7-21-2015

Statistical Shrinkage Methods for Classification, Prediction, and Feature Extraction Using Genomewide Gene Expression Data and Small Sample Sizes

Behrouz Madahian

STATISTICAL SHRINKAGE METHODS FOR CLASSIFICATION, PREDICTION, AND FEATURE EXTRACTION USING GENOMEWIDE GENE EXPRESSION DATA AND SMALL SAMPLE SIZES

by

Behrouz Madahian

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

Major: Mathematical Sciences

The University of Memphis

August 2015

Acknowledgements

I Would like to take this opportunity to thank those who made this thesis possible. First of all, I want to thank you my advisor Dr. Lih Yuan Deng for his excellent guidance and persistent patience toward me during my PhD at University of Memphis. The confidence he put in me has been extremely important to complete my studies. He has always been readily available for my many questions which have guided my studies. I am also extremely grateful to my supervisor Dr. Ramin Homayouni. I had the honor of working with Dr. Ramin Homayouni for the last 5 years from the time I started my Masters in Bioinformatics. He has always given me freedom to explore new ideas, learn new techniques, and taught me how to think critically and work across multi-disciplinary frameworks which helped me mature quickly. I am extremely fortunate to have him as my mentor. His highly dynamic research team has helped me deepen my knowledge and experience in several different areas of research and expertise. His continual support and encouragement helped me through my studies. I would also like to thank Dr. Dale Bowman, Dr. Hongmei Zhang and Dr. Su Chen for serving as my committee members and for their valuable comments and suggestions. I would like to thank all instructors in the division of Statistics for offering very valuable courses making learning enjoyable. I would also like to thank my family for supporting me in my studies which without their help this work was not possible.

Abstract

Madahian, Behrouz. PhD. The University of Memphis. August, 2015. Statistical Shrinkage Methods for Classification, Prediction, and Feature Extraction Using Genomewide Gene Expression Data and Small Sample Sizes. Major Professor: Dr. Lih Yuan Deng.

With advent of new technologies, more data is being collected than ever before. Data is pouring in from every conceivable direction: from operational and transactional systems, from Micro array experiments and Genome Wide Association Studies, from inbound and outbound customer contact points, from mobile media and the Web to mention a few. Researchers and investigators in many fields are faced with the problem of identifying important effects among thousands of variables in high dimensional datasets. This process often results in non or weekly identified effects. Nowadays a common problem when processing data sets with large number of variables compared to small sample sizes is to estimate the parameters associated with each variable. When the number of variables far exceeds the number of samples, the parameter estimation becomes very difficult. The attempt to find important variables deriving different phenomena based on single variable analysis is more likely to not give a comprehensive picture due to complexity of the phenomena and presence of several predictors with potentially significant effects. Thus, methods based on single variable analysis are too simple to give a comprehensive picture of phenotype architecture. Therefore, more statistically challenging models which are able to accommodate simultaneous analysis of a large number of variables despite small

sample sizes are essential in these cohorts. In this thesis, we developed several novel methods for sample classification, prediction and feature extraction in cohorts with large number of variables compared to small sample sizes using Bayesian shrinkage methods as well as non-parametric methods such as Support Vector Machines and Random Forests. We utilized Generalized Double Pareto and Double Exponential prior distributions on parameters in Bayesian Generalized Linear Models setting. These distributions have a spike at zero shrinking the parameters towards zero which imposes sparsity in the model. We utilized Markov Chain Monte Carlo (MCMC) method based on Gibbs sampling algorithm to estimate the parameters. The models were applied to Microarray data sets such as prostate cancer, leukemia, and breast cancer cohorts. In order to obtain more robust results 50 resampling on train and test data was performed and average performance of the models in 50 runs were reported. We investigated the classification accuracy, feature extraction ability, and prediction ability of the models. Based on our findings, the Bayesian hierarchical models developed obtain high classification accuracy as well as result in more cohesive variable sets compared to other common methods used for the same purpose. We show that using few predictors obtained from our models, we achieve higher performance compared to other competitive methods. We also investigated the use of literature to aid the selection of initial predictors used in the model. Our finding suggests that even though in some instances use of literature will result in better prediction and classification, this is not unanimously true and in some cases it results in poorer performance. This is mainly due to the fact that literature based predictor

sets can be weak signals in the data set at hand as well as our information about the variables deriving different phenomena based on literature is not fully complete. Ideally, we would like to use literature to tune and prioritize signals directly coming from the experiment. To this end, we developed a literature aided sparse Bayesian Generalized linear model that uses literature information a priori to guide the choice of hyper parameters and amount of shrinkage imposed in the model. The developed model not only achieves high classification accuracy, sensitivity, and specificity but also, results is substantially more relevant genesets which turns out to explain the underlying mechanisms of phetotypes better.

TABLE OF CONTENTS

List of Tables

List of Figures

Chapter 1

Introduction

The hereditary material in humans and almost all other organisms is stored in DNA or deoxyribonucleic acid [2]. Four basic molecular units called nucleotides form linear double-stranded polymer called DNA. These four base pairs are adenine, thymine, cytosine, and guanine. For simplicity they are called A, T, C, and G bases respectively. DNA bases pair up with each other with strict base pairing rule: A pairs with T with 2 hydrogen bonds and C pairs with G with 3 hydrogen bonds. Genetic information is stored in sequence of nucleotides. Genes are specific sequences of DNA that provide instructions for several activities in the cells. The coding parts of DNA sequence determine what the purpose of the gene is and the non-coding sequence determines when the gene is expressed. When a gene becomes active, in a process called transcription an RNA copy of the gene's information is created. In RNA the 3 base pairs A,C,G are the same as the ones in DNA and instead of T it has Uracil (U). In RNA, A pairs with U and C pairs with G as in DNA. RNA molecule has 3 different types:

- Messenger RNA (mRNA): it contains genetic information needed to make proteins.
- Transfer RNA (tRNA): performs a role in protein production in the cell by transferring protein building block (amino acids) to the protein synthesis machinery.
- Ribosomal RNA (rRNA): it is the RNA component of ribosome (protein synthetic machinery).

The process of conversion from mRNA to protein is called translation. The abundance of corresponding RNA for each gene determines the levels of gene expression which is generally an indicator of the amount of protein produced [22]. Since not all the genes are active at the same time, the study of expressed genes under different conditions such as different cancers or treatment has proved to be very effective in casting light on gene disease associations.

DNA Microarray

DNA microarray technology enables scientist to examine several thousands of genes at the same time. A microarray is made up of thousands of precisely placed nucleotides called probes on a small piece of glass. Each probe contains different DNA oligonucleotide sequence that is complementary to the mRNA of interest. The DNA oligonucleotide is immobilized on the microarray surface using photolithography or spotting techniques. Generally, mRNA is reverse transcribed to generate a more stable molecule called complementary DNA or cDNA. After this process is done, the cDNA molecules are labeled with fluorescents dyes. The cDNA molecules bind to the probes that are complementary to their sequence by hydrogen bonds. Then the array is washed and scanned by confocal scanners. The intensity of the lights emitted is used to determine the amounts of mRNA which is the surrogate to the gene expression values and the amount of proteins produced.

Proteins (such as enzymes, hormone receptors to mention a few) are functional units of cells. Some examples of cellular activities performed by proteins includes but are not limited to cell differentiation, response to

environmental stimuli, cell division (mitosis), and cell death (apoptosis). Since proteins are functional machinery of the cells and the amount and type of the proteins produced in the cells are determined by the genotype of the cell, expression of genes determine the phenotypes of cells and organisms. That means that different organisms and tissues can perform their specific functions through expression of different genes.

Using DNA Microarray to Help Diagnostics and Therapeutics

DNA Microarray provides facilities for researchers to learn more about different types of diseases such as study of cancer. In the past, scientists have classified different types of cancers based on the morphology of the organs in which the tumor develops. Microarray technology provides invaluable means for studying diseases based on patterns of gene expression in tumor cells which has opened new channels for diagnostics and innovative therapeutics. By using microarrays, design of targeted treatment strategies towards specific types of cancers has become possible. Furthermore, by examining the gene activity differences in normal and tumor cells, treated and untreated tumor cells, scientists will be able to understand exactly how different therapies affect tumors which potentially can lead to more effective treatments.

Gene Expression Analysis

One of the most important applications of DNA Microarrays is based on gene expression analysis. Estimation of the level of expression of several thousands of genes for the sample of cells have been made possible by the use of DNA microarrays. In gene expression analysis, molecular signature of the tissue

is obtained by allowing the RNA obtained from the tissue hybridize on the DNA Microarray. This information may help to perform better disease classification, guide choice of therapy, and identify new therapeutic targets.

Sample Classification Based on Gene Expression Analysis

The classification of different tumor types is of major importance in cancer diagnostics and new therapeutic discoveries [2, 3, 71]. A disease like cancer is fundamentally a malfunction of genes [3]. It is known that cancer classification based on gene expression data provides the key information for addressing fundamental problems pertaining to diagnostics of cancer and discovery of new drugs. Diagnostics and discrimination of sample types based on gene expression data has the potential to provide reliable and accurate cancer classification. Many studies have shown the superior diagnostic performance of cancer classification based on gene expression data compared to traditional methods based on morphology and clinical appearance [79, 80, 100, 105].

A variety of techniques have been developed which utilize gene expression data for cancer classification. Some of these methods include but are not limited to Bayesian analysis [16, 59], support vector machine (SVM) [36, 84, 92], self-organizing maps [63], k-nearest neighbor (KNN) [77, 106], and ensemble methods [44, 83]. Most of these methods are based on selecting a subset of these genes as biomarkers and then performing cancer classification based on these genes. Principal component analysis (PCA) has been used for the analysis of gene expression data [78]. PCA enables researchers to reduce the dimension and thus complexity of the data and explain the variation in the data based on first

few principal components. PCA is especially useful for visualization and clustering of the samples based on their gene expression data [78]. In many of the approaches, the variables are assumed fixed, but in many cases where the predictor variables are random, such as gene expression data, assumptions can be made that result in the same formulation as in fixed case [74]. One such assumptions is a joint multivariate normal distribution for response and predictors, other is an analysis of response conditioned upon the random predictors. For the remaining discussion we will assume an appropriate assumption has been made.

Binary and Multi-category Classification Problems

DNA microarray technology shifted the scale of genomics research by providing capabilities to study several thousands of genes at the same time in a single experiment. DNA microarray measures the relative amount of mRNA. Transcriptional changes reflect the status of disease including cancers and thus gene expression profiles can be used in classification of different types of cancer [69]. Binary classification problems deal with situations where phenotypes have two possible categories. For instance, in cancer studies, gene expression profiles can be used for classifying the samples into normal and tumor tissues. When the phenotype under study has more than two categories, the multinomial classification problem exists. Some classification algorithms naturally permit the use of more than two classes while others are binary algorithms.

In several applications, the multi-class classification is reduced to several binary classification problems. One strategy is training a single classifier per class by considering samples of that class as positive and other samples as negative

samples [10]. Another approach creates $\binom{k}{2}$ binary classifiers for a k-way multi class problem. For each binary classifier, two sample types are used to train the model to predict each of the two. At testing, all the classifiers are applied to each sample and the class that has the most number of asignments is predicted by the combined classifier [10]. Several methods have been developed that permit the use of more than two categories of outcome such as logistic regression, and Random Forests. In most of these methods, the probability of belonging to each category of outcomes is predicted for each sample.

Cancer Classification Challenges and Shortcoming of Current Methods

Even though the DNA microarrays have made simultaneous monitoring of thousands of gene expressions possible, sample sizes remain small, most of them have less than 100 samples. On the other hand, the number of genes-attribute space- is enormous. Each observation has thousands of genes associated with it. Assume we mapped the samples in the attribute space, then the samples will be very sparse in the high dimensional space. Most classification algorithms are not powerful enough to deal with datasets with this kind of characteristics. Thus, applying standard classification methods to such data will result in several problems. High dimensionality and small sample size may give rise to overfitting. Additionally, having so many genes results in expensive computation time. Therefore developing an effective classification algorithm based on gene expression data is not an easy task [103].

Another challenge arises from the presence of noise in the gene expression data. The noise can be categorized into technical and biological noise [8]. The

noises introduced at various stages of data preparation is called technical noise. The noise introduced by genes that are not relevant to the cancer classes is called biological noise-most of the genes are not related to the cancer under study. The presence of noise coupled with small sample size makes accurate classification of tumor types very difficult [8]. The majority of genes in gene expression data analysis are not related to the phenotype under study, dealing with these huge number of irrelevant genes, which comprise a disproportionate number of attributes in gene expression dataset, provides another challenge. In most gene expression studies, the number of relevant genes comprise a small portion of the total number of genes. Additionally, the presence of irrelevant genes reduces the discriminating power of those relevant genes. Extracting these genes from the pool of several thousands of genes is a big challenge.

The fourth challenge arises from the fact that classification accuracy is not the only goal in cancer classification. Biological relevancy is another appealing criterion to most biologists. Biological information revealed during the process can help in further gene function discovery [103]. Therefore, classifiers that not only produce high classification accuracy but provide insight into biology are desirable. In order to highlight those variables that are most relevant to certain phenomena, it is necessary to develop an approach to weed out unimportant variables.

To tackle this problem, several approaches based on the idea of single variable analysis at a time have been proposed including: the t-test [21], a regression modeling approach [87], mixture model approach [66] and non-parametric methods [91]. The shortcoming of all these methods is that they are all univariate

variable selection methods. However, most complex phenomena are polygenic; a single variable analysis can only detect a very small portion of variation and, also, may not be powerful for identifying weaker associations [7]. In addition, it is very common for different variables to interact with each other to form a complex network of interactions, which cannot be characterized from individual analyses.

Thus, the need for new methods which are able to analyze large number of variables becomes more obvious. Set based approaches to finding significant variable have the following advantages to single variable analyses. First, by inferring associations over sets of related variables, they can potentially decrease uncertainty around variables and false positive. Second, the insights into the functional links provided, facilitates interpretation of results. The last but not least, they can potentially uncover a significant pattern distributed over multiple variables while the changes in individual variables have a small effect providing a much better framework to investigate architecture of complex diseases. In order to address limitations that come with single variable analysis methods, lots of research has focused on the development of various approaches for simultaneous analysis of multiple variables [53, 94, 98].

In linear regression framework, least square method is used to obtain estimate of parameters. The ordinary least square estimates obtained are not quite satisfactory mainly due to poor accuracy of prediction resulting from high variances of estimates and poor performance when the dataset at hand contains large number of variables with small sample size [88]. Often, one would like to establish a smaller subset which offers the strongest effect and discriminating

power. It is believed that prediction accuracy can be improved by setting the parameters associated with unimportant variables to zero and thus obtaining more accurate prediction for significant variables [88]. Traditionally, by using forward selection, backward elimination, and stepwise selection a subset of predictors in a regression framework is obtained. However, these approaches are computationally expensive and unstable even when the number of predictors is not large [7]. Furthermore, in this setting there are thousands of variables compared to small sample size at hand which can result in over fitting and can fail to identify important predictors. Thus, the data structure makes it impossible to use traditional multivariate regression for analysis [48]. Researchers have used logistic regression extensively when the response variable is binary and multi-category. But for the data structure explained above, procedures incorporated into the software packages to obtain maximum likelihood estimates of parameters will become computationally intensive and sometimes intractable. In addition, the maximization process may not converge to the maximum likelihood estimates and predictors may have large estimated variances resulting in poor prediction accuracy [70].

There has been a great effort to develop methods that are able to analyze lots of variables simultaneously by inducing sparseness in the model while highlighting the relevant variables. Least Absolute Shrinkage and Selection Operator (LASSO) work by Tibshirani in 1996 drew much attention to the area [88]. There exists a rich literature discussing methods to analyze the LASSO and related approaches [45, 104, 107, 108]. After the work of [89] and [27],

Bayesian approach to the same problem gained interest. A Bayesian LASSO was proposed by park and Casella (2008) and Hans (2009), [37, 67]. However, these procedures may cause over-shrinking of large coefficients due to the relatively light tails of the double exponential prior thus introducing major bias. Using a normal-Jeffreys prior which has heavier tails than the double exponential distribution, small coefficients may shrink to zero while minimally shrinking large coefficients and thus obtaining better results. However use of this prior has no meaning from an inferential aspect as it results in an improper posterior [5]. An alternative class of hierarchical priors were proposed that uses Bayesian adaptive Lasso with non-convex penalization [90]. However, it lacks simple analytic form.

Armagan et. al (2011) proposed the Generalized Double Pareto (GDP) prior distribution [5] . The properties of this distribution that makes it appealing include: having a spike at zero alongside student-t like tails, a simple analytic form and yielding a proper posterior. In addition, it resembles the double exponential density in the neighborhood of zero and has heavier tails compared to double exponential, remedying unwanted bias resulting from over shrinkage of parameters toward zero [5].

Classification Accuracy and Biological Relevance

Another challenging problem in analyzing gene expression data is the fact that identification of a set of biologically relevant markers with high predictive power remains difficult. Several machine learning algorithms have been used for cancer classification with promising results. However, majority of machine learning algorithms are geared toward obtaining the highest classification

accuracy and do not take into account the biological relevance of the markers obtained. Thus, in the majority of applications markers found do not convey meaningful biological information and are merely good classifiers. Thus, a machine learning schema that is able to bridge classification accuracy and biological relevance will be of high merit to the community and can potentially result in deeper understanding of mechanisms involved.

GCAT is a web-based tool that determines the functional coherence of gene sets by performing latent semantic analysis of Medline abstracts [96]. In GCAT, each gene –document was generated by concatenation of all titles and abstracts of the Medline. After gene-document was collected, latent Semantic Analysis (LSA) is used to calculate the gene-gene similarity matrix. LSA is a variant of the vector space model that reduces the dimensions of the matrix by applying Singular Value Decomposition (SVD) so that genes can be compared more conceptually [39]. Thus, LSA allows extraction of both explicit and implicit gene relationships from the literature. In a vector space Model, the semantic structure of a document is represented as a vector in word space and the degree of similarity between documents is calculated by the cosine of the angles between document vectors [39, 96]. Given any set of genes, GCAT calculates the cosine distribution of the gene set compared with that of a random gene set. More specifically, Fisher's Exact test is used to determine if the number of gene relationships above the cosine value 0.6 is significantly different from that which is expected by chance. The p-value obtained from this procedure is called Literature derived p-value (Lpv) [96]. Small Lpv values indicate that the input gene list are

functionally cohesive as opposed to random set of genes. In what follows we use GCAT to assess the biological relevance of set of markers obtained from our model.

Dissertation Outline

In this thesis we integrated double exponential prior and Generalized Double Pareto prior into the Bayesian Generalized Linear Models framework to induce sparseness in situations with the number of parameters to be predicted far exceeding the number of samples. In Chapter 2, we develop a hierarchical Bayesian Generalized Linear Model for binary response situations using Generalized Double Pareto prior on model parameters. In chapter 3, we develop a sparse Bayesian multinomial model that can handle multi-category response variables that are ordinal in nature. The model is applied and tested on a prostate cancer progression data set [90]. In chapter 4, we extend the model developed in chapter 2 in order to handle situations with ordinal response variables in Generalized Linear Models framework. This model was tested on a prostate cancer stages data set. Resampling techniques were used in order to remove the bias caused by the choice of training and test samples.

We performed 50 resamplings on the training and test samples and the average accuracy of the model across 50 runs was reported. We investigated the effect of literature aided initial input variable list on model performance in chapter 5. In chapter 6, we developed a literature aided sparse Bayesian generalized linear model that incorporates literature information to guide choice of hyper-parameters and amount of shrinkage imposed in the model thus bridging

predictive power and biological relevance of markers obtained. Chapter 7 includes discussion on future work.

Chapter 2

A Bayesian Approach for Inducing Sparsity in Generalized Linear Models Using Generalized Double Pareto Prior

Abstract

Identification of marker genes for classification of samples using microarray or RNAseq expression data remains challenging. In these settings, the data sets often contain a large number of variables (genes) and a relatively small number of samples which may render the variable selection process unstable. In addition, single variable analysis methods are too simple to give a comprehensive picture of the molecular mechanisms underlying complex phenotypes. Therefore, methods are needed to shrink the number of variables (induce sparsity) to avoid over-fitting, while accommodating simultaneous analysis of a large number of genes despite small sample sizes. The Generalized Double Pareto (GDP) prior is used to induce sparsity in a Bayesian generalized linear model setting. The GDP distribution has a spike at zero like the double exponential density, but has a Student t-like tail which helps remedy over-shrinkage of signals toward zero. In this study, a fully Bayesian hierarchical model was developed in order to facilitate Gibbs sampling. The GDP model was evaluated using three published datasets on leukemia and breast cancer. For each experiment, we randomly divided the samples into training and test groups. For each data set, using the top 10 genes, the GDP model achieved higher classification sensitivity (0.91-1.0) than the double exponential model (0.86-1.0). Interestingly, we found that the GDP model

identified marker genes with high literature derived functional cohesion. The top 100 genes identified by the GDP model had a literature p-value ranging from 2.06E-7 to 4.49E-24 for the three data sets, compared to the double exponential model (1.1E-3 to 3.13E-6). We conclude that the Bayesian sparse model with GDP prior results in better classification and more functionally relevant marker genes.

Introduction

High throughput expression studies are commonly used to identify genes that contribute mechanistically to a phenotype or provide biomarkers for classification of samples related to a given phenotype. A major challenge in analysis of gene expression is that relatively few samples are analyzed with respect to many thousands of gene expression variables. To address this problem, several approaches based on the idea of single variable analysis have been proposed [22, 66, 86, 91]. The shortcoming of all these methods is that they are univariate gene selection methods. However, most complex traits are polygenic so that a single variable analysis can only detect a very small portion of covariants and may not be powerful enough to identify weaker effects [7]. In order to address limitations that come with single variable analysis methods, recent efforts have focused on the development of various approaches for simultaneous analysis of multiple variables [53, 94, 98].

Traditionally, by using forward selection, backward elimination, and stepwise selection a subset of predictors in a regression framework is obtained. However, these approaches are computationally expensive and unstable even when the

number of predictors is small, [7, 48]. Researchers have used logistic regression extensively when the response variable is binary. But for the data structure explained above, procedures used to obtain maximum likelihood estimates of parameters will become computationally intensive and sometimes intractable. In addition, the maximization process may not converge to the maximum likelihood estimates and predictors may have large estimated variances which results in poor prediction accuracy [70].

It was previously proposed that prediction accuracy can be improved by setting the unimportant parameters associated with variables to zero and thus obtaining more accurate prediction for significant variables, [88]. There has been a great effort to develop methods that are able to analyze many variables simultaneously by inducing sparseness in the model while highlighting the relevant variables. Least Absolute Shrinkage and Selection Operator (LASSO) work by Tibshirani in 1996 drew much attention to the area. There exists a rich literature discussing methods to analyze the LASSO and related approaches [45, 104, 107, 108]. After the work of [27, 89] Bayesian approach to the same problem gained interest. A Bayesian LASSO was proposed recently by [67] and [37]. However, these procedures may cause over-shrinking of large coefficients due to the relatively light tails of the double exponential prior, and thus may introduce major bias. Using normal-Jeffreys prior which has heavier tails than double exponential distribution, we would be able to shrink small coefficients to zero while minimally shrinking large coefficients and thus obtaining better results. However, it has no meaning from an inferential aspect as it leads to an improper posterior, [5].

An alternative class of hierarchical priors proposed in [33] that uses Bayesian adaptive Lasso with non-convex penalization. However, it lacks simple analytic form. In [5], authors proposed the Generalized Double Pareto (GDP) prior distribution with application to continuous outcomes. The properties of this distribution that makes it appealing include: having a spike at zero alongside student-t like tails, simple analytic form and yielding proper posterior. In addition, it resembles double exponential density in the neighborhood of zero and has heavier tails compared to double exponential remedying unwanted bias resulting from over shrinkage of parameters toward zero [5].

In this article, we integrated a GDP prior into the Bayesian generalized linear models framework to induce sparseness in situations where the number of parameters to be predicted far exceeds the number of samples. The model developed can be used to analyze binary phenotypes. In step one, we derive the fully conditional distributions for all parameters in a multi-level hierarchical model in order to perform the fully Bayesian treatment of the problem. In the second step, the Markov Chain Monte Carlo (MCMC) method based on Gibbs sampling algorithm developed in step one is used to estimate all the parameters [30, 31]. The model shows a great flexibility to fit many variables at the same time. We apply our method to a leukemia dataset, [32], and two breast cancer tumor data sets [18] and [93]. The goal of the study is three-fold: Identification of a small number of genes having the greatest discriminating power in order to allow researchers to quickly focus on the most promising candidates for diagnostics and therapeutics, using the developed model to obtain the probability of each sample

belonging to one of the categories of phenotype, and obtaining high classification accuracy. In addition, we expect to not only achieve the above goals but also identify more biologically relevant genes to the phenotype under study.

Methods

In many different fields of science such as gene expression analysis continuous outcome variables are not very common and most often we are faced with dichotomous or multi-level response variables. In these situations, the simple linear regression model which is designed for analyzing models with continuous outcome variables is not appropriate. Generalized linear models (GLM) provide a way to handle these situations. Consider a situation with binary response. Let $y_1, y_2, ..., y_n$ represent the observed response variables in which 'n' is the number of observations (samples). Here y_i can take on 0 or 1 if for example the sample is normal or cancer respectively. In the case of gene expression analysis, gene expression levels are measured for each sample and we let w_{ij} represent the expression level of gene j in the i^{th} sample. In the context of GLM, nonlinear link functions are used to associate the nonlinear, non-continuous response variable to the linear predictor $\boldsymbol{w}_i{}^T\boldsymbol{\theta}$ in which $\boldsymbol{\theta}$ is a 1*p vector of $\boldsymbol{\theta}$=$[\theta_1, .., \theta_p]^T$ associated with covariate vector $\boldsymbol{w}_i$=$[w_{i1}, .., w_{ip}]^T$. Let H represent this link function. The GLM model can be represented as [1, 62]:

$$H(E(y_i)) = H(P(y_i = 1)) = \boldsymbol{w}_i{}^T\boldsymbol{\theta} \tag{2.1}$$

In this formula, w_i is the vector of covariates for individual i. We used logistic link function which corresponds to logistic regression [62].

$$H(P(y_i = 1)) = H(P_i) = ln\left(\frac{P_i}{1 - P_i}\right) \; ; \quad where \;\; P_i = P(y_i = 1) \qquad (2.2)$$

In order to be able to find the posterior distributions of parameters, we need to integrate the likelihood function multiplied by joint prior distributions of all parameters. However, this approach will result in an intractable integration. As explained in [1], in order to be able to set up the Gibbs sampler, we introduce 'n' independent latent variables $l_1, l_2, ..., l_n$ defined as $l_i = \boldsymbol{w}_i^T\boldsymbol{\theta} + e_i$. We assume logistic distribuion on error temrs, $F(e_i) = \frac{1}{1+e^{-e_i}}$, to obtain logistic regression. In order to be able to set up the Gibbs sampler, we approximate the logistic distribution on the latent variables with t-distribution defined as $l_i \sim t_v(\boldsymbol{w}_i{}^T\boldsymbol{\theta})$. The reason for choosing t-distribution is that logistic distribution has heavy tails and normal distribution does not provide a good approximation. Hence, we used student-t distribution with v degrees of freedom on latent variables to provide a better approximation for distribution on latent variables. We treat the degrees of freedom as unknown and estimate it alongside other parameters. It should be noted that this distribution is a non-central t-distribution with v degrees of freedom and non-centrality parameter $\boldsymbol{w}_i{}^T\boldsymbol{\theta}$. The following relationship is established between response and corresponding latent variable.

$$y_i = \begin{cases} 1 & \text{if} \quad l_i \geq 0 \\ 0 & \text{Otherwise} \end{cases}$$

This way the response and latent variables are linked in binary outcome situations. This approach connects the logistic regression for y_i to a linear

regression model for the latent variable l_i, [1]. The probability of each sample belonging to the category 1 can be calculated as follows.

$$p(y_i = 1) = p(l_i \geq 0) = p(e_i \geq -\boldsymbol{w}_i^T \boldsymbol{\theta}) = p(e_i < \boldsymbol{w}_i^T \boldsymbol{\theta}) = \frac{1}{1 + e^{-\boldsymbol{w}_I^T \boldsymbol{\theta}}}$$

Bayesian Hierarchical Model and Prior Distributions

In order to sample l_i from $t_v({\boldsymbol{w}_i}^T \boldsymbol{\theta})$, we use the following hierarchical model which is equivalent to sampling from the corresponding t-distribution [34]. This two-level hierarchical form is easier to work with both analytically and computationally compared to the original form of the t distribution [34]. This two level hierarchical distribution enables us to obtain closed forms for fully conditional posterior distributions of parameters.

$$l_i|\Lambda_i, \boldsymbol{\theta} \sim N({\boldsymbol{w}_i}^T \boldsymbol{\theta}, \frac{1}{\Lambda_i}); \;\; \Lambda_i \sim Gamma(\frac{v}{2}, \frac{v}{2}) \tag{2.3}$$

Here gamma distribution is defined as $\pi(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$. We put independent generalized double Pareto priors on all θ s. This prior is defined as follows, [5].

$$f(\theta|\zeta, \rho) = \frac{1}{2\zeta} * (1 + \frac{|\theta|}{\rho\zeta})^{-(1+\rho)} \; ; \; \rho, \zeta > 0 \tag{2.4}$$

Letting $\theta_j \sim GDP(\zeta = \frac{\delta}{\rho}, \rho)$ independently, the joint distribution of θs is defined as follows [5].

$$\pi(\boldsymbol{\theta}) = \prod_{j=1}^{p} [\frac{1}{2\frac{\delta}{\rho}} * (1 + \frac{|\theta_j|}{\delta})^{-(1+\rho)}] \tag{2.5}$$

GDP prior can be represented as a scale mixture of normal distributions leading to computational simplifications that makes Gibbs sampling feasible [5]. The

$GDP(\frac{\delta}{\rho}, \rho)$ prior is equivalent to the following hierarchical representation [5].

$$\theta_j|\tau_j \sim N(0, \tau_j);\ \tau_j \sim Exp(\frac{{\lambda_j}^2}{2});\ \lambda_j \sim Gamma(\rho, \delta) \tag{2.6}$$

The hyper parameters ρ and δ control the shape of the GDP distribution and thus the amount of shrinkage induced [5]. As δ increases the distribution becomes flatter and variance increases. As ρ increases the tails of distribution becomes lighter, variance becomes smaller, and the distribution becomes more peaked [5]. Thus, large values of ρ may cause unwanted bias for large signals and stronger shrinkage for noise-like signals while larger values of δ flattens the distribution and we may lose the ability to shrink noise-like signals [5]. In the absence of information on hyper parameters one can either set them to default values ($\rho = \delta = 1$) or choose a hyper prior distribution and let data speak about the values of these hyper parameters. We adopt the following prior distributions for these parameters.

$$\pi(\rho) = \frac{c}{(1 + c\rho)^2}; c > 0 \Rightarrow median(\rho) = \frac{1}{c} \tag{2.7}$$

$$\pi(\delta) = \frac{c'}{(1 + c'\delta)^2}; c > 0 \Rightarrow median(\delta) = \frac{1}{c'} \tag{2.8}$$

The priors on ρ and δ correspond to generalized Pareto priors with location parameter 0, shape parameter 1, and scale parameters c^{-1} and c'^{-1} respectively. As mentioned in the above formula, c and c' determine the location of the median of the distribution of parameters ρ and δ . For sampling purposes, we do the following transformations that leads to uniform prior distribution for the new

parameters (the proof is given in appendix 1) [5].

$$u_1 = \frac{1}{1+c\rho}; \qquad u_2 = \frac{1}{1+c'\delta} \tag{2.9}$$

Defining the parameters as above, the hierarchical representation of the model is as follows. $l_i|\Lambda_i, \boldsymbol{\theta} \sim N\left(\boldsymbol{w}_i{}^T\boldsymbol{\theta}, \frac{1}{\Lambda_i}\right)$, $\Lambda_i \sim Gamma\left(\frac{v}{2}, \frac{v}{2}\right)$, $\theta_j \sim N\left(0, \tau_j\right)$, $\tau_j \sim Exp\left(\frac{{\lambda_j}^2}{2}\right)$, $\lambda_j \sim Gamma\left(\rho, \delta\right)$, and we use non-informative uniform prior on v. Using the above mixture representation for the parameters and defining the prior distributions, we obtain following fully conditional posteriors that lead to a straightforward gibbs sampling algorithm. The derivation of fully conditional posterior distributions is presented in appendix1.

$$l_i|- \sim TN\left(\boldsymbol{w}_i{}^T\boldsymbol{\theta}, \frac{1}{\Lambda_i}\right) \tag{2.10}$$

In equation 2.10, 'TN' stands for truncated normal distribution, l_i is sampled from truncated normal distribution with parameters defined above. Point of truncation is zero and in each iteration of the Gibbs sampling, each l_i is sampled from above the truncation point if corresponding y_i is 1 and it will be sampled from the below the truncation point otherwise.

$$\boldsymbol{\theta}|- \sim MVN\left(\left[W^T\Lambda W + T^*\right]^{-1} W^T\Lambda\boldsymbol{L}, \left[W^T\Lambda W + T^*\right]^{-1}\right) \tag{2.11}$$

The normal distribution defined above is a multivariate normal distribution with mean vector and variance covariance matrix as specified. Where, $T^*_{p*p} = diag({\tau_1}^{-1}, ..., {\tau_p}^{-1})$, $\Lambda_{n*n} = diag(\Lambda_1, ..., \Lambda_n)$, and W is the n*p design matrix

in which w_{ij} represents expression level of gene j in the i^{th} sample.

$$\tau_j{}^{-1}|- \sim Inv - Ggaussian\left(\sqrt{\frac{\lambda_j{}^2}{\theta_j{}^2}}, \lambda_j{}^2\right) \tag{2.12}$$

In equation 2.12, Inv-Gaussian denotes inverse Gaussian distribution with location $\sqrt{\frac{\lambda_j{}^2}{\theta_j{}^2}}$ and scale $\lambda_j{}^2$. Each λ_j and Λ_j are sampled according to equation 2.13 and 2.14 respectively.

$$\lambda_j|- \sim Gamma\left(\rho + 1, |\theta_j| + \delta\right); j = 1, .., p \tag{2.13}$$

$$\Lambda_r|- \sim Gamma\left(\frac{v+1}{2}, \frac{1}{2}\left[(l_r - \boldsymbol{w}_r{}^T\boldsymbol{\theta})^2 + v\right]\right); r = 1, .., n \tag{2.14}$$

The fully conditional distributions for v, u_1, and u_2 are represented in equations 2.15 to 2.17 [5].

$$v|- \propto \left[\prod_{i=1}^{n} \Lambda_i{}^{\frac{v}{2}-1} exp\left(\frac{-v\Lambda_i}{2}\right)\right] * \left[\prod_{n=1}^{n} \frac{\frac{v}{2}^{\frac{v}{2}}}{\Gamma(\frac{v}{2})}\right] \tag{2.15}$$

$$u_1|- \propto \left(\frac{1-u_1}{cu_1}\right)^p * \prod_{j=1}^{p}\left(1 + \frac{|\theta_j|}{\delta}\right)^{-(\frac{1-u_1}{cu_1}+1)} \tag{2.16}$$

$$u_2|- \propto \left(\frac{c'u_2}{1-u_2}\right)^p * \prod_{j=1}^{p}\left(1 + \frac{c'u_2}{1-u_2}|\theta_j|\right)^{-(1+\rho)} \tag{2.17}$$

As we can see, the fully conditional distributions of v, u_1, and u_2 do not have closed form and thus we adopt the following embedded giddy gibbs sampling to sample from v, ρ, and δ [5, 75]. On a grid of k values $(v_1, v_2, ..., v_k)$ representing values of degrees of freedom considered, we perform the following procedure [5, 75].

- Calculate the weights as $r_i = \pi(v_i|-)$ according to formula 2.15.
- Normalize the weights ${r_i}^N = \frac{r_i}{\sum_{i=1}^{k} r_i}$
- Sample one value from $(v_1, v_2, ..., v_k)$ with probabilities $\left({r_1}^N, {r_2}^N, ..., {r_k}^N\right)$.

On a grid of values in interval $(0, 1)$ we use the same procedure to sample one value from u_1 and u_2 to use in the current iteration of Gibbs sampling. The only difference is that at the end of the procedure we transform u_1 and u_2 back to ρ and δ using $\rho = \frac{1}{c}\left[\frac{1}{u_1} - 1\right]$ and $\delta = \frac{1}{c'}\left[\frac{1}{u_2} - 1\right]$ respectively. The concise description of the Gibbs sampling algorithm explained above is represented in figure 1.

Datasets

The model was evaluated using one leukemia dataset [32] and two different breast cancer data sets [18] and [93]. The Golub leukemia data set included bone marrow or peripheral blood samples from 72 patients with either acute lymphoblastic leukemia (ALL) or acute myeloid leukemia (AML). The gene expression levels for 7129 human genes were measured for this cohort. For our study, this dataset was randomly split into a training group of 38 samples containing 27 ALL and 11 AML samples and a test group of 34 samples containing 20 ALL and 14 AML samples [32]. The Chin breast cancer data set contains gene expression profiles in 118 primary breast tumors (28 basal-like and 90 non-basal like) from a cohort of patients treated according to the standard of care between 1989 and 1997 [18]. The dataset was randomly divided into two training and test groups such that each group contains equal number of basal-like and non-basal samples. The Wang breast cancer data set contains gene

expression profiles in 249 breast tumors (43 basal-like samples and 206 non-basal) from patients with lymph-node negative breast cancer who were treated during 1980–1995, but who did not receive systemic neo-adjuvant or adjuvant therapy [93]. The dataset was randomly divided into two training and test

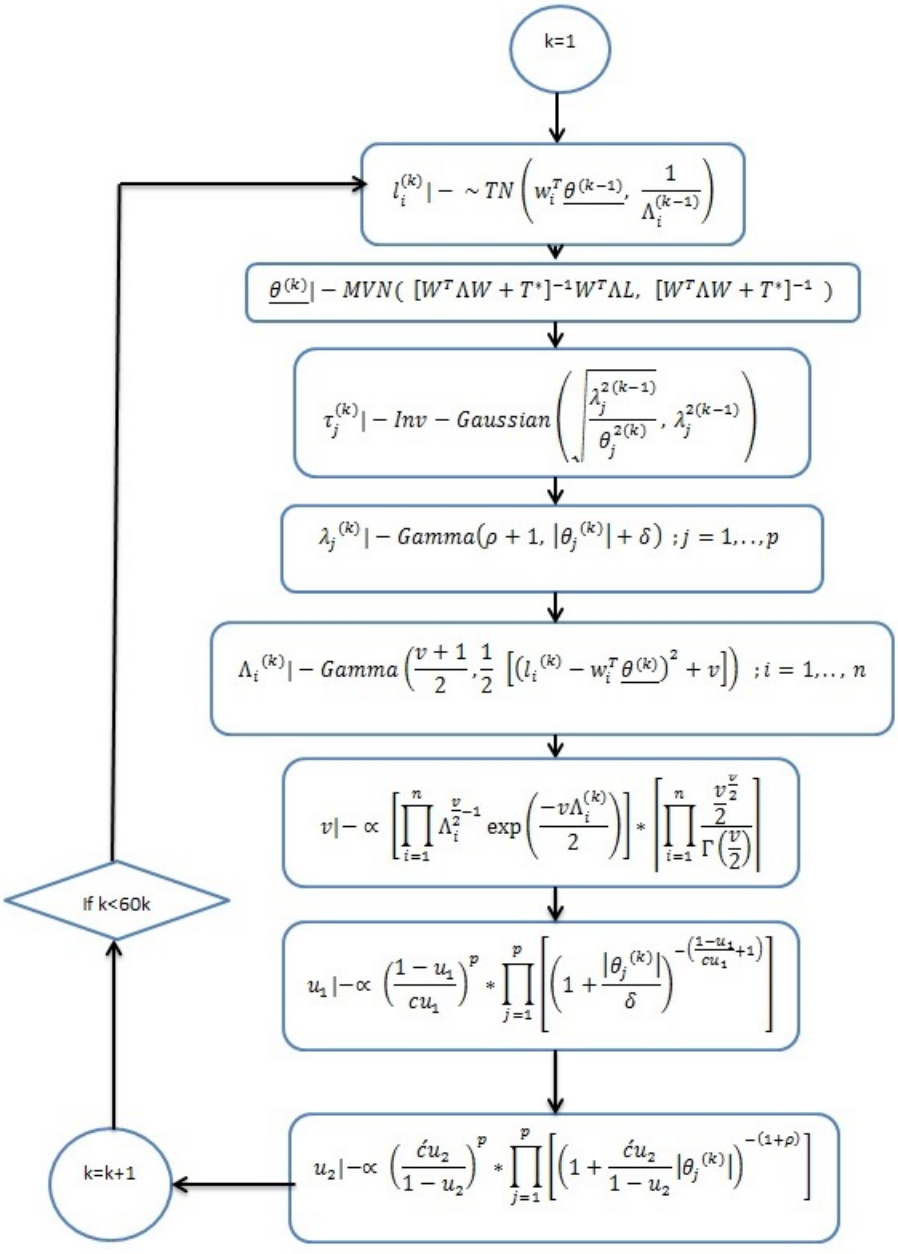

Figure 1: Gibbs sampling algorithm for model with Generalized Double Pareto prior and binary response.

groups such that each group contained equal sample number of basal-like and non-basal samples.

Results

For each data set, we used top five hundred differentially expressed genes as input to our model. For each of the train groups, the Gibbs sampler is run for 60,000 iterations and we discard the first 20,000 samples as burn in. In order to sample hyper parameters ρ and δ, we set c and c' to 1 to achieve the standard behavior of GDP prior [5]. Genes were selected based on posterior mean of θ associated with each gene. Figure 2 represents posterior mean of θ s for the 500 genes input to the model for the Golub data set. While some noise like signals are reduced toward zero, other signals stand out which turn out to be biologically more relevant to AML and ALL. Additionally, we obtained another sparse Bayesian Generalized linear model by imposing double exponential prior on θs (SBDE), [37, 56, 57, 67]. We used our model for class prediction of AML and ALL samples on the leukemia data set and basal-like and non-basal tumor samples in the breast cancer data sets. For example, the probability of a new sample being ALL was calculated as follows.

$$P(y_{new} = 1) = \frac{1}{1 + exp\left(-\boldsymbol{w}_i{}^T \hat{\boldsymbol{\theta}}\right)} \tag{2.18}$$

In this formula, $\hat{\boldsymbol{\theta}}$ is the posterior mean of θ s obtained for each train group and w_i is the vector of gene expression values associated with the corresponding $\hat{\theta}$ s used for prediction. Using only the first top ten genes-obtained from training the

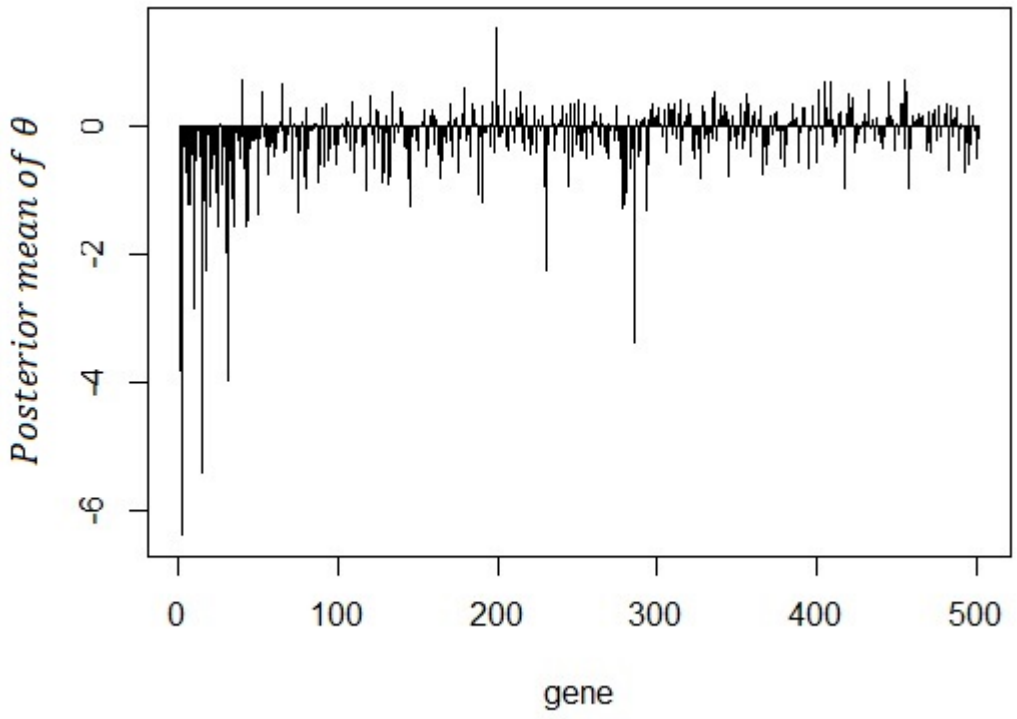

Figure 2: Posterior mean of θ asociated with each gene.

model on train groups ,we analyzed percent correct classification (accuracy), sensitivity, and specificity as measures of evaluation of the model. Sensitivity and specificity are statistical measures that evaluate performance of binary classifiers [4]. Sensitivity measures the proportion of actual positives (e.g. basal-like) that are identified by the model to be positive (e.g. basal-like) and specificity measures the proportion of negatives (e.g. non-basal) that are correctly classified as negative (e.g. non –basal). As a measure of the robustness of the model, we switched train and test groups for each data set and run the model again and obtained the classification results on the new test groups. Table 1 and 2 show the classification results on the test groups for each data set obtained using our model compared to SBDE. Also, the sensitivity and specificity of the

Table 1: Classification accuracy, sensitivity, and specificity for test groups.

DataSet	Model	Accuracy	Sensitivity	Specificity
Golub	GDP	0.941	1	0.86
Golub	SBDE	0.912	0.95	0.86
Gray	GDP	0.949	0.84	1
Gray	SBDE	0.966	0.86	1
Wang	GDP	0.976	0.91	0.99
Wang	SBDE	0.952	0.86	0.97

Table 2: Classification Accuracy, Sensitivity, and specificity for test group. Train and test groups switched

DataSet	Model	Accuracy	Sensitivity	Specificity
Golub	GDP	0.921	0.89	0.91
Golub	SBDE	0.895	0.85	0.91
Gray	GDP	0.949	0.79	1
Gray	SBDE	0.932	0.71	1
Wang	GDP	0.92	0.92	0.93
Wang	SBDE	0.911	0.71	0.95

classification when train and test groups were switched are show in Table 2 for test groups.

Latent Semantic Analysis(LSA) is a technique in natural language processing used for analyzing the relationships between a set of documents and the terms they contain by producing a set of concepts related to documents and terms [39]. Medline is the premier bibliographic database for biomedicine supported by national library of medicine. GCAT is a web-based tool that determines the functional coherence of gene sets by performing latent semantic analysis of Medline abstracts [96]. In GCAT, each gene –document was generated by concatenation of all titles and abstracts of the Medline. After gene-document was collected, latent Semantic Analysis (LSA) is used to calculate the gene-gene similarity matrix. LSA is a variant of the vector space model that reduces the

Table 3: GCAT top 100 genes' p-values for the GDP model compared to model with double exponential prior (SBDE).

DataSet	T-test	GDP	SBDE
Golub	0.19	2.1E-7	0.0015
Gray	3.1E-5	2.9E-8	0.0011
Wang	7.44E-40	4.49E-24	3.13E-6

dimensions of the matrix by applying Singular Value Decomposition (SVD) so that genes can be compared more conceptually [39]. Thus, LSA allows extraction of both explicit and implicit gene relationships from the literature. In a vector space Model, the semantic structure of a document is represented as a vector in word space and the degree of similarity between documents is calculated by the cosine of the angles between document vectors [39, 96]. Given any set of genes, GCAT calculates the cosine distribution of the gene set compared with that of a random gene set. More specifically, Fisher's Exact test is used to determine if the number of gene relationships above the cosine value 0.6 is significantly different from that which is expected by chance. The p-value obtained from this procedure is called Literature derived p-value (Lpv) [96]. Small Lpv values indicate that the input gene list are functionally cohesive as opposed to random set of genes.

We utilized GCAT to obtain the literature p-value for the top 100 genes obtained from t-test, SBDE, and GDP model. The results of the analysis is shown in Table 3. As we can see the p-values obtained under our model is highly more significant compared to SBDE which indicates that our model results in more biologically relevant genes compared to SBDE and t-test.

Discussion

Microarray gene expression technology continues to be used to obtain more understanding of mechanisms of human diseases, develop classifiers for prediction of poor versus good outcomes, and to detect relevant signals amidst a large body of noises [25, 65]. These information can be used for tailoring the treatments towards individuals [13, 72]. Gene expression studies usually measure several thousands of genes across the entire genome for few number of samples. Statistical modeling becomes challenging as the familiar “large p small n situations” arises. Identification of biologically relevant markers as well as ability to classify samples are of high interest among the community. Previous studies have shown that the correct selection of subsets of genes from microarray data is key for accurate classification of disease phenotypes, [17, 23].

In order to highlight those covariates that are most relevant to certain phenotype, it is necessary to develop an approach to weed out unimportant covariates [102]. Models that induce sparsity in terms of number of covariates in the model are of interest in order to obtain reliable and accurate predictions by learning classifiers [52]. It has been shown that majority of informative markers may not be highly differentially expressed and thus models that use very light-tailed priors are prone to the danger of losing biologically valuable information contained in these markers [47].

The key contribution of this work is to utilize a Generalized Double Pareto prior and develop a sparse Bayesian hierarchical Generalized linear model that can

accommodate binary phenotypes, obtain high classification accuracy, and identify biologically relevant genes at the same time. In our model, while shrinking small effects toward zero and producing sparse solutions, the over shrinkage problem caused by using light-tailed priors would be remedied by the heavier tails obtained via mixing over the hyper parameters using GDP prior [5]. We applied the model to the leukemia data set [32], and two breast cancer data sets [18] and [93]. We used the model to do prediction of sample type on the test datasets. The Bayesian set up enables us to assign the samples to one of the categories in a coherent way. Classification accuracy, sensitivity, and specificity were used as measures of model performance. As shown in table 1 and 2, the model developed obtains high classification accuracy, sensitivity and, specificity in all 3 data sets and outperforms the SBDE model in all cases except 1.

In order to test robustness of the model, we switched training and test data set and trained the model on the new train dataset and performed classification on the new training and test data sets. In this case the model results in better classifying accuracy in all three data sets which is in accordance with the results obtained in the first analysis. GCAT literature p-value of the top 100 genes obtained from the model represents the biological relevance of markers obtained [96]. Our model results in more significant literature based p-values which indicates that more biologically relevant genes are obtained using our model compared to SBDE. In conclusion, using the GDP prior in a Bayesian generalized linear models frame work we were able to achieve high classification accuracy and obtain biologically relevant marker genes to the outcomes in each experiment.

There exists a possibility of utilizing Metropolis–Hastings algorithm instead of Griddy Gibbs sampling algorithm employed to sample hyper-parameters v, u_1, u_2. Metropolis Hastings is a Markov chain Monte Carlo (MCMC) method for obtaining a sequence of random samples from a probability distribution for which direct sampling is difficult [31]. The Metropolis–Hastings algorithm can draw samples from any probability distribution $P(x)$, provided you can compute the value of a function $f(x)$ which is close to the density of P. On the other hand, most simple rejection sampling methods suffer from the dimensionality, where the probability of rejection increases exponentially as a function of the number of dimensions [31]. Metropolis Hastings algorithm is only useful when you can find a suitable "jumping" density which is "similar" (close) to its target density to avoid excessively slow mixing [31]. This is a difficult task, especially for high-dimensional space. In addition, the metropolis algorithm within each iteration on the last part of the MCMC procedure would dramatically increase the running time of the MCMC process.

In future, we plan to incorporate literature information into the prior distributions in order to design literature informed priors that would potentially enable us to obtain machine learning models with high classification accuracy which provide very enriched set of markers with high biological relevance to the phenotype under study. This potential development which could bridge the gap between classification accuracy and biological relevance will be of high merit to the community and can potentially result in deeper understanding of mechanisms involved. The model developed here should be extendable to datasets with

multi-level response variables. In chapter 3 and chapter 4, we explore the development of sparse bayesian generalized linear models to address multi-category response situations.

Chapter 3

Application of Sparse Bayesian Generalized Linear Model to Gene Expression Data for Classification of Prostate Cancer Subtypes

Abstract

A major limitation of expression profiling is caused by the large number of variables assessed compared to relatively small sample sizes. In this study, we developed a multinomial Probit Bayesian model which utilizes the double exponential prior to induce shrinkage and reduce the number of variables in the model. A hierarchical Sparse Bayesian Generalized Linear Model (SBGLM) was developed in order to facilitate Gibbs sampling which takes into account the progressive nature of the response variable. The method was evaluated using a published dataset (GSE6099) which contained 99 prostate cancer cell types in four different progressive stages [90]. Initially, 398 genes were selected using ordinal logistic regression with a cutoff value of 0.05 after Benjamini and Hochberg FDR correction. The dataset was randomly divided into training (N=50) and test (N=49) groups such that each group contained equal number of each cancer subtype. In order to obtain more robust results we performed 50 re-samplings of the training and test groups. Using the top ten genes obtained from SBGLM, we were able to achieve an average classification accuracy of 85% and 80% in training and test groups, respectively. To functionally evaluate the model performance, we used a literature mining approach called Geneset Cohesion Analysis Tool [96]. Examination of the top 100 genes produced an

average functional cohesion p-value of 0.007 compared to 0.047 and 0.131 produced by classical multi-category logistic regression and Random Forest approaches, respectively. In addition, 96% of the SBGLM runs resulted in a GCAT literature cohesion p-value smaller than 0.047. Taken together, these results suggest that sparse Bayesian Multinomial Probit model applied to cancer progression data allows for better subclass prediction and produces more functionally relevant gene sets.

Introduction

As data collection technologies evolve, the number of variables which can be measured in experiments increase. For example, modern microarray experiments can measure the expression levels of several thousand genes simultaneously. Since the number of samples is typically much smaller than the number of variables, it is challenging to identify important genes among the large amount of data points [15]. Many univariate analysis approaches have been applied to select important genes from microarray experiments such as t-test [21], regression modeling [87], mixture model [66] and non-parametric methods [24, 91]. However, since most complex traits are polygenic, a single variable analysis can only detect a very small portion of the relevant variation and may not be powerful enough to identify weaker interactions between the variables [7].

In order to address limitations of single variable analysis methods, several approaches have been developed for simultaneous analysis of multiple variables [53, 94, 98]. In linear regression framework, the least square method is used to obtain estimate of parameters. The ordinary least square estimates

obtained are not quite satisfactory mainly due to poor accuracy of prediction resulting from high variances of estimates, the large number of variables with respect to small sample size, and the error in variables [88]. It is preferred to select a smaller subset of variables, sometimes referred to as feature selection, which offer the strongest effect and discriminating power. A standard method used to improve the parameter estimation, prediction, and classification is subset selection and its variants such as backward elimination, forward and stepwise selections. These methods are all discrete processes and can be highly inconsistent, meaning that a small change in the data can result in very different models [45, 48, 70, 88, 104, 107]. In addition, these approaches are computationally expensive and unstable when sample sizes are much smaller than the number of variables [49, 88]. Moreover in this setting, over-fitting is a major concern and may result in failure to identify important predictors. Thus, the data structure of typical microarray experiments makes it difficult to use traditional multivariate regression analysis [7]. Given the aforementioned drawbacks, several groups have developed methods to simultaneously analyze a large number of variables [26, 49, 64, 101, 107]. It has been proposed that prediction accuracy can be improved by setting the parameters associated with unimportant variables to zero and thus obtaining more accurate prediction for the significant variables [88].

Various methods such as K-nearest neighbor classifiers [24], linear discriminant analysis [99], and classification trees [24] have been used for multi-class cancer classification and discovery [14, 20, 73]. However in all these methods, gene selection and classification are treated as two distinct steps that

can limit their performance. One alternative to deal with these situations is using Generalized Linear Models (GLM) [37, 58, 60, 62]. Researchers have used GLM methodology extensively when the response variable is not continuous. But for typical microarray experiments, procedures to obtain maximum likelihood estimates of parameters will become computationally intensive and sometimes intractable. In addition the maximization process may not converge to the maximum likelihood estimates and predictors may have large estimated variances which results in poor prediction accuracy [70]. In order to avoid over-fitting and improve model accuracy, models which impose sparsity in terms of variables (genes) are desirable [88]. Least Absolute Shrinkage and Selection Operator (LASSO) is a well-known method for inducing sparseness in the model while highlighting the relevant variables [45, 88, 95, 107]. A Bayesian LASSO method was proposed by [37, 67] in which double exponential prior is used on parameters in order to impose sparsity in the model. In this article, we integrate double exponential prior distribution into the Bayesian generalized linear model framework to induce sparseness in situations where the number of parameters to be predicted exceeds the number of samples. The model developed can be used to analyze multi-category phenotypes such as progressive stages of cancer. In step one, we derive the fully conditional distributions for all parameters in a multi-level hierarchical model in order to perform the fully Bayesian treatment of the problem. In the second step, the Markov Chain Monte Carlo (MCMC) method [30, 31] based on Gibbs sampling algorithm is used to estimate all the parameters. This model takes into account the ordinal nature of the response

variable. We applied and evaluated our model to a publicly available prostate cancer progression dataset [90]. The goals of the study are to test if a hierarchical Sparse Bayesian Generalized Linear Model (SBGLM) can: 1) Identify a smaller number of genes with high discriminating power; 2) Obtain high classification accuracy; 3) Identify more biologically relevant genes related to the phenotype under study.

Methods

In many biomedical research applications, dichotomous or multi-level outcome variables are desired. In these situations, the simple linear regression model which is designed for continuous outcome variables is not appropriate due to heteroscedasticity and non-normal errors. Furthermore, there is no guarantee that the model will predict legitimate responses (e.g. 1, 2, 3, and 4 in polytomous response variable with 4 levels). Generalized linear models (GLM) provide a way to address these situations [58, 60, 62]. Let $[y_i, w_{i1}, ..., w_{ip}]_{i=1}^{n}$ represent n observations in which the response variables y_i can take values 1, 2, 3, .., k where k is the number of categories of the ordinal response variable. In addition, let $(w_{i1}, .., w_{ip})$ represent the value of variable 1 to variable p in observation $'i'$. In the case of gene expression analysis, gene expression levels are measured for each sample and w_{ij} represents expression level of gene j in i^{th} sample. We implemented GLM for ordinal response in Bayesian framework by utilizing link functions and careful introduction of latent variables [1]. In Bayesian framework, the joint distribution of all parameters is proportional to the likelihood multiplied by the joint prior distributions on the parameters. More specifically, in Bayesian

Multinomial Probit Model, the likelihood function is defined as in formula (3.1) in which π_{ij} is the probability that sample i is from j^{th} category where j ranges from 1 to k and k is the number of ordinal categories of response variable [1]. In formula 3.1, $I(y_i = j)$ is an indicator function having value one if the y_i is in category j and zero otherwise. It should be noted that each observation contributes one value in the inner product to the equation (3.1) since the indicator function returns value of zero if j is not equal to the category of outcome for the sample.

$$L(\boldsymbol{\pi}|\boldsymbol{y}) = \prod_{i=1}^{n} \left[\prod_{i=1}^{k} \left(\pi_{ij}^{I(y_i=j)} \right) \right] \tag{3.1}$$

In order to be able to find the posterior distributions of parameters, we must integrate the likelihood function multiplied by joint prior distributions of all parameters. However, this approach will lead to an intractable integration. As explained in [1], in order to be able to set up the Gibbs sampler and incorporate regression parameters into the model, we introduce 'n' independent latent variables $l_1, l_2, ..., l_n$ defined as $l_i = \boldsymbol{w}_i^T \boldsymbol{\theta} + e_i$ with $e_i \sim N(0, 1)$ [1]. In this formula, $\boldsymbol{w}_i^T$ is the vector of gene expressions for individual i and $\boldsymbol{\theta} = (\theta_1, .., \theta_p)$ are parameters associated with variable 1 to variable p. The following relationship is established between response variable and its corresponding latent variable [1].

$$y_i = \begin{cases} 1 & iff \qquad -\infty = \gamma_1 \leq l_i < \gamma_2 \\ 2 & iff \qquad 0 = \gamma_2 \leq l_i < \gamma_3 \\ : & \\ k & iff \qquad \gamma_k \leq l_i < \gamma_{k+1} = \infty \end{cases} \tag{3.2}$$

In order to insure that the thresholds are identifiable, following the guidelines of [1], we fix γ_2 at zero and γ_1 and γ_{k+1} are defined according to equation above. In the context of GLM, we use nonlinear link functions to associate the nonlinear, non-continuous response variable to the linear predictor $w_i^T\boldsymbol{\theta}$. Using the relations defined above, the probability of each sample being in category j (j=1, 2, .., k) is derived in equation 3.3 in which Φ represents cumulative distribution function of standard normal distribution and π_{ij} is the probability of sample i being from category j [1].

$$\zeta_{ij} = P(y_i \leq j) = P(l_i \leq \gamma_{j+1}) = P(e_i + \boldsymbol{w}_i^T\boldsymbol{\theta} \leq \gamma_{j+1}) = \Phi(\gamma_{j+1} - w_i^T\boldsymbol{\theta}) \ ; \ \pi_{ij} = \zeta_{ij} - \zeta_{ij-1} \tag{3.3}$$

In this way, the linear predictor $\boldsymbol{w}_i^T\boldsymbol{\theta}$ is linked to the multi-category response variable y_i. The function that links the linear predictor to the response variable is called a link function and in the multinomial Probit model, this link function is cumulative distribution of standard normal density as defined above [1, 60].

Bayesian Hierarchical model and prior distributions

A sparse Bayesian ordinal Probit model was implemented which takes into account the ordinal nature of cancer progression stages and can accommodate

large number of variables. We used independent double exponential prior distributions on θ_j as follows [7, 67]. It should be noted that θ_j is the parameter associated with gene j. This prior distribution has a spike at zero and light tails which enables us to incorporate sparsity in terms of number of variables used in the model [7, 107].

$$\pi(\theta_j|\lambda) = \frac{\sqrt{\lambda}}{2} e^{-\sqrt{\lambda}|\theta_j|} \tag{3.4}$$

The double exponential distribution can be represented as scale mixture of normal with an exponential mixing density [7, 37, 67, 107]. This hierarchical representation will be used in order to be able to set up the Gibbs sampler [7, 67, 107].

$$\frac{\sqrt{\lambda}}{2} e^{-\sqrt{\lambda}|\theta_j|} = \int_0^\infty \frac{1}{\sqrt{2\pi\eta_j}} e^{-\frac{\theta_j^2}{2\eta_j}} * \frac{\lambda}{2} e^{-\frac{\lambda}{2}*\eta_j} d\eta_j \tag{3.5}$$

Having $l_i \sim N\left(\boldsymbol{w}_i^T \boldsymbol{\theta}, 1\right)$, the following hierarchical prior distribution is used on parameters associated with gene 1 to gene p [7].

$$\theta_j|\eta_j \sim N\left(0, \eta_j\right) \quad ; \quad \eta_j \sim Exp\left(\frac{\lambda}{2}\right) \tag{3.6}$$

Defining the parameters as above, the hierarchical representation of the model is as follows. $l_i|\boldsymbol{\theta} \sim N(\boldsymbol{w}_i{}^T \boldsymbol{\theta}, 1)$, $\theta_j|\eta_j \sim N(0, \eta_j)$, and $\eta_j \sim Exp\left(\frac{\lambda}{2}\right)$. We also assume uniform priors on thresholds and we will find their fully conditional posterior distribution alongside other parameters. Using the above mixture representation for the parameters and defining prior distributions, we obtain the following fully conditional posterior distributions that will be used in a simple Gibbs

sampling algorithm.

$$l_i|- \sim DTN(\boldsymbol{w}_i^T\boldsymbol{\theta}, 1) \tag{3.7}$$

In formula 3.7, DTN stands for doubly truncated normal distribution. For observation 'i' with $y_i = r$, l_i must be sampled from normal distribution defined above truncated between γ_r and γ_{r+1} in each iteration of the algorithm.

$$\boldsymbol{\theta}|- \sim MVN\left(\left[W^TW + T^{-1}\right]^{-1}W^T\boldsymbol{L}, \left[W^TW + T^{-1}\right]^{-1}\right) \tag{3.8}$$

Fully conditional posterior distribution of vector of model parameters is multivariate normal distribution with mean vector and variance covariance matrix as specified where $T = diag(\eta_1, \eta_2, .., \eta_p)$. In 3.8, W is the $n * p$ matrix in which w_{ij} represents expression level of gene j in i^{th} sample and p is the number of genes (variables) in the model and $L = [l_1, l_2, \ldots, l_n]^T$ and 'n' is the number of samples. The fully conditional distribution of hyper-parameters $\eta_j^{-1}, j = 1, .., p$ are inverse-Gaussian distribution with location $\frac{\sqrt{\lambda}}{|\theta_j|}$ and scale λ. In each iteration of the Gibbs sampling, η_j^{-1} is sampled from the inverse gaussian distribution defined in equation 3.9.

$$\eta_j^{-1}|- \sim inv - Gaussian\left(\frac{\sqrt{\lambda}}{|\theta_j|}, \lambda\right) \tag{3.9}$$

In the case of multinomial response, we assign independent uniform priors to thresholds and thus the fully conditional distribution for thresholds is uniform distribution and we need to sample them in each iteration of Gibbs sampling alongside other parameters in the model [1].

$$\gamma_s|- \propto \prod_{i=1}^{n}\left[I(y_i = s-1) * I(\gamma_{s-1} \leq l_i < \gamma_s) + I(y_i = s) * I(\gamma_s \leq l_i < \gamma_{s+1})\right] \tag{3.10}$$

As explained in [1] , the conditional posterior distribution of γ_s can be seen to be $Uniform(\delta_1, \delta_2)$ in which $\delta_1 = max\left[max_i\left[l_i|y_i = s-1\right], \gamma_{s-1}\right]$ and $\delta_2 = min\left[min_i\left[l_i|y_i = s\right], \gamma_s\right]$, [1]. It should be noted that $I()$ is indicator function and its value is one if its argument is true and is zero otherwise. Figure 3 represents the Gibbs sampling algorithm workflow in a coherent way.

Dataset and feature selection

The method was applied to a published dataset on prostate cancer progression downloaded from Gene Expression Omnibus at NCBI (GSE6099) [90]. The data set contains gene expression values for 20,000 probes and 101 samples corresponding to five prostate cancer progressive stages (subtypes): Benign, prostatic intraepithelial neoplasia (PIN), Proliferative inflammatory atrophy (PIA), localized prostate cancer (PCA), and metastatic prostate cancer (MET) [90]. Since there were only two samples for PIA, we removed these samples from further analysis. Probes with missing in more than 10 percent of the samples were removed from the data set. For the remaining probes, the missing values were imputed by using the mean value of the probe across samples with non-null values. Before applying our model to this data set, for each gene we performed logistic regression for ordinal response. This method enables us to take into account the ordinal nature of response variable in the analysis and provides a gene list to be used as input to the model. Genes were ranked based on the p-value associated with the hypothesis $H_0 : \theta_i = 0$ from the most significant to least significant. θ_i is the parameter associated with gene i. We performed Benjamini and Hochberg FDR correction [9]. An FDR cutoff value of

0.05 resulted in a list of 398 genes. Thus, the input to our model was 398 variables (genes) for 99 samples corresponding to four different prostate cancer subtypes. The Gibbs sampling algorithm was implemented in R software and the program ran for $60k$ iterations and the first $20k$ was discarded as burn-in.

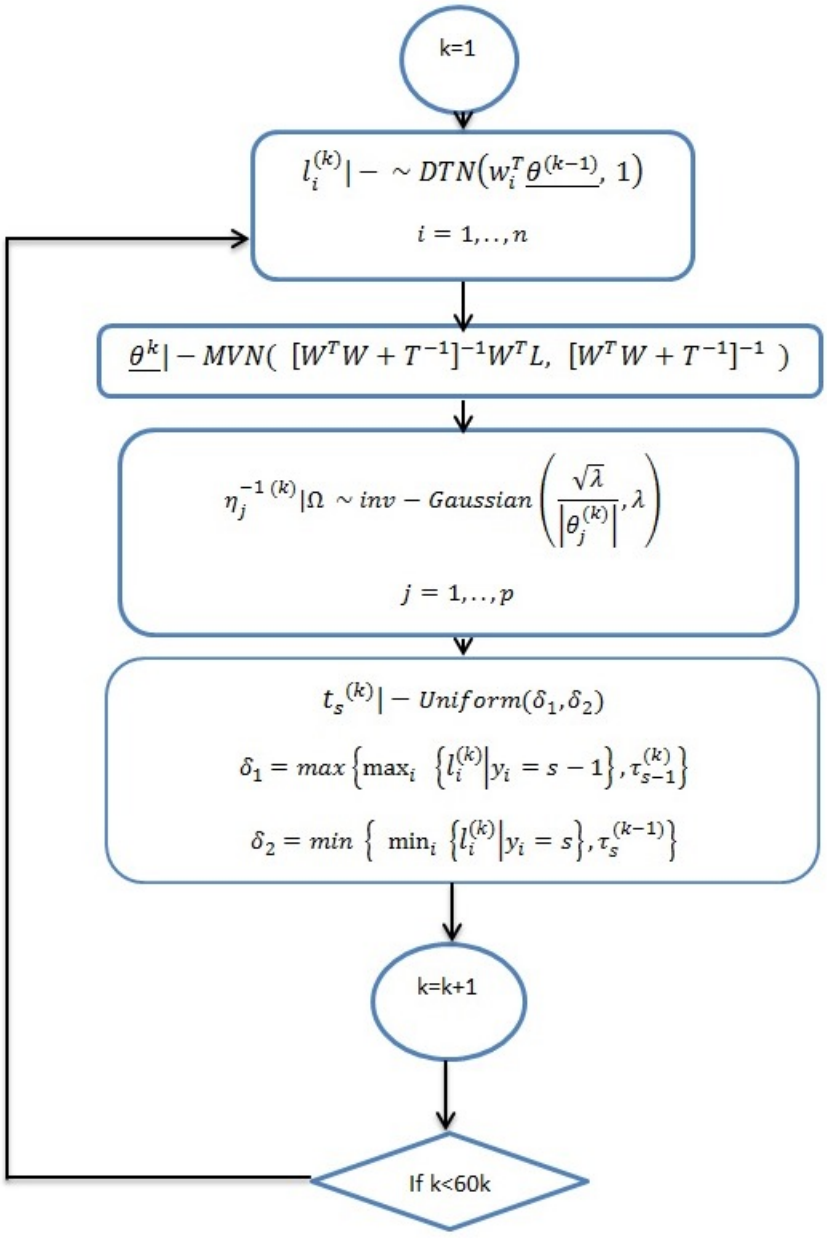

Figure 3: Gibbs sampling algorithm flowchart for sparse Bayesian Generalized Linear model utilizing Double Exponential prior and multinomial response.

Evaluation

The dataset was randomly divided into training (N=50) and test (N=49) groups such that each group contained an equal number of prostate cancer subtypes, Benign, PIN, PCA and MET. Taking benign as an example, there are 34 benign samples, we randomly divide it into two groups one half is used as part of the training and the other half is hold out to be used as part of the test set for model evaluation. The same procedure is repeated for PIN, PCA, and MET to obtain complete train and test sets each having equal number of each subtype. Genes were ranked based on posterior mean of parameters and the top 10 or 50 genes obtained from the model were used for classification. In order to make the model more robust we performed 50 re-samplings on selection of training and test groups and re-ran the model. The average performance of SBGLM was compared to two well-known classification methods: Support Vector Machine (SVM) and Random Forrest. SVM was implemented in R software using Kernlab library [42] and Random Forest was implemented in R using randomForest library [51].

Results

Figure 4 shows an example of the mean of posterior distribution of θ s associated with 398 genes in a single run of SBGLM. We used the top 10 or 50 genes to test the classification accuracy of the SBGLM on 50 resampled training and test groups. Each training and test group had an equal number of the four prostate cancer subtypes: Benign, prostatic intraepithelial neoplasia (PIN), localized prostate cancer (PCA), and metastatic prostate cancer (MET). We found

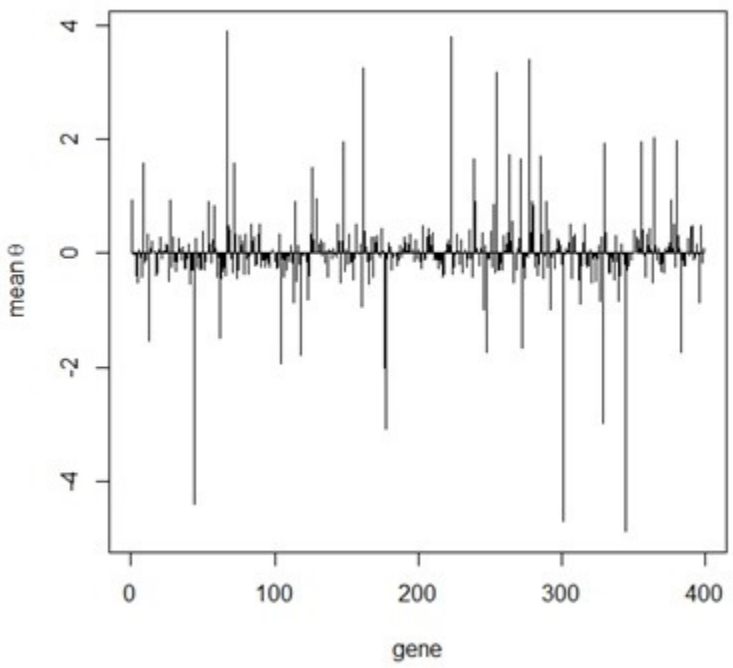

Figure 4: Posterior mean of θ asociated with gene 1 to gene 398.

that the average overall classification accuracy of the SBGLM was 80.4 and 82.3 percent when using 10 and 50 marker genes, respectively (Table 4). The performance of SBGLM approach was compared to two well-known classification methods, SVM and Random Forest [12] when using top 10 or top 50 genes from 398 input genes. We found that the overall accuracy of SBGLM was substantially better than SVM and was comparable, albeit slightly lower, to Random Forrest when using either 10 or 50 marker genes.

It is important to note that the feature selection for SVM and Random Forests was based on the p-values of the ordinal linear regression model (top 10 and top 50 from the 398 input genes). These results indicate that a small subset of the 398 input genes is better for predicting prostate cancer progression. Next, we examined the performance of SBGLM with regard to classifying the different

Table 4: Overall average accuracy of SBGLM, SVM and Random Forest using 10 and 50 marker genes.

Model	P=10	P=50
SBGLM	80.4(0.06)	82.3(0.063)
SVM	53.6(2.7)	0.67(3.04)
Random Forest	83(1.6)	84.6(2)

Table 5: Average classification accuracy of prostate cancer subtypes in the test group using SBGLM, SVM and Random Forest with 10 marker genes.

Sample Type	SBGLM	SVM	Random Forest
Benign	95.1(6)	84.4 (5.3)	91.1(4.5)
PIN	61.7(2.8)	9(7.2)	61.4(1.9)
PCA	86.9(1.1)	37.4(9)	86.7 (2.1)
MET	56(3.2)	55.3 (1.2)	82.8(7.3)

subtypes of prostate cancer in comparison to SVM and Random Forrest (Table 5). When using 10 marker genes, SBGLM classified all four subtypes of prostate cancer more accurately than SVM, and it performed better than Random Forrest for classifying Benign, PIN, and PCA. Interestingly however, when using 50 marker genes, SBGLM performed better than Random Forrest at classifying Benign, PIN and MET(Table 6). These results indicate that the performance of SBGLM is comparable to Random Forrest in classifying subtypes of prostate cancer, although the results for both methods are sensitive to the number of selected marker genes. Since the results of SBGLM were comparable to Random Forrest, we next asked if SBGLM gene rankings were more or less relevant to the biological mechanisms associated with prostate cancer progression. As a first step in evaluating the biological relevance for the top ranked genes in the models, we used a literature based method called GeneSet Cohesion Analysis Tool (GCAT) [96].

Table 6: Average classification accuracy of prostate cancer subtypes in the test group using SBGLM, SVM and Random Forest with 50 marker genes.

Sample Type	SBGLM	SVM	Random Forest
Benign	99.6(1.9)	90.1(1.7)	96.8(1.3)
PIN	53.4(1.4)	38.2(8.2)	52(1.1)
PCA	65.4(7.2)	45.8(6.2)	84.8(5.4)
MET	95.4(6.3)	81.8(1.6)	83.6(7.09)

GCAT is a web-based tool that determines the functional coherence p-values of gene sets based on latent semantic analysis of Medline abstracts [96]. The literature derived p-value is obtained by comparing distribution of gene similarities for the gene set to the one obtained for a randomly selected genes from the whole genome [96]. The small Lpv is an indication of functional cohesion of gene set. Table 7 shows the average GCAT literature derived p-values (LPv) for the top 100 genes obtained from 50 runs of SBGLM and Random Forrest as well as the top 100 genes based on the p-value rank ordering of single gene analysis using ordinal logistic regression. We found that on average, SBGLM produced more functionally cohesive gene list (LPv = 0.007) compared to classical logistic regression (LPv= 0.047) and Random Forest (LPv=0.131). Notably, 96 percent of the runs had smaller LPv than 0.047, produced by initial p-value ranking. Based on these results, we conclude that although SBGLM produces comparable classification accuracy as Random Forrest, it identifies more biologically relevant gene markers.

Discussion

Complex diseases and biological processes are caused by interaction of multiple genes (gene products). Hence, current approaches which rely on single

Table 7: Literature based functional cohesion p-values (LPv) of the top 100 genes obtained from three different models.

Model	GCAT P-value
SBGLM	0.007 (0.001)
Classical Logistic regression	0.047
Random Forest	0.131(0.07)

variable analysis have limited utility in understanding molecular mechanisms and identification of genetic biomarkers for classification of diseases [14, 21, 73]. Moreover, most genomic approaches collect data for a much larger set of gene variables compared to the number of samples being investigated. Therefore, highly regularized approaches, such as penalized regression models, are needed to identify non-zero coefficients, enhance model predictability and avoid over-fitting [107]. Lastly, continuous response variables which are a requirement of linear regression methods are not applicable to response variables (phenotypes) that are dichotomous or polytomous. To address these limitations, we developed a sparse Bayesian multinomial model and evaluated its performance using prostate cancer gene expression data. We found that the SBGLM classification accuracy of prostate cancer subtypes were comparable to Random Forrest. However, SBGLM identified more biologically relevant gene sets (Table 7).

Based on these results, we posit that SBGLM may be a better approach to simultaneously identify marker genes for classifications as well as gaining insights into the molecular mechanisms of the phenotype under investigation. Interestingly, using fewer genes, SBGLM had very good discrimination

performance for classifying benign (99.6% accuracy) versus metastatic prostate cancer (95.4% accuracy), but the model discrimination was weaker for PIN and PCA (Table 5). These results are consistent with the previous observation that PIN and PCA share markedly similar expression signatures [90]. We found that increasing the number of marker genes to 50 does not improve discrimination between PIN and PCA, suggesting that different molecular mechanisms may underlie the progression of PIN to PCA.

Chapter 4

A Robust Bayesian Approach for Inducing Sparsity in Generalized Linear Models with Multi-Category Response

Abstract

The dimension and complexity of gene expression data obtained from microarrays has created challenging data analysis problems. Specifically, large number of genes to be analyzed compared to small number of samples is a major limitation in expression profiling. This issue has attracted attention to shrinkage and estimation methods. In this study, We utilized the Generalized Double Pareto (GDP) prior to induce sparsity in Bayesian generalized linear models setting. GDP while has a spike at zero like the double exponential density, it also has a Student t-like tail behavior which helps us remedy over shrinkage of signals toward zero and thus offers more robustness properties. A hierarchical Sparse Bayesian Generalized Linear Model using GDP prior (SBGG) was developed in order to facilitate Gibbs sampling which takes into account the progressive nature of the response variable. Bayesian computation is straightforward via the simple Gibbs sampling algorithm developed. The method was evaluated using a published dataset (GSE6099) which contained 99 prostate cancer cell types in four different progressive stages. Initially, 398 genes were selected using ordinal logistic regression with a cut-off value of 0.05 after Benjamini and Hochberg FDR correction. The dataset was randomly divided into training (N=50) and test (N=49) groups such that each group contained equal number of each cancer subtype. In

order to obtain more robust results we performed 50 re-samplings of the training and test groups. We were able to achieve an average classification accuracy of 86% and 82.5% in training and test groups, respectively using only the top ten genes obtained from SBGG. We functionally evaluated the model performance by using a literature mining approach called Geneset Cohesion Analysis Tool. Examination of the top 100 genes produced an average functional cohesion p-value of 2.0E-4 compared to 0.007, 0.047, and 0.131 produced by Sparse Bayesian Generalized Linear Model obtained by imposing double exponential prior on parameters (SBGDE), classical multi-category logistic regression, and Random Forest approaches, respectively. In addition, 100 percent of the SBGG runs resulted in a GCAT literature cohesion p-value smaller than 0.047. Based on our results, we conclude that the Bayesian Multinomial Generalized Linear model applied to cancer progression data results in better subclass prediction and produces more functionally relevant gene sets.

Background

Genomic research has benefited from microarray technology as a high throughput discovery tool. In modern microarray experiments, expression levels of several thousand of genes are measured across small number of samples (usually less than100). The dimension and complexity of gene expression data obtained from microarrays creates challenging data analysis problems. One of the major challenges is related to the nature of microarray experiments having substantially smaller number of samples compared to tens of thousands of variabes. This is due to the fact that the very small sample size makes it very

challenging to identify important genes among the pool of large number of genes at hand [15]. Several statistical methods in univariate and multivariate analysis frameworks have been developed to address this problem. Some of the univariate analysis approaches applied to selection of important genes from microarray experiments include t-test [46], regression modelling [50], mixture model [66] and non-parametric methods [90, 91]. However, single gene analysis is unable to identify weaker associations especially for complex polygenic phenotypes for which the relevant variation is distributed across several variables [7]. In order to address limitations of single variabe analysis methods, Several approaches for simultaneous analysis of multiple variabes have been developed [53, 94, 98].

One of these classical techniques is linear regression. In a linear regression framework, the least square method is used to obtain estimate of parameters. In cohorts with large number of variabes compared to much smaller sample size, parameter estimates based on ordinary least squares have high variances which results in poor prediction accuracy [88]. Feature selection that can result in set of genes with strongest effect and discriminating power is of high interest. Variable selection in regression framework namely backward elimination, forward selection, and stepwise selection have been used as a standard method to improve parameter estimation and prediction. One of the shortcomings of these methods is that these are discrete processes which are very sensitive to the changes in the data at hand. That is, a minor change in data can result in very different models [49, 88, 107]. Additionally, the computational complexity of these approaches when the number of variabes are very large makes them less

attractive for gene expression analysis [49, 88]. Moreover in this setting, over-fitting is a major concern and may result in failure to identify important predictors. Thus, the data structure of typical microarray experiments makes it difficult to use traditional multivariate regression analysis [7].

Several groups have developed methods in an attempt to overcome these drawbacks [49, 53, 94, 98, 108]. Various methods such as K-nearest neighbour classifiers [90], linear discriminant analysis [99], and classification trees [90] have been used for multi-class cancer classification and discovery [14, 20, 73]. However, gene selection and classification are treated as two separate steps which can limit their performance. One alternative to deal with these situations is using Generalized Linear Models (GLM) [58, 60, 62]. Scientist in many different fields are faced with traits that are categorical such as normal and cancerous tissues in case of binary traits and study of stages of cancer progression which can have multiple categories. For situations with categorical phenotypes, researchers have used GLM methodology for data analysis, prediction, and classification. For typical microarrays, due to extensively large number of variabes, the maximum likelihood estimates of parameters will become computationally intensive and sometimes intractable. Additionally, since the sample size is much smaller than number of variabes, the maximum likelihood estimates may have large estimated variances and thus result in poor prediction accuracy. The last but not least, maximization process may not converge to maximum likelihood estimates [94]. It has been proposed that prediction accuracy can be improved by setting the unimportant variabes to zero and thus obtaining

more accurate prediction for the significant variabes [88].

In order to avoid over-fitting and improve model accuracy, models which impose sparsity in terms of variables (genes) are desirable [88]. Least Absolute Shrinkage and Selection Operator (LASSO) is a well-known method for inducing sparseness in the model while highlighting the relevant variables [45, 88, 107]. A Bayesian LASSO method was proposed by [37, 67] in which double exponential prior is used on parameters in order to impose sparsity in the model. However, these procedures may cause over-shrinking of large coefficients due to the relatively light tails of the double exponential prior thus introducing major bias [58]. Using normal-Jeffreys prior which has heavier tails than double exponential distribution, we would be able to shrink small coefficients to zero while minimally shrinking large coefficients and thus obtaining better results. However it has no meaning from an inferential aspect as it leads to an improper posterior [5]. An alternative class of hierarchical priors proposed in [14] that uses Bayesian adaptive Lasso with non-convex penalization. However, it lacks simple analytic form. In [5] authors proposed the Generalized Double Pareto (GDP) prior distribution. The properties of this distribution that makes it appealing include: having a spike at zero alongside student-t like tails, simple analytic form and yielding proper posterior. In addition, it resembles double exponential density in the neighborhood of zero and has heavier tails compared to double exponential remedying unwanted bias resulting from over shrinkage of parameters toward zero [5]. In this article, for the first time- to the best of our knowledge- we integrate GDP prior into the Bayesian generalized linear models framework dealing with

multi-category ordinal response variables to induce sparseness in situations with number of parameters to be predicted far exceeding the number of samples. The model developed can be used to analyze multi-category phenotypes such as progressive stages of cancer. In step one, we derive the fully conditional distributions for all parameters in a multi-level hierarchical model in order to perform the fully Bayesian treatment of the problem.

In the second step, the Markov Chain Monte Carlo (MCMC) method [30, 31] based on Gibbs sampling algorithm is used to estimate all the parameters. This model takes into account the progressive levels of the response variable. We applied and evaluated our model to a publicly available prostate cancer progression dataset [90]. Our study has 3 goals, testing if the model developed can :1)result in a smaller subset of genes with high discriminating power, 2) obtaining high classification accuracy; 3) in addition to above goals, we aim at finding more biologically relevant genes related to phenotype under study compared to competitive methods.

Methods

Let $y_1, y_2, .., y_n$ represent the observed response variables which can take values 1, 2, 3, .., k where k is the number of categories of the ordinal response variable. In addition, let w_{ij} represent the value of variabe 'j' in sample 'i'. In the case of gene expression analysis, gene expression levels are measured for each sample and w_{ij} represents expression level of gene j in i^{th} sample. We implemented GLM for ordinal response in Bayesian framework by utilizing logistic link function and careful introduction of latent variables [1]. In Bayesian framework

joint distribution of all parameters is proportional to likelihood multiplied by prior distributions on the parameters. The generic form of likelihood function for Bayesian Multinomial model was represented in chapter 3 formula (3.1).

As explained in [1], in order to be able to set up the Gibbs sampler and incorporate regression parameters into the model, we introduce $'n'$ independent latent variables $l_1, l_2, .., l_n$ defined as $l_i = \boldsymbol{w}_i^T \boldsymbol{\theta} + e_i$ and $F(e_i) = \frac{1}{1+e^{-e_i}}$ [55]. In this formula, $\boldsymbol{w}_i^T$ is the vector of gene expressions for individual i and $\boldsymbol{\theta} = (\theta_1, .., \theta_p)$ are parameters associated with variable 1 to p respectively. The relationship between response variable and the corresponding latent variables are explained in chapter 3 formula (3.2) [1]. In order to insure that the thresholds are identifiable, following the guidelines of [1] we fix γ_2 at zero and γ_1, and γ_{k+1} are defined according to equation (3.2). In the context of GLM, we use nonlinear link functions to associate the nonlinear, non-continuous response variable to the linear predictor $w_i^T \boldsymbol{\theta}$, [1, 62]. It should be noted that logistic distribution has heavy tails and thus normal distribution does not provide a good approximation and hence we used student-t distribution with v degrees of freedom on latent variables. We treat the degrees of freedom as unknown and estimate it alongside other parameters. Using the relations defined above, the probability of each sample being in category $j (j = 1, 2, .., k)$ is derived in equation (4.1) in which π_{ij} is the probability of sample i being from category j [1].

$$\zeta_{ij} = P(y_i \leq j) = P(l_i \leq \gamma_{j+1}) = P(\boldsymbol{w}_i^T \boldsymbol{\theta} + e_i \leq \gamma_{j+1}) = \frac{1}{1 + e^{-(\gamma_{j+1} - \boldsymbol{w}_i^T \boldsymbol{\theta})}} \;\; ; \;\; \pi_{ij} = \zeta_{ij} - \zeta_{ij-1} \tag{4.1}$$

In this way, the linear predictor, $\boldsymbol{w}_i^T\boldsymbol{\theta}$, is linked to the multi-category response variable y_i. The function that links the linear predictor to the response variable is called a link function and in the multinomial Logistic model, this link function is cumulative distribution of standard Logistic density as defined above [1, 58, 60]

Prior distributions and Bayesian set up

A sparse Bayesian ordinal logistic model was implemented which takes into account the ordinal nature of cancer progression stages and can accommodate large number of variabes. In order to sample l_i from $t_v(\boldsymbol{w}_i^T\boldsymbol{\theta})$, we use the hierarchical model represented in chapter 2 formula (2.3) which is equivalent to sampling from the corresponding t-distribution [62]. This two-level hierarchical form is easier to work with both analytically and computationally compared to the original form of the t distribution [62]. We put independent generalized double Pareto priors on all θ s as represented in formula (2.4) [5]. This prior distribution has a spike at zero and light tails which enables us to incorporate sparsity in terms of number of variabes used in the model [5]. WE put independent GDP prior on all parameters as $\theta_j \sim GDP(\zeta = \frac{\delta}{\rho}, \rho)$ independently. The joint distribution of θs was obtained in chapter 2 formula (2.5).

GDP prior can be represented as a scale mixture of normal distributions leading to computational simplifications that makes Gibbs sampling feasible. The $GDP(\frac{\delta}{\rho}, \rho)$ prior is equivalent to hierarchical representation presented in formula (2.6) [5]. The hyper parameters ρ and δ control the shape of the GDP distribution and thus the amount of shrinkage induced. As δ increases the distribution becomes flatter and variance increases. As ρ increases the tails of distribution

becomes lighter, variance becomes smaller, and the distribution becomes more peaked. Thus, large values of ρ may cause unwanted bias for large signals and stronger shrinkage for noise-like signals while larger values of δ flattens the distribution and we may lose the ability to shrink noise-like signals. As mentioned in [5], by increasing ρ and δ at the same rate the variance remains constant but tails of the distribution becomes lighter converging to Laplace density in limit. This can lead to over-shrinkage of coefficients that are away from zero. In the absence of information on hyper parameters one can either set them to default values $(\rho = \delta = 1)$ or choose a hyper prior distribution and let data speak about the values of these hyper parameters.

We adopt the prior distributions defined in chapter 2 formulas (2.7) and (2.8) for these parameters. The priors on ρ and δ correspond to generalized Pareto priors with location parameter 0, shape parameter 1, and scale parameters c^{-1} and c'^{-1} respectively. As mentioned in the formula (2.7) and formula (2.8), c and c' determine the location of the median of the distribution of parameters ρ and δ . For sampling purposes, we use the transformations presented in formula (2.9) that lead to uniform prior distribution for the new parameters [5]. Defining the parameters as above, the hierarchical representation of the model is as follows. $l_i|\lambda_i, \boldsymbol{\theta} \sim N\left(\boldsymbol{w}_i{}^T\boldsymbol{\theta}, \frac{1}{\Lambda_i}\right)$, $\Lambda_i \sim Gamma\left(\frac{v}{2}, \frac{v}{2}\right)$, $\theta_j \sim N(0, \tau_j)$, $\tau_j \sim Exp\left(\frac{{\lambda_j}^2}{2}\right)$, $\lambda_j \sim Gamma\left(\rho, \delta\right)$, and we put noninformative uniform prior on v. Using the above mixture representation for the parameters and defining the prior distributions, we obtain following conditional posteriors that lead to a straightforward gibbs sampling algorithm.

$$l_i|- \sim DTN(\boldsymbol{w}_i{}^T\boldsymbol{\theta}, \frac{1}{\Lambda_i}) \tag{4.2}$$

In formula (4.2), DTN stands for doubly truncated normal distribution with mean $\boldsymbol{w}_i^T\boldsymbol{\theta}$ and variance $\frac{1}{\Lambda_i}$. For observation 'i' with $y_i = r$, l_i must be sampled from normal distribution defined above truncated between γ_r and γ_{r+1} in each iteration of the algorithm. In each iteration of Gibbs sampling procedure, $\boldsymbol{\theta}$ is sampled from the multivariate normal distribution with mean vector and variance covariance matrix as derived in equation (2.11). The fully conditional posterior distribtion for parameters $[\tau_j^{-1}]_{j=1}^n$ is Inverse Gaussian distribution defined in equation (2.12). In each iteration of the Gibbs sampling, each λ_j and Λ_j is sampled according to equation (2.13) and (2.14) respectively. The fully conditional distributions for v,u_1, and u_2 are represented in equations (2.15)-(2.17). As explained in chapter 2, the fully conditional distributions of v, u_1, and u_2 (formula 2.15-2.17) do not have closed form and thus we adopt the following embedded giddy gibbs sampling to sample from v, ρ, and δ [5, 75]. On a grid of k values $(v_1, v_2, ..., v_k)$ representing all possible values of degrees of freedom we perform the following procedure.

- Calculate the weights as $r_i = \pi(v_i|-)$ according to formula 15.
- Normalize the weights $r_i{}^N = \frac{r_i}{\sum_{i=1}^k r_i}$
- Sample one value from $(v_1, v_2, ..., v_k)$ with probabilities $(r_1{}^N, r_2{}^N, ..., r_k{}^N)$.

On a grid of values in interval $(0, 1)$ we use the same procedure to sample one value from u_1 and u_2 to use in the current iteration of Gibbs sampling. The only difference is that at the end of the procedure we transform u_1 and u_2 back to ρ and

δ using $\rho = \frac{1}{c}[\frac{1}{u_1} - 1]$ and $\delta = \frac{1}{c'}[\frac{1}{u_2} - 1]$ respectively. In the case of multinomial response, we assign independent uniform priors to thresholds and thus the fully conditional distribution for thresholds is uniform distribution and we need to sample them in each iteration of Gibbs sampling alongside other parameters in the model [1]. The fully conditional distribution on thresholds is represented in formula (3.10). Using equation (3.2), and (3.9), the conditional posterior distribution of γ_s can be seen to be $Uniform(\delta_1, \delta_2)$ in which $\delta_1 = max[max_i[l_i|y_i = s-1], \gamma_{s-1}]$ and $\delta_2 = min[min_i[l_i|y_i = s], \gamma_s]$. It should be noted that $I()$ is indicator function and its value is one if its argument is true and is zero otherwise. The Gibbs sampling procedure is explained in the flowchart represented in figure 6.

Dataset and Feature Selection

The method was applied to a published dataset on prostate cancer progression downloaded from Gene Expression Omnibus at NCBI (GSE6099) [90]. The data set contains gene expression values for 20,000 probes and 101 samples corresponding to five prostate cancer progressive stages (subtypes): Benign, prostatic intraepithelial neoplasia (PIN), Proliferative inflammatory atrophy (PIA), localized prostate cancer (PCA), and metastatic prostate cancer (MET) [90]. Since there were only two samples for PIA, we removed these samples from further analysis. Probes with null values in more than 10% of the samples were removed from the data set. For the remaining probes, the null values were imputed by using the mean value of the probe across samples with non-null values. Before applying our model to this data set, for each

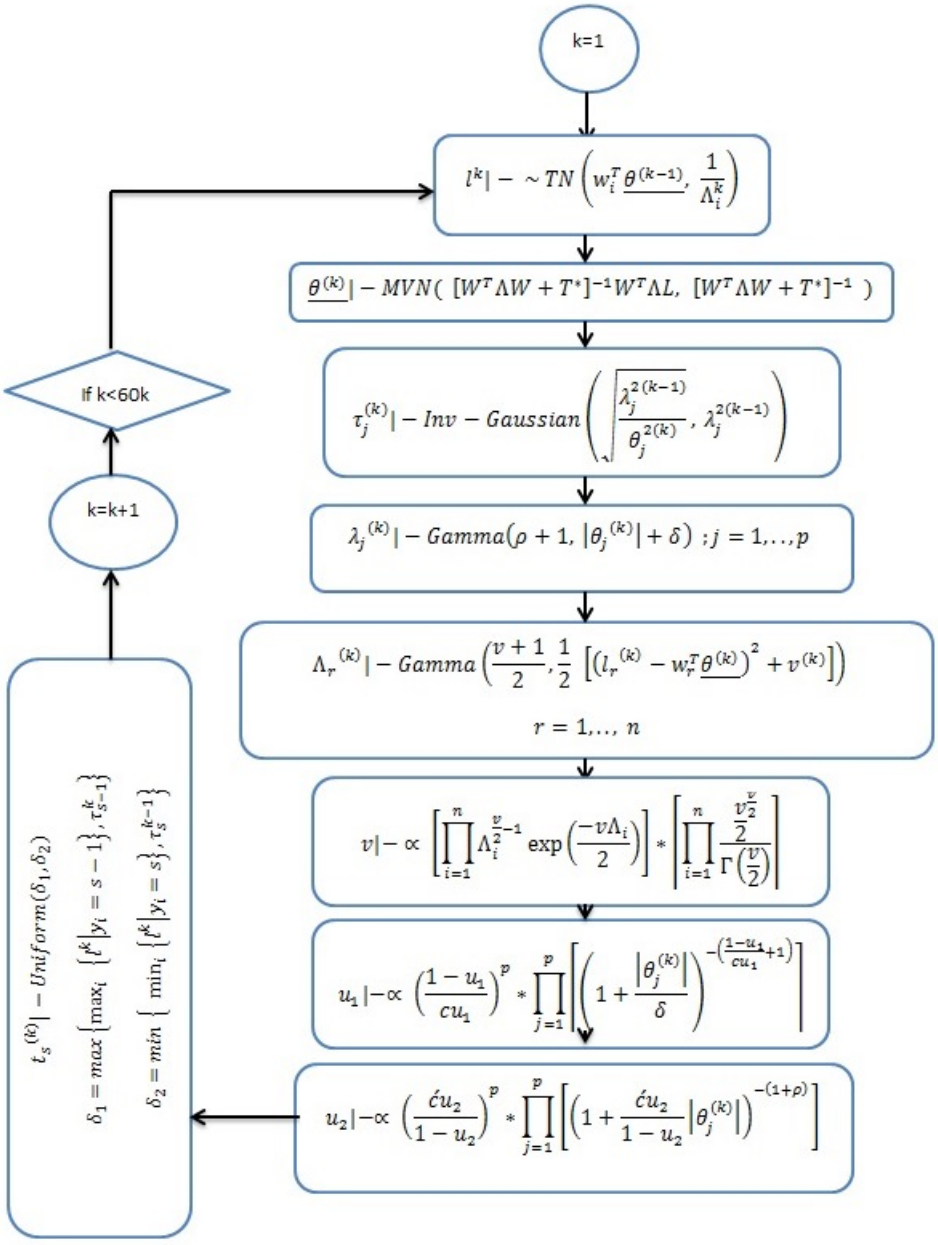

Figure 5: Gibbs sampling procedure for SBGG model.

gene we performed logistic regression for ordinal response.

This method enables us to take into account the ordinal nature of response variable in the analysis and prepare a gene list used as input to the model. Genes were ranked based on the p-value associated with the hypothesis $H_0 : \theta_i = 0$ from

the most significant to least significant. In here θ_i is the parameter associated with gene i. We performed Benjamini and Hochberg FDR correction [9]. An FDR cut-off value of 0.05 resulted in a list of 398 genes. Thus, the input to our model was 398 variabes (genes) for 99 samples corresponding to four different prostate cancer subtypes. The Gibbs sampling algorithm was implemented in R software and the program ran for $60k$ iterations and the first $20k$ was discarded as burn-in.

Simulation and Cross validation procedure

The dataset was randomly divided into training (N=50) and test (N=49) groups such that each group contained an equal number of prostate cancer subtypes Benign, PIN, PCA and MET. Genes were ranked based on posterior mean of parameters and the top 10 or 50 genes obtained from the model were used for classification. In order to make the model more robust, we performed 50 re-samplings on selection of training and test groups and re-ran the model. The average performance of SBGG was compared to three well-known classification methods: Support Vector Machine (SVM), Random Forrest, and Sparse Bayesian Generalized Linear Model obtained imposing double exponential prior (SBGDE) on parameters. SVM was implemented in R software using Kernlab library [42] and Random Forest was implemented in R using randomForest library [51], We implemented the SBGDE according to [57, 58] in R software.

Results

Figure 6 shows an example of the mean of posterior distribution of θ s associated with 398 genes in a single run of SBGG. We used the top 10 or 50 genes to test the classification accuracy of the SBGG on 50 resampled training

and test groups. In order to have a balanced data set, each training and test group had an equal number of the four prostate cancer subtypes: Benign, prostatic intraepithelial neoplasia (PIN), localized prostate cancer (PCA), and metastatic prostate cancer (MET). We found that the average overall classification accuracy of the SBGG was 82.5% and 94.2% when using 10 and 50 marker genes, respectively (Table 8). Three well known classification methods namely,

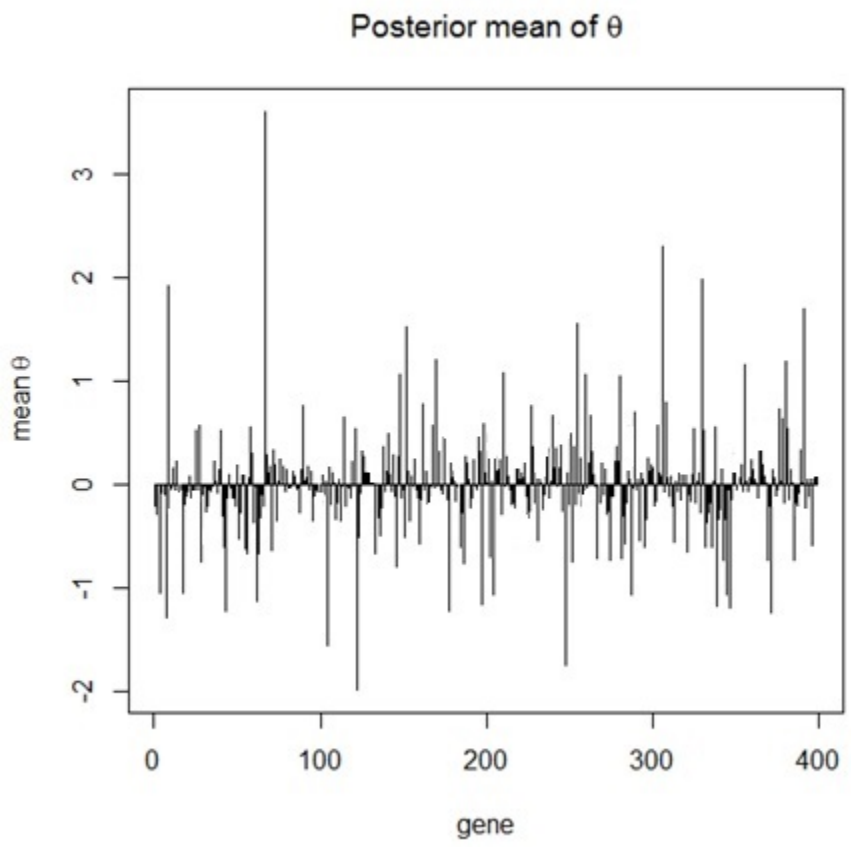

Figure 6: Posterior mean of θ asociated with gene 1 to gene 398 obtained from Gibbs Sampling.

Random Forest [12], Support Vector Machine(SVM) [43], and SBGDE [56] were implemented and the classification results were compared to our model. We found that the overall accuracy of SBGG was substantially better than SVM and SBGDE

Table 8: Overall average accuracy of SBGLM, SVM and Random Forest using 10 and 50 marker genes.

Model	P-10	P-50
SBGG	82.5(0.55)	94.9(3.08)
SBGDE	80.4(0.06)	82.3 (0.063)
SVM	53.6(2.7)	67(3.04)
Random Forest	83(1.6)	84.6(2)

when using top 10 and top 50 genes for classification. Table 8, shows that when using 10 marker genes, Random Forest performs slightly better than SBGG (0.5% higher average classification accuracy). However when using 50 marker genes, SBGG achieves measurably higher classification accuracy than Random Forest.

It is important to note that the feature selection for SVM and Random Forests was based on the p-values of the ordinal linear regression model (top 10 and top 50 from the 398 input genes). These results indicate that a small subset of the 398 input genes is better for predicting prostate cancer progression. Next, we examined the performance of SBGG with regard to classifying the different subtypes of prostate cancer in comparison to SVM, Random Forrest, and SBGDE (Table 9, and Table 10). When using 10 marker genes, SBGG classified all four subtypes of prostate cancer more accurately than SVM, and it outperformed SBGG for classifying PIN, PCA, and MET. It also performed better than Random Forrest for classifying PIN, and PCA. Interestingly however, when using 50 marker genes, SBGG performed substantially better than SVM in classifying all tumor subtypes and outperformed SBGDE in classifying PIN and PCA samples. SBGDE performed slightly better that SBGG for classifying benign samples using 50 marker genes. Comparison of classification results to Random Forrest shows that

Table 9: Average classification accuracy of prostate cancer subtypes in the test group using SBGG, SBGDE, SVM, and Random Forrest models with 10 marker genes.

Sample Type	SBGG	SBGDE	SVM	Random Forest
Benign	89.4(6.1)	95.1(6)	84.4(5.3)	91.1(4.5)
PIN	62.5(1.6)	61.7(2.8)	9 (7.2)	61.4(1.9)
PCA	98.7(0.7)	86.9(1.1)	37.4(9)	86.7(2.1)
MET	59.4(2.06)	56(3.2)	55.3(1.2)	82.8(7.3)

Table 10: Average classification accuracy of prostate cancer subtypes in the test group using SBGG, SBGDE, SVM, and Random Forrest models with 50 marker genes.

Sample Type	SBGG	SBGDE	SVM	Random Forest
Benign	95.4(3.07)	99.6(1.9)	90.1(1.7)	96.8(1.3)
PIN	80.6(0.08)	53.4(1.4)	38.2(8.2)	52(1.1)
PCA	98.9(1.9)	65.4(7.2)	45.8(6.2)	84.8(5.4)
MET	96.8(4.6)	95.4(6.3)	81.8(1.6)	83.6(7.09)

SBGG outperforms Random Forest in all categories except benign for which Random Forest achieves slightly better accuracy.

These results indicate that the performance of SBGG is comparable to Random Forrest in classifying subtypes of prostate cancer and slightly better, although the results for both methods are sensitive to the number of selected marker genes. Since the results of SBGG were comparable to Random Forrest, we next asked if SBGG gene rankings were more or less relevant to the biological mechanisms associated with prostate cancer progression. As a first step in evaluating the biological relevance for the top ranked genes in the models, we used a literature based method called GeneSet Cohesion Analysis Tool (GCAT) [96]. GCAT is a web-based tool that determines the functional coherence p-values of gene sets based on latent semantic analysis of Medline abstracts [96]. Table 11 shows the average GCAT literature derived p-values (LPv) for the top

Table 11: Literature based functional cohesion p-values (LPv) of the top 100 genes obtained from three different models.

Sample Type	Lpv
SBGG	2.0E-4(1.7E-5)
SBGDE	0.007(0.001)
Classical Logistic Regression	0.047
Random Forest	0.131(0.07)

100 genes obtained from 50 runs of SBGG, Random Forrest, SBGDE as well as the top 100 genes based on the p-value rank ordering of single gene analysis using ordinal logistic regression. We found that on average, SBGG produced more functionally cohesive gene list (LPv = 2.0E-4) compared to SBGDE (LPv=0.007), classical logistic regression (LPv= to 0.047) and Random Forest (LPv=0.131). Notably, 100% of the runs had smaller LPv than 0.047, produced by single gene analysis using classical logistic regression p-value ranking. The Literature p-value for the median run was 4.50E-06 compared to 1.90E-04 for SBGDE and 2.85E-02 For Random Forest.

Discussion

Complex disease and biological processes are polygenic and caused by interaction of multiple gene products. Hence, single gene analysis approaches utility is limited in understanding complex molecular mechanisms and identification of genetics biomarkers for classification of diseases [21, 31, 73]. Additionally, large number of genes collected compared to small number of samples in microarray experiments makes the data analysis, feature extraction and prediction quite challenging. In the situations that we are faced with fat datasets with $p >> n$, highly regularized approaches, such as penalized

regression models, are needed to identify non-zero coefficients, enhance model predictability and avoid over-fitting [70]. The Bayesian Lasso which is a Bayesian version of L_1 penalized regression is such one of the most popular techniques.

However, this procedure inherits the problem of over-shrinking of large coefficients due to the relatively light tails of the double exponential prior and may miss some of the important factors in the model. Recently, the Generalized Double Pareto (GDP) prior distribution was proposed as an alternative to induce sparseness in situations when we are faced with large number of variabes compared to sample size [5]. The authors applied the proposed method in the normal linear regression model framework. This prior has a simple analytic form, yields a proper posterior and possesses appealing properties, including a spike at zero, Student t-like tails, and a simple characterization as a scale mixture of normals leading to a straightforward Gibbs sampler for posterior inferences that makes Bayesian shrinkage estimation and regularization feasible [5]. Utilizing this prior in a more general framework of generalized linear models, we presented a Bayesian hierarchical model to simultaneously fit and estimate all variabes in p » n situations. While shrinking small effects toward zero and producing sparse solutions, the over shrinkage problem caused by using light-tailed priors would be remedied by the heavier tails obtained via mixing over the hyper parameters. We developed a sparse Bayesian multinomial model and evaluated its performance using prostate cancer gene expression data. We employ latent variables which are distributed as student-t distribution to account for heavy tails of logistic distribution to specialize the model to a regression model.

We fit the model in a fully Bayesian approach, employing the MCMC algorithm to generate posterior samples from the joint posterior distribution, which can be used to make various posterior inferences. The Bayesian algorithm developed treats all parameters as unknown including hyper parameters associated with the GDP hierarchical representation and generates their posterior samples alongside other parameters. We used the model to do prediction of tumor type on the test dataset. The Bayesian set up enables us to assign the tumors to one of the categories in a coherent way. In addition, we obtain the probability of each tumor belonging to one of the categories that is much more meaningful than hard rules of assignment that use 0 or 1 to correspond to being in a special category or not. Also, we use small number of genes to do the prediction which simplifies the experimental procedure.

We compared the model performance to three well known models: random forests, SVM, and SBGDE. The average classification accuracy of SBGG using 10 marker genes was higher than SBGDE and SVM and was only 0.5% lower that Random Forest. However, when using 50 marker genes it outperforms all the three other methods (Table 9 and Table 10). We found that the SBGG classification accuracy of prostate cancer subtypes were comparable to Random Forrest when using 10 marker genes for classification and it outperforms Random Forest in 3 out of four categories when we used 50 marker genes. Additionally, it outperforms SBGDE in 3 out of 4 categories when using 10 marker genes for classification and 3 out of 4 categories when using 50 marker genes. Furthermore, SBGG identified more biologically relevant gene sets (Table 11).

Based on these results, we posit that SBGG may be a better approach to simultaneously identify marker genes for classifications as well as gaining insights into the molecular mechanisms of the phenotype under investigation. Interestingly, using fewer genes, SBGG had very good discrimination performance for classifying benign (89.4% accuracy) versus PCA (98.7% accuracy), but the model discrimination was weaker for PIN and MET (Table 2). These results are consistent with the previous observation that PIN and PCA share markedly similar expression signatures [90]. Random Forests are an ensemble method for classification that has been shown to have good performance in many bioinformatics applications. However, Random Forrest is prone to over-fitting in datasets with noisy classification tasks. In addition, it is very hard to interpret the classifications made by Random Forests. Furthermore, if data contain categorical variables with different number of levels, Random Forest favors variables with more levels, making the variable importance measures unreliable [11].

Conclusion

It is important to note that the classification accuracy of all three models were compared using a selected set of 398 genes which were obtained based on p-value of single gene analysis using an ordinal regression model. Hence, this biases the initial gene selection process. It is possible that some biologically relevant genes to the prostate cancer progression might have been missed by this analysis due to low signal. One way to perform an initial gene selection could be to consider gene pathway information as described previously by others [81].

Chapter 5

Evaluation of literature aided variable selection in classification and feature prioritization

One of the fundamental tasks in biomedical research is analysis of gene-disease association. A major source to achieve this goal is examination of Microarray data due to relatively low cost. Since a large number of genes are typically involved with possible interactions, the association of genes with diseases is very complicated. Methods employed for gene ranking(gene prioritization) are based on statistical or knowledge based approaches to find genes most likely associated with a given disease [41, 76]. Technical and biological variability are the main causes of the noisy nature of gene expression data which makes their analysis complex. This makes prediction methods aimed at obtaining gene disease associations often less that adequate [68]. Even with reliable gene expression data, statistical analysis of that data remains largely challenging [97]. Normally, gene expression data are ranked by the strength of the signal compared across disease and control tissues. Several studies have aimed at comparing the results of multiple studies of the same genes and have found little correlation between results [29]. Several factors contribute to this issue including individual variation, different gene activation cycles, and variations in protocols used to prepare the tissues [25]. Biomedical literature can be used as an informative way to obtain the relevance of genes to different diseases. One caveat of this method is that many genes from poorly studied organisms are not

well defined in literature. Additionally, a comprehensive summarization of the literature attached to genes of different organisms is a challenging task [28].

Another fundamental task in biomedical research is regarding cancer classification. Cancer research is one of the major areas in medical fields. It is clear that prediction of different tumor types with high accuracy offers the advantage of providing better treatment and reducing toxicity in the patients. In the past, cancer classification has majorly been based on morphological and clinical experiments. It has been reported that these methods have limited diagnostic ability due to their several limitations [6]. Gene expression data can provide the key for addressing the fundamental issues relating to cancer diagnostics and drug discovery [54]. The main two aspects of cancer classification is classification accuracy and ability to reveal meaningful gene information. The high dimensionality of gene expression data(tens of thousands of genes) compared to very small sample sizes (usually below 100) makes cancer classification a daunting task. Another issue is that most genes are irrelevant to the cancer classification task at hand. Additionally, it is very common for highly differentially expressed genes not to be relevant to the disease under study [35, 54]. Some researchers proposed to perform gene selection prior to applying cancer classification methods. This step will help in reducing the data size and thus improving the running time of the classification algorithm.

Additionally, another issue concerning cancer classification is statistical significance versus biological relevance of cancer classifiers [54]. Most cancer classification methods available are from statistical and machine learning area

ranging from parametric methods such as generalized linear models to nonparametric methods such as nearest neighbor analysis and support vector machines [54]. One common aspect of most of these methods used in cancer classification is that the primary goal of authors is classification accuracy and they are less concerned with biological relevance [54]. This is due to the fact that most classifiers are built based on strong signals or differentially expressed genes across sample types. However, the majority of cancer related genes might not be highly differentially expressed and thus the classifiers obtained based on differentially expressed genes across different sample types might not reveal biologically relevant genes. Gene selection methods that are based on signal strength and differentially expressed genes, choose genes that are highly differentially expressed across different tissue types, i.e. cancer and normal tissue that might not necessarily be related to cancer.

GeneIndexer

Literature information can be used to select biologically relevant genes from gene expression data in order to build cancer classifiers. This method can potentially be very helpful in order to highlight the biological relevance of classifiers built based on gene expression data. Here we investigate a very famous gene ranking method based on biological literature called Gene Indexer [39]. Gene Indexer utilizes Latent Semantic Indexing (LSI), a vector space model for information retrieval, to automatically identify conceptual gene-gene and gene-disease relationships from titles and abstracts of MEDLINE citations [39]. This method has proved to identify gene-keyword and gene-gene relationships

with high average accuracy. Additionally, this method is able to obtain implicit relationships between genes and keywords that proves very helpful in identifying conceptual relationships [39].

Front-end Gene Selection Using Gene Indexer

We used Gene Indexer to obtain the input gene list to our models and other classifiers built in previous chapters. For the leukemia data set of Golub et al used in chapter 2 [32], the “leukemia” keyword was used to obtain literature correlation of genes with leukemia. Genes were ranked based on their literature correlation and the top 500 genes were used as input to the models. This data set consisted of 72 samples obtained from Golub et al. [32]. In the Golub data set, the bone marrow or peripheral blood samples from 72 patients with either acute lymphoblastic leukemia (ALL) or acute myeloid leukemia (AML) were collected. The gene expression levels for 7129 human genes were measured for this cohort. We then extract these gene expressions from the data set and use them as input to the classifiers. The data set was randomly divided into train and test data sets of size 37 and 35 respectively. The training data sets contained 24 ALL and 13 AML samples and test data set contains 23 ALL and 12 AML samples. Genes were ranked based on the posterior mean of parameters and the top 10 genes obtained from the model were used for classification. In order to make the model more robust, this process is repeated 50 times and the average classification accuracy on training and test samples are reported.

The second data set used was the prostate cancer progression dataset downloaded from Gene Expression Omnibus at NCBI (GSE6099) [90]. The data

set contains gene expression values for 20,000 probes and 99 samples corresponding to four prostate cancer progressive stages (subtypes): Benign, prostatic intraepithelial neoplasia (PIN), localized prostate cancer (PCA), and metastatic prostate cancer (MET) [90]. For this data set, we used the “prostate cancer” keyword to obtain literature correlation of genes with “prostate cancer”. The gene expressions corresponding to the top 500 genes were extracted from the data set and used as input to the classifiers. The dataset was randomly divided into training (N=50) and test (N=49) groups such that each group contained an equal number of prostate cancer subtypes Benign, PIN, PCA and MET. Genes were ranked based on posterior mean of the parameters and the top 10 genes obtained from the model were used for classification. In order to make the model more robust we performed 50 re-samplings on selection of training and test groups and re-ran the model. The average performance of SBGG was compared to three well-known classification methods: Support Vector Machine (SVM) [85], Random Forrest [40], and Sparse Bayesian Generalized Linear Model obtained imposing double exponential prior (SBGDE) on parameters. SVM was implemented in R software using Kernlab library [42] and Random Forest was implemented in R using randomForest library [51], We implemented the SBGDE according to [57] in R software.

Classifiers based on GeneIndexer and signal strength input gene lists

Table 12 represents the classification accuracy obtained for the binary classifiers built for the leukemia data set. In this table, SBGG represents Bayesian Generalized model built using Generalized double Pareto prior, “SBGDE” is the

Table 12: Average Classification accuracy, Sensitivity, and specificity for test groups.

Model	Gene Selection	Accuracy	Sensitivity	Specificity
SBGG	Diff expression	94.1(3.05)	0.95(0.04)	0.93(0.029)
SBGG	GeneIndexer	87.3(6.6)	0.9(0.054)	0.79(0.073)
SBGDE	Diff expression	91.2(10.8)	0.95(0.12)	0.86(0.098)
SBGDE	GeneIndexer	81(7.09)	0.85(0.052)	0.71(0.09)
SVM	Diff expression	63(13)	0.7(0.16)	0.5(0.11)
SVM	GeneIndexer	69.1(8)	0.75(0.03)	0.57(0.11)
Random Forest	Diff expression	93(4.2)	0.9(0.036)	0.93(0.048)
Random Forest	GeneIndexer	88(2.7)	0.85(0.019)	0.93(0.035)

Table 13: Average classification accuracy of prostate cancer subtypes in the test group using SBGG, SBGDE, SVM, and Random Forest.

Sample Type	Gene Selection	SBGG	SBGDE	SVM	Random Forest
Benign	Diff expression	89.4(6.1)	95.1(6)	84.4(5.3)	91.1(4.5)
Benign	GeneIndexer	72(2.9)	51(10.04)	83(1.2)	88(1.08)
PIN	Diff expression	62.5(1.6)	61.7(2.8)	9.0(7.2)	61.4(1.9)
PIN	GeneIndexer	73(6.3)	66 (5.1)	13(8.6)	60.3(2.3)
PCA	Diff expression	98.7(0.7)	86.9(1.1)	37.4(9)	86.7(2.1)
PCA	GeneIndexer	87(1.4)	69(2.3)	37.4(5.2)	84.7(1.7)
MET	Diff expression	59.4(2.06)	56 (3.2)	55.3(1.2)	82.8(7.3)
MET	GeneIndexer	48(7.2)	44 (2.5)	37 (6)	54(9.3)

classifier build using Double exponential prior, SVM is the classifier based on support vector machines and the last classifier is build based on Random Forests. Table 13 represent the classification accuracy of multi-category classifiers built on prostate cancer progression data set. The values in this table show the average classification accuracy on the test data set for multi-category GDP model (SBGG), sparse Bayesian Generalized linear model developed using double exponential prior (SBGE), Support Vector Machine model (SVM), and Random Forest model. The input gene selection is done in two different scenarios: the p-value rank ordering of single gene analysis using ordinal logistic regression and top Genes obtained from GeneIndexer.

Results and Discussion

Table 12 represents the classification accuracy, sensitivity, and specificity measures for the sparse Bayesian Generalized linear model based on generalized double Pareto prior (SBGG), sparse Bayesian Generalized linear model based on double exponential prior (called SBGDE), Support Vector Machine model (SVM), and the model based on Random Forests (Random Forest). As we can see the classifiers built based on input gene list obtained from differential expression p-value ranking obtain higher classification accuracy, sensitivity, and specificity than the rankings based on GeneIndexer in the majority of the classifiers. In the GDP model, the classification accuracy and sensitivity and specificity is close for the two paradigms. For the SVM model, the support vector machine obtains higher classification accuracy, sensitivity, and specificity when using Gene Indexer as input gene selection. For the Random Forest model the classification accuracy in both paradigms are very close to each other. All in all the classifiers built on highly differentially expressed genes obtain higher classification accuracy compared to the Gene Indexer input gene list counterparts.

Table 13 represents the average classification accuracy for the multi-category classifiers built on prostate cancer progression data sets. The four models developed are the Sparse Bayesian Generalized Linear model based on generalized double Pareto prior(SBGG), Sparse Bayesian Generalized Linear model based on double exponential prior (SBGDE), the Support Vector Machine model, and Random Forest model. For the Benign, sample type the models that

are built based on input gene list obtained using single gene analysis using ordinal logistic regression outperform the counterparts built upon the input gene list obtained from Gene Indexer. For the PIN sample type, the models based on Gene Indexer outperform their counter parts for SBGG, SBGDE, SVM and the classification accuracy is a bit lower for the Random Forest model but comparable for the two scenarios. For the PCA sample type, the models based on ordinal logistic regression input gene list outperform the gene Indexer input gene list models in 3 out of four models namely, SBGG, SBGDE, and Random Forest and it has slightly lower classification accuracy for the SVM model. For the MET sample types, the models based on ordinal logistic regression input gene lists outperform all the models built based on Gene Indexer input gene list.

Even though the classifiers built based on the GeneIndexer gene selection paradigm come close in classification accuracy, sensitivity, and specificity to their counterparts for some of the classifiers, in majority of the classifiers, the gene selection based on signal strength results in higher accuracy, sensitivity, and specificity. One of the main reason for this phenomena is that the majority of cancer related genes are not highly differentially expressed across different tissue types which lowers their ability to be highly powerful predictors. On the other hand, gene selection methods that are based on signal strength and differentially expressed genes choose genes that are highly differentially expressed across different tissue types, i.e. cancer and normal tissue. However, most of these genes can be housekeeping genes or genes that are differentially expressed during different cell cycles that might not be necessarily related to cancer.

Thus, the classifiers obtained based on differentially expressed genes across different sample types might not reveal biologically relevant genes. Thus, even though these models obtain higher classification accuracy, they suffer from the fact that they do not obtain comprehensive biological relevance in the set of predictors obtained. Reduction in uncertainty due to technical and biological variability through more comprehensive and unified tissue preparation and experimentation can bridge the gap between classification accuracy and biological relevance in cancer analytics and obtain more informative machine learning models in cancer diagnostic and therapeutics. Another important issue is that using an input gene list solely based on current literature hugely biases the downstream results due to the fact that it ignores the signals coming out of the experiment. Ideally, we would want signals coming from the experiment have greater weight but have a technique to prioritize and tune these signals based on biological information from literature.

In chapter 6, we develop a literature aided Sparse Bayesian Generalized Linear Model (LSBGG) which can incorporate literature information in the form of prior knowledge in tuning the prior distribution imposed on parameters. This way we are able to take into account the biological relevance of markers to guide the amount of shrinkage imposed in the model. Thus, we would be able to potentially bridge the gap between classification accuracy and biological relevance and obtain a set of markers which have high diagnostic capability based on more biological relevance to phenotype under study.

Chapter 6

Development of Literature Aided Bayesian Sparse Generalized Linear Model- Bridging Classification Accuracy and Biological Relevance

Abstract

Gene expression profiling has two major limitations that offset their statistical performance. Firstly, large numbers of variables are assessed compared to relatively small sample sizes. Secondly, identification of a set of biologically relevant markers with high predictive power remains difficult. Several machine learning algorithms have been used for cancer classification which are geared toward obtaining high classification accuracy and do not take into account the biological relevance of the markers obtained. Thus, in the majority of applications, markers found do not convey meaningful biological information and are merely good classifiers. A machine learning schema that is able to bridge classification accuracy and biological relevance will be of high merit to the community and can potentially result in deeper understanding of the mechanisms involved. In this study, we developed a Literature aided Sparse Bayesian Generalized Linear model which utilizes Generalized Double Pareto prior (LSBGG) to induce shrinkage in terms of the number of variabes. Additionally, instead of uninformed hyper parameters for the prior distributions, we adopt a literature informed approach to adjust the hyper parameters based on a marker's biological relevance to the phenotype under study. This will aid us in controlling shrinkage imposed on genes based on their biological relevance.

The method was applied to the leukemia data set of Golub et al. (1999). The data set was randomly divided into train and test samples of sizes 37 and 35 respectively and classification performance on the test group was evaluated. We performed 50 resamplings on the training and test groups. The top 500 highly differentially expressed genes obtained were used for the modeling step. Using the top 10 genes obtained from our model, we were able to achieve 96% average classification accuracy. Additionally average sensitivity and specificity of 97% and 93% was achieved across the 50 runs. The model without incorporation of biological information (SBGG) achieves averages of 87%, 92%, and 83% accuracy, sensitivity, and specificity respectively. Additionally, There were 41 genes common in all 50 runs for the literature aided model compared to only 6 common genes for the model with uninformed choice of hyper parameters. This results suggest that the literature model results in more consistent results with significantly higher biological relevance. Taken together, these results suggest that the literature informed Sparse Bayesian Generalized Linear Model applied to leukemia data sets allows for better subclass prediction based on more functionally relevant gene sets.

Introduction

The ability of cost-efficient gene expression analysis brings the possibility of studying the relationship between complex traits or diseases and genes across the entire human genome. Microarray studies usually include tens of thousands of genes assayed for a few number of experimental units [19]. The widely applied methods for analyzing gene expression data are based on single marker analysis

in which the association of each gene to the traits are analyzed independently [21, 66, 87, 91]. However, these methods are not capable of capturing variance present in the polygenic phenotypes arising from several variabes in a complex system [7]. Due to this limitation, simultaneous analysis of genes has received more attention recently [53, 94, 98]. There are two main challenges in developing methods for simultaneous analysis of genes in gene expression data. Firstly, the large disparity between the number of variables and the number of observations in the model reduces the accuracy of the prediction and selection.

Another challenge in gene expression analysis is identification of a set of biologically relevant markers with high predictive power. For example several machine learning algorithms have been used for cancer classification which are geared toward obtaining highest classification accuracy and do not take into account the biological relevance of the markers obtained. Thus, in majority of applications markers found do not convey meaningful biological information and are merely good classifiers. Thus, a machine learning schema that is able to bridge classification accuracy and biological relevance will be of high merit to the community and can potentially result in deeper understanding of the mechanisms involved. This is especially crucial when the goal of data analysis is the identification of highly accurate but small panels of biomarkers with potential clinical utility [82]. For large-scale problems with p»n, in linear regression, there is a mass of literature in both frequentist and Bayesian framework. Frequentists methods impose constraints on the size of the coefficients known as penalization.

The most popular one is the L_1 norm penalty called Lasso introduced by Tibshirani [88].

A commonly used method for imposing shrinkage in the Bayesian framework is achieved by imposing prior distributions centered around zero [5, 7, 34, 37, 49, 67]. In the majority of these methods the rate of shrinkagemay not be desirable due to the fact that the same rate of shrinkage is imposed on all parameters and all the coefficients are shrunk with the same rate. Literature based association of markers to the trait are not taken into account by these models which does not allow the capture of the comprehensive picture of disease phenotype. Due to this limitation, only partial information is gained from the biological stand point. Thus, a machine learning schema that is able to bridge classification accuracy and biological relevance can potentially result in deeper understanding of mechanisms involved.

A more desirable penalization method would be one that incorporates literature information into the prior distribution by imposing different rates of shrinkage, is obtained by adjusting the shape of the prior distribution on parameters. In this study, we developed a literature aided Bayesian Shrinkage Generalized Linear model which utilizes Generalized Double Pareto prior (LSBGG) to induce shrinkage in terms of the number of variabes. Instead of uninformed hyper parameters for the prior distributions, we adopt a literature informed approach to adjust the hyper parameters based on the marker's biological relevance to the phenotype under study. This will aid us in guiding shrinkage imposed on genes based on their biological relevance. This way we are

able to impose different degrees of shrinkage adjusted based on literature association of marker to the phenotype under study. We applied our method to the leukemia data set of Golub et al. [32].

Methods

Let $y_1, y_2, ..., y_n$ represent the observed response variables in which $'n'$ is the number of observations (samples). Here, y_i can take on 0 or 1 if for example the sample is normal or cancer respectively. In the case of gene expression analysis, gene expression levels are measured for each sample and we let w_{ij} represent the expression level of gene j in the i^{th} sample. We use a logistic link function introduced in formula (2.1) and (2.2) to associate the probability of belonging to one of the categories to the linear combination of variabes. As explained in [1], in order to be able to set up the Gibbs sampler, we introduce $'n'$ independent latent variables $l_1, l_2, ..., l_n$ with $l_i \sim t_v(\boldsymbol{w}_i{}^T\boldsymbol{\theta})$ where $l_i \geq 0$ if $y_i = 1$ and $l_i < 0$ if $y_i = 0$. This approach connects the logistic regression for y_i to a linear regression model for the latent variable l_i, [1]. It should be noted that the logistic distribution has heavy tails and thus the normal distribution does not provide a good approximation. Hence, we used student-t distributions with v degrees of freedom on latent variables, $l_i \sim t_v(\boldsymbol{w}_i{}^T\boldsymbol{\theta})$ [61]. We fix the degrees of freedom at 9 as the t-distribution with 9 degrees of freedom closely approximates logistic distribution [61].

Prior Distributions and Hyper Parameter Settings

In order to sample l_i from $t_v(\boldsymbol{w}_i{}^T\boldsymbol{\theta})$, we use the hierarchical model of formula (2.3) which is equivalent to sampling from the corresponding t-distribution. This

two-level hierarchical form is easier to work with both analytically and computationally compared to the original form of the t distribution [34]. This two level hierarchical distribution enables us to obtain closed forms for fully conditional posterior distributions on parameters. We put independent generalized double Pareto priors on all θ s as defined in formula (2.4) [5]. As mentioned in chapter 2, GDP prior can be represented as a scale mixture of normal distributions leading to computational simplifications that makes Gibbs sampling feasible. The $GDP\left(\frac{\delta_j}{\rho_j}, \rho_j\right)$ prior is equivalent to the following hierarchical representation [5].

$$\theta_j|\tau_j \sim N(0, \tau_j);\ \tau_j \sim Exp\left(\frac{{\lambda_j}^2}{2}\right);\ \lambda_j \sim Gamma(\rho_j, \delta_j) \tag{6.1}$$

The hyper parameters ρ_j and δ_j control the shape of the GDP distribution and thus the amount of shrinkage induced. As δ_j increases the distribution becomes flatter and variance increases. As ρ_j increases the tails of distribution becomes lighter, variance becomes smaller, and the distribution becomes more peaked. Thus, large values of ρ_j may cause unwanted bias for large signals and stronger shrinkage for noise-like signals while larger values of δ_j flattens the distribution and we may lose the ability to shrink noise-like signals. Here, we use literature information to guide the choice of hyper parameters ρ_j and δ_j. We divided the genes into 5 bins using quantiles of literature correlation of genes to 5 groups (0-20 percentile, 21-40 percentile, 41-60 percentile, 61-80 percentile, and 81-100 percentile). Genes on the higher percentiles have higher correlation to the cancer query. We set $\rho_j = 1$ if the literature correlation for a gene is located in the highest

bin and set $\rho_j = 1.4, 1.6, 1.8, 2$ if the gene is located in $61-80$, $40-60$, $21-40$, and $0-20$ percentiles respectively and we set $\delta_j = 1$. This way we are able to incorporate biological knowledge and control the amount of shrinkage imposed on each gene based on the association of the gene to the phenotype under study. Figure 7 represents this choice of prior distributions on parameters. Figure 8

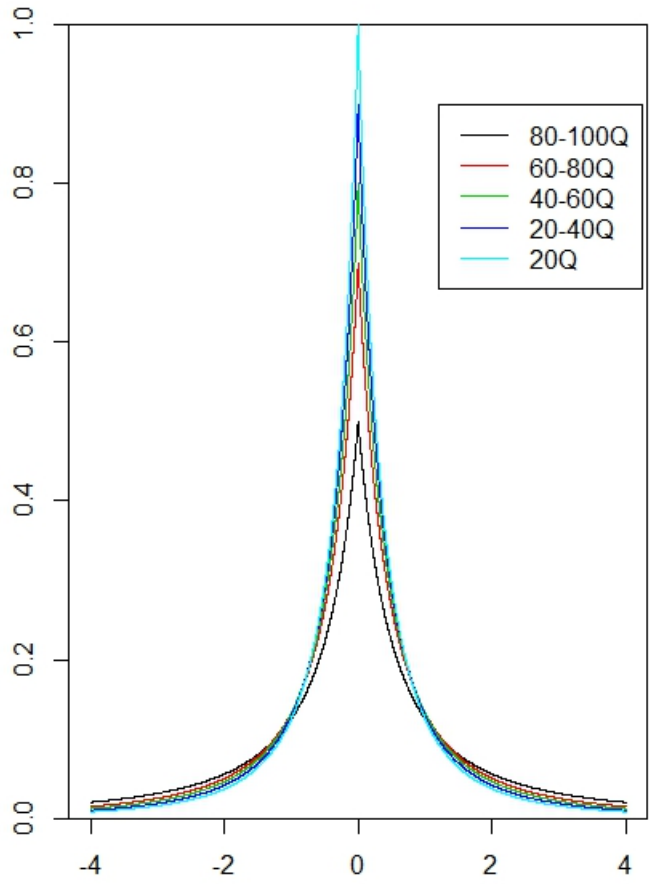

Figure 7: Literature based GDP Prior.

shows a zoomed-in view of tail behavior of these distribution in order to demonstrate the tail behaviors more clearly.

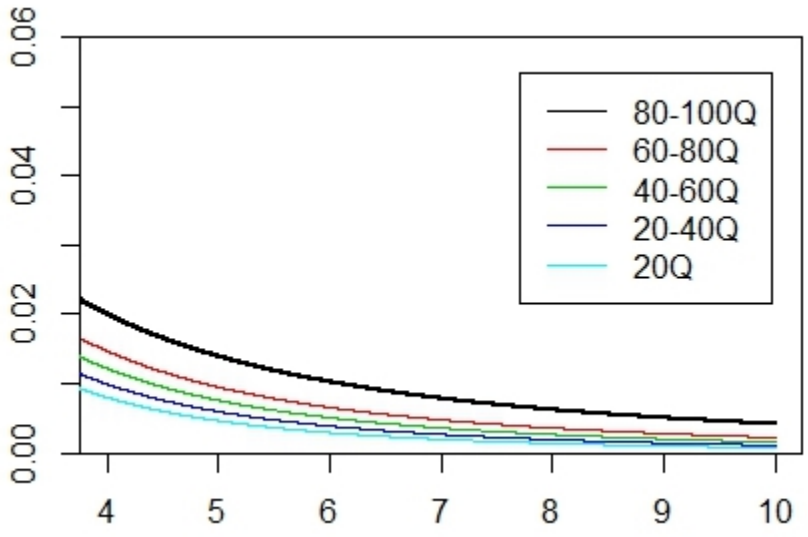

Figure 8: Tail behavior for literature based GDP Prior.

Fully Conditional Posterior Distributions

Defining the parameters as above, the hierarchical representation of the model is as follows. $l_i|\lambda_i, \boldsymbol{\theta} \sim N\left(\boldsymbol{w}_i{}^T\boldsymbol{\theta}, \frac{1}{\Lambda_i}\right)$, $\Lambda_i \sim Gamma\left(\frac{v}{2}, \frac{v}{2}\right)$, $\theta_j \sim N\left(0, \tau_j\right)$, $\tau_j \sim Exp\left(\frac{{\lambda_j}^2}{2}\right)$, $\lambda_j \sim Gamma\left(\rho_j, \delta_j\right)$. Using the above mixture representation for the parameters and defining the prior distributions, we obtain conditional posteriors as derived in formulas (2.10)- (2.12) for l_i, $\boldsymbol{\theta}$, and τ_j^{-1} respectively leading to straight forward Gibbs sampling algorithm. Each λ_j is sampled according to the fully conditional distribution defined in formula (6.2) and Λ_j is sampled according to equation (2.14).

$$\lambda_j|- \sim Gamma\left(\rho_j + 1, |\theta_j| + \delta_j\right); j = 1, .., p \tag{6.2}$$

The Gibbs sampling algorithm is concisely represented in the Figure 9 as a flowchart.

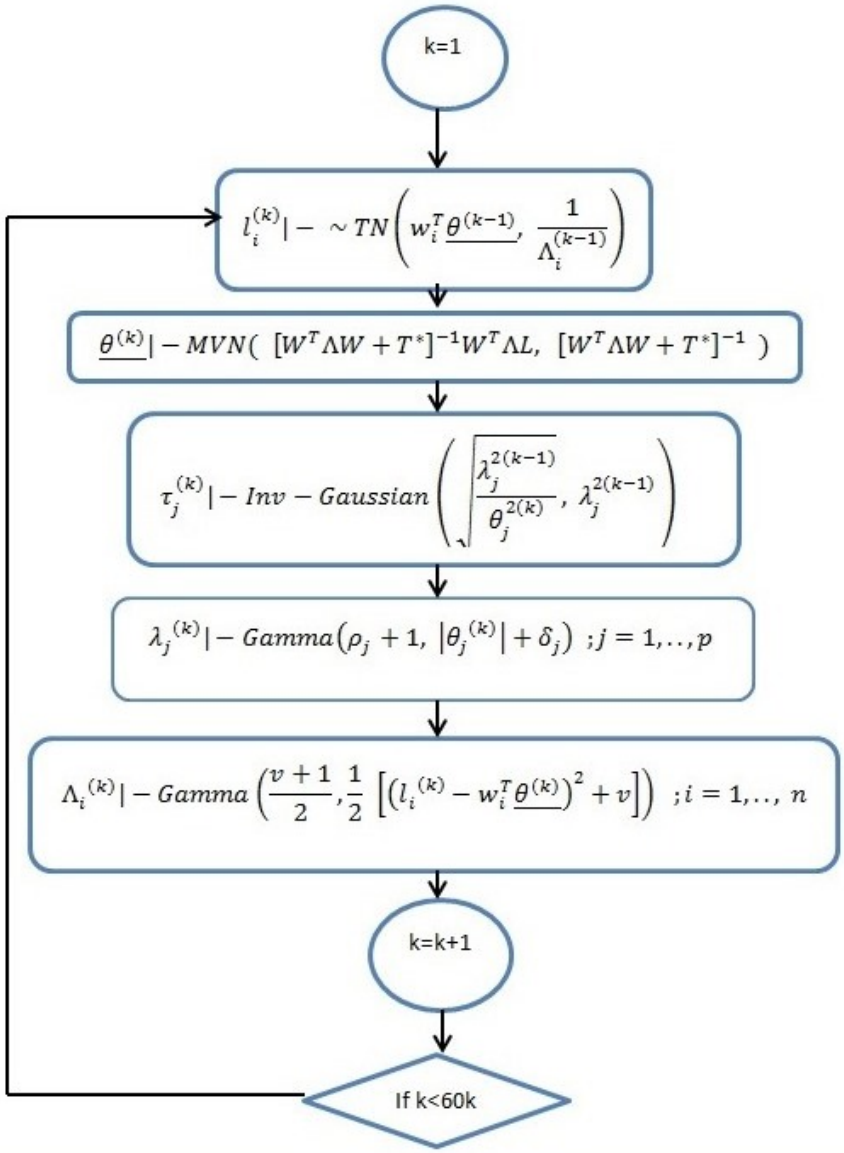

Figure 9: LSBGG. flow chart representing Gibbs sampling algorithm.

Application

We apply our model to the leukemia data set of Golub et al. [32]. In the Golub data set, the bone marrow or peripheral blood samples from 72 patients with either acute lymphoblastic leukemia (ALL) or acute myeloid leukemia (AML) were collected. The gene expression levels for 7129 human genes were measured for

this cohort. The dataset is randomly split into training group of 37 samples containing 24 ALL samples and 13 AML samples and test group of 35 samples containing 23 ALL and 12 AML samples. The model is trained on the train data set and tested on the test data set and accuracy, sensitivity, and specificity of the model is reported. We used the top 500 genes based on p-value rankings of differentially expressed genes for downstream analysis. The literature correlation of these genes to leukemia is obtained using GeneIndexer. GeneIndexer is a commercially available software used to classify and prioritize genes based on functional information in the biomedical literature. It mines for explicit and implicit relationships, and finds an association between the genes and keywords [39].

Results

The Gibbs Sampler was run for $60k$ iteration and the first $20k$ is discarded as burn in. Genes were ranked based on posterior mean of θ associated with each gene. A plot of posterior mean of θ associated with genes is represented in Figure 10. Using top 5, 10, 20, 30, 40, and 50 genes obtained from the model, we

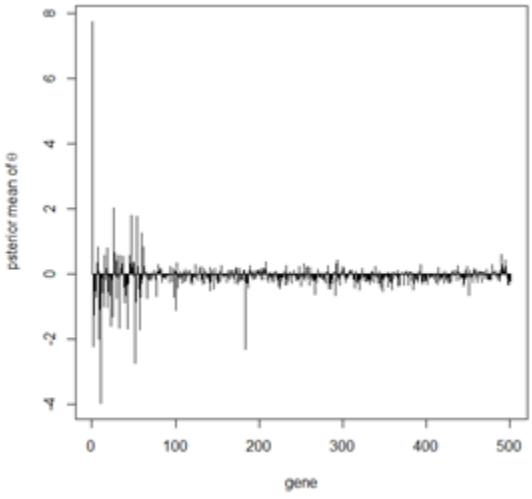

Figure 10: LSBGG. Posterior mean of θ asociated with each gene.

evaluate classification accuracy, sensitivity and specificity on test data sets across 50 runs. The model performance was compared to the set up in which both ρ_j and δ_j are set to 1 for all genes and no biological information was incorporated into the model. These results are represented in Table 14. Each row represents model performance using different numbers of genes for classification. We used, 5, 10, 20, 30, 40, and 50 genes and analyzed the model performance. The LSBGG model outperforms the SBGG model in average accuracy, sensitivity and specificity regardless of number of genes used for model evaluation.

Table 14: Classification Accuracy, Sensitivity, and Specificity Analysis.

		LSBGG			SBGG	
#genes	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
5	0.93(0.054)	0.94(0.08)	0.92(0.1)	0.84(0.096)	0.87(0.15)	0.78(0.24)
10	0.96(0.031)	0.97(0.039)	0.93(0.074)	0.89(0.08)	0.92(0.11)	0.83(0.208)
20	0.96(0.035)	0.98(0.039)	0.93(0.081)	0.92(0.056)	0.96(0.05)	0.85(0.17)
30	0.96(0.03)	0.98(0.026)	0.94(0.07)	0.94(0.042)	0.97(0.0476)	0.89(0.12)
40	0.97(0.03)	0.99(0.023)	0.93(0.075)	0.94(0.0408)	0.97(0.048)	0.89(0.117)
50	0.97(0.027)	0.99(0.021)	0.94(0.066)	0.95(0.031)	0.97(0.044)	0.92(0.088)

As it is obvious from the Table 14 , the model with literature informed choice of hyper parameters results in better classification accuracy across different number of genes chosen for classification. The classification accuracies reported are the average results across 50 runs.

Next, we examined the top hundred genes obtained from the model across 50 runs to obtain the number of common genes and examine the consistency of results obtained in different runs of the model. The number of common genes was 41 for our literature aided model compared to 6 for the other scenario. This results suggest that the literature model results in more consistent gene sets based on biologically relevant genes. This is strong indication of generalizability of the

model for potential clinical use in diagnostics and therapeutics.

Simulation Study

In this section, we performed a simulation study with 6 different scenarios and evaluate the performance of the literature aided model in each setting. We simulated two different data sets of sizes 30 (15 cases and 15 controls) and 50 (25 cases and 25 controls). For each data set, we simulated 20 gene expression values, assuming the first 5 genes are differentially expressed and the rest are not. We also assumed that the first 5 genes are biologically relevant to the response variable. Genes that are differentially expressed are randomly sampled from $N(\mu_1, \sigma_1)$ and the rest are randomly sampled from $N(\mu_2, \sigma_2)$. We set $\mu_1 = 3, \mu_2 = 0, \sigma_1 = \sigma_2 = 1$. We first elaborate on the data set with 30 samples in 3 different scenarios. We randomly divided the data into train (N=15) and test(N=15). The model was trained on the train data set and its performance was evaluated using the test data. The Gibbs sampling algorithm was run for 40k iteration and the first 20k will be discarded as burn in. In the first scenario, we assumed the first 5 genes are biologically relevant to response and the rest are not. We put GDP distribution on gene j $(j = 1, .., 20)$ with parameters specified as $(\rho_j = 1, \delta_j = 1; j = 1, .., 5)$ and $(\rho_j = 2, \delta_j = 1; j = 6, .., 20)$. These two distributions are shown in figure below. In the second scenario, we did not assume any biological information in the model and we put GDP distribution with parameters $(\rho_j = 1, \delta_j = 1; j = 1, .., 20)$ on parameters associated with each gene. In scenario 3, we assume the first 5 genes are biologically relevant to response variable however these associations were mis-specified and assigned randomly (we

randomly assign 5 genes to be our biologically relevant genes).

We randomly divided the date set into train and test samples 50 times and the average model performance and associated standard deviations on the test samples across 50 runs for all three scenarios are presented in Table 15. We used different number of genes (P represents the number of genes used for classification) in order to be able to evaluate model performance across different number of genes for classification. As we can see, in scenario 1 in which the

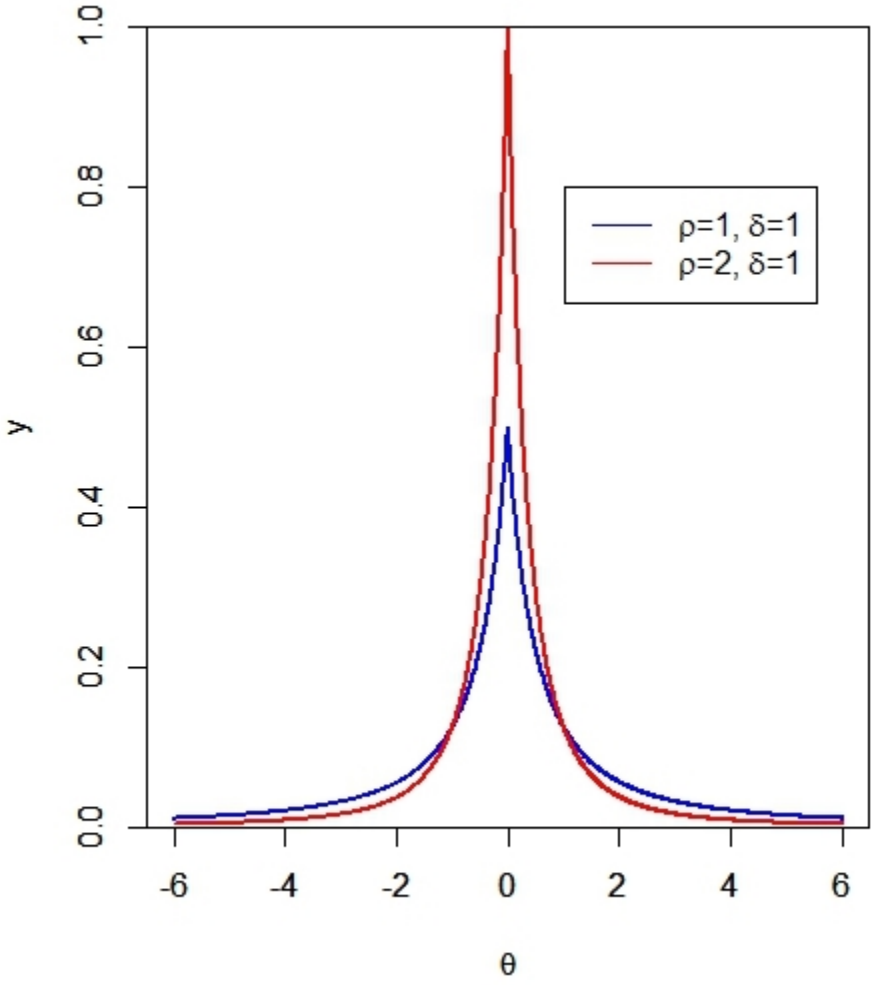

Figure 11: pecification of GDP prior distributions used in simulation study.

biological relevance of genes is correctly specified and used as prior information,

Table 15: Simulaton study: Classification Accuracy, Sensitivity, and Specificity Analysis, N=30 (associated standard deviations are represented in parentheses) .

	P=5			P=10			P=20		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
Scenario1	0.96(0.07)	0.94(0.08)	0.98(0.067)	0.98(0.055)	0.99(0.06)	0.96(0.042)	0.98(0.036)	0.97(0.052)	0.99(0.028)
Scenario2	0.88(0.084)	0.92(0.09)	0.84(0.06)	0.89(0.07)	0.91(0.06)	0.88(0.07)	0.92(0.06)	0.96(0.08)	0.88(0.03)
Scenario3	0.75(0.12)	0.68(0.095)	0.82(0.13)	0.806(0.1)	0.75(0.076)	0.87(0.08)	0.853(0.08)	0.905(0.11)	0.8(0.063)

the model obtained highest performance. The model which does not incorporate prior information achieved the second best performance. It is interesting to note that when the biological relevance of markers are miss-specified, the model performance went down. This is due to over-shrinkage imposed on true signals.

For data set with 50 samples, we use the same procedure to generate the data set, randomly dividing the data set into 25 samples for training and 25 for testing. We repeated scenario 1 – scenario 3. Table 16 represent the model performance evaluation in these scenarios.

Table 16: Simulaton study: Classification Accuracy, Sensitivity, and Specificity Analysis , N=50 (associated standard deviations are represented in parentheses).

	P=5			P=10			P=20		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
Scenario1	0.93(0.07)	0.97(0.09)	0.9(0.08)	0.945(0.06)	0.92(0.08)	0.97(0.03)	0.97(0.026)	0.955(0.03)	0.98(0.034)
Scenario2	0.903(0.08)	0.87(0.10)	0.94(0.08)	0.9(0.068)	0.92(0.05)	0.89(0.07)	0.93(0.04)	0.91(0.05)	0.954(0.04)
Scenario3	0.79(0.15)	0.75(0.13)	0.83(0.11)	0.81(0.09)	0.76(0.11)	0.85(0.08)	0.86(0.06)	0.88(0.07)	0.845(0.06)

Consistent with the results obtained previously, the model in which the biological information are correctly identified and incorporated into the model achieved higher classification accuracy, sensitivity, and specificity compared to the model without biological information as well as the model with miss-specified biological information. Taken together based on the simulation study, we argue that the literature aided model with correct specification of biological information achieves highest performance compared to the model with miss-specified

biological information and model without use of biological information a priori.

Simulation Study part 2

In this study we added noise to the variables(gene expressions) to see how robust the model performance is in the presence of noise. We explored several scenarios. We add/ subtract 5, 10, 20, 30, 40, and 50 percent of the actual variable measured to the variable as random noise. For instance, assume we want to induce 5% noise to the data. Let x_{ij} represents gene expression value for gene 'j' in sample 'i'. In the new data set, we replace x_{ij} by $x_{ij} * U(0.95, 1.05)$ where U stands for a uniform distribution. We do this procedure for all genes in the dataset. It should be noted that we use runif function in R to generate random uniform numbers. The Golub data set was used for this study. The biological relevance of genes was incorporated into the model by adjusting the shape of the GDP distribution as described in section 6.5 and figure 7. The Gibbs sampling algorithm was run for 40k iterations and the first 20k was discarded as burn-in. The data is devided randomly into training and test samples according to chapter 5 section 2. We performed 50 resampling on the training and test data sets. The average classification accuracy, sensitivity, and specificity and associated standard deviations using top 10 marker genes obtained from the model are represented in the table 17. Based on table above, as the amount of noise in the data increases the model accuracy decreases. However, this decrease is not dramatic and for example the model accuracy is still above 90% in the presence of 20% noise. This demonstrate that the methodology is reasonably robust to the presence of noise in the system.

Table 17: Simulaton study part 2: Average classification accuracy, Sensitivity, Specificity and associated standard deviations (in parentheses).

	Accuracy	Sensitivity	Specificity
5% noise	0.943(0.05)	0.95(0.042)	0.927(0.062)
10% noise	0.935(0.055)	0.942(0.07)	0.91(0.068)
20% noise	0.903(0.07)	0.92(0.082)	0.88(0.076)
30% noise	0.86(0.074)	0.84(0.06)	0.9(0.08)
40% noise	0.815(0.08)	0.805(0.06)	0.842(0.1)
50% noise	0.79(0.11)	0.8(0.085)	0.78(0.13)

Chapter 7

Conclusions and Future Work

In many disciplines, such as gene expression analysis and genome-wide association studies, values of a large number of variables are measured simultaneously. Thus, it is very common to have a disproportionate number of variables compared to small sample sizes. In order to highlight those variables that are most relevant to certain phenotypes, it is necessary to develop an approach to weed out unimportant variables. Most complex diseases are caused by multiple effects and thus a single variable analysis can only detect a very small portion of variation and may not be powerful enough for identifying weaker associations [7]. In the situations that we are faced with fat datasets with $p \gg n$, highly regularized approaches are needed to identify non-zero coefficients, enhance model predictability and avoid over-fitting [38].

To address these limitations, we developed several Bayesian methods using different specialized priors that impose sparsity in terms of number of variables in the model. Using a double exponential prior on parameters we developed a sparse model in a Generalized Linear Model framework (SBGDE). This model can be used for classification of cancer progression stages. We evaluated the performance of the model using a publicly available data set on prostate cancer progression. Using the top 10 genes and top 50 genes obtained from the model we compared average classification accuracy and class-specific classification accuracy to well-known machine learning methods such as Support Vector

Machine (SVM) and Random Forest. The model outperforms SVM in all categories and has comparable performance, albeit slightly lower, to Random Forests. However, SBGDE identified more biologically relevant gene sets compared to the other methods investigated.

The double exponential prior has light tails when compared to GDP that can cause over-shrinkage of parameters towards zero, which may impose unwanted bias. In order to address this problem, we investigated another prior distribution with more tail robustness property. Recently, the Generalized Double Pareto (GDP) prior distribution was proposed as an alternative to induce sparseness in situations when we are faced with a large number of variables compared to sample size [5]. The authors applied the proposed method in the normal linear regression model framework. This prior has a simple analytic form, yields a proper posterior and possesses appealing properties, including a spike at zero, Student t-like tails, and a simple characterization as a scale mixture of normal distributions leading to a straightforward Gibbs sampler for posterior inferences that makes Bayesian shrinkage estimation and regularization feasible [5].

Utilizing this prior in a more general framework of generalized linear models, we presented a sparse Bayesian hierarchical model that can incorporate a large number of variables compared to small sample sizes. While shrinking small effects toward zero and producing sparse solutions, the over shrinkage problem caused by using light-tailed priors would be remedied by the heavier tails. Using the GDP prior, we develop a sparse Bayesian generalized linear model (SBGG). We evaluated the performance of the model using the leukemia data set of Golub

et. al. [32]. Sensitivity, specificity, and classification accuracy measures were used to evaluate the model. It is interesting that we found that SBGG outperforms SBGDE and obtains higher classification accuracy and sensitivity and specificity. We also obtained the GCAT literature p-value of top 100 genes obtained from the model [96]. The SBGG results in more significant literature based p-values which indicates that this model gives more biologically relevant genes compared to SBGDE.

Additionally, we extended the SBGG model further to encompass data sets with multi-category ordinal response. We developed a sparse Bayesian multinomial model and evaluated its performance using prostate cancer gene expression data. We compared the model performance to three models: Random Forests, SVM, and SBGG. We found that the SBGG classification accuracy of prostate cancer subtypes were comparable to Random Forrest when using 10 marker genes for classification and it outperforms Random Forest in 3 out of four categories when we used 50 marker genes. Additionally, it outperforms SBGDE in 3 out of 4 categories when using 10 marker genes for classification and 3 out of 4 categories when using 50 marker genes. Furthermore, SBGG identified more biologically relevant gene sets. We next asked if SBGG gene rankings were more or less relevant to the biological mechanisms associated with prostate cancer progression. In order to evaluate the biological relevance for the top ranked genes in the models, we used a literature based method called GeneSet Cohesion Analysis Tool (GCAT) [96]. GCAT is a web-based tool that determines the functional coherence p-values of gene sets based on latent semantic analysis of

Medline abstracts [96]. The average GCAT literature derived p-values (LPv) for the top 100 genes obtained from 50 runs of SBGG, Random Forrest, SBGDE as well as the top 100 genes based on the p-value rank ordering of single gene analysis using ordinal logistic regression. We found that on average, SBGG produced more functionally cohesive gene list (LPv = 2.0E-4) compared to SBGDE (LPv=0.007), classical logistic regression (LPv= to 0.047) and Random Forest (LPv=0.131). Notably, 100% of the runs had smaller LPv than 0.047, produced by single gene analysis using classical logistic regression p-value ranking. The Literature p-value for the median run was 4.50E-06 compared to 1.90E-04 for SBGDE and 2.85E-02 For Random Forest. Based on these results, we posit that SBGG may be a better approach to simultaneously identify marker genes for classifications as well as for gaining insights into the molecular mechanisms of the phenotype under investigation compared to the other three methods.

It is important to note that the initial gene set input to the model for the binary and multi category situations are selected based on single gene analysis paradigm. Hence, this could bias the initial gene selection process. Gene selection methods that are based on signal strength and differentially expressed genes choose genes that are highly differentially expressed across different tissue types, i.e. cancer and normal tissue. However, most of these genes can be housekeeping genes or genes that are differentially expressed during different cell cycles that might not be necessarily related to cancer. It is possible that some biologically relevant genes to the phenotype might have been missed by this analysis due to low signal.

Literature information can be used to select biologically relevant genes from gene expression data in order to build cancer classifiers. These methods can be very helpful in order to improve the biological relevance of classifiers built based on gene expression data. Here, we investigated a very famous gene ranking method based on biological literature called Gene Indexer [39]. Gene Indexer utilizes Latent Semantic Indexing (LSI), a vector space model for information retrieval, to automatically identify conceptual gene-gene and gene-disease relationships from titles and abstracts of MEDLINE citations [39]. LSI method has proved to identify gene-keyword and gene-gene relationships with high average accuracy. Additionally, this method is able to obtain implicit relationships between genes and keywords that proves very helpful in identifying conceptual relationships [39]. The genes obtained based on literature were used for classification using SBGG, SBGDE, SVM, and Random Forest. The results were compared to the same models applied to input genes obtained from tests of differential expression p-values.

For the binary response situation, we used leukemia data set for evaluating our hypothesis. In SBGG model, the classification accuracy and sensitivity and specificity were very close for the two paradigms. The SVM model obtained higher classification accuracy, sensitivity, and specificity when using Gene Indexer as input gene selection. For the Random Forest model the classification accuracy in both paradigms are very close to each other. In conclusion, the classifiers built on signal strength obtain higher classification accuracy compared to the Gene Indexer input gene list counterparts.

Next, we evaluated the literature based input gene list in the multi-category response situations using prostate cancer progression data set [90]. The average classification accuracy for the multi-category classifiers built on prostate cancer progression data sets were compared. For the Benign sample type, the models that are built based on input gene list from differential expression test outperform the counterparts built upon input gene list obtained from Gene Indexer. For the PIN sample type, the models based on Gene Indexer outperform their counter parts for SBGG, SBGDE, SVM and the classification accuracy is a bit lower for the Random Forest Model but comparable for the two scenarios. For the PCA sample type, the models based on highly differentially expressed genes outperform the Gene Indexer input gene list models in 3 out of four models namely, SBGG, SBGDE, and Random Forest and it has slightly lower classification accuracy for the SVM model. For the MET sample types, the models based on differentially expressed input gene lists outperform all the models built based on Gene Indexer input gene list.

Even though the classifiers built based on Gene Indexer gene selection paradigm come close in classification accuracy, sensitivity, and specificity to their counterparts for some of the classifiers, in the majority of the classifiers, the gene selection based on signal strength results in higher accuracy, sensitivity, and specificity. One of the main reasons for this phenomena is that the majority of cancer related genes are not highly differentially expressed across different tissue types which lowers their ability to be highly powerful predictors. On the other hand, gene selection methods that are based on signal strength and differentially

expressed genes, choose genes that are highly differentially expressed across different tissue types, i.e. cancer and normal tissue. However, housekeeping genes and genes highly differentially expressed in different cell cycles are obtained from these methods that might not be related to cancer. Thus, the classifiers obtained based on differentially expressed genes across different sample types might not reveal biologically relevant genes. Even though these models obtain higher classification accuracy, they suffer from the fact that they do not obtain comprehensive biological relevance in the set of predictors obtained. Using input gene list solely based on current literature biases the downstream results due to the fact that it ignors the signals observed in the experiment. Ideally, we would want to let signals observed have an impact on the results but have a technique to prioritize and tune these signals based on biological information from literature. Bridging the gap between classification accuracy and biological relevance will be of high merit to the community and can potentially result in deeper understanding of mechanisms involved.

In chapter 6, we developed a literature aided sparse Bayesian generalized Linear model which utilizes Generalized Double Pareto prior (LSBGG) to induce shrinkage in terms of number of variables. Instead of uninformed hyper parameters for the prior distributions, we adopt a literature informed approach to adjust the hyper parameters based on marker's biological relevance to the phenotype under study. This will aid us in controlling shrinkage imposed on genes based on their biological relevance. Using the top 10 genes obtained from our model, we were able to achieve 95.5% average classification accuracy.

Additionally average sensitivity and specificity of 97.2% and 92.9% was achieved across 50 runs. The model without incorporation of biological information (SBGG) achieves average 88.6%, 91.8%, and 82.5% accuracy, sensitivity, and specificity respectively. The LSBGG model demostrated superior performance consistently regardless of the number of genes used for classification (Table 14). Additionally, There were 41 genes common in all runs for the literature aided model compared to 6 genes for the model with uninformed choice of hyper parameters. This results suggest that the literature aided model produces more consistent results with significantly higher biological relevance. Taken together, these results suggest that literature informed Sparse Bayesian Generalized Linear Model applied to leukemia data sets allows for better subclass prediction based on more functionally relevant gene sets.

There exists a possibility of utilizing Metropolis–Hastings algorithm instead of Griddy Gibbs sampling algorithm employed to sample hyper-parameters v, u_1, u_2 in chapter 2. Metropolis Hastings is a Markov chain Monte Carlo (MCMC) method for obtaining a sequence of random samples from a probability distribution for which direct sampling is difficult [31]. The Metropolis–Hastings algorithm can draw samples from any probability distribution $P(x)$, provided you can compute the value of a function $f(x)$ which is close to the density of P. On the other hand, most simple rejection sampling methods suffer from the dimensionality, where the probability of rejection increases exponentially as a function of the number of dimensions [31]. Metropolis Hastings algorithm is only useful when you can find a suitable “jumping” density which is “similar” (close) to its target density to avoid

excessively slow mixing [31]. This is a difficult task, especially for high-dimensional space. In addition, the metropolis algorithm within each iteration on the last part of the MCMC procedure would dramatically increase the running time of the MCMC process.

In future, we plan to extensively investigate the LSBGG model performance across several different cancer cohorts. Additionally, we plan to investigate the possible development of an effective model to translate biological information to choice of hyper parameters. Furthermore, it is possible to evaluate performance of models developed using pathway driven feature selection methods while considering more complex variance-covariance matrix structures which takes into account gene-gene interactions. In addition to these potentially exciting new developments, further development is possible by considering survival time data frameworks.

1. Albert, J., Chib, S.: **Bayesian analysis of binary and polychotomous response data**. *Journal of American Statistical Association* 1993, **88**: 669–679.

2. Alizadeh, A., Eisen, M., Eric, D., Lossos, I., Rosenwald, A., Boldrick, J., Sabet, H., Tran, T., Yu, X., Powell, J., Yang, L., Marti, G., Moore, T., Hudson, J., Lu, L., Lewis, D., Tibshirani, R., Sherlock, G., Chan, W., Greiner, T., Weisenburge, D., Armitage, J., Warnke, R., Levy, R., Wilson, W., Grever, M., Byrd, J., Botstein, D., Brown, P., Staudt, L.: **Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling**. *Nature* 2000, **403**: 503–511.

3. Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., Levine, A.: **Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays**. *Proceedings of the National Academy of Sciences* 1999, **96**: 6745–6750.

4. Altman, D., Bland, J.: **Diagnostic tests. 1: Sensitivity and specificity**. *Journal of American Statistical Association* 1994, **38(6943)**: 1552.

5. Armagan, A., Dunson, D., Lee, J.: **Generalized double Pareto shrinkage**. *Statistica Sinica* 2011. doi:arXiv:1104.0861

6. Azuaje, A.: **Interpretation of genome expression patterns: computational challenges and opportunities**. *IEEE Engineering in Medicine and Biology* 2000, **1**: 26–41.

7. Bae, K., Mallick, B.: **Gene Selection Using a Two-Level Hierarchical Bayesian Model**. *Bioinformatics* 2004, **20**: 3423–3430.

8. Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M., Yakhini, Z.: **Tissue classification with gene expression profiles**. *Journal of Molecular Biology* 2000, **7**: 559–5583.

9. Benjamini, Y., Hochberg, Y.: **Controlling the false discovery rate: a practical and powerful approach to multiple testing**. *Journal of the Royal Statistical society* 1995, **B 57**: 289–300.

10. Bishop, C.: **Pattern Recognition and Machine Learning**. Springer, NewYork 2006

11. Boulesteix, A., Janitza, S., Kruppa, J., König, I.: **Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics**. *WIREs Data Mining Knowledge Discovery* 2012, **2**: 493–507.

12. Breiman, L.: **Random Forests**. *Machine Learning* 2001, **45(1)**: 5–32.

13. Butte, A.: **The use and Analysis of microarray data", Nature Review Drug Discovery**. *Journal of National Cancer Institute* 2002, **1)**: 951–960.

14. Calvo, A., Xiao, N., Kang, J., Best, C., Leiva, I., Emmert-Buck, M., Jorcyk, C., Green, J.: **Alterations in gene expression profiles during prostate cancer progression: functional correlations to tumorigenicity and down-regulation of selenoprotein-P in mouse and human tumors**. *Cancer Research* 2002, **62(18)**: 5325–5335.

15. Cao, J., Zhang, S.: **Measuring statistical significance for full Bayesian methods in microarray analyses**. *Bayesian Analysis* 2004, **5(2)**: 413–427.

16. Cawley, G., Talbot, N.: **Gene selection in cancer classification using sparse logistic regression with Bayesian regularization**. *Bioinformatics* 2006, **22**: 2348–2355.

17. Chang, C., Wang, J., Zhao, C., Fostel, J., Tong, W., Bushel, P., Deng, Y., Pusztai, L., Symmans, W., Shi, T.: **Maximizing biomarker discovery by minimizing gene signatures**. *BMC Genomics* 2011, **12(suppl)**: 5–6.

18. Chin, K., DeVries, S., Fridlyand, J., Spellman, P., Roydasgupta, R., Kuo, W., Lapuk, A., Neve, R., Qian, Z., Ryder, T., Chen, F., Feiler, H., Tokuyasu, T., Kingsley, C., Dairkee, S., Meng, Z., Chew, K., Pinkel, D., Jain, A., Ljung, B., Esserman, L., Albertson, D., Waldman, F., Gray, J.: **Genomic and transcriptional aberrations linked to breast cancer pathophysiologies**. *Cancer Cell* 2006, **10(6)**: 529–541.

19. Cordero, F., Botta, M., Calogero, R.: **Microarray data analysis and mining approaches**. *Briefings in Functional Genomics and Proteomics* 2008, **4**: 265–281.

20. Dalgin, G., Alexe, G., Scanfeld, D., Tamayo, P., Mesirov, J., Ganesan, S., DeLisi, C., Bhanot, G.: **Portraits of breast cancer progression**. *BMC Bioinformatics* 2007, **8**: 291.

21. Devore, J., Peck, R.: **Statistics: The Exploration and Analysis of Data**. Duxbury Press, Pacific Grove, CA 1997

22. Devore, J., Peck, R.: **Statistics: The Exploration and Analysis of Data**. Duxbury, Pacific Grove CA 1997

23. Ding, C., Peng, H.: **Minimum redundancy feature selection from microarray gene expression data**. *Journal of Bioinformatics and Computational Biology* 2005, **3**: 185–205.

24. Dudoit, S., Fridlyand, J., Speed, T.: **Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data**. *Journal of American Statistical Association* 2002, **97**: 77–87.

25. Dupuy, A., Simon, R.: **Critical Review of Published Microarray Studies for cancer outcome and guidelines on statistical analysis and reporting**. *Journal of National Cancer Institute* 2007, **99(2)**: 147–157.

26. Fan, J., Li, R.: **Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties**. *Journal of the American Statistical Association* 2001, **96**: 1348–1360.

27. Figueiredo, M.: **Adaptive sparseness for supervised learning**. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2003, **25**: 1150–1159.

28. Fontaine, J., Priller, F., Barbosa-Silva, A., Andrade-Navarro, M.: **Genie: literature-based gene prioritization at multi genomic scale**. *Nucleic Acid research* 2011, **39**

29. Fortunel, N., Otu, H., Ng, H., Chen, J., Mu, X., Chevassut, T., Li, X., Joseph, M., Bailey, C., Hatzfeld, J., Hatzfeld, A., Usta, F., Vega, V., Long, P., Libermann, T., Lim, B.: **Comment on Stemness: Transcriptional Profiling of Embryonic and Adult Stem Cells and A Stem Cell Molecular Signature**. *Science* 2003, **302(5644**

30. Gelfand, A., Smith, A.: **Sampling-based approaches to calculating marginal densities**. *Journal of American Statistical Association* 1990, **88**: 881–889.

31. Gilks, W., Richardson, S., Spiegelhalter, D.: **Markov Chain Monte Carlo in Practice**. Chapman and Hall, London 1996

32. Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C., E, L.: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring**. *Science* 1999, **286**: 531–537.

33. Griffin, J., Brown, P.: **Bayesian adaptive lassos with non-convex penalization**. *Technical report IMSAS, University of Kent* 2007

34. Griffin JE, B.P.: **Alternative prior distributions for variable selection with very many more variables than observations**. *Technical Report* 2005

35. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: **Gene selection for cancer classification using support vector machines**. *Machine Learning* 2000, **46**

36. Guyon I, B.S.V.V. Weston J: **Gene selection for cancer classification using support vector machines**. *Machine Learning* 2002, **46**: 389–422.

37. Hans, C.: **Bayesian lasso regression**. *Biometrika* 2009, **96**: 835–845.

38. Hastie, T., Tibshirani, R., Friedman, J.: **High-dimensional Problems: P > N. The Elements of Statistical Learning**. Springer, New York 2009

39. Homayouni, R., Heinrich, K., Wei, L., Berry, M.: **Gene clustering by latent semantic indexing of MEDLINE abstracts**. *Bioinformatics* 2005, **21**

40. Hsueh, H., Zhou, D., Tsai, C.: **Random forests-based differential analysis of gene sets for gene expression data**. *Gene* 2013, **518(1)**: 179–186.

41. Kann, M.: **Advances in translational bioinformatics**. *Briefings in Bioinformatics* 2010, **11(1)**: 96–110.

42. Karatzoglou, A., Smola, A., Hornik, K., Zeileis, A.: **kernlab - An S4 Package for Kernel Methods in R**. *Journal of Statistical Software* 2004, **11(9)**: 1–20.

43. Karatzoglou, A., Smola, A., Hornik, K., Zeileis, A.: **kernlab - An S4 Package for Kernel Methods in R**. *Journal of Statistical Software* 2004, **11(9)**: 1–20.

44. Kim, E., Kim, S., Ashlock, D., Nam, D.: **MULTI-K: Accurate classification of microarray subtypes using ensemble k-means clustering**. *BMC Bioinformatics* 2009, **10**: 26.

45. Knight, K., Fu, W.: **Asymptotics for Lasso-Type Estimators**. *The Annals of Statistics* 2000, **28**: 1356–1378.

46. Lai, L., Reinders, M., J Van't Veer, L., Wessels, L.: **A Comparison of Univariate and Multivariate Gene Selection Techniques for Classification of cancer data sets**. *BMC Bioinformatics* 2006, **7**: 235.

47. Laurent, G., Shtokalo, D., Tackett, M., Yang, Z., Vyatkin, Y., Milos, P., Seilheimer, B., McCaffrey, T., Kapranov, P.: **On the importance of small changes in RNA expression**. *Methods* 2013, **63(1)**: 18–24.

48. Li, J., Das, K., Fu, G., Li, R., Wu, R.: **The Bayesian lasso for genome-wide association studies**. *Bioinformatics* 2011, **27**: 516–523.

49. Li, J., Das, K., Fu, G., Li, R., Wu, R.: **The Bayesian lasso for genome-wide association studies**. *Bioinformatics* 2011, **27(4)**: 516–523.

50. Li, Y., Liang, M., Zhang, Z.: **Regression Analysis of Combined Gene Expression Regulation in Acute Myeloid Leukemia**. *PLOS Computational Biology* 2014. doi:10.1371/journal.pcbi.1003908

51. Liaw, A., Wiener, M.: **Classification and Regression by randomForest**. *R News* 2002, **2(3)**: 18–22.

52. Liu, Q., Sung, A., Chen, Z., Liu, J., Huang, X.: **Feature Selection and Classification of MAQC-II Breast Cancer and Multiple Myeloma Microarray Gene Expression Data**. *PLoS ONE* 2009, **4(12)**: 8250.

53. Logsdon, B., Hoffman, G., Mezey, J.: **A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis**. *Nature Genetics* 2010, **11**: 1–13.

54. Lu, Y., Han, J.: **Cancer classification using gene expression data**. *Information Systems* 2003, **28**

55. Lynch, S.M.: **Introduction to Applied Bayesian Statistics and Estimation for Social Scientists**. Springer, New York 2007

56. Madahian, B., Faghihi, U.: **A Fully Bayesian Sparse Probit Model for Text Categorization**. *Open Journal of Statistics* 2014, **4**: 611–619.

57. Madahian, B., Deng, L., Homayouni, R.: **Application of Sparse Bayesian Generalized Linear Model to Gene Expression Data for Classification of Prostate Cancer Subtypes**. *Open Journal of Statistics* 2014, **2**: 518–526.

58. Madsen, H., Thyregod, P.: **Introduction to General and Generalized Linear Models**. Chapman and Hall/CRC, London 2011

59. Marttinen, P., Myllykangas, S., Corander, J.: **Bayesian clustering and feature selection for cancer tissue samples**. *BMC Bioinformatics* 2009, **10**: 90.

60. McCullagh, P., Nelder, J.: **Generalized Linear Models**. Chapman and Hall, London 1989

61. MUDHOLKAR, S.M., GEORGE, E.O.: **A remark on the shape of the logistic distribution**. *Biometrika* 1978, **65**: 667–668.

62. Nelder, J., Wedderburn, R.: **Generalized Linear Models**. *Journal of the Royal Statistical society* 1972, **135(3)**: 370–384.

63. Newman, A., Cooper, J.: **AutoSOME: a clustering method for identifying gene expression modules without prior knowledge of cluster number**. *BMC Bioinformatics* 2010, **11**: 117.

64. Nott, D., Leng, C.: **Bayesian Projection Approaches to Variable Selection in Generalized Linear Models**. *Computational Statistics and Data Analysis* 2010, **54(12)**: 3227–3241.

65. Novianti, P., Roes, K., Eijkemans, M.: **Evaluation of Gene Expression Classification Studies: Factors Associated with Classification Performance**. *PLoS ONE* 2014, **9(4)**: 96063.

66. Pan, W.: **A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments**. *Bioinformatics* 1996, **18**: 546–554.

67. Park, T., Casella, G.: **The Bayesian lasso**. *Journal of American Statistical Association* 2008, **103**: 681–686.

68. Pearson, T., Manolio, T.: **How to interpret a genome-wide association study**. *Journal of the American Medical Association* 2008, **229(11)**: 1335–1344.

69. Perou, C., Jeffrey, S., Van de Rijn, M., Rees, C., Eisen, M., Ross, D., Pergamenschikov, A., Williams, C., SX, Z., JC, L.: **Distinctive gene expression patterns in human mammary epithelial cells and breast cancers**. *Proceedings of National Academy of Science* 1999, **96**: 9212–9217.

70. Pike, H., Smith, P.: **Bias and Efficiency in Logistic Analysis of Stratified Case-Control Studies**. *American Journal of Epidemiology* 1980, **9**: 89–95.

71. Pomeroy, S., Tamayo, P., Gaasenbeek, M., Sturla, L., Angelo, M., McLaughlin, M., Kim, J., Goumnerova, L., Black, P., Lau, C., Allen, J., Zagzag, D., Olson, J., Curran, T., Wetmore, C., Biegel, J., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovitzky, G., Louis, D., Mesirov, J., Lander, E., Golub, T.: **Prediction of central nervous system embryonal tumour outcome based on gene expression**. *Nature* 2002, **415**: 436–442.

72. Puszatri, L., Symmans, F., Van de Vijver, M.: **Development of Pharmacogenimic markers to select prospective chemotherapy for breast cancer**. *Breast Cancer* 2005, **12**: 73–85.

73. Pyon, Y., Li, J.: **Identifying Gene Signatures from Cancer Progression Data Using Ordinal Analysis**. *BIBM* 2009, **8**: 136–141.

74. Rencher, A.C.: **Multivariate Statistical Inference and Applications**. Wiley & Sons, New York 1998

75. Ritter, C., Tanner, M.: **Facilitating the Gibbs sampler: The Gibbs stopper and the griddy-Gibbs sampler**. *Journal of American Statistical Association* 1992, **97**: 861–868.

76. safnat, G., Jasch, D., Misra, A., Choong, M., Lin, F., Coiera, E.: **Gene Based Association with literature based enrichment**. *Journal of Biomedical Informatics* 2014, **49**: 221–226.

77. Singh, D., Febbo, P., Ross, K., Jackson, D., Manola, J., Ladd, C., Tamayo, P., Renshaw, A., D'Amico, A., Richie, J., Lander, E., Loda, M., Kantoff, P., Golub, T., Sellers, W.: **ene expression correlates of clinical prostate cancer behavior**. *Cancer Cell* 2002, **1**: 203–209.

78. Song, J., Ren, Y., Yan, F.: **Classification for high-throughput data with an optimal subset of principal components**. *Computational Biology and Chemistry* 2009, **33**: 408–413.

79. Statnikov, A., Aliferis, C., Tsamardinos, I., Hardin, D., Levy, S.: **A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis**. *Bioinformatics* 2005, **21**: 631–643.

80. Su, Z., Hong, H., Perkins, R., Shao, X., Cai, W., Tong, W.: **Consensus analysis of multiple classifiers using non-repetitive variables: Diagnostic application to microarray gene expression data**. *Computational Biology and Chemistry* 2007, **31**: 48–56.

81. Subramanian, A., Tamayo, P., Mootha, V., Mukherjee, S., Ebert, B., Gillette, M., Paulovich, A., Pomeroy, S., Golub, T., Lander, S., Mesirov, P.: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. *Proceedings of National Academy of Science* 2005, **102**: 15545–15550.

82. Swan, B., upper, B., Sczyrba, A., Lauro, F., Martinez-Garcia, M., Gonzáleze, J., Luo, H., Wright, J., Landry, Z., Hanson, N., Thompsona, B., Poultona, N., Schwientek, P., Acinas, S., Giovannoni, S., Moran, M., Hallam, S., Cavicchiolic, R., Woyke, T., Stepanauskas, R.: **Microarray data analysis and mining approaches**. *Proceedings of National Academy of Sciences* 2013, **110(28)**: 11463–11468.

83. Tan, A., Naiman, D., Xu, L., Winslow, R., Geman, D.: **Simple decision rules for classifying human cancers from gene expression profiles**. *Bioinformatics* 2005, **21**: 3896–3904.

84. Tang, L., Du, W., Fu, H., Jiang, J., Wu, H., Shen, G., Yu, R.: **New variable selection method using interval segmentation purity with application to blockwise kernel transform support vector machine classification of high-dimensional microarray data**. *Journal of Chemical Information and Modeling* 2009, **49**: 2002–2009.

85. Terrence, S., Cristianini, N., Duffy, N., Bednarski, D., Schummer, M., Haussler, D.: **Support vector machine classification and validation of cancer tissue samples using microarray expression data**. *Bioinformatics* 2000, **16(10)**: 906–914.

86. Thomas, J., Olson, J., Tapscott, S., Zhao, L.: **An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles**. *Genome Research* 2001, **11**: 1227–1236.

87. Thomas, J., Olson, J., Tapscott, S., Zhao, L.: **An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles**. *Genome Research* 2001, **11**: 1227–1236.

88. Tibshirani, R.: **Regression Shrinkage and Selection via the Lasso**. *Journal of the Royal Statistical Society* 1996, **58**: 267–288.

89. Tipping, M.: **Sparse Bayesian learning and the relevance vector machine**. *Journal of Machine Learning Research* 2001, **1**: 211–244.

90. Tomlins, S., Mehra, R., Rhodes, D., Cao, X., Wang, L., Dhanasekaran, S., Kalyana-Sundaram, S., Wei, J., Rubin, M., Pienta, K., Shah, R., Chinnaiyan, A.: **Integrative Molecular Concept Modeling of Prostate Cancer Progression**. *Nature Genetics* 2007, **39**: 41–51.

91. Troyanskaya, O., Garber, M., Brown, P., Botstein, D., Altman, R.: **Nonparametric methods for identifying differentially expressed genes in microarray data**. *Bioinformatics* 2002, **18**: 1454–1461.

92. Wang, L., Zhu, J., Zou, H.: **Hybrid huberized support vector machines for microarray classification and gene selection**. *Bioinformatics* 2008, **24**: 412–419.

93. Wang, Y., Klijn, J., Zhang, Y., Sieuwerts, A., Look, M., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M., Yu, J., Jatkoe, T., Berns, E., Atkins, D., Foekens, J.: **Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer**. *lancet* 2005, **365**: 671–679.

94. Wu, T., Chen, Y., Hastie, T., Sobel, E.: **Genome-wide association analysis by lasso penalized logistic regression**. *Bioinformatics* 2009, **25**: 714–721.

95. Xu, H., Caramanis, C., Mannor, S.: **Robust Regression and Lasso**. *IEEE Transactions On Information Theory* 2010, **56(7)**: 3561–3574.

96. Xu, L., Furlotte, N., Lin, Y., Heinrich, K., Berry, M., George, E., Homayouni, R.: **Functional Cohesion of Gene Sets Determined by Latent Semantic Indexing of PubMed Abstracts**. *PLoS ONE* 2011, **6(4)**: 18851.

97. Xu, L., Cheng, C., George, E., Homayouni, R.: **Literature aided determination of data quality and statistical significance threshold for gene expression studies**. *BMC Genomics* 2012, **13(suppl 8)**: 23.

98. Yang, J., Benyamin, B., McEvoy, B., Gordon, S., Henders, A., Nyholt, D., Madden, P., Heath, A., Martin, N., Montgomery, G., Goddard, M., Visscher, P.: **Common SNPs explain a large proportion of the heritability for human height**. *Nature Reviews Genetics* 2010, **42**: 565–569.

99. Ye, J., Li, T., Xiong, T., Janardan, R.: **Using Uncorrelated Discriminant Analysis for Tissue Classification with Gene Expression Data**. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 2004, **1(4)**: 181–190.

100. Yeoh, E., Ross, M., Shurtleff, S., Williams, W., Patel, D., Mahfouz, R., Behm, F., Raimondi, S., Relling, M., Patel, A., Cheng, C., Campana, D., Wilkins, D., Zhou, X., Li, J., Liu, H., Pui, C., Evans, W., Naeve, C., Wong, L., Downing, J.: **Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling**. *Cancer Cell* 2002, **1**: 133–143.

101. Yi, N., Xu, S.: **Bayesian LASSO for Quantitative Loci Mapping**. *Genetics* 2008, **179(2)**: 1045–1055.

102. Yi, N., Xu, X., Yang, X., Mallick, H.: **Multiple Comparisons in genetic association studies: a hierarchical modeling approach**. *Statistical Applications in Genetics and Molecular Biology* 2014, **13(1)**: 35–48.

103. Ying, L., Jiawei, H.: **Cancer classification using gene expression data**. *Information Systems* 2003, **28**: 243–268.

104. Yuan, M., Lin, Y.: **Efficient Empirical Bayes Variable Selection and Estimation in Linear Models**. *Journal of American Statistical Association* 2005, **100**: 1215–1225.

105. Zhang, H., Yu, C., Singer, B.: **Cell and tumor classification using gene expression data: construction of forests**. *Proceedings of National Academy of Science* 2003, **100**: 4168–4172.

106. Zhang, J., Deng, H.: **Gene selection for classification of microarray data based on the Bayes error**. *BMC Bioinformatics* 2007, **8**: 370.

107. Zou, H.: **The Adaptive Lasso and Its Oracle Properties**. *Journal of American Statistical Association* 2006, **101**: 1418–1429.

108. Zou, H., Li, R.: **One-step sparse estimates in non-concave penalized likelihood models**. *The Annals of Statistics* 2008, **36(4)**: 1509–1533.

Appendix 1 In this we derive fully conditional posterior distribution of parameter for the models in chapter 2, chapter 4, and chapter 6. In these chapters Generalized Double Pareto Prior is utilized for impossing sparsity in the models.

Derivation of transformations used on parameters ρ and δ

Let ρ and δ have the following distributions.

$$\pi(\rho) = \frac{c}{(1+c\rho)^2}; c > 0$$

$$\pi(\delta) = \frac{c'}{(1+c'\delta)^2}; c' > 0$$

Define the new variables u_1 and u_2 as follows:

$$u_1 = \frac{1}{1+c\rho}; \qquad u_2 = \frac{1}{1+c'\delta}$$

Using simple inverse method technique we can see that u_1 and u_2 are uniformly distributed. Here, we show the process for u_1.

$F(\rho) = \frac{1}{1+c\rho}$ is the cdf of the pdf $\pi(\rho) = \frac{c}{(1+c\rho)^2}$. We know that for $Y = F(X)$ has a $U(0,1)$.

SBGG: Deriving fully conditional distributions for parameters used in Gibbs sampling.

Let the matrices Λ and T^* be diagonal matrices define as defined as $\Lambda = \text{diag}(\Lambda_1, \cdots, \Lambda_n)$ and $T^* = \text{diag}(\tau_1^{-1}, ...\tau_p^{-1})$. Using the prior specifications in chapter 2, we obtain the following joint distribution.

$$\pi(\boldsymbol{\theta}, \boldsymbol{L}|\boldsymbol{y}) \propto$$

$$\prod_{i=1}^{n} [\ [I(y_i = 1) * I(l_i > 0) + I(y_i = 0) * I(l_i \leq 0)]\ *\ \Lambda_i^{1/2}\ *\ exp\left(\frac{\left(l_i - \boldsymbol{w}_i^T\boldsymbol{\theta}\right)^2}{\frac{-2}{\Lambda_i}}\right)\ *$$

$$\pi(v)\ *\ \left[\frac{\frac{v}{2}^{\frac{v}{2}}}{\Gamma\left(\frac{v}{2}\right)}{\Lambda_i}^{\frac{v}{2}-1}\ *\ exp(\frac{-v}{2} * \Lambda_i)\right]\ ]\ *\ \left[\prod_{j=1}^{p}\frac{1}{\sqrt{\tau_j}}\right] exp(\frac{-1}{2}\boldsymbol{\theta}^T T^*\boldsymbol{\theta})\ *\ \left[\prod_{j=1}^{p}\lambda_j^2\right]\ *$$

$$\left[exp\left(\frac{-1}{2}\sum_{j=1}^{p}\lambda_j^2\tau_j\right)\right]\ *\ \left[\prod_{j=1}^{p}\lambda_j^{\rho-1}\right]\ *\ exp\left(-\delta\sum_{j=1}^{p}\lambda_j\right)\ *\ \pi(u_1)\ *\ \pi(u_2)$$

Now we need to obtain fully conditional distributions for all the parameter in the model. In what follows, in each subsection we derive these fully conditional distributions step by step.

Fully Conditional Posterior Distribution for θ

It can be seen that the fully conditional distribution on $\boldsymbol{\theta}$ is proportional to the following.

$$\boldsymbol{\theta}|- \propto\ \left[\prod_{i=1}^{n} exp(\frac{\Lambda_i * (l_i - \boldsymbol{w}_i^T\boldsymbol{\theta})^2}{-2})\right] * exp\left(\frac{-1}{2}\boldsymbol{\theta}^T T^*\boldsymbol{\theta}\right)$$

Next, we show that this fully conditional distribution is multivariate normal and obtain the corresponding mean vector and variance covariance matrix needed in order to be able to sample these parameters in each iteration of Gibbs sampling.

$$\boldsymbol{\theta}|- \propto\ exp\left(\sum_{i=1}^{n}\frac{1}{-2}(l_i - \boldsymbol{w}_i^T\boldsymbol{\theta})^T\Lambda_i(l_i - \boldsymbol{w}_i^T\boldsymbol{\theta})\right)\ *\ exp\left(\frac{-1}{2}\boldsymbol{\theta}^T T^*\boldsymbol{\theta}\right)$$

$$\boldsymbol{\theta}|- \propto\ exp\left(\frac{1}{-2}\left[(\boldsymbol{L} - W\boldsymbol{\theta})^T\Lambda\left(\boldsymbol{L} - W\boldsymbol{\theta}\right)) + \ (\boldsymbol{\theta}^T T^*\boldsymbol{\theta})\right]\right)$$

$$\boldsymbol{\theta}|- \propto\ exp\left(\ \frac{1}{-2}\left[\ \boldsymbol{L}^T\Lambda\boldsymbol{L} - 2\boldsymbol{\theta}^T W^T\Lambda\boldsymbol{L} + \boldsymbol{\theta}^T W^T\Lambda W\boldsymbol{\theta} + \boldsymbol{\theta}^T T^*\boldsymbol{\theta}\ \right]\ \right)$$

$$\boldsymbol{\theta}|- \propto\ exp\left(\ \frac{1}{-2}\left[\ \boldsymbol{\theta}^T(W^T\Lambda W + T^*)\boldsymbol{\theta} - 2\boldsymbol{\theta}^T W^T\Lambda\boldsymbol{L}\ \right]\ \right)$$

$$\boldsymbol{\theta}|- \propto$$

$$exp\left(\ \frac{1}{-2}\left[\ \left(\boldsymbol{\theta} - (W^T\Lambda W + T^*)^{-1}W^T\Lambda\boldsymbol{L}\right)^T\ \left(W^T\Lambda W + T^*\right)\ \left(\boldsymbol{\theta} - (W^T\Lambda W + T^*)^{-1}W^T\Lambda\boldsymbol{L}\right)\ \right]\right)$$

Therefore, the Fully conditional distribution on $\boldsymbol{\theta}$ ismultivariate normal

distribution with following specification.

$$\boldsymbol{\theta}|- \sim MWN\left[\ \left(W^T\Lambda W + T^*\right)^{-1} W^T\Lambda \boldsymbol{L}\ ,\ \left(W^T\Lambda W + T^*\right)^{-1}\ \right]$$

Fully Conditional Posterior for τ_j

Before delving into derivation of fully conditional posterior for τ_j, we introduce inverse Gaussian distribution for matter of consistancy. Let $x \sim Inv - Gaussian(\mu, \sigma)$. The pdf of x is defined as follows.

$$f(x) = \left[\frac{\sigma}{2\pi x^3}\right]^{\frac{1}{2}} * exp\left[\frac{-\sigma(x-\mu)^2}{2\mu^2 x}\right]$$

based on the joint distribution demonstrated early in this appendix, the fully conditional posterior for τ_j is proportional to the following.

$$\tau_j|- \propto \frac{1}{\tau_j^{\frac{1}{2}}} * exp\left[-\frac{1}{2}\left(\frac{\theta_j^2}{\tau_j} + \lambda_j^2\tau_j\right)\right]$$

In order to be able to effectively sample τ_j in each iteration of Gibbs sampling, we need to obtain the closed form of the distribution and obtain the equations defining mean and variance of this distribution. In what follows, the details of the process taken is explained step by step.

Let $K = \frac{1}{\tau_j}$ then we have:

$$g(k) = f(\tfrac{1}{k}) * \tfrac{1}{k^2} \propto \frac{1}{k^{\frac{-1}{2}}} * exp\left[-\tfrac{1}{2}(\frac{\theta_j^2}{\frac{1}{k}} + \lambda_j^2 * \tfrac{1}{k})\right]\ *\ \tfrac{1}{k^2}$$

$$g(k) \propto \frac{1}{k^{\frac{3}{2}}} * exp\left[-\tfrac{1}{2}(k\theta_j^2 + \tfrac{\lambda_j^2}{k})\right]$$

$$g(k) \propto \frac{1}{k^{\frac{3}{2}}} * exp\left[-\tfrac{1}{2}\left(\tfrac{\theta_j^2 k^2 + \lambda_j^2}{k}\right)\right]$$

$$g(k) \propto \frac{1}{k^{\frac{3}{2}}} * exp\left[-\tfrac{1}{2}\theta_j^2\left(\frac{k^2 + \frac{\lambda_j^2}{\theta_j^2}}{k}\right)\right]$$

$$g(k) \propto \frac{1}{k^{\frac{3}{2}}} * exp\left[-\frac{1}{2}\theta_j^2\left(\frac{k^2+\frac{\lambda_j^2}{\theta_j^2}-\frac{2k\lambda_j}{\theta_j}+\frac{2k\lambda_j}{\theta_j}}{k}\right)\right]$$

$$g(k) \propto \frac{1}{k^{\frac{3}{2}}} * exp\left[-\frac{1}{2}\theta_j^2\frac{(k-\left(\frac{\lambda_j^2}{\theta_J^2}\right)^{\frac{1}{2}})^2}{k}\right]$$

$$g(k) \propto \frac{1}{k^{\frac{3}{2}}} * exp\left[-\frac{1}{2}\theta_j^2\left(\frac{\lambda_j^2}{\theta_j^2}\right)\ \frac{\left(k-\left(\frac{\lambda_j^2}{\theta_J^2}\right)^{\frac{1}{2}}\right)^2}{k*\frac{\lambda_j^2}{\theta_j^2}}\right]$$

$$g(k) \propto \frac{1}{k^{\frac{3}{2}}} * exp\left[-\lambda_j^2\ \frac{\left(k-\left(\frac{\lambda_j^2}{\theta_J^2}\right)^{\frac{1}{2}}\right)^2}{2\frac{\lambda_j^2}{\theta_j^2}*k}\right]$$

$$k|- = \tau_j^{-1}|- \sim Inv-Gaussian\left(\mu=\left(\frac{\lambda_j^2}{\theta_j^2}\right)^{\frac{1}{2}}\ ,\ \sigma=\lambda_j^2\right)$$

$$\tau_j^{-1}|- \sim Inv-Gaussian\left(\mu=\sqrt{\frac{\lambda_j^2}{\theta_j^2}}\ ,\ \sigma=\lambda_j^2\right)$$

Fully Conditiona Posteriorl Distribution for λ_j

Here, we show that fully conditional distribution on λ_j is gamma distribtution and obtain the parameters associated with it.

$$\pi(\tau_j,\lambda_j|-) = \pi(\tau_j|\lambda_j,-) * \pi(\lambda_j|-)$$

$$\pi(\tau_j,\lambda_j|-) \propto \frac{\lambda_j}{\tau_j^{\frac{1}{2}}} * exp\left[-\frac{1}{2}\left(\frac{\theta_j^2}{\tau_j}+\lambda_j^2\tau_j\right)\right]\ *\ \lambda_j^{\rho+1-1}exp\left(-\delta\lambda_j\right)$$

$$\pi(\tau_j,\lambda_j|-) \propto$$

$$\frac{\lambda_j}{\tau_j^{\frac{1}{2}}}\ *\ exp\left[-\frac{1}{2}\left(\frac{\theta_j^2}{\tau_j}+\lambda_j^2\tau_j-2\theta_j^2\left(\frac{\lambda_j^2}{\theta_j^2}\right)^{\frac{1}{2}}\right)\right]\ *\ exp\left[-\theta_j^2(\frac{\lambda_j^2}{\theta_j^2})^{\frac{1}{2}}\right]\ *\ \lambda_j^{\rho+1-1}exp\left[-\delta\lambda_j\right]$$

$$\pi(\lambda_j|-) = \frac{\pi(\tau_j,\lambda_j|-)}{\pi(\tau_j|\lambda_j,-)}$$

Using the kernel obtained for $\pi(\tau_j|\lambda_j,-)$ in previous section we obtain:

$$\pi(\lambda_j|-) \propto exp\left[-\theta_j^2 * \frac{|\lambda_j|}{|\theta_j|}\right] * \lambda_j^{\rho+1-1}exp\left[-\delta\lambda_j\right]$$

$$\pi(\lambda_j|-) \propto exp\left[-\theta_j^2 * \frac{\lambda_j}{|\theta_j|} - \delta\lambda_j\right] * \lambda_j^{\rho+1-1}$$

$\pi(\lambda_j|-) \propto exp\left[-\lambda_j(|\theta_j|+\delta)\right] * \lambda_j^{\rho+1-1}$

$\lambda_j|- \sim Gamma\left(\rho+1, |\theta_j|+\delta\right)$

Fully Conditional Distribution for Λ_r

It can be easily seen that the fully conditional distribution on Λ_J is gamma distributions and the parameters associated with it are specified below.

$\Lambda_r|- \propto \Lambda_r^{\frac{1}{2}} exp\left[-\frac{(l_r - w_r^T\theta)^2}{2} * \Lambda_r\right] * \Lambda_r^{\frac{v}{2}-1} exp\left[-\frac{v}{2}\Lambda_r\right]$

$\Lambda_r|- \sim Gamma\left[\frac{v+1}{2}, \frac{1}{2} * \left((l_r - w_r^T\theta)^2 + v\right)\right]$

The Fully Conditional Distributions for u_1, and u_2

Having each θ_j $GDP(\frac{\delta}{\rho}, \rho)$ independently, the joint distribution of θs is as follows: $\pi(\boldsymbol{\theta}) = \prod_{j=1}^{p}\left[\frac{1}{2\frac{\delta}{\rho}} * \left(1 + \frac{|\theta_j|}{\delta}\right)^{-(1+\rho)}\right]$

We put prior distributions as defined in equations (2.7) and equation (2.8). Transformations defined in equation (2.9) is used which results in uniform priors for new variables $u_1 and u_2$. Using the results from [5] the following posterior distributions are obtained for u_1 and u_2.

$u_1|- \propto \left(\frac{1-u_1}{cu_1}\right)^p * \prod_{j=1}^{p}\left(1 + \frac{|\theta_j|}{\delta}\right)^{-(\frac{1-u_1}{cu_1}+1)}$

$u_2|- \propto \left(\frac{c'u_2}{1-u_2}\right)^p * \prod_{j=1}^{p}\left(1 + \frac{c'u_2}{1-u_2}|\theta_j|\right)^{-(1+\rho)}$

Appendix D: Deriving fully conditional distributions for parameters used in Gibbs sampling.

The fully conditional distributions for the Sparse Bayesian model developed using double exponential prior developed in chapter3 are explored in this appendix. Let T be a diagonal matrix defined as $T = \text{diag}(\eta_1, \cdots, \eta_p)$.

Using the prior distributions as defined in chapter 3, we obtain the following joint distribution.

$$\pi(\boldsymbol{\theta}, \boldsymbol{l}|\boldsymbol{y}) \propto \prod_{i=1}^{n} \quad \left[\sum_{j=1}^{k} [I(y_i = j) * I(\gamma_j < l_i \leq \gamma_{j+1})\right] \; exp\left(\frac{(l_i - w_i^T\boldsymbol{\theta})^2}{-2}\right) \; *$$

$$\left[\prod_{j=1}^{p} \frac{1}{\sqrt{\eta_j}}\right] \; * \; exp\left(\frac{-1}{2}\boldsymbol{\theta}^T T^{-1}\boldsymbol{\theta}\right) \; * \; exp\left(-\frac{\lambda}{2}\sum_{j=1}^{p}\eta_j\right)$$

Fully Conditional Posterior Distribution for model parameters θ

based on the joint distribution obtained above, the fully conditional distribution on $\boldsymbol{\theta}$ is as follows.

$$\boldsymbol{\theta}|\Omega \propto \; \left[\prod_{i=1}^{n} exp(\frac{(l_i - w_i^T\boldsymbol{\theta})^2}{-2})\right] * exp\left(\frac{-1}{2}\boldsymbol{\theta}^T T^*\boldsymbol{\theta}\right)$$

We need to obtain closed form for this fully conditional distribution and obtain the mean parameter and variance covariance matrix associated with it.

$$\boldsymbol{\theta}|\Omega \propto \; exp\left[\sum_{i=1}^{n} \frac{1}{-2}(l_i - w_i^T\boldsymbol{\theta})^T(l_i - w_i^T\boldsymbol{\theta})\right] \; * \; exp\left(\frac{-1}{2}\boldsymbol{\theta}^T T^*\boldsymbol{\theta}\right)$$

$$\boldsymbol{\theta}|\Omega \propto \; exp\left(\frac{1}{-2}[(\boldsymbol{L} - W\boldsymbol{\theta})^T(\boldsymbol{L} - W\boldsymbol{\theta}) + \; (\boldsymbol{\theta}^T T^*\boldsymbol{\theta})]\right)$$

$$\boldsymbol{\theta}|\Omega \propto \; exp\left(\; \frac{1}{-2}[\; \boldsymbol{L}^T\boldsymbol{L} - 2\boldsymbol{\theta}^T W^T\boldsymbol{L} + \boldsymbol{\theta}^T W^T W\boldsymbol{\theta} + \boldsymbol{\theta}^T T^{-1}\boldsymbol{\theta} \;]\; \right)$$

$\boldsymbol{\theta}|\Omega \propto \ exp\left(\frac{1}{-2}[\ \boldsymbol{\theta}^T(W^TW + T^{-1})\boldsymbol{\theta} - 2\boldsymbol{\theta}^T W^T \boldsymbol{L}\ ] \right)$

$\boldsymbol{\theta}|\Omega \propto$

$exp\left(-\frac{1}{2}[\boldsymbol{\theta} - (W^TW + T^{-1})^{-1}W^T\boldsymbol{L}]^T(W^TW + T^{-1})[\boldsymbol{\theta} - (W^TW + T^{-1})^{-1}W^T\boldsymbol{L}]\right)$

Based on these results, the fully conditional distribution on parameters $\boldsymbol{\theta}$ is

multivariate normal distribution with parameters specifications as defined below.

$\boldsymbol{\theta}|\Omega \sim MWN\left[\ (W^TW + T^{-1})^{-1}W^T\boldsymbol{L}\ ,\ (W^TW + T^{-1})^{-1}\ \right]$

Fully Conditional Posterior for η_j

$\eta_j|\Omega \propto \frac{1}{\eta_j^{\frac{1}{2}}} * exp\left[-\frac{1}{2}(\frac{\theta_j^2}{\eta_j} + \lambda_j^2\eta_j)\right]$

We need to obtain closed form of the fully conditional distribution on η_j in order to

be able to sample these parameters efficiently in each iteration of Gibbs sampling.

Let $Z = \frac{1}{\eta_j}$ then we have:

$G(z) = P(Z \leq z) = P(\frac{1}{\eta_j} \leq z) = P(\eta_j \geq \frac{1}{z}) = 1 - F(\frac{1}{z})$

$g(z) = f(\frac{1}{z}) * \frac{1}{z^2}$

$g(z) \propto \frac{1}{z^{\frac{3}{2}}} * exp\left[-\frac{1}{2}(\frac{z^2\theta_j^2 + \lambda}{z})\right]$

$g(z) \propto \frac{1}{z^{\frac{3}{2}}} * exp\left[-\frac{1}{2}\theta_j^2(\frac{z^2 + \frac{\lambda}{\theta_j^2}}{z})\right]$

$g(z) \propto \frac{1}{z^{\frac{3}{2}}} * exp\left[-\frac{1}{2}\theta_j^2(\frac{z^2 + \frac{\lambda}{\theta_j^2} - \frac{2z\lambda}{\theta_j} + \frac{2z\lambda}{\theta_j}}{z})\right]$

$g(z) \propto \frac{1}{z^{\frac{3}{2}}} * exp\left[-\frac{1}{2}\theta_j^2\frac{(z - \frac{\sqrt{\lambda}}{|\theta_j|})^2}{z}\right]$

$$g(z) \propto \frac{1}{z^{\frac{3}{2}}} * exp\left[-\frac{1}{2}\theta_j^2 * \frac{\lambda}{\theta_j^2}\frac{(z-\frac{\sqrt{\lambda}}{|\theta_j|})^2}{z\frac{\lambda}{\theta_j^2}}\right]$$

$$z|\Omega \propto Inv - Gaussian\left(\frac{\sqrt{\lambda}}{|\theta_j|}, \lambda\right)$$

Fully Conditional Distributions for γ_s

$\gamma_s|\Omega \propto \prod_{i=1}^{n}\left[I(y_i = s-1) * I(\gamma_{s-1} \leq l_i < \gamma_s) + I(y_i = s) * I(\gamma_s \leq l_i < \gamma_{s+1})\right]$

Using equation 3.2, and 3.10, and based on the results in [1], the conditional posterior distribution of γ_s can be seen to be $Uniform(\delta_1, \delta_2)$ in which $\delta_1 = max\left[max_i[l_i|y_i = s-1], \gamma_{s-1}\right]$ and $\delta_2 = min\left[min_i[l_i|y_i = s], \gamma_s\right]$. It should be noted that $I()$ is indicator function and its value is one if its argument is true and is zero otherwise [1]. This argument is based on the results presented in [1].