(\mathbf{I}_n+h\mathbf{1}_n\mathbf{1}_n')^{-1}\mathbf{X}_{\boldsymbol{\Xi}_{:,r}}+\mathbf{H}_{\boldsymbol{\Xi}_{:,r}}^{-1}$, $\mathbf{N}=\mathbf{X}_{\boldsymbol{\Xi}_{:,r}}'(\mathbf{I}_n+h\mathbf{1}_n\mathbf{1}_n')^{-1}\mathbf{Z}_{:,r}$ and $\mathbf{J}=\mathbf{Z}_{:,r}'(\mathbf{I}_n+h\mathbf{1}_n\mathbf{1}_n')^{-1}\mathbf{Z}_{:,r}$. In summary, integrating out $\mathbf{B}$ yields

$$
\begin{aligned}
& p(\mathbf{Z}, \mathbf{A}, \boldsymbol{\Xi}, \mathbf{k} | \mathbf{X}, \mathbf{Y}) \\
&= \int p(\mathbf{Z}, \mathbf{A}, \mathbf{B}, \boldsymbol{\Xi}, \mathbf{k} | \mathbf{X}, \mathbf{Y}) d\mathbf{B} \\
&= \prod_{r=1}^{s} \int p(\mathbf{Z}_{:,r}, \alpha_r, \mathbf{B}_{\boldsymbol{\Xi}_{:,r},r}, \boldsymbol{\Xi}_{:,r}, \mathbf{k} | \mathbf{X}, \mathbf{Y}) d\mathbf{B}_{\boldsymbol{\Xi}_{:,r},r} \qquad \text{(F.5)} \\
&= \prod_{r=1}^{s} \int (2\pi)^{p/2} \left| \mathbf{H}_{\boldsymbol{\Xi}_{:,r}} \right|^{-1/2} \\
&\quad \cdot \exp\left\{ -\frac{1}{2\sigma_r^2} \left[(\mathbf{B}_{\boldsymbol{\Xi}_{:,r},r} - \mathbf{W}_{\boldsymbol{\Xi}_{:,r}}^{-1} \mathbf{N})' \mathbf{W}_{\boldsymbol{\Xi}_{:,r}} (\mathbf{B}_{\boldsymbol{\Xi}_{:,r},r} - \mathbf{W}_{\boldsymbol{\Xi}_{:,r}}^{-1} \mathbf{N}) + \mathbf{J} - \mathbf{N}' \mathbf{W}_{\boldsymbol{\Xi}_{:,r}} \mathbf{N} \right] \right\} d\mathbf{B}_{\boldsymbol{\Xi}_{:,r},r} \\
&\propto \prod_{r=1}^{s} \left[\left(\left| \mathbf{W}_{\boldsymbol{\Xi}_{:,r}} \right| \left| \mathbf{H}_{\boldsymbol{\Xi}_{:,r}} \right| \right)^{-1/2} \exp\left\{ -\frac{\mathbf{J} - \mathbf{N}' \mathbf{W}_{\boldsymbol{\Xi}_{:,r}} \mathbf{N}}{2\sigma_r^2} \right\} \right],
\end{aligned}
$$

where $\mathbf{J} - \mathbf{N}' \mathbf{W}_{\boldsymbol{\Xi}_{:,r}} \mathbf{N} = \mathbf{Z}'_{:,r} \mathbf{P}^{*-1}_{\boldsymbol{\Xi}_{:,r}} \mathbf{Z}_{:,r} = \mathbf{Z}'_{:,r} \mathbf{P}^{-1}_{\boldsymbol{\Xi}_{:,r}} \mathbf{Z}_{:,r}$ and $\mathbf{P}^{*-1}_{\boldsymbol{\Xi}_{:,r}} = (\mathbf{I}_n + h\mathbf{1}_n\mathbf{1}'_n)^{-1} - (\mathbf{I}_n + h\mathbf{1}_n\mathbf{1}'_n)^{-1} \mathbf{X}_{\boldsymbol{\Xi}_{:,r}} \cdot [\mathbf{X}'_{\boldsymbol{\Xi}_{:,r}} (\mathbf{I}_n + h\mathbf{1}_n\mathbf{1}'_n)^{-1} \mathbf{I}_n]^{-1} \mathbf{X}_{\boldsymbol{\Xi}_{:,r}} + \mathbf{H}^{-1}_{\boldsymbol{\Xi}_{:,r}}]^{-1} \mathbf{X}'_{\boldsymbol{\Xi}_{:,r}} (\mathbf{I}_n + h\mathbf{1}_n\mathbf{1}'_n)^{-1}$. The last matrix takes the simple form $\mathbf{P}_{\boldsymbol{\Xi}_{:,r}} = \mathbf{I}_n + h\mathbf{1}_n\mathbf{1}'_n + \mathbf{X}_{\boldsymbol{\Xi}_{:,r}} \mathbf{H}_{\boldsymbol{\Xi}_{:,r}} \mathbf{X}'_{\boldsymbol{\Xi}_{:,r}}$, since $\mathbf{P}^{*-1}_{\boldsymbol{\Xi}_{:,r}} \mathbf{P}_{\boldsymbol{\Xi}_{:,r}} = \mathbf{I}_n$.

From Equation (F.5) $p(\mathbf{Z} | \boldsymbol{\Xi}, \mathbf{X}) \sim \prod_{r=1}^{s} MVN(\mathbf{0}, \sigma_r^2 \mathbf{P}_{\boldsymbol{\Xi}_{:,r}})$, which is a multivariate normal distribution. From it we can easily calculate the full conditional distribution of $\mathbf{Z}$ given the rest,

$$
\begin{aligned}
p(\mathbf{Z} | \boldsymbol{\Xi}, \mathbf{k}, \mathbf{X}, \mathbf{Y}) &\sim \prod_{r=1}^{s} \left[MVN(\mathbf{0}, \sigma_r^2 \mathbf{P}_{\boldsymbol{\Xi}_{:,r}}) \right] \prod_{i=1}^{n} \mathbb{1}(\mathbf{Z}_{i,:} \in G_i) \qquad \text{(F.6)} \\
&\propto \prod_{r=1}^{s} \left[\left| \mathbf{P}_{\boldsymbol{\Xi}_{:,r}} \right|^{-1/2} \exp\left\{ -\frac{1}{2\sigma_r^2} \left[\mathbf{Z}'_{:,r} \mathbf{P}^{-1}_{\boldsymbol{\Xi}_{:,r}} \mathbf{Z}_{:,r} \right] \right\} \right] \prod_{i=1}^{n} \mathbb{1}(\mathbf{Z}_{i,:} \in G_i).
\end{aligned}
$$

where G_i is given by Equation (7.6).

Using Equations (F.1) and (F.6) it is true that

$$
\begin{aligned}
& p(\boldsymbol{\Xi} | \mathbf{Z}, \mathbf{k}, \mathbf{X}, \mathbf{Y}) \\
&\propto \prod_{r=1}^{s} \left[p(\boldsymbol{\Xi}_{:,r}) p(\mathbf{Z}_{:,r} | \boldsymbol{\Xi}_{:,r}) \right] \\
&= \prod_{r=1}^{s} \left[p(\boldsymbol{\Xi}_{:,r}) \int p(\mathbf{Z}_{:,r} | \alpha_r, \mathbf{B}_{\boldsymbol{\Xi}_{:,r},r}, \boldsymbol{\Xi}_{:,r}) p(\alpha_r) p(\mathbf{B}_{\boldsymbol{\Xi}_{:,r},r} | \boldsymbol{\Xi}_{:,r}) d\alpha_r d\mathbf{B}_{\boldsymbol{\Xi}_{:,r},r} \right] \\
&\propto \prod_{r=1}^{s} \left[p(\boldsymbol{\Xi}_{:,r}) \left(\left| \mathbf{H}_{\boldsymbol{\Xi}_{:,r}} \right| \left| \mathbf{W}_{\boldsymbol{\Xi}_{:,r}} \right| \right)^{-1/2} \left| \mathbf{P}_{\boldsymbol{\Xi}_{:,r}} \right|^{-1/2} \right],
\end{aligned}
$$

where the proportionality constant does not depend on $\boldsymbol{\Xi}_{:,r}$ and $\left| \mathbf{H}_{\boldsymbol{\Xi}_{:,r}} \right| \left| \mathbf{W}_{\boldsymbol{\Xi}_{:,r}} \right| = \left| \mathbf{H}_{\boldsymbol{\Xi}_{:,r}} \mathbf{W}_{\boldsymbol{\Xi}_{:,r}} \right| = \left| \mathbf{I}_n + \mathbf{X}_{\boldsymbol{\Xi}_{:,r}} \mathbf{H}_{\boldsymbol{\Xi}_{:,r}} \mathbf{X}'_{\boldsymbol{\Xi}_{:,r}} (\mathbf{I}_n + h\mathbf{1}_n\mathbf{1}'_n)^{-1} \right|$.

This fully conditional density of k_d given the rest is calculated easily as

we saw at the end of the Appendix B.

Bibliography

Abramovich, F. and Grinshtein, V. (2010). Map model selection in Gaussian regression. *Electronic Journal of Statistics*, 4, 932–949.

Ai-Jun, Y. and Xin-Yuan, S. (2010). Bayesian variable selection for disease classification using gene expression data. *Bioinformatics*, 26(2), 215–222.

Aijun, Y., Xinyuan, S., and Yunxian, L. (2013). Multi-class classification via Bayesian variable selection with gene expression data. *Statistical Diagnostics for Cancer: Analyzing High-Dimensional Data*, pp. 75–92.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pp. 199–213. Springer.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422), 669–679.

Antoniadis, A., Lambert-Lacroix, S., and Leblanc, F. (2003). Effective dimension reduction methods for tumor classification using gene expression data. *Bioinformatics*, 19(5), 563–570.

Archer, K. and Williams, A. (2012). L1 penalized continuation ratio models for ordinal response prediction using high-dimensional datasets. *Statistics in Medicine*, 31(14), 1464–1474.

Bae, K. and Mallick, B. K. (2004). Gene selection using a two-level hierarchical Bayesian model. *Bioinformatics*, 20(18), 3423–3430.

Baragatti, M. and Pommeret, D. (2012). A study of variable selection using g-prior distribution with ridge parameter. *Computational Statistics & Data Analysis*, 56(6), 1920–1934.

Barber, D. (2012). *Bayesian reasoning and machine learning*. Cambridge University Press.

Barbieri, M. M. and Berger, J. O. (2004). Optimal predictive model selection. *Annals of Statistics*, pp. 870–897.

Berger, J. O. and Molina, G. (2005). Posterior model probabilities via path-based pairwise priors. *Statistica Neerlandica*, 59(1), 3–15.

Bornn, L., Gottardo, R., and Doucet, A. (2010). Grouping priors and the bayesian elastic net. *arXiv preprint arXiv:1001.4083*.

Bottolo, L., Chadeau-Hyam, M., Hastie, D. I., Langley, S. R., Petretto, E., Tiret, L., Tregouet, D., and Richardson, S. (2011). ESS++: a C++ objected-oriented algorithm for Bayesian stochastic search model exploration. *Bioinformatics*, 27(4), 587–588.

Bottolo, L. and Richardson, S. (2010). Evolutionary stochastic search for Bayesian model exploration. *Bayesian Analysis*, 5(3), 583–618.

Brown, P. J., Vannucci, M., and Fearn, T. (1998a). Bayesian wavelength selection in multicomponent analysis. *Journal of Chemometrics*, 12(3), 173–182.

Brown, P. J., Vannucci, M., and Fearn, T. (1998b). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(3), 627–641.

Brown, P. J., Vannucci, M., and Fearn, T. (2002). Bayes model averaging with selection of regressors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3), 519–536.

Cavanaugh, J. E. and Neath, A. A. (1999). Generalizing the derivation of the Schwarz information criterion. *Communications in Statistics-Theory and Methods*, 28(1), 49–66.

Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, 85(2), 347–361.

Chipman, H., George, E. I., McCulloch, R. E., Clyde, M., Foster, D. P., and Stine, R. A. (2001). The practical implementation of Bayesian model selection. *Lecture Notes-Monograph Series*, 38, 65–134.

Chopin, N. (2011). Fast simulation of truncated Gaussian distributions. *Statistics and Computing*, 21(2), 275–288.

Chu, W. and Keerthi, S. S. (2007). Support vector ordinal regression. *Neural Computation*, 19(3), 792–815.

Conteduca, V., Sansonno, D., Ingravallo, G., Marangi, S., Russi, S., Lauletta, G., and Dammacco, F. (2012). Barrett's esophagus and esophageal cancer: an overview. *International Journal of Oncology*, 41(2), 414–424.

Cowles, M. K. (1996). Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models. *Statistics and Computing*, 6(2), 101–111.

Dawid, A. P. (1981). Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika*, 68(1), 265–274.

Dayhoff, M. O. and Schwartz, R. M. (1978). A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*. Citeseer.

de Jonge, P. J. F., van Blankenstein, M., Grady, W. M., and Kuipers, E. J. (2013). Barrett's oesophagus: epidemiology, cancer risk and implications for management. *Gut*, 61(1), 191–202.

Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2000). Bayesian variable selection using the Gibbs sampler. *Biostatistics Basel*, 5, 273–286.

Dettling, M. (2004). BagBoosting for tumor classification with gene expression data. *Bioinformatics*, 20(18), 3583–3593.

Dobra, A. (2009). Variable selection and dependency networks for genomewide data. *Biostatistics*, 10(4), 621–639.

Fearn, T. (2012). An introduction to Bayesian statistics, part 4: An analysis of the two-group classification problem. *NIR News*, 23(6), 20–21.

Fearn, T., Brown, P. J., and Besbeas, P. (2002). A Bayesian decision theory approach to variable selection for discrimination. *Statistics and Computing*, 12(3), 253–260.

Finney, D. (1947). Probit analysis, 1952.

Fitzgerald, R. C., di Pietro, M., Ragunath, K., Ang, Y., Kang, J.-Y., Watson, P., Trudgill, N., Patel, P., Kaye, P. V., Sanders, S., et al. (2014). British Society of Gastroenterology guidelines on the diagnosis and management of Barrett's oesophagus. *Gut*, 63(1), 7–42.

Foreman, L. (2016). *Infrared spectroscopy as a clinical diagnostic method for detection of (pre-) cancerous areas of the oesophagus*. PhD thesis, University College London.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 1.

Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3), 397–416.

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., and Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10), 906–914.

Galimberti, G., Soffritti, G., and Di Maso, M. (2012). Classification trees for ordinal responses in R: the rpartScore package. *Journal of Statistical Software*, 47(1), 1–25.

Ganganwar, V. (2012). An overview of classification algorithms for imbalanced datasets. *International Journal of Emerging Technology and Advanced Engineering*, 2(4), 42–47.

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4), 1360–1383.

George, E. and Foster, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika*, 87(4), 731–747.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423), 881–889.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, 7(2), 339–373.

Geweke, J. (1991). 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*, pp. 571–578. Citeseer.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439), 531–537.

Gupta, M. and Ibrahim, J. G. (2007). Variable selection in regression mixture modeling for the discovery of gene regulatory networks. *Journal of the American Statistical Association*, 102(479), 867–880.

Gutiérrez, P. A., Perez-Ortiz, M., Sanchez-Monedero, J., Fernández-Navarro, F., and Hervas-Martinez, C. (2016). Ordinal regression methods: survey and experimental study. *IEEE Transactions on Knowledge and Data Engineering*, 28(1), 127–146.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *SIGKDD Explorations Newsletter*, 11(1), 10–18.

Hand, D. J. and Till, R. J. (2001). A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine Learning*, 45(2), 171–186.

Hastie, T., Tibshirani, R., and Friedman, J. J. H. (2001). *The elements of statistical learning*. Springer New York.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55–67.

Holmes, C. C. and Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, 1(1), 145–168.

Hothorn, T., Hornik, K., Strobl, C., Zeileis, A., and Hothorn, M. T. (2015). Package 'party'. *Package Reference Manual for Party Version 0.9-998*, 16, 37.

Hurvich, C. and Tsai, C. (1995). Model selection for extended quasi-likelihood models in small samples. *Biometrics*, 51(3), 1077.

Imai, K. and van Dyk, D. A. (2005). A Bayesian analysis of the multinomial probit model using marginal data augmentation. *Journal of econometrics*, 124(2), 311–334.

Imori, S., Yanagihara, H., and Wakaki, H. (2011). General formula of bias-corrected AIC in generalized linear models.

Janitza, S., Tutz, G., and Boulesteix, A.-L. (2014). Random forests for ordinal response data: prediction and variable selection. Technical Report 174, Ludwig-Maximilians-University Munich, Department of Statistics.

Japkowicz, N. and Shah, M. (2011). *Evaluating learning algorithms: a classification perspective*. Cambridge University Press.

Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Proceedings of the Cambridge Philosophical Society*, 31(02), 203–222.

Jiao, X. and van Dyk, D. A. (2015). A corrected and more efficient suite of MCMC samplers for the multinomal probit model. *arXiv preprint arXiv:1504.07823*.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430), 773–795.

Kerkhof, M., Van Dekken, H., Steyerberg, E., Meijer, G., Mulder, A., De Bruine, A., Driessen, A., Ten Kate, F., Kusters, J., Kuipers, E., et al. (2007). Grading of dysplasia in Barrett's oesophagus: substantial interobserver variation between general and gastrointestinal pathologists. *Histopathology*, 50(7), 920–927.

Kohn, R., Smith, M., and Chan, D. (2001). Nonparametric regression using linear combinations of basis functions. *Statistics and Computing*, 11(4), 313–322.

Kotecha, J. H. and Djuric, P. M. (1999). Gibbs sampling approach for generation of truncated multivariate Gaussian random variables. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, 1999*, volume 3, pp. 1757–1760. IEEE.

Kotti, E., Manolopoulou, I., and Fearn, T. (2015). Bayesian variable selection in the probit model with mixture of nominal and ordinal responses. In *Workshop Autonomous Citizens: Algorithms for Tomorrow's Society*.

Kotti, E., Manolopoulou, I., and Fearn, T. (2016a). Bayesian variable selection for a mixture of nominal and ordinal responses. In *13th International Society for Bayesian Analysis World Meeting (ISBA)*.

Kotti, E., Manolopoulou, I., and Fearn, T. (2016b). Bayesian variable selection for mixture of nominal and ordinal responses via an indicator matrix. In *2nd UCL Conference on the Theory of Big Data*. http://www.ucl.ac.uk/bigdata-theory/wp-content/uploads/2016/01/Kotti.pdf.

Kotti, E., Manolopoulou, I., and Fearn, T. (2016c). Hierarchical Bayesian variable selection in the probit model with mixture of nominal and ordinal

responses. In *2016 IEEE Workshop on Statistical Signal Processing (SSP)*, pp. 576–580. IEEE.

Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, 60, 65–81.

Kwon, D., Tadesse, M. G., Sha, N., Pfeiffer, R. M., and Vannucci, M. (2007). Identifying biomarkers from mass spectrometry data with ordinal outcome. *Cancer Informatics*, 3, 19–28.

Lamnisos, D., Griffin, J. E., and Steel, M. F. (2009). Transdimensional sampling algorithms for Bayesian variable selection in classification problems with many more variables than observations. *Journal of Computational and Graphical Statistics*, 18(3), 592–612.

Lamnisos, D., Griffin, J. E., and Steel, M. F. (2012). Cross-validation prior choice in Bayesian probit regression with many covariates. *Statistics and Computing*, 22(2), 359–373.

Lamnisos, D., Griffin, J. E., and Steel, M. F. (2013). Adaptive Monte Carlo for Bayesian variable selection in regression models. *Journal of Computational and Graphical Statistics*, 22(3), 729–748.

Lao-Sirieix, P., Roy, A., Worrall, C., Vowler, S. L., Gardiner, S., and Fitzgerald, R. C. (2006). Effect of acid suppression on molecular predictors for esophageal cancer. *Cancer Epidemiology Biomarkers & Prevention*, 15(2), 288–293.

Lee, K. E., Sha, N., Dougherty, E. R., Vannucci, M., and Mallick, B. K. (2003). Gene selection: a Bayesian variable selection approach. *Bioinformatics*, 19(1), 90–97.

Linardakis, M. and Dellaportas, P. (2003). Assessment of athens's metro passenger behaviour via a multiranked probit model. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52(2), 185–200.

Madigan, D., York, J., and Allard, D. (1995). Bayesian graphical models for discrete data. *International Statistical Review*, 63(2), 215–232.

McCullagh, P. (1984). Generalized linear models. *European Journal of Operational Research*, 16(3), 285–292.

McCulloch, R. and Rossi, P. E. (1994). An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics*, 64(1), 207–240.

McManus, D. T., Olaru, A., and Meltzer, S. J. (2004). Biomarkers of esophageal adenocarcinoma and Barrett's esophagus. *Cancer Research*, 64(5), 1561–1569.

Miller, A. (2002). *Subset selection in regression*. CRC Press.

Minkova, L. D. and Omey, E. (2014). A new Markov binomial distribution. *Communications in Statistics-Theory and Methods*, 43(13), 2674–2688.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404), 1023–1032.

Montgomery, E. (2005). Is there a way for pathologists to decrease interobserver variability in the diagnosis of dysplasia? *Archives of Pathology & Laboratory Medicine*, 129(2), 174–176.

Muenz, L. R. and Rubinstein, L. V. (1985). Markov models for covariate dependence of binary sequences. *Biometrics*, 41(1), 91–101.

Muller, P., Parmigiani, G., and Rice, K. (2006). FDR and bayesian multiple comparisons rules.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT Press.

Nandram, B. and Chen, M.-H. (1996). Reparameterizing the generalized linear model to accelerate gibbs sampler convergence. *Journal of Statistical Computation and Simulation*, 54(1-3), 129–144.

National Institute for Health and Clinical Excellence (2014). Epithelial radiofrequency ablation for Barrett's oesophagus. In *Interventional Procedure Guidance [IPG344]*.

Old, O., Moayyedi, P., Love, S., Roberts, C., Hapeshi, J., Foy, C., Stokes, C., Briggs, A., Jankowski, J., Barr, H., et al. (2015). Barrett's oesophagus surveillance versus endoscopy at need study (BOSS): protocol and analysis plan for a multicentre randomized controlled trial. *Journal of Medical Screening*, 22(3), 158–164.

Park, M. Y. and Hastie, T. (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4), 659–677.

Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, 103(482), 681–686.

Pasquini, C. (2003). Near infrared spectroscopy: fundamentals, practical aspects and analytical applications. *Journal of the Brazilian Chemical Society*, 14(2), 198–219.

Pazzani, M., Merz, C., Murphy, P., Ali, K., Hume, T., and Brunk, C. (1994). Reducing misclassification costs. In *Proceedings of the Eleventh International Conference on Machine Learning*, pp. 217–225. Morgan Kauffmann.

Peng, B., Zhu, D., Ander, B. P., Zhang, X., Xue, F., Sharp, F. R., and Yang, X. (2013). An integrative framework for Bayesian variable selection with informative priors for identifying genes and pathways. *PloS one*, 8(7), e67672.

Peruggia, M. (2003). Model selection and multimodel inference: A practical information-theoretic approach. *Journal of the American Statistical Association*, 98(463), 778–779.

Piccarreta, R. (2008). Classification trees for ordinal variables. *Computational Statistics*, 23(3), 407–427.

Plackett, R. L. (1950). Some theorems in least squares. *Biometrika*, 37(1/2), 149–157.

Quaroni, L. and Casson, A. G. (2009). Characterization of barrett esophagus and esophageal adenocarcinoma by fourier-transform infrared microscopy. *Analyst*, 134(6), 1240–1246.

Rabinovitch, P. S., Longton, G., Blount, P. L., Levine, D. S., and Reid, B. J. (2001). Predictors of progression in Barrett's esophagus III: baseline flow cytometric variables. *The American Journal of Gastroenterology*, 96(11), 3071–3083.

Richardson, S., Bottolo, L., and Rosenthal, J. S. (2010). Bayesian models for sparse regression analysis of high dimensional data. *Bayesian Statistics*, 9, 539–569.

Robert, C. P. (1995). Simulation of truncated normal variables. *Statistics and Computing*, 5(2), 121–125.

Russu, A., Malovini, A., Puca, A. A., and Bellazzi, R. (2012). Stochastic model search with binary outcomes for genome-wide association studies. *Journal of the American Medical Informatics Association*, 19(e1), e13–e20.

Saadi, H., Liquet, B., Chadeau-Hyam, M., Bottolo, L., and Richardson, S. (2016). R2GUESS: a GPU-based R package for a Bayesian variable selection model accommodating multivariate responses. *Journal of Statistical Software*, 69(2), 1–32.

Savitzky, A. and Golay, M. J. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8), 1627–1639.

Schisterman, E. F., Moysich, K. B., England, L. J., and Rao, M. (2003). Estimation of the correlation coefficient using the bayesian approach and its applications for epidemiologic research. *BMC medical research methodology*, 3(1), 1.

Scholkopft, B. and Mullert, K.-R. (1999). Fisher discriminant analysis with kernels. *Neural Networks for Signal Processing IX*, 1, 41–48.

Seber, G. (2000). F 1984. Multivariate observations.

Sha, N., Vannucci, M., Tadesse, M. G., Brown, P. J., Dragoni, I., Davies, N., Roberts, T. C., Contestabile, A., Salmon, M., Buckley, C., et al. (2004). Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage. *Biometrics*, 60(3), 812–819.

Shaheen, N. J., Falk, G. W., Iyer, P. G., and Gerson, L. B. (2016). ACG clinical guideline: diagnosis and management of Barretts esophagus. *The American Journal of Gastroenterology*, 111(1), 30–50.

Shimamura, K., Ueki, M., Kawano, S., and Konishi, S. (2016). Bayesian generalized fused lasso modeling via neg distribution. *arXiv preprint arXiv:1602.04910*.

Strimenopoulou, F. and Brown, P. J. (2008). Empirical Bayes logistic regression. *Statistical Applications in Genetics and Molecular Biology*, 7(2).

Sun, B.-Y., Li, J., Wu, D. D., Zhang, X.-M., and Li, W.-B. (2010). Kernel discriminant learning for ordinal regression. *IEEE Transactions on Knowledge and Data Engineering*, 22(6), 906–910.

Tachmazidou, I., Johnson, M. R., and De Iorio, M. (2010). Bayesian variable selection for survival regression in genetics. *Genetic Epidemiology*, 34(7), 689–701.

Talhouk, A., Doucet, A., and Murphy, K. (2012). Efficient Bayesian inference for multivariate probit models with sparse inverse correlation matrices. *Journal of Computational and Graphical Statistics*, 21(3), 739–757.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1), 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1), 91–108.

Trevisan, J., Angelov, P. P., Carmichael, P. L., Scott, A. D., and Martin, F. L. (2012). Extracting biological information with computational analysis of Fourier-transform infrared (FTIR) biospectroscopy datasets: current practices to future perspectives. *Analyst*, 137(14), 3202–3215.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44(1), 92–107.

Witten, D. and Tibshirani, R. (2011). Penalized classification using Fisher's linear discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5), 753–772.

Witten, D. and Witten, M. D. (2015). Package penalizedlda.

Woodbury, M. A. (1950). Inverting modified matrices. *Memorandum Report, Statistical Research Group, Princeton University.*

Yang, Y., Wainwright, M. J., and Jordan, M. I. (2015). On the computational complexity of high-dimensional bayesian variable selection. *arXiv preprint arXiv:1505.07925.*

Zagari, R. M., Fuccio, L., Wallander, M.-A., Johansson, S., Fiocca, R., Casanova, S., Farahmand, B. Y., Winchester, C. C., Roda, E., and Bazzoli, F. (2008). Gastro-oesophageal reflux symptoms, oesophagitis and Barrett's oesophagus in the general population: the loiano–monghidoro study. *Gut*, 57(10), 1354–1359.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno De Finetti*, 6, 233–243.

Zellner, A. and Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. *Trabajos de Estadística y de Investigación Operativa*, 31(1), 585–603.

Zhang, Y. (2013). Epidemiology of esophageal cancer. *World Journal of Gastroenterology*, 19(34), 5598–5606.

Zhou, X., Wang, X., and Dougherty, E. (2006). Multi-class cancer classification using multinomial probit regression with Bayesian gene selection. *IEEE Proceedings Systems Biology*, 153(2), 70–78.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476), 1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320.

Zou, H. and Zhang, H. H. (2009). On the adaptive elastic net with a diverging number of parameters. *Annals of Statistics*, 37(4), 1733–1751.

Zucknick, M. and Richardson, S. (2008). MCMC methods for gene expression profiling via Bayesian variable selection. *Lifestat 2008*.