1-1-2013

Selection and Clustering for Disease Associated Genetic Variants

Yubo Zou
University of South Carolina - Columbia

Selection and Clustering for Disease Associated Genetic Variants

by

Yubo Zou

Bachelor of Science
East China Normal University 1999

Master of Science
East China Normal University 2002

Doctor of Philosphy
Memorial University 2006

Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosphy in

Biostatistics

The Norman J. Arnold School of Public Health

University of South Carolina

2013

Accepted by:

Hongmei Zhang, Major Professor

James Hussey, Committee Member

Wilfried Karmaus, Committee Member

Jianjun Hu, Committee Member

Lacy Ford, Vice Provost and Dean of Graduate Studies

Acknowledgments

I would like to express my great appreciation to my advisor, Dr. Hongmei Zhang, for her careful guidance, patience, insight and invaluable suggestions throughout my PH. D. program. Without her help, this work would not have been possible. My deepest thanks to her is not only for her help of my academic research and kind encouragement, but also for her great effort to help me improving my writing skills in statistics.

Many thanks go to Dr. James Hussey, Dr. Wilfrad Karmaus and Dr. Jianjun Hu as members of my dissertation committee. I am sincerely grateful to Drs. Tim Hanson, Ian Dryden, Xiangzhen Huang, James Hardin, Bo Cai, Hrishikesh Chakraborty who taught me varies of statistics/biostatistics courses. I also thank them for their encouragement and nice advice.

I would like to take this opportunity to thank the Department of Epidemiology and Biostatistics, Drs. Robert McKeown and James Hussey, the department chairs, for providing me with the tuition supplement, necessary facilities for research fellowship and travel funds for conferences. Thanks go to all staff members at the department for their help. I am also grateful to the School of Graduate Studies, the School of Public Health at the University of South Carolina for travel funds support.

I deeply thank my parents, brothers and parents-in-law. Their understanding and emotional support have been inspiring me to complete my studies.

Finally, a special thanks to my wife Jiajia, her strength, patience and encouragement are most important for me, specifically when I encounter difficulties. She gives me helpful suggestions and helps me to overcome those frustrations.

Abstract

Epigenetics is the study of chemical reactions, which are orchestrated for the development and maintenance of an organism. Genetic or epigenetic variants (GEVs) encompass different types of genetic measures, such as Deoxyribonucleic acid (DNA) methylation at different CpG sites, expression level of genes, or single nucleotide polymorphisms (SNPs). With the development of technology, huge amount of genetic and epigenetic information is produced. However, the rich information potentially brings in challenges in data analyses. Thus it is necessary to reduce the dimension of data to improve efficiency. In my dissertation, I will focus on two directions of dimension reduction: variable selection and clustering.

The first project on dimension reduction was motivated by an epigenetic project aiming to identifying GEVs that are associated with a health outcome. Due to the potential non-linear interaction between GEVs, we designed a backward variable selection procedure to select informative GEVs. It is built upon a reproducing kernel-based method for evaluating the joint effect of a set of GEVs, e.g, a set of CpG sites. These GEVs may interact with each other in an unknown and complex way. Simulation studies indicate that the selection method is robust to different types of interaction effects, linear or non-linear. We demonstrate the method using two data sets with the first data selecting important SNPs that are associated with lung function and the second identifying important CpG sites such that their methylation is jointly associated with active smoking measured by cotinine levels.

The second project was motivated by the potential heterogeneity in clusters identified by existing methods. Traditional approaches focus on the clustering of either

subjects or (response) variables. However, clusters formed through these approaches possibly lack homogeneity. To improve the quality of clusters, we propose a clustering method through joint clustering. Specifically, the variables are first clustered based on the agreement of relationships (unknown) between variable measures and covariates of interest, and then subjects within each variable cluster are further clustered to form refined joint clusters. A Bayesian method is proposed for this purpose, in which a semi-parametric model is used to evaluate any unknown relationship between variables and covariates of interest, and a Dirichlet process is utilized in the process of second-step subjects clustering. The proposed method has the ability to produce homogeneous clusters composed of a certain number of subjects sharing common features on the relationship between some (response) variables and covariates. Simulation studies are used to examine the performance and efficiency of the proposed method. The method is then applied to DNA methylation measures of multiple CpG sites.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Epigenetics is the study of chemical reactions, which are orchestrated for the development and maintenance of an organism. Those reactions switch parts of the genome off and on at strategic times and locations. For example, in a differentiated cell, only 10% to 20% of the genes are active, and different sets of active genes make a skin cell different from a brain cell. Late in the life, a wide variety of environmental factors play a role in shaping the epigenome. Social interactions, physical activity, diet and other inputs can trigger changes in gene expression allowing body cells to respond dynamically to the outside world. These changes would occur throughout a whole life of anyone.

Epigenetic changes can modify the activation of certain genes, but not the sequence of DNA. Most epigenetic changes only occur within the course of one individual organism lifetime, but if gene disactivation occurs in a sperm or egg cell that results in fertilization, then some epigenetic changes can be transferred to the next generation, a process called epigenetic inheritance. Epigenetic inheritance goes against the idea that inheritance happens only through the DNA code that passes from parent to offspring. It means that a parent's experiences, in the form of epigenetic tags, can be passed down to future generations. Proving epigenetic inheritance is not always straightforward. Researchers should consider larger genomes, and several generations directly exposed to the same environ mental conditions at the same

time (i.e. for a pregnant female, three generations, mother, fetus, reproductive cells, at once are exposed to the same environmental conditions).

There are varieties of heritable epigenetic inheritance, and the DNA methylation, which is the modifications to the histone proteins around which DNA is wrapped when it is condensed in the cell nucleus, is one of the most popular. The addition of a – CH_3 group to the DNA sequence in places where a cytosine is followed by a guanine nucleotide is often called CpG site. With the development of biotechnology, measuring DNA methylation throughout an individual's genome has become more feasible. Therefore, researchers and practitioners have begun investigating the role of DNA methylation in disease development. It has been proven that DNA methylation is associated with cancer (Esteller [2007], Lujambio et al. [2008]), and cancer is frequently one of the hottest topics in public health. The commonly used DNA methylation assays are the GoldenGate and Infinium Methylation assays produced by Illumina, which measures DNA methylation for more than 450K CpG sites. The 450K Illumina arrays produce the intensity of methylated signal over the sum of methylated and unmethylated signal intensities. As for other high dimensional data, efficiently analyzing the data is a challenge task.

It has been a long history of studying Single-Nucleotide Polymorphisms (SNPs), which causes changes in the DNA sequence occurring when a single nucleotide in the genome differs between members of a biological species or paired chromosomes in an human. SNPs of humans can affect how humans develop diseases and respond to pathogens, drugs, vaccines, and are also thought to be the key factor in realizing the personalized medicine (Carlson [2008]). Furthermore, the genome-wise association studies in the past decade provided an opportunity to identify potential genetic causes of diseases.

Given the amount of info presented in genome-wide DNA methylation and geology data, reducing the dimension of data and identifying important genetic and

epigenetic factors are important components in studies related to DNA methylation and SNPs. It is critical to effectively detect CpG sites or SNPs that are important or informative with respect to disease risk or certain outcome. Identification of genetic or epigenetic variants (GEVs) that are associated with disease risk could help public health practitioners for disease intervention. This can be achieved by the following two ways:

1. We can perform variable selections to exclude the non-informative variables directly. With the technological development, a large numbers of GEVs are allowed to be deployed in studies.

2. Clustering GEVs and subjects will identify important and unimportant GEVs by use of association-based clustering.

1.2 Existing Methods for Variables Selection

Several commonly used variable selection methods exist for subset selection: choosing a subset of candidate variables to use in the final model. This is often done by backwards, forwards, and stepwise selection. In backwards selection, we start with a model containing all variables under consideration. Variables are then dropped one-by-one according to a predetermined criterion, usually that the partial t-test p-value is nonsignificant. This continues until all variables remaining in the model have a significant p-value. Forward selection is the inverse: adding variables one-by-one if they meet the pre-specified criterion. Stepwise selection is a hybrid of these two approaches, in which a variable can be added or dropped at each step.

Penalty functions are used in several different approaches for variable selections. In the regularization framework, various penalty functions are used to perform variable selection by putting relatively large penalties on small coefficients. Let us consider

the linear regression

$$y_i = \sum_{j=0}^{p} \beta_j x_{ij} + \varepsilon_i,$$

where y_i is the response, x_{ij} are the p predictors, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T$ is the vector of coefficients, and $\varepsilon \sim N(0, \sigma^2)$ is the independent noise. By minimizing residual sum of square (RSS), the ordinary least square (OLS) produces best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$. When constraint is given as $\sum_{j=1}^{p} |\beta_j|^2 \leq t; t > 0$, it is the ridge regression (first introduced by Hoerl and Kennard [1970]). Frank and Friedman [1993] expanded to bridge regression by generalizing the constraint to

$$L_q(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\beta_j|^q \leq t \text{ with } q > 0.$$

Knight and Fu [2000] mentioned that the bridge regression can perform the variable selection only when $q \leq 1$. If $q = 1$, the bridge regression corresponds to the least absolute shrinkage and selection operator (LASSO, Tibshirani [1996]), and it tends to produce a sparsity of nonzero coefficients in the final model. Tibshirani sought to improve the method because subset selection had lower-than desirable accuracy. The LASSO method does parameter estimation and variable selection simultaneously and can be applied to generalized linear models. However, it was shown that the LASSO produces biased estimates for the large coefficients (Liu and Wu [2007], Zou [2006]).

To improve the LASSO, Zou [2006] proposed adaptive LASSO by introducing weights in the constraints, i.e.

$$\sum_{j=1}^{p} w_j |\beta_j| \leq t \text{ and } \boldsymbol{w} = \frac{1}{|\widehat{\boldsymbol{\beta}}|^\gamma}, \gamma > 0,$$

where $\widehat{\boldsymbol{\beta}}$ is a root n-consistent estimator to $\boldsymbol{\beta}^*$ ($\widehat{\boldsymbol{\beta}}$(ols) is one of the choice). Zou not only pointed out the flaws of LASSO, also gives necessary conditions under which the LASSO works (see Theorem 1 in Zou [2006]). As noted in Zou [2006], the method of adaptive LASSO truly enjoys the oracle property, in that it assures convergent model with the convergent estimated of coefficients.

Because of its convexity, the L_1 penalty (LASSO or similar) is a popular choice for variable selection. However, it produces biased estimates for the large coefficients as mentioned above. The L_0 penalty,

$$L_0(\boldsymbol{\beta}) = \sum_{j=1}^{p} I(\beta_j \neq 0)$$

is another attractive approach for variable selection because it directly penalizes the number of nonzero coefficients. But the optimization involved is discontinuous and nonconvex, and therefore it is very challenging to implement and the results are not stable (non-identifiable). The support vector machine (SVM) was proposed for classification (Vapnik [1999]), and it is commonly implemented in the field of machine learning. The idea of the standard SVM is to search for the optimal separating hyperplane with maximum separation between two classes, which is closely related to L_2 penalty (Cristianini and Shawe-Taylor [2000]). Bradley and Mangasarian [1998] studied the SVM using the L_1 penalty. The first approach they used was based on constructing a plane in which a weighted sum of distances of misclassified points to the plane is minimized. The second approach used two parallel bounding planes in n-dimensional space R^n and attempted to push the two planes as far apart as possible, which improved generalization for the linear problems. Zhu et al. [2004] preferred the L_1 SVM and argued that the L_1 SVM have advantage when there are redundant noise variables. They proposed an algorithm for calculating the solution path of the L_1 SVM as a function of its tuning parameter. The method has the ability to select relevant variables and ignore redundant variables and does not suffer from the noise inputs as much as L_2 SVM.

Liu and Wu [2007] proposed a new penalty that combines the L_0 and L_1 penalties, and implement the new penalty by developing a global optimization algorithm using mixed integer programming (MIP). They showed that the new penalty retains the advantages of both the L_0 and L_1 penalties, and demonstrated that it can deliver

better variable selection than the L_1 penalty while yielding a more stable model than the L_0 penalty.

As pointed out by Fan and Li [2001], the L_q penalty functions do not simultaneously satisfy the mathematical conditions for unbiasedness, sparsity, and continuity. Fan and Li [2001] proposed the smoothly clipped absolute deviation (SCAD) penalty to overcome the drawbacks of the L_q penalty, and is defined as

$$p_\lambda^{SCAD}(\beta_j) = \begin{cases} \lambda|\beta_j| & |\beta_j| \leq \lambda, \\ -\left(\frac{|\beta_j|^2 - 2a\lambda|\beta_j| + \lambda^2}{2(a-1)}\right) & \lambda < |\beta_j| \leq a\lambda, \\ \frac{(a+1)\lambda^2}{2} & |\beta_j| > a\lambda, \end{cases}$$

for $a > 2$. The SCAD penalty is continuously differentiable on $(-\infty, 0) \bigcup (0, +\infty)$ but singular at 0, and can produce sparse set of solution and approximately unbiased coefficients for large coefficients. This threshold rule involves two unknown parameters λ and a, and the best pair (λ, a) could be obtained using two dimensional grids search using some criteria like cross validation methods. Based on simulation studies, Fan and Li [2001] suggested $a = 3.7$ is a good choice for various problems. They further argued that the performance of various variable selection problems do not improve significantly with a different selection of a.

Besides the above summarized frequentist approaches, there are varieties of studies in the Bayesian framework. Unlike the frequentist framework, Bayesian methods identify the most plausible models in stead of a single model. For example, in linear regression

$$y_i = \sum_{j=0}^{p} \beta_j x_{ij} + \varepsilon_i,$$

indicator variables $\delta_j, j = 0, \ldots, p$ need to inserted, and the regression could be expressed as

$$y_i = \sum_{j=0}^{p} \delta_j \beta_j x_{ij} + \varepsilon_i.$$

Since $\delta_j \in \{0, 1\}$, there are 2^{p+1} models in the model space. Gibbs based variable selection (GVS) (Kuo and Mallick [1998]) specify the hierarchical structure

$$P(\boldsymbol{\delta}, \boldsymbol{\beta}) = P(\boldsymbol{\beta}|\boldsymbol{\delta})P(\boldsymbol{\delta}).$$

The variables are included in the final model if they have a high posterior probability. The stochastic search variable select (SSVS) was first introduced by George and McCulloch [1993], and extended for multivariate case by Brown et al. [1998]. The SSVS method specify a mixture prior for $\boldsymbol{\beta}$ as

$$P(\beta_j|\delta_j) = (1 - \delta_j)N(0, \tau^2) + \delta_j N(0, g\tau^2),$$

where the first density is centered around 0 with small variance. The method could give identifiability for β_j and δ_j, but the fixed prior parameters (τ^2 and $g\tau^2$) are data dependent.

Yi et al. [2003] proposed a Bayesian method for identifying multiple quantitative trait loci in experimental design based on the stochastic search variable selection (SSVS) which was introduced by George and McCulloch [1993]. The SSVS method was firstly developed for linear models and then adopted for generalized linear models (George and McCulloch [1997]), log-linear models (Ntzoufras et al. [2000]), and multivariate regression models (Brown et al. [1998]). The advantages for SSVS method is that the dimensionality is kept constant across all possible models by limiting the posterior distribution of nonsignificant terms in a small neighborhood of zero instead of removing them from the model. So, the SSVS method can be easily implemented via the Gibbs sampler, and evaluate each variable effect on the dependent variable. Theo and Mike [2004] used a quantitative trait loci mapping by variance components method that performs a Bayesian integration over zero, one, two and more quantitative trait loci models to realize the variable selection. The information from all traits are used simultaneously, together with the linkage disequilibrium information to improve the power and precision of quantitative trait loci mapping.

Xu [2003] used a Bayesian method under the random regression coefficient model by taking Jeffrey's prior (normal with mean zero and a unique variance for each gene effect) for simultaneously evaluating quantitative traits loci effects associated with markers of the entire genome. Xu stated that the approach is analogous to the Bayesian method of Meuwissen et al. [2001] for BLUP prediction of gene effects in ourbred populations. Zhang and Xu [2005] also adopted the Jeffrey's prior by developing a penalized likelihood method.

Figueiredo [2003] proposed a Bayesian approach to sparse regression and variable selection, where the advantage is the absence of parameters controlling the degree of sparseness. Figueiredo [2003] built a hierarchical Bayes interpretation of the Laplacian Prior as a normal/independent distribution firstly. Then a Jeffreys' noninformation second-level hyperprior was built which expressed scale invariance. At last, an EM algorithm was applied to estimate the parameters. Tibshirani [1996] suggested that LASSO estimates can be interpreted as posterior mode estimates when the regression parameters have independent and identical Laplace prior. Park and Casella [2008] considered a full Bayesian analysis using a conditional Laplace prior of the form

$$P(\boldsymbol{\beta}|\sigma^2) = \sum_{j=1}^{p} \frac{\lambda}{2\sigma} e^{-\lambda|\beta_j|/\sigma},$$

and the noninformative scale-invariant marginal prior $P(\sigma^2) = \dfrac{1}{\sigma^2}$. The importance of conditioning on σ^2 is to ensure the posterior being unimodal. Yi and Xu [2008] proposed several Bayesian hierarchical models for mapping multiple quantitative trait loci that simultaneously fit and estimate all possible genetic effects associated with all markers across the entire genome. The prior (exponential and scaled inverse-χ^2) distributions for the genetic effects that are scale mixtures of normal distributions with mean 0 and unknown effect-specific variances are used. The exponential prior results in a Bayesian version of the LASSO model, while the scaled inverse-χ^2 prior leads to the Student's t model. Yi and Xu [2008] fit the models in a fully Bayesian approach

by employing the MCCM sampling, and they not only gave the point estimates but also interval estimates of all parameters.

Green [1995] introduced the reversible Markov chain samplers that jump between parameter subspaces of differing dimensionality. The reversible jump MCMC is flexible and entirely constructive, which have wide applicability in variable selection for inference of model having dimension that is not fixed. This method was then used by Sillanpää and Bhattacharjee [2006], Lunn et al. [2006].

For the nonlinear (nonparametric) regression model

$$y_i = f(\boldsymbol{x}_i) + \varepsilon_i,$$

where $f(\cdot)$ is the regression function, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$ is a p-dimension predictors, and $\varepsilon \sim N(0, \sigma^2)$ is the random error, the popular model is smoothing splines analysis of variance (SS-ANOVA),

$$f(\boldsymbol{x}_i) = b + \sum_{j=1}^{p} f_j(x_{ij}) + \sum_{j<k} f_{jk}(x_{ij}, x_{ik}) + \cdots , \tag{1.1}$$

where b is constant, f_j are main effects, f_{jk} are two-way interactions, and so on. The SS-ANOVA provides the generalized additive model, and gives the framework for nonparametric function estimation. Using penalized least squares, with penalty being the sum of squared norms or terms in 1.1, has been a popular approach to estimating SS-ANOVA. In order to determine which variables should be included in the model, Gu [1992] introduced using cosine diagnostics as model checking tools after model fitting in Gaussian regression. Chen [1993] proposed interaction spline models via SS-ANOVA and developed a non-standard test procedure to select variables.Antoniadis and Fan [2001] proposed a group SCAD approach for regularization in wavelets approximation. Gunn and Kandola [2002] investigated a sparse kernel approach. Zhang et al. [2004] proposed a likelihood basis pursuit approach to model selection and estimation in the SS-ANOVA for exponential families. They expanded

each nonparametric component function in 1.1 as a linear combination of a large number of basis functions and applied an L_1 penalty to the coefficients of all the basis functions. However, the sparsity in coefficients improves but not guarantees sparsity in SS-ANOVA components, and a separate model selection has to be performed after model fitting.

Later, Lin and Zhang [2006] introduced a new approach based on penalized least squares with the penalty functional being the sum of component norms rather than the sum of squared component norms, and they called it as component selection and smoothing operator (COSSO). Also, the author proved the existence of the COSSO estimate and the rate of convergence is $n^{-d/(2d+1)}$, where d is the order of smoothness of the components. They pointed out that COSSO reduces to the LASSO when COSSO formulation is used in linear models, and COSSO is a nontrivial extension of the LASSO (in linear models) to multivariate nonparametric models because the penalty used in COSSO is the sum of component norms. They also gave an alternative formulation of the COSSO for efficient computation.

Meier et al. [2009] proposed a penalized least-squares estimator for variable selection and estimation in a nonparametric additive model in which the numbers of zero and nonzero f_j's may both be larger than n. With probability approaching 1, their method selects a set of f_j's containing all the additive components whose distance from zero in a certain metric exceeds a specified threshold. However, the model-selection consistency was not established.

Ravikumar et al. [2009] applied penalty on the L_2 norm of the nonparametric components, as well as the mean value of the components to ensure identifiability. The method required that the eigenvalues of a design matrix (formed from the basis functions for the nonzero components) be bounded away from zero and infinity. However, it is not clear whether this condition holds in general.

Wang et al. [2007] considered the variable selection for varying coefficient mod-

els by combining smoothing spline method with the SCAD procedure, where the time-varying coefficients were represented in terms of B-spline basis functions. They proposed a penalized estimation procedure to select the sets of basis functions, while the penalty produced sparse solutions by thresholding small estimates to zero and provided unbiased estimates for large coefficients. Their method is similar to the group least angle regressions (LARS) of Efron et al. [2004] or group LASSO of Yuan and Lin [2006]. Wang and Xia [2009] considered the group LASSO and SCAD methods in varying coefficient models with a fixed number of coefficients and covariates. Their method is readily implementable to all combinations of other shrinkage methods and nonparametric smoothing methods including the one-step sparse estimator of Zou and Li [2008]. And Bach [2008] established the model-selection consistency under conditions that are considerably complicated.

Semiparametric regression models retain the flexibility of non parametric models and the explanatory power of generalized linear models, and have been extensively studied. The general form of a semiparametric model could be written as

$$y_i = \alpha + \boldsymbol{\beta}' \boldsymbol{x}_i + \varepsilon_i, \tag{1.2}$$

where y_i is the response variable, $\boldsymbol{x}_i$ is a p-dimensional covariate vector, α is an unspecified baseline function, $\boldsymbol{\beta}$ is a vector of unknown regression coefficients, and ε_i is a mean zero noise as usual. Model 1.2 does not require parameterize the baseline function which normally is complicated in practice. Because of the advantage mentioned above, model 1.2 and its variations have been studied thoroughly. Fan and Li [2004] studied the estimation and model selection for semiparametric model in logitudinal data. They proposed two simple, reliable and effective estimation procedures for regression coefficients. Firstly, the difference-based estimator (DBE) provides a simple and good initial estimate of $\boldsymbol{\beta}$ and does not rely on any smoothing techniques. Then, the estimator is refined by the profile lease squares estimator depending on a choice of smoothing parameter. Finally, a wealth of bandwidth selection techniques

is used for model-selection. The author also established the asymptotic normality of the profile least squares, and derived a consistent standard error formula by applying the sandwich formula. The estimation of the baseline function α was also proposed by using local polynomial regression.

Li and Liang [2008] proposed a class of variable selection procedure s for the parametric component of the generalized varying-coefficient partially linear model (GVCPLM). They studied the asymptotic properties of the estimate $\boldsymbol{\beta}$, illustrated the dependence of the rate of convergence and regularization parameters. To select significant variables in the nonparametric component, the author extended generalized likelihood ratio tests (GLRT) by Fan et al. [2001] from fully nonparametric models to semiparametric models. They showed that the limiting null distribution of the semiparametric GLRT does not rely on unknown nuisance parameter and it follows a chi-square distribution with a diveraging degrees of freedom, which allows using asymptotic chi-square distribution or bootstrap method to obtain critical values for the GLRT easily.

The existing methods are not appropriate for GEVs, and lack of the ability to catch the complicated interactions between large amount of available GEVs. We propose a selection method build into semi-parametric models which has ability to account for the complex interaction.

1.3 Bayesian Methods

Bayesian method is based on a mathematical handling of uncertainty, which relies on the Bayes rule

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)},$$

where x might be a data point and θ some model parameters. $P(\theta)$ is the probability of θ and mostly refereed as prior. It represents the prior probability of model parameters before observing any data. $P(x|\theta)$ is the probability of x conditioned on θ and

refereed as likelihood. $P(\theta|x)$ is the posterior probability of θ. $P(x)$ is an integral over all values of θ of the product $P(x|\theta)P(\theta)$ and can be regarded as a normalising constant to ensure that $P(\theta|x)$ is a proper density. Therefore, the Bayes rule is mostly expressed as

$$P(\theta|x) \propto P(x|\theta)P(\theta).$$

The Bayes method has advantage of providing confidence intervals on parameters and probability values on hypotheses that are more in line with commonsense interpretations. It also provides a way of learning from data to update beliefs. The Bayesian approaches are initially proposed by Bayes and Laplace in the 18th century and further developed by modern statisticians in the 20th century, especially after the development of computer-intensive sampling methods of estimation.

Bayesian inference has closely relationship with the sampling-based estimation methods. The Markov chain Monte Carlo (MCMC) methods are a class of algorithms for sampling from probability distributions based on constructing a Markov chain that has the desired distribution as its stationary distribution, which provides estimates of density or probabilities relating to the parameter (Smith and Gelfand [1992]). Monte Carlo posterior summaries typically include posterior means and variances of the parameters. The posterior mean can be shown to be the best estimate of central tendency for a density under a squared error loss function ([Robert, 2012, Chapter 3]), while the posterior median is the best estimate when absolute loss is used.

A $100(1-\alpha)\%$ credible interval for parameter θ is any interval $[a, b]$ of values that has probability $1-\alpha$ under the posterior density of θ. The most common credible interval is the equal-tail credible interval. When MCMC sampling are used, the $(1-\alpha)\%$ confidence interval is estimated using the $\frac{\alpha}{2}$ and $\left(1-\frac{\alpha}{2}\right)$ quantiles of the sampled output $\{\theta^{(t)}, t = B+1, \ldots, T\}$ where superscript (t) denotes the sample at the t^{th} iteration, B is the number of burn-in iteration, and T is the total number of iterations. Another form of credible interval is the $100(1-\alpha)\%$ highest

probability density (HPD) interval, such that the density for every point inside the interval exceeds that for every point outside the interval, and is the shortest possible $100(1-\alpha)\%$ credible interval.

For a given function on the parameter $\Delta = \Delta(\theta)$, the posterior means and variances of such functions obtained from MCMC samples are estimates of the integrals

$$\begin{aligned} E[\Delta(\theta)|x] &= \int \Delta(\theta) P(\theta|x) d\theta, \\ \text{var}[\Delta(\theta)|x] &= \int \Delta^2(\theta) P(\theta|x) d\theta - [E(\Delta(\theta)|x)]^2 \\ &= E(\Delta^2(\theta)|x) - [E(\Delta(\theta)|x)]^2. \end{aligned}$$

Such expectations, density or probability estimates may sometimes be obtained analytically for conjugate analyses, or they can be approximated analytically by expanding the relevant integral (Tierney et al. [1987]). However, such approximations are less good for posteriors that are not approximately normal, or there are multimodalities, and also become impractical for complex multiparameter problems and random effects models. The MCMC technique is relatively straightforward for applications involving sampling from one or more chains after convergence to a stationary distribution which approximates the posterior $P(\theta|x)$. With the increasing of observations or the number of parameters, the required number of iterations to reach stationarity will tend to increase. There are other factors affect the number of iterations, i.e. the number of levels in a hierarchical model, choosing nonlinear rather than a simple linear regression.

The Metropolis-Hastings algorithm (Hastings [1970]) is the baseline for MCMC schemes that simulate a Markov chain $\theta^{(t)}$ with the posterior $P(\theta|x)$ as its stationary distribution. The Metropolis-Hastings algorithm updates the state θ from a distribution for θ with probabilities $P(\theta)$ by using a proposal distribution $g(\theta, \theta^*)$ as follows:

1. Draw a candidate state, θ^* according to the proposal distribution $g(\theta, \theta^*)$.

2. Compute the acceptance probability

$$a(\theta^*, \theta) = \min\left\{1, \frac{g(\theta, \theta^*)}{g(\theta^*, \theta)} \frac{P(\theta^*)}{P(\theta)}\right\} \tag{1.3}$$

3. With probability $a(\theta^*, \theta)$, set the new state θ' to θ^*. Otherwise, let θ' be the same as θ.

The Metropolis-Hastings algorithm keeps P invariant. As pointed out by Hastings [1970], if $\boldsymbol{\theta}$ is multidimensional, proposal distributions that change only one component of $\boldsymbol{\theta}$ are often used. Updates based on several such proposals can be combined in order to obtain an ergodic Markov chain that will converge to P. If the proposal density is symmetric, i.e. $g(\theta^*, \theta) = g(\theta, \theta^*)$, then the Metropolis-Hastings algorithm reduces to the algorithm developed by Metropolis et al. [1953]. The Metropolis-Hastings algorithm works most successfully when the proposal density matches , at least approximately, the shape of the target density $P(\theta|x)$.

The Gibbs sampler (Casella and George [1992], Gelfand and Smith [1990], Gilks et al. [1993]) is a special componentwise Metropolis-Hastings algorithm where the proposals are accepted with probability 1. The Gibbs sampler was originally introduced by Geman and Geman [1984] for Bayesian image reconstruction. Repeated sampling from Gibbs sampler generates an autocorrelated sequence of numbers that, subject to regularity conditions, eventually 'forgets' the staring values $\theta^{(0)}$ and converges to a stationary sampling distribution $P(\theta|x)$.

The full conditional densities may be obtained from the joint density $P(\theta, x) = P(x|\theta)P(\theta)$ and in many cases reduce to standard densities from which sampling is straightforward. When non-standard densities occurred, the Metropolis-Hastings algorithm will be adopted.

There are many unresolved problems around the assessment of convergence of MCMC sampling technique. Some statisticians think that a single long chain is adequate to explore the posterior density, provided allowance is made for dependence

in the samples (Bos [2004], Geyer [1992]. This single long run may be adequate for straightforward problems. The method of Raftery and Lewis [1992] provided an estimate of the number of MCMC iterations required to achieve a specified accuracy of the estimated quantiles of parameters or functions. As pointed out by Draper [2000], the Raftery-Lewis diagnosis include the minimum number of iterations needed to estimate the specified quantile to the desired precision which is a lower bound and may tend to be conservative. Geweke et al. [1991] procedure by adopting chi-square tests considers different portions of MCMC samples to determine whether they can be considered as coming from the same distribution. Geweke's procedure particularly compares the initial and final portions of a chain (e.g. the first 10% and the last 50%) with the tests using sample means and asymptotic variances.

By contrast, many practitioners prefer to use two or more parallel chains with diverse starting values to ensure full coverage of the sample space of the parameters. They agree to use single long run as a preliminary to obtain inputs to multiple chains. Convergence of multiple chains can be assessed using Gelman and Rubin [1992, 1996] scale reduction factors. The variation of samples $\theta_j^{(t)}$ within the j^{th} chain $(j = 1, \ldots, J)$ is defined as

$$w_j = (\theta_j^{(t)} - \bar{\theta}_j)^2/(T-1),$$

where T is the number of iterations after an initial burn-in B iterations. The variation within chains w is the average of the w_j, and the variation between-chain is measured by

$$b = \frac{T}{J-1}\sum_{j=1}^{J}(\bar{\theta_j} - \bar{\theta})^2.$$

Gelman and Rubin [1992, 1996] proposed to compare variation in the sampled parameter values within and between chains.

Another statistic used for multiple-chain convergence is the Brooks-Gelman-Rubin statistic (Brooks and Gelman [1998]), which is a ratio of parameter interval lengths.

The length of the $100(1-\alpha)\%$ interval for parameter θ is obtained for each of the J chains, which provides J within-chain interval lengthes with mean I_U. The same $100(1-\alpha)\%$ interval I_P for the pooled TJ samples is also obtained. Then the ratio I_P/I_U shold converge to 1 if there is convergent mixing over different chains. Brooks and Gelman [1998] also proposed a multivariate version of the original Gelman-Rubin ratio.

The parameter samples from the MCMC technique are correlated, which may affect the convergence in MCMC sampling. The extent of correlation depends on factors like the parameterisation, and the complexity of the model. The analysis of autocorrelation in the sequences of MCMC samples leads to an application on the time series. Also, the correlation between parameters within the parameter set tends to delay convergence and increase the dependence between successive iterations. Zuur et al. [2002] mentioned to center predictor variables, while Robert and Mengersen [1999] suggested a reparameterisation of discrete normal mixtures to improve convergence.

1.4 Existing Methods for Clustering

Variable selection can identify the important or informative variables contributing to the outcome variable. The clustering analysis can further reduce the dimension of the data and make it easier for researchers analyzing data. The clustering problem has attracted much attention in the past few decades. Traditional approaches focus on the clustering of either subjects or (response) variables. However, clusters formed through these approaches possibly lack homogeneity. To cluster objects, there are a varieties of methods:

1. The connectivity based clustering, which is also known as hierarchical clustering, is based on distance between objects. The core of the method is the determination of distance function, and the popular choices include single-linkage clustering, complete linkage clustering, and unweighted pair group method with

arithmetic mean (Sibson [1973], Defays [1977]). These methods are fairly easy to understand, but the results are not always easy to use.

2. Centroid based clustering, which needs to find centers of each cluster and decide the partition. The most popular algorithm is called K-means (MacQueen et al. [1967], Lloyd [1982]), and the common approach is to search only for approximate solutions. The biggest drawback of this algorithm is the number of clusters needs to be pre-specified.

3. Density based clustering is based on areas of high density then the remainder of the data set. The most popular density based clustering method is DBSCAN (Ester et al. [1996]). The method cluster points that satisfy a density criterion but not the distance thresholds. The method could form a cluster of any shape, which is much different from other approaches. And, the complexity of DBSCAN is fairly low. The disadvantages of density based clustering are requiring of density drop to detect cluster borders, and lack of ability to detect intrinsic cluster structures.

4. Distribution based clustering partitions the objects most likely from the same distribution. One method is to utilize the Dirichlet process (Neal [2000]), which models the data set with a number of Gaussian distributions that are initialized randomly and whose parameters are iteratively optimized to fit better to the data set. This algorithm can not only provide the clusters, also capture correlation and dependence of attributes. However, users need to choose appropriate model to optimize.

In order to perform a 2-way clustering, bi-clustering (or co-clustering, 2-mode clustering) was introduce by Hartigan [1972]. It was developed to allow simultaneous clustering of the rows and columns of a matrix, and generated subset of rows which exhibit similar behavior across a subset of columns, or vice versa. The bi-clustering

was applied to clustering microarray data, and varieties of different algorithms and methods were developed to process the bi-clustering. Madeira and Oliveira [2004], Prelić et al. [2006] compared lots of the bi-cluster algorithms. Kaiser and Leisch [2008] contributed an R package `biclust` (R Core Team [2012]) which adopted the plaid model (introduced by Lazzeroni and Owen [2002], and improved by Turner et al. [2005]), Bimax algorithm (Prelić et al. [2006]), Xmotifs algorithm (Murali and Kasif [2003]), algorithms by Cheng and Church [2000], and Kluger et al. [2003]. However, the bi-clustering considered the grouping of rows and columns separated, and finally overlapped the rows and columns clustering together to form the 2-way clustering. It ignored the possible association between rows and columns, and could result inaccurate grouping.

All aforementioned clustering methods focus on data description without considering the assumption of variables with external variables. Now we will review the model-based clustering. The cluster analysis can be based on probability models (see Bock [1996], Bock et al. [1998] for a survey). This probabilistic approach provided insight into when the data conform to the model, and led to the development of new clustering methods. It has been shown that some of the non-model-based clustering methods are approximate estimation methods for certain probability models. For example, standard k-means clustering and Ward's method (Ward Jr [1963]) are equivalent to known procedures for approximately maximizing the multivariate normal classification likelihood when the covariance matrix is the same for each component and proportional to the identity matrix.

Finite mixture models have been proposed and investigated in the context of clustering (Scott and Symons [1971], Duda et al. [1973], Binder [1978]). Given data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ being independent multivariate observations, the likelihood for a mixture model with C component is

$$L(\theta_1, \ldots, \theta_g; \tau_1, \ldots, \tau_C | \boldsymbol{y}) = \prod_{i=1}^{n} \sum_{k=1}^{C} \tau_k f_k(\boldsymbol{y}_i | \theta_k),$$

where f_k and θ_k are the density and parameters of the kth component in the mixture and τ_k is the probability that an observation belongs to the kth component ($\tau_k \geq 0; \sum_{k=1}^{C} \tau_k = 1$). Often, f_k is the multivariate normal density parameterized by mean $\boldsymbol{\mu}_k$ and covariance matrix Σ_k. Data generated by mixtures of multivariate normal densities clustered into groups centered at the mean $\boldsymbol{\mu}_k$. The shape, volume and orientation of the clusters are controlled by the covariances Σ_k, which may also be parameterized to impose cross-clusters constraints. Common instances include $\Sigma_k = \lambda I$, where all clusters are of the same size, and only one parameter is required to characterize the covariance structure of the mixture; $\Sigma_k = \Sigma$ constant across clusters, where $d(d+1)/2$ parameters are needed if the data are d-dimensional (Friedman and Rubin [1967]); and unrestricted Σ_k, where $C(d(d+1)/2)$ parameters are required (Scott and Symons [1971]).

Banfield and Raftery [1993] generalized the cross-cluster constraints in multivariate normal mixtures by proposing covariance matrices through eigenvalue decomposition in the form $\Sigma_k = \lambda_k D_k A_k D_k'$, where D_k is the orthogonal matrix of eigenvectors, A_k is a diagonal matrix whose elements are proportional to the eigenvalues, and λ_k is an associated constant of proportionality. The idea lies on treating λ_k, A_k, and D_k as independent sets of parameters and allowing parameters to be the same or vary among clusters. The parameters control geometric shape of each clusters in the $d-$dimensional space. The approach is the generalization of the work by Murtagh and Raftery [1984], in which covariance matrices are restricted to be the same across clusters. There are many other parsimonious parameterizations of covariances matrices, i.e. the intraclass correlation or one-factor model by Jöreskog [1970], where all of the off-diagonal elements of the correlation matrix are equal; autoregressive and other parameterizations common in time series (Box et al. [2011]), and models common in geostatistics in which covariances are functions of distance in either Euclidean (Journel and Huijbregts [1978]) or deformed space (Sampson and Guttorp [1992]).

To identify clusters, the expectation-maximization (EM) algorithm has been apopted by (Dempster et al. [1977], McLachlan and Krishnan [2007]) by treating cluster assignment of each subject as missing values. Denote by $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iC})$ a vector of latent indicators composed of one 1 and $C-1$ zeros, i.e.

$$z_{ik} = \begin{cases} 1 & \text{if } \boldsymbol{x}_i \text{ belongs to group } k \\ 0 & \text{otherwise;} \end{cases}$$

denoting the cluster assignment of subject i. Consequently, $\boldsymbol{x}_i = (\boldsymbol{y}_i, \boldsymbol{z}_i)$ are the "complete data" in the frequentist framework. Assuming that each $\boldsymbol{z}_i$ is independently and identically distributed according to a multinomial distribution of one draw from C groups with probabilities $\tau_1 \ldots, \tau_C$. The EM algorithm alternates between two steps, an "E-step", in which the conditional expectation of the complete-data log-likelihood given the observed data and the current parameter estimates is computed, and an "M-step", where parameters that maximize the expected log-likelihood from the E step are calculated. Under some mild regularity conditions, EM algorithm can converge to a local maximum of the observed data likelihood (Dempster et al. [1977], Boyles [1983], Wu [1983], McLachlan and Krishnan [2007]). For the clustering EM mixture model, assume the density of an observation $\boldsymbol{y}_i$ given $\boldsymbol{z}_i$ is given by $\prod_{k=1}^{k} f_k(\boldsymbol{y}_i|\theta_k)^{z_{ik}}$, the complete data log-likelihood is

$$l(\theta_k, \tau_k, z_{ik}|\boldsymbol{x}) = \sum_{i=1}^{n}\sum_{k=1}^{C} z_{ik} \log\Big(\tau_k f_k(\boldsymbol{y}_i|\theta_k)\Big).$$

The E step of the EM algorithm for mixture models is given by

$$\widehat{z}_{ik} \leftarrow \frac{\widehat{\tau}_k f_k(\boldsymbol{y}_i|\widehat{\theta}_k)}{\sum_{j=1}^{C} \widehat{\tau}_k f_j(\boldsymbol{y}_j|\widehat{\theta}_j)}.$$

The M step involves maximizing the log-likelihood in terms of τ_k and θ_k with z_{ik} fixed at the values computed in the E step, $\widehat{z}_{ik}$. The EM estimation for mixture models has a number of limitations: 1) the rate of convergence can be slow; 2) the EM algorithm for multivariate normal mixtures breaks down when the covariance

associated with one or more components is singular or nearly singular, and it may fail or even give inaccurate results. Celeux and Govaert [1992] proposed a variant of EM called classification EM, in which z_{ik} are converted to a discrete classification before performing the M step. The method is equivalent to standard k-means clustering (MacQueen et al. [1967], Hartigan and Wong [1979]).

There are many recent advances in clustering made in a Bayesian framework, which allows simultaneous estimation of many interdependent parameters in complex models. As mentioned in Latch et al. [2006], there are three most widely used Bayesian clustering methods, PARTITION (Dawson and Belkhir [2001]), STRUCTURE (Pritchard et al. [2000], Falush et al. [2003], Pritchard and Wen [2003]), and BAPS (Corander et al. [2003, 2004, 2006]). All three methods use the minimization of the Hardy-Weinberg and linkage disequilibrium, which would result incorrectly grouping individuals from different, randomly-mating populations into a common population.

Dawson and Belkhir [2001] estimated the number of clusters in a sample by employing a Markov chain Monte Carlo method to generate an estimate of the posterior distribution of the sample partition. They assumed that individuals are pure ancestry. The STRUCTURE method employed an ad hoc approach for inferring the number of clusters. The author of STRUCTURE method derived the posterior distribution of the number of clusters from separate MCMC chains, each with a different fixed value of the number of clusters. The BAPS uses a greedy stochastic optimization algorithm (Fletcher [2013]) to search for the most likely number of clusters. Unlike the PARTITION, both STRUCTURE and BAPS allow individuals to be of mixed ancestry.

One important issue arising in applying clustering is determination of the number of clusters. McLachlan and Basford [1988] proposed to use of resampling to determine the number of clusters in model-based clustering. Banfield and Raftery [1992] derived

an approximation to the integrated likelihood based on the classification likelihood, but it subsequently showed performing less well then BIC. Later, several other approximations were proposed to the integrated likelihood and the performance of these criteria were compared by Biernacki and Govaert [1999].

Another tool for clustering is the Dirichlet process mixture models that has been a popular approach of identifying latent classes and can explain the dependencies observed between variables, and it became computationally feasible with the development of Markov chain methods from the posterior distribution of the parameters. The Dirichlet process could dynamically determine the number of clusters and simultaneously obtain the clustering assignment on observations. The general form the Dirichlet process mixture model could be express as

$$\begin{aligned} y_i|\boldsymbol{\theta}_i &\sim F(\boldsymbol{\theta}_i) \\ \boldsymbol{\theta}_i|G &\sim G \\ G &\sim \mathrm{DP}(G_0, \alpha) \end{aligned} \tag{1.4}$$

where $\boldsymbol{\theta}_i$ is the set of model parameters. "$X \sim S$" means "X has the distribution S". The distribution from which the y_i are drawn is a mixture of distribution s of the form $F(\boldsymbol{\theta})$ with th mixing distribution over $\boldsymbol{\theta}$ being G. The prior of the mixing distribution is a Dirichlet process (Ferguson [1973]), with concentration parameter α and base distribution G_0. Additionally, the distributions F and G_0 will depend on extra hyperparameters for specific model not mentioned in the above general form of Dirichle process mixture model, and those hyperparameters along with concentration parameter α will be given priors at a higher level.

The Gibbs sampling can easily be implemented for models based on conjugate prior distributions. However, if non-conjugate priors are used, straightforward Gibbs sampling requires performing difficult numerical integration. West and Escobar [1993] introduced to use a Monte Carlo approximation to the integral, and in many contexts

pointed out likely relative large error. MacEachern and Müller [1998] proposed an exact approach for handling non-conjugate priors that uses a mapping from a set of auxiliary parameters. The introduced "no gaps" and "complete" algorithms in MacEachern and Müller [1998] are widely applicable, but were mentioned somewhat inefficient (Neal [2000]). Walker and Damien [1998] introduced a different auxiliary variable method to some Dirichlet process mixture models. However, Waker and Damien's method is unsuitable for general use, additionally requires the computation of a difficult integral.

Since the realization of Dirichlet process are discrete with probability one, the Dirichlet process mixture models can be viewed as countably infinite mixtures (Ferguson [1973]). In Blackwell and MacQueen [1973], the author expressed the Dirichlet process as

$$\boldsymbol{\theta_i}|\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_{i-1} \quad \sim \quad \frac{1}{i-1-\alpha}\sum_{j=1}^{i-1}\delta(\boldsymbol{\theta}_j)+\frac{\alpha}{i-1-\alpha}G_0,$$

where $\delta(\boldsymbol{\theta})$ is the distribution concentrated at the single point $\boldsymbol{\theta}$, $pR+(1-p)S$ represented the distributions that is the miture of R and S with proportions p and $1-p$, respectively. Since the independency and exchangeable of data, by treating that i is the last of the observation, the above Dirichlet process could be written as

$$\boldsymbol{\theta}_i|\boldsymbol{\theta}_{-i} \quad \sim \quad \frac{1}{n-1-\alpha}\sum_{j\neq i}\delta(\boldsymbol{\theta}_j)+\frac{\alpha}{n-1-\alpha}G_0,$$

where n is the number of observation, $\boldsymbol{\theta}_{-i}$ denotes all $\boldsymbol{\theta}_j$'s for $j\neq i$. By specifying conjugate prior on the Dirichlet process, and obtaining direct Gibbs sampling, Escobar [1994], Escobar and West [1995] produced an ergodic Markov chain. However, the convergence to the posterior distribution may be slow, and sampling thereafter may be inefficient.

By taking into account of the similarity of observations within each cluster, Bush and MacEachern [1996], West and Escobar [1993] proposed a more efficient method to only draw one posterior for each cluster, which improved lots on the efficiency.

If conjugate prior are used, the steps could be even reduced by integrating out the parameters in the posterior to obtain the clustering assignment quick. This is the algorithm presented by MacEachern [1994]. For non-conjugate priors, MacEachern and Müller [1998] proposed a framework that allows auxiliary values drawn from G_0 to be included to define a valid Markov chain. The "no gap" algorithm was developed and can be applied to any model for which we can sample from G_0, and compute the likelihood. Because of assigning an observation to a newly created mixture component, the "no gaps" algorithm is a little inefficient in the algorithm mechanism. MacEachern [1994] also proposed an algorithm based on a "complete" mapping. It requires maintaining n values of model parameters for each observation, which is inefficient when the $k \ll n$ where k is the number of clusters.

Neal [2000] proposed three algorithms to handle non-conjugate priors cases by using Metropolis-Hastings updates. The first two algorithms used more than one metropolis-Hastings update for each of the clustering assignment of each observation. As pointed by Neal himself, it is difficult to say which behavior is better. However, since small values of concentration parameter α is small (around one) are often used, we might wonder whether al algorithm that could consider the creation of a new component more often might be more efficient. In Neal [2000], the author also proposed another algorithm by combining the Metropolis-Hastings updates and partial Gibbs sampling updates. A new algorithm by the addition of m auxiliary parameters, similar to that of MacEachern and Müller [1998], was also presented in Neal [2000]. The difference is that the auxiliary parameters are regarded as existing only temporarily, which allows more flexibility in constructing algorithms. When $m = 1$, the algorithm closely resembles the "no gaps" algorithm of MacEachern and Müller [1998]. As $m \to \infty$, the algorithm approaches the behavior of algorithm of Bush and MacEachern [1996], West and Escobar [1993]. Since the flexibility of this algorithm, Jara et al. [2011], Jara [2007] adopted it into the R (R Core Team [2012]) package

`DPpackage`.

In the Dirichlet process mixture model, the concentration parameter α plays an important role. It controlled the probability of new cluster during calculation. A larger α might suggest more clusters in the final stage, while smaller α corresponds fewer clusters. Dorazio et al. [2008] pointed that the posterior inferences appeared sensitive to the choice of priors of α. The Gibbs samples of α depends only on n, the number of observations, and C, the number of clusters, which allows the distribution of α to exert considerable influence on the posterior. This prior sensitivity in estimating α has drew attention by varieties of research groups (Liu [1996], McAuliffe et al. [2006]). Liu [1996] suggested an empirical Bayes approach wherein posterior inferences are computed while conditioning on a maximum likelihood estimate (MLE) of α. The Theorem 4 of Liu [1996] established that the MLE of α satisfies an equivalence between the conditional prior and conditional posterior means of C, which is a random variable for the number of distinct clusters. Liu [1996] also proved that

$$E(C|n, \widehat{\alpha}, \text{data}) = E(n, \widehat{\alpha}) = \sum_{i=1}^{n} \frac{\widehat{\alpha}}{\widehat{\alpha} + i + 1},$$

where $\widehat{\alpha}$ denotes the MLE of α. The computation steps are summarized below:

1. Use Gibbs sampling to generate a sample of the posterior conditioned on a fixed value of α. McAuliffe et al. [2006] suggested the initial value of α to be a guess between $1/\log n$ and $n/\log n$.

2. The posterior mean $E(C|n, \widehat{\alpha}, \text{data})$ is approximated by averaging the values of C in the posterior sample, $\bar{C} = \dfrac{1}{R}\sum_{r=1}^{R} C^{(r)}$, where $C^{(r)}$ is the simulated value of C in the rth draw of a total of R posterior sample.

3. Compute the value of α that satisfies $\bar{C} = \sum_{i=1}^{n} \alpha/(\alpha + i + 1)$.

4. Repeat Steps 1–2 using the value of α obtained in Step 3 until convergence.

For the method, the MLE of α is obtained by alernating between inference and estimation steps, while inferences for all other parameters are computed by conditioning on $\widehat{\alpha}$. Some other analysts adopted a prior for α centered at $\widehat{\alpha}$ (Jara et al. [2011]).

The Dirichlet process mixture model is a good choice for clustering. It does not require specify the number of clusters, while some other methods may require to fix the number of clusters before inference. Dirichlet process determines the number of clusters dynamically based on the structure of data. However, it only handles clusters of observations. Other methods of clustering are focusing on grouping observations. The bi-clustering could handle the two way data and group both horizontally (rows) and vertically (columns). But the bi-clustering is a non-model based method, and it restricts to the data only and not considers the model. Furthermore, the bi-clustering does the job by rows and columns separately, then overlaps the two sets of clusters to form the two way clustering, which ignores the dependency between rows and columns. Here we will propose a joint clustering model that considers the correlation between the rows and columns. Additionally, the model involves predictors.

1.5 Outline

The dissertation is organized as follows. In Chapter 2, we will propose a linear mixed model based on reproducing kernel to reveal the unknown relationship between outcomes and predictors. The previous work under the kernel machine framework is reviewed and then a mixed model based on the kernel will be proposed. Algorithm to select the informative variables are given, and followed by a simulation study to demonstrate the performance of the proposed method, and comparison against some existing methods. Finally, the proposed method will be applied to two real data sets and give the inferences.

In Chapter 3, we will propose a joint clustering model to group both the outcome

and observations. We will incorporate the penalized splines to reveal the non-additive relation between outcomes and predictors. The clustering of the dependent variables will be realized firstly. Then within each identified dependent cluster, the observations are further clustered by adopting Dirichlet process mixture model. Number of observations clusters is determined dynamically by Dirichlet process itself. The detailed steps to perform the joint clustering are then presented, and followed by a simulation study to verify the performance of the proposed model. Then, we will study a real data set by applying the proposed method.

In Chapter 4, we will discuss the possible shortage of the proposed linear mixed model for variable selection, and joint clustering model. Some interesting directions for further research on the two topics are presented.

Chapter 2

Variable Selection

The work in this chapter is motivated by a collaborative project aiming to test the effect of a set genetic or epigenetic variants (GEVs) and select important variants that are associated with disease risk.

2.1 Introduction

The present work is motivated by a collaborative project aiming to test the effect of a set of genetic or epigenetic variants (GEVs) and select important variants that are associated with a health outcome of interest. With the technological developments, large numbers of genetic or epigenetic variants are allowed to be deployed. The genome-wide association studies have been a popular tool and emphasized testing the effect of individual variants via linear or generalized linear models. Although this individual analysis has been proven useful in identifying disease-susceptibility variants for breast cancer (Easton et al. [2007], Hunter et al. [2007]), prostate cancer (Yeager et al. [2007], Gudmundsson et al. [2007]), and type 2 diabetes (Sladek et al. [2007], Scott et al. [2007]), this mode of analysis lack the ability to detect the significance of a large set of variants and complex interactions between the variants. To test the significance of a variants set (set analysis), Liu et al. [2007], Wu et al. [2010] proposed an approach built upon reproducing kernels (Kimeldorf and Wahba [1970], O'Sullivan et al. [1986]). This type of analysis has the ability to handle a large number of variants and takes into account the effects of interaction of any unknown form between the variants. It is considered that GEVs do not function individually, rather, they work in

collaboration with others to manifest a disease condition. Thus the reproducing kernel based set analysis method is an appropriate choice for the examination of GEVs' set effect on health outcomes. However, this set analysis relies on the correct prior selection of the variants, and the resulting conclusion can be misleading if the prior selection lacks proper justification (He et al. [2012]). Furthermore, this method is not able to identify which variants are the true contributors. Selecting important genetic variants such as single nucleotide polymorphisms (SNPs) or epigenetic variants, e.g., deoxyribonucleic acid (DNA) methylation of a set of CpG sites, is critical to disease intervention.

Due to the possibly complex and usually unknown form of association between GEVs and an outcome, existing variable selection methods applied in linear or non-linear models may not be applicable to the selection of GEVs. Most methods focused on selecting variables in linear models. For instance, the method built upon noncon-cave penalized likelihood with a smoothly clipped absolute deviation (SCAD) penalty introduced by Fan and Li [2001], Li and Liang [2008], the method built upon ridge regression by Frank and Friedman [1993], the least absolute shrinkage and selection operator (LASSO) by Tibshirani [1996], and the closely related approach adaptive LASSO by Zou [2006]. Recently, some methods for feature selections in non-linear models are developed (Rech et al. [2001], Radchenko and James [2010], Castle and Hendry [2012]). These methods are generally built on splines or Taylor series expansions and may have difficulty in describing complex interaction effects. In addition, they are not appropriate for discrete variables such as single nucleotide polymorphisms (SNPs). In the area of machine learning, variable selection in semi-parametric models constructed using reproducing kernels has been discussed in Rosasco et al. [2010], although it is also limited to continuous variables and requires intensive computing. Support vector machines recursive feature elimination (Guyon et al. [2002]), another machine learning technique commonly used in gene selection, is a non-parametric ap-

proach built upon discriminant analyses, which is suitable to binary outcomes. The non-parametric approach random forest (RF) introduced by Breiman [2001], on the other hand, is more flexible. Outcome variables and predictors can be of any type, continuous or categorical. Importance values in RF, calculated essentially based on distance reduction of each node in a tree, are used to determine if a variable is important enough to be kept. However, it is unclear if this approach performs better than other variable selection approaches. Besides the frequentist approaches, there exist various methods and algorithms for variable selection in the Bayesian framework i.e. O'Hara and Sillanpää [1970]; for instance, the stochastic search variable selection method built upon a mixture prior distribution for the regression coefficients by George and McCulloch [1993] and methods utilizing g-priors (Zellner [1986], Smith and Kohn [1996], Liang et al. [2008]). However, these methods are also constructed for linear models with continuous variables.

In this chapter, we propose a simple method to select variables through set analyses that utilizes reproducing kernels to evaluate the relationship that is possibly non-linear and complex between the independent variables or predictors and the dependent variable (Liu et al. [2007], Wu et al. [2010]). Unlike previous non-linear variable selection methods, the proposed selection procedure can be used to select categorical variables, such as SNPs, and continuous variables, such as methylated CpG sites that are potentially associated with a disease. The remaining of the chapter is organized as follows. The Method section introduces the model in a reproducing kernel framework, discusses some popular choices of kernels, and presents a detailed procedure of selecting variable in reproducing kernels. We demonstrate and evaluate the performance of the proposed method in the Results section through simulations. To illustrate the method, we apply the proposed approach to two data sets: one to identify CpG sites that are potentially associated with active smoking, and the other to select SNPs that are possibly associated with lung function. We summarize our

methods and findings in the Conclusion section.

2.2 The Model

Suppose we observe a vector of responses $\boldsymbol{y}_{n\times 1}$, a matrix of GEVs $\boldsymbol{z}_{n\times g}$ whose joint (overall) effect is of interest (e.g., DNA methylation in a pathway, or SNPs), and a covariate matrix $\boldsymbol{X}_{n\times q}$. Here n is the sample size, g is the number of GEVs, and q is the number of covariates. We assume that the mean of the response is modeled as

$$E(y_i|\boldsymbol{X}_i, \boldsymbol{z}_i) = f^{-1}\{\boldsymbol{X}_i\boldsymbol{\beta} + h(\boldsymbol{z}_i)\},$$

where $h(\cdot)$ is an unknown function, and $\boldsymbol{\beta}_{q\times 1}$ describes the additive linear effects of q covariates $\boldsymbol{X}$. Define $\boldsymbol{h}(\boldsymbol{z})_{n\times 1}$ to be a vector of unknown functions evaluating the joint effect of g GEVs ($\boldsymbol{z}$) that is possibly non-linear and may involve complex interactions between $\boldsymbol{z}$; $\boldsymbol{h}(\boldsymbol{z})$ can be modeled parametrically or non-parametrically. Function $f(\cdot)$ is a known link function. For instance, $f(\cdot)$ being the identity function results in a partially linear model and the inverse of a probit function gives a probit regression model. In this work, we take the identity link and $\boldsymbol{y}$ being continuous. Denote by ϵ_i the random error between y_i and $E(y_i|\boldsymbol{X}_i, \boldsymbol{z}_i)$ and assume $\epsilon_i \sim N(0, \sigma^2)$. Our goal is to select a subset of GEVs that have legitimate contributions to the joint effect and exclude variables with no contributions.

As noted above, we allow the g variables $\boldsymbol{z}$ to have a complex (interaction) effect on the response variable. In practice, this is particularly true among genes or epigenes functioning in the same pathway. To this end, we incorporate reproducing kernels into the modeling process in appreciation of their ability to describe any underlying unknown patterns and the ability of handling large number of variables as in Liu et al. [2007], Wu et al. [2010]. Specifically, we represent $h(\cdot)$ using a kernel function $K(\cdot, \cdot)$. By the Mercer's theorem (Mercer [1909], Cristianini and Shawe-Taylor [2000]), under some regularity conditions, the kernel function $K(\cdot, \cdot)$ specifies a unique function

space $\mathcal{H}$ spanned by a particular set of orthogonal basis functions. The orthogonality is defined with respect to L_2 norm. Following the Mercer's theorem, any function $h(\cdot)$ in the function space $\mathcal{H}$ can be represented as a linear combination of reproducing kernels as in Cristianini and Shawe-Taylor [2000], González-Recio et al. [2008],

$$h(\boldsymbol{z}_i) = \sum_{k=1}^{n} K(\boldsymbol{z}_i, \boldsymbol{z}_k)\alpha_k = \boldsymbol{K}_i'\boldsymbol{\alpha},$$

where $\boldsymbol{\alpha} = (\alpha_k, k = 1, \cdots, n)'$ is a vector of unknown parameters and $\boldsymbol{K}_i'$ is the ith row of kernel matrix $\boldsymbol{K}$. Defining $h(\cdot)$ non-parametrically as above has two major advantages in that it can handle large number of covariates and can capture potentially complex interaction between variables $\boldsymbol{z}$ via the specified kernel function.

The kernel function $K(\cdot, \cdot)$ determines the space of functions used to approximate the function $h(\cdot)$. Using appropriate kernel functions for different types of data has the potential to increase the efficiency of the estimating process. For GEVs that are continuous such as gene expression or DNA methylation levels, there are two commonly used kernel functions, the polynomial kernel and the Gaussian kernel. A dth polynomial kernel is defined as

$$K(\boldsymbol{z}_i, \boldsymbol{z}_j) = (\boldsymbol{z}_i^T\boldsymbol{z}_j + \rho)^d,$$

where ρ and d are tuning parameters. The dth polynomial kernel corresponds to models with dth order polynomials including the cross product terms. For example, when $d = 1$, the corresponding model is a linear regression that only has main effects. A Gaussian kernel is in the form of

$$K(\boldsymbol{z}_i, \boldsymbol{z}_j) = \exp(-\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2/\rho),$$

with $\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2 = \sum_{k=1}^{g}(z_{ik} - z_{jk})^2$ and ρ being a tuning parameter. The functionality of Gaussian kernels is similar to that of exponential and Laplacian kernels (see Shawe-Taylor and Cristianini [2004], Zhang and Gan [2012]). All these kernels are constructed for continuous variables. A Gaussian kernel approaches to a first order

polynomial kernel when ρ approaches infinity. Gaussian kernels are more flexible to underlying joint effects of $h(\boldsymbol{z}_i)$.

For GEVs that are discrete such as SNPs, the identity-by-state allele sharing (IBS) kernel by Mountain and Cavalli-Sforza [1997], Wessel and Schork [2006] and the exponential kernel built upon similarity matrix by González-Recio et al. [2008] are commonly used. In genetics, IBS kernel is defined based on the agreement of alleles between subjects. Denote by g the number of loci, $\boldsymbol{z}_{ik}$ and $\boldsymbol{z}_{jk}$ the genotypes of individuals i and j, respectively, at the kth locus ($k = 1, \ldots, g$); $s^k_{i,j}(z_{ik}, z_{jk})$ is a function mapping the genotype variants, for individuals i and j at locus k. It has a value of 0 if individuals i and j are homozygous for different alleles (e.g., $z_{ik} = AA$ and $z_{jk} = TT$), a value of 1 if they share one allele (e.g., $z_{ik} = AA$ and $z_{jk} = AT$), and a value of 2 if they share both alleles (e.g., $z_{ik} = AA$ and $z_{jk} = AA$). The IBS kernel is constructed as an average of agreement between subjects i and j,

$$K(\boldsymbol{z}_i, \boldsymbol{z}_j) = \sum_{k=1}^{g} s^k_{i,j}(z_{ik}, z_{jk})/(2g).$$

Allele frequencies can also be included as a weighting parameter to improve the kernel, and the weighted IBS kernel is defined as

$$K(\boldsymbol{z}_i, \boldsymbol{z}_j; \boldsymbol{w}) = \sum_{k=1}^{g} \frac{1}{\sqrt{w_k}} s^k_{i,j}(z_{ik}, z_{jk})/(2g),$$

where w_k can be the minor allele frequency for the k^{th} SNP in the SNP set. Similarity-based exponential kernel was first introduced by González-Recio et al. [2008]. The frequencies, denoted by f_{ks}, at locus k of genotype s (with $s = 1, 2$, or 3 being one of the three possible genotypes, eg., AA, AT, or TT) is calculated first. Then, initialize the $(i, j)^{th}$ entry of the similarity matrix $S_{ij} = 0$ and update it as the following

$$\text{if} \begin{cases} z_{ik} = z_{jk} \Rightarrow \text{subscore} = \text{subscore} \times f_{ks} \\ \\ z_{ik} \neq z_{jk} \Rightarrow S_{ij} = S_{ij} + \text{subscore}; \quad \text{subscore} = 1. \end{cases}$$

for k from 1 to g. The similarity-based exponential kernel is expressed as

$$K(\boldsymbol{z}_i, \boldsymbol{z}_j) = \exp(-S_{ij}).$$

Comparing the IBS kernel and the exponential kernel, the IBS kernel is more sensitive to the underlying differences in genetic variants between subjects.

In this article, we consider the Gaussian kernel for continuous variants such as gene expressions or DNA methylation, due to its flexibility and its ability in modeling complex functions (Liu et al. [2007]), and the IBS kernel for discrete variants such as SNPs because of its robustness to vanish matrix singularity.

2.3 Parameter Estimation and Score Tests for τ

To estimate parameter $\boldsymbol{\beta}$ and evaluate the joint effect of GEV variants $\boldsymbol{z}$, a penalized least squares method with L_2 penalty on $h(\boldsymbol{z}_i)$ is proposed by Liu et al. [2007], denoted as the least squares kernel machine (LSKM). As shown in Liu et al. [2007], the estimating process is equivalent to maximizing a likelihood function constructed through a linear mixed model by treating $h(\boldsymbol{z}_i)$ as a random effect and assuming

$$\{h(\boldsymbol{z}_i), i = 1, \cdots, n\} \sim \boldsymbol{N}(\boldsymbol{0}, \tau K),$$

with τ being a regularization parameter. Parameter τ evaluates the joint effect of $\boldsymbol{z}$, the main focus of this method.

To infer $\boldsymbol{\beta}$, τ, and σ^2, the method of restricted maximum likelihood (REML) is preferred due to its unbiasness property on estimating variance parameters. The likelihood function under REML is written as (see Liu et al. [2007])

$$l_R = -\frac{1}{2}\log|V(\boldsymbol{\theta})| - \frac{1}{2}\log|X^T V^{-1}(\boldsymbol{\theta})X| - \frac{1}{2}(\boldsymbol{y} - X\boldsymbol{\beta})^T V^{-1}(\boldsymbol{\theta})(\boldsymbol{y} - X\boldsymbol{\beta}), \quad (2.1)$$

where $V = \sigma^2 I + \tau K$, the variance of $\boldsymbol{y}$. Vector $\boldsymbol{\theta}$ is a collection of parameters, $\boldsymbol{\theta} = (\tau, \rho, \sigma^2)^T$ for continuous data, and $\boldsymbol{\theta} = (\tau, \sigma^2)^T$ for discrete data due to different

choices of kernels. Estimation of $\boldsymbol{\theta}$ proceeds by maximimizing the likelihood function (2.1). It is easy to derive that the estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (X^T V^{-1} X)^{-1} X^T V^{-1} \boldsymbol{y}.$$

For the parameters in $\boldsymbol{\theta}$, a numerical approach such as the Newton Raphson method is needed to maximize the likelihood and infer the parameters.

Following the estimate of τ is to test

$$\begin{aligned} H_0 &: \tau = 0 \\ H_1 &: \tau > 0, \end{aligned}$$

that is, whether the genetic or epigenetic variants $\boldsymbol{z}$ as a whole unit contribute significantly to the outcome of interest. Liu et al. [2007], Zhang and Lin [2003] proposed a score test built upon residuals. The score statistic of τ under H_0 can be written as

$$Q_\tau(\hat{\boldsymbol{\beta}}, \hat{\sigma^2}, \cdot) - \mathrm{tr}\{P_0 K\},$$

where $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma^2}$ are the maximum likelihood estimates of β and σ^2, respectively, under the null linear model

$$\begin{aligned} y_i &= \boldsymbol{x}_i \boldsymbol{\beta} + e_i \\ P_0 &= I - X(X^T X)^{-1} X^T, \end{aligned}$$

"$\cdot$" denotes any parameters unique to a kernel, and

$$Q_\tau(\boldsymbol{\beta}, \sigma^2, \cdot) = \frac{1}{2\sigma^2} (\boldsymbol{y} - X\boldsymbol{\beta})^T K (\boldsymbol{y} - X\boldsymbol{\beta}).$$

Under H_0, the distribution of $Q(\cdot)$ is approximated by a scaled χ^2 distribution $\kappa \chi^2_\nu$

with

$$\begin{aligned}
\kappa &= \tilde{I}_{\tau\tau}/2\tilde{e}, \\
\tilde{\nu} &= 2\tilde{e}^2/\tilde{I}_{\tau\tau}, \\
\tilde{I}_{\tau\tau} &= I_{\tau\tau} - I_{\tau\sigma^2} I^{-1}_{\sigma^2\sigma^2} I_{\tau\sigma^2}, \\
I_{\tau\tau} &= \mathrm{tr}\{P_0 K\}^2/2, \\
I_{\tau\sigma^2} &= \mathrm{tr}\{P_0 K P_0/2\}, \\
I_{\sigma^2\sigma^2} &= \mathrm{tr}\{P_0^2\}/2, \\
\tilde{e} &= \mathrm{tr}\{P_0 K\}/2.
\end{aligned}$$

It is worthy noting that, when genetic or epigenetic variants $\boldsymbol{z}$ are continuous such as gene expression levels or methylation measures, tuning parameter ρ in the Gaussian kernel is not estimable under H_0 simply because the kernel matrix K disappears under H_0. To take this into consideration, in our selection process, we take ρ as the value estimated using REML based on the consideration that unimportant variables do not significantly contribute to the joint effect of $\boldsymbol{z}$.

2.4 GEV Selection within the Reproducing Kernel Framework

In this section, we propose a procedure to select important GEVs based on their contribution to the joint effect of $\boldsymbol{z}$ on an outcome of interest. The selection procedures will achieve two sequential goals. The first goal is to select a set of variants showing a significant joint effect, and the second is to, based on the selected set of variants, further exclude unimportant (redundant) ones and identify a parsimonious set of GEVs. The procedure for the first goal is referred as initial selection steps, and for the second goal we denote it as selection refinement steps. They are backward variable selection procedures and outlined below:

1. Initial Selection: Identify a subset of GEVs in $\boldsymbol{z}$, $\boldsymbol{z}_{sub}$, showing a significant

joint effect.

a) Start from the full model with g GEVs, i.e. each $\boldsymbol{z}_i$ is a g-dimensional vector; calculate the p-value using the score test, and denote it as p_0. Denote by α the significance level. If $p_0 > \alpha$, we prefer the null hypothesis $\tau = 0$. This implies that either all these g variants are non-informative or some variants are informative but the variation brought in by non-informative variants prevents the rejection of H_0.

b) For $k = 1, \ldots, g$, remove the kth GEV, i.e., each $\boldsymbol{z}_i$ is now a $(g-1)$-dimensional vector; calculate the p-value using the score test, and denote it as $p_{k,-}$. If $p_{k,-} < p_0$ for some $k = 1, \ldots, g$, the model using fewer variants provides "better" fit than if using more variants. Remove the sth variant from the model, where $s = \text{argmin}_k(p_{k,-})$.

c) The model now consists of $g - 1$ variants. Let $g = g - 1$ and go to step 1(b) until $p_{s,-} < \alpha$ and denote the set of selected variables as $\boldsymbol{z}_{sub}$, which is a set of GEVs showing a significant set effect on the outcome of interest. Denote the p-value of $\boldsymbol{z}_{sub}$ as $p_{\boldsymbol{z}_{sub}}$.

2. Selection Refinement: Identify a parsimonious set of GEVs in $\boldsymbol{z}_{sub}$.

a) Find the estimation $\widehat{\tau}$ and its $(1-\alpha)100\%$ confidence interval C by using the g variants in $\boldsymbol{z}_{sub}$, and go further for the selection of parsimonious variants.

b) For $k = 1, \ldots, g$, remove the kth variants; calculate the p-value $p_{k,-}$ and $(1-\alpha)100\%$ confidence interval $C_{k,-}$; find the estimation $\widehat{\tau}_k$. If $p_{k,-} < p_{\boldsymbol{z}_{sub}} < \alpha$ and $C \cap C_{k,-} \neq \emptyset$, we potentially have a comparable model by using fewer variants. Remove the sth variant from the model, where $s = \text{argmin}_k(|\widehat{\tau} - \widehat{\tau}_k|)$.

c) Let $g = g - 1$ and go to step 2(b) until no $p_{k,-} < p_{\boldsymbol{z}_{sub}} < \alpha$ or until $C \cap C_{k,-} = \emptyset$.

d) The remaining variant(s) is (are) considered to be important to the response variable y and form a parsimonious set of GEVs showing a significant joint effect.

The above selection method follows, in spirit, the generic selection process such as the stepwise selection process (Peduzzi et al. [1980]) or forward selection process (May et al. [2008]) proposed earlier for non-parametric models. The type I error rate of the proposed selection process is $\alpha - \alpha^2 + \Delta < \alpha$ with $\Delta < \alpha^2$, implying a conservative selection process. To show this, let's consider one-sided tests. If the kth variant is removed, then we will have $p_{k,-} < \alpha$ and $C \cap C_{k,-} \neq \emptyset$. In this case, the type I error is the true value of τ is actually outside the overlapped region of the two intervals. The probability of this error occurring (the type I error rate) is $\alpha(1 - \alpha) + \delta\alpha < \alpha$, where $\delta < \alpha$. Similar results can be drawn for two-sided tests. Thus the selected parsimonious set of variants are expected to be informative and such informativity can be a possible consequence of complex GEV effects that parametric models may not be able to describe. It is worthy noting that the initial selection and refinement steps do not conflict with the traditional backward selection techniques. After initial selection, GEVs that contribute to the outcome of interest as a whole unit will be identified. However, it is possible that some GEVs do not play a role in determining the strength of the joint effect, and thus are redundant. This was the motivation of the refinement procedure to identify these redundant GEVs and remove them. The idea of this procedure is that a GEV should be excluded if its inclusion does not cause any significant change in overall effect estimation but results in reduction in statistical significance, i.e., resulting larger p-values in the test of joint effect. As seen in the simulation studies below, the proposed approach outperforms several existing parametric methods that assume a specific format of the association

between a response variable and a set of independent variables and also does better than a method built upon random forest.

2.5 Simulation Study

In this section, via simulations, we evaluate the performance of the proposed variable selection procedures using continuous $\boldsymbol{z}$ variables. The results from discrete variables are expected to be comparable. We consider the following three models, each representing a different type of contribution from $g = 12$ variables in $\boldsymbol{z}_i$:

1. $E(y_i) = x_i + 3\log(z_{i1})\cos(z_{i2}) + 2\exp(z_{i3})$
2. $E(y_i) = x_i + 3(z_{i1} - z_{i2})^2 + 2z_{i3}$
3. $E(y_i) = x_i + 2(z_{i1} - z_{i2}) + 3z_{i3}$

where x_i is generated from $N(0, 4)$, z_{il} from uniform $(0, \frac{g}{2l})$ for $l = 1, \cdots, 12$. The random error is assumed to be normally distributed with mean 0 and variance 1. Model 1 represents regression models with non-linear effects of z_{il} including main effect and interactions. Both models 2 and 3 are linear regression models but model 2 includes an additive interaction effect of z_1 and z_2. For each model, we generate 1000 Monte Carlo (MC) replicates, each with the same size n. We choose three sample sizes, $n = 100, 200, 400$. The significance level is set at $\alpha = 0.1$ in the hypothesis testing. Unlike categorical $\boldsymbol{z}$'s, for continuous $\boldsymbol{z}$ variables, in order to perform the hypothesis test of $\tau = 0$, we need an estimate of tuning parameter ρ, which is taken as the estimate from data using the REML method discussed in the Method section. To summarize our findings, we record the proportions of correct selection (all important variables are selected and unimportant ones are all excluded), over selection (all important variables plus at least one unimportant variable), under selection (a subset of important variables and no unimportant variables), and partial selection (a subset

of important variables plus at least one unimportant variable). Note that the sum of these proportions can be less than 1, for instance, when no important variables are selected. The average model size is also recorded. The algorithm is programmed in R (R Core Team [2012]) and the R codes are available to users with interest in the methods.

Table 2.1: Variable selection summary for different methods ($n = 200$). *"Avg.size" is the average model size. Other numerical values are proportions of selection among 1000 MC iterations. Model 1,* $E(y_i) = x_i + 3\log(z_{i1})\cos(z_{i2}) + 2\exp(z_{i3})$*; Model 2,* $E(y_i) = x_i + 3(z_{i1} - z_{i2})^2 + 2z_{i3}$*; Model 3,* $E(y_i) = x_i + 2(z_{i1} - z_{i2}) + 3z_{i3}$.

Model		Proposed	AIC	BIC	LASSO	ALASSO	RF-based
1	Correct selection	0.452	0	0	0.006	0.009	0.011
	Over selection	0.046	0.05	0	0.183	0.006	0.206
	Under selection	0.407	0	1	0.276	0.915	0.757
	Partial selection	0.095	0.95	0	0.535	0.070	0.026
	Avg. size	2.750	5.966	1	4.301	1.412	3.079
2	Correct selection	0.911	0	0	0.026	0.129	0
	Over selection	0.009	0.049	0	0.604	0.068	0
	Under selection	0.078	0	1	0.041	0.684	0.999
	Partial selection	0.002	0.951	0	0.329	0.119	0.001
	Avg. size	2.931	5.964	1	5.475	2.427	2.001
3	Correct selection	0.842	0	0	1	0.888	0.121
	Over selection	0.146	0.081	0	0	0.112	0.005
	Under selection	0.009	0	0	0	0	0.873
	Partial selection	0.003	0.919	1	0	0	0
	Avg. size	3.166	9.351	3	3	3.136	2.132

Due to the limited literature on variable selection in semi-parametric models, we compare the proposed procedure with methods of AIC backward selection, BIC

backward selection, LASSO, and adaptive LASSO (ALASSO) (Tibshirani [1996], Zou [2006]) applied to linear regressions evaluating the association of $\boldsymbol{y}$ with the variables X and $\boldsymbol{z}$. ALASSO is applied to linear additive models and with the feature of enjoying the oracle property (see Zou [2006]), that is, the method will correctly select the model as if the correct submodel were known. ALASSO thus serves as a benchmark for model 3. The earlier developed variable selection method LASSO (Tibshirani [1996]), on which ALASSO is built, is also included in our comparison. Besides variable selection methods in parametric models, the feature selection method random forest (RF) built upon regression trees (Breiman [2001]) is considered as well and the R package `randomForest` (Liaw and Wiener [2002]) is implemented to obtain, for each variable, an averaged node impurity reduction due to split of the variable over all trees, where a node impurity is evaluated by mean square errors. To select important variables, a segmented regression line is fitted to the descending node impurity reductions and variables at or before the first changing point are treated as important ones. Using changing points from segmented regressions to identify important variables is motivated by the idea of using scree plots to select components in principal components analyses. Our simulations indicate that using scree plots is overly conservative in selecting variables and has the tendency to exclude important variables. To make our procedure comparable with these existing methods, we focus on our selection to achieve parsimonious sets of variables. In the real data applications, we include an illustration of the first goal in variable selection, that is, selecting a significant set of variables allowing the existence of possibly redundant ones in the set.

Results from 1000 MC replicates based on $n = 200$ are summarized in Table 2.1. We observe that the proposed method performs well for linear and nonlinear associations, and the average selected number of important variables is close to the actual model size. Although the LASSO and ALASSO do slightly better than our approach in the linear situation (model 3), they perform poorly when the associations are

non-linear. This is consistent with the origination of the two methods, which are developed for linear models. We also observe that LASSO tends to over select variables as indicated by the results from model 2, which is essentially a linear regression model with an interaction effect of z_1 and z_2. This observation of over selection is consistent with previous finding (Horowitz and Huang [2010]). The AIC and BIC selections do not choose the important variables often. As expected, AIC has the tendency to select a large number of variables as indicated by average model sizes, while BIC has the tendency to under select with model sizes much smaller than the truth. For the random forest-based (RF-based) approach, we first illustrate its selection using one data set generated from model 1. Figure 2.1 displays the sorted node impurity reductions. After fitting a segmented regression, the first changing point is at 2.093 indicating that the first two variables (variables 2 and 3) are deemed as important variables. Note that only variable 3 would be selected if we used scree plot. For all the 1000 MC replicates, as seen in Table 2.1, the RF-based method severely under selects variables regardless of linear or non-linear associations between $\boldsymbol{y}$ and $\boldsymbol{z}$.

Table 2.2: Correctness rates (Proportions of correct selection among 1000 MC simulations from different methods with respect to different numbers of GEVs. *Model,* $E(y_i) = x_i + 3\log(z_{i1})\cos(z_{i2}) + 2\exp(z_{i3})$, $n = 200$.

	% correct					
Number of GEVs	Proposed	AIC	BIC	LASSO	ALASSO	RF-based
25	0.400	0	0	0.008	0.123	0.019
50	0.293	0	0	0.004	0.128	0.018
100	0.161	0	0	0.001	0.098	0.014

We further evaluate the performance of the method with respect to the number of GEVs and sample size with focus on the most important criterion, proportion of correct selection. To evaluate the impact of number of variables on the correctness rate of variable selection, we use model 1 because non-linear associations are common

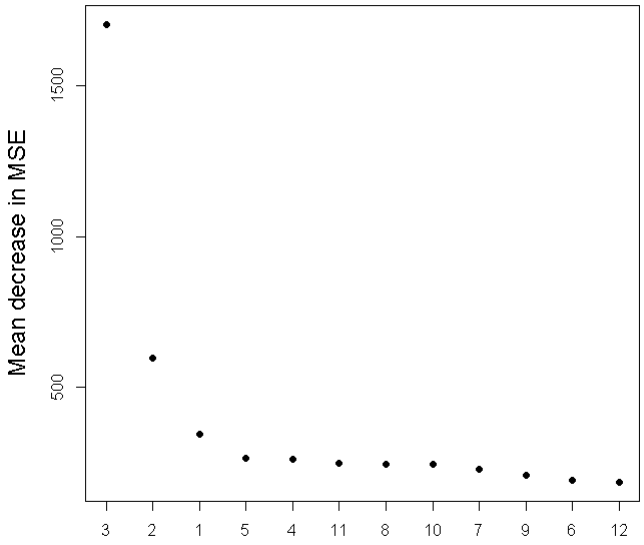

Figure 2.1: Illustration of segment regression on node impurity reductions obtained from the random forest method. *Model* $E(y_i) = x_i + 3\log(z_{i1})\cos(z_{i2}) + 2\exp(z_{i3})$, $n = 200$. *A segment regression with one change point is fitted to the data and he change point is at 2.093.*

in genetic and epigenetic studies. The sample size is taken as $n = 200$ and we consider a larger number of GEVs, $g = 25, 50$ and 100. Note that even when $g = 25$, the number of possible main effect terms plus the number of all possible two-way interactions exceed $n = 200$. The results are listed in Table 2.2. When the number of GEVs increasing, the proportion of correct selection decreases, but still higher than the correctness rates from the competing methods. On the other hand, we expect such impact will be diminished when we have more observations. To examine the impact of sample size on the correctness rate of variable selection, we compare the selection results based on $n = 200$ with those using $n = 100$ and $n = 400$ observations, where g is kept at $g = 12$. The results are listed in Table 2.3. When the sample size is small, a strong trend of over selection is observed. However, it decreases quickly accompanied by a quick increase in correctness rate as the sample size increases (Figures 2.2, 2.3, and 2.4). The partial selection rates are low consistently in all choices of sample

Table 2.3: Simulation results of different sample sizes for the proposed method. *"Avg.size" is the average model size. Other numerical values are proportions of selection among 1000 MC iterations. Model 1,* $E(y_i) = x_i + 3\log(z_{i1})\cos(z_{i2}) + 2\exp(z_{i3})$*; Model 2,* $E(y_i) = x_i + 3(z_{i1} - z_{i2})^2 + 2z_{i3}$*; Model 3,* $E(y_i) = x_i + 2(z_{i1} - z_{i2}) + 3z_{i3}$.

Model	Type	$n = 100$	$n = 200$	$n = 400$
1	Correct selection	0.031	0.452	0.639
	Over selection	0.948	0.046	0.197
	Under selection	0	0.407	0.072
	Partial selection	0.021	0.095	0.092
	avg. size	9.792	2.75	3.695
2	Correct selection	0	0.911	0.941
	Over selection	0.983	0.009	0.02
	Under selection	0	0.078	0.039
	Partial selection	0.017	0.002	0
	avg. size	11.587	2.931	2.989
3	Correct selection	0.467	0.842	0.889
	Over selection	0.516	0.146	0.106
	Under selection	0.007	0.009	0.005
	Partial selection	0.010	0.003	0
	avg. size	5.096	3.166	3.117

sizes, indicating the method's reluctance to exclude important variables. Overall, with sufficient sample size, the proposed method has the ability to effectively identify the significant variables that contribute to the dependent variable.

2.6 Real Data Analysis

We apply the proposed methods to two data sets to identify important genetic and epigenetic variants. One data set is composed of forced vital capacity (FVC) measures of lung function of 680 subjects and 13 SNPs that are potentially associated with lung function. Among these 13 SNPs, 6 are in chitinase 3-like 1 ("CHI3L1") gene and 7

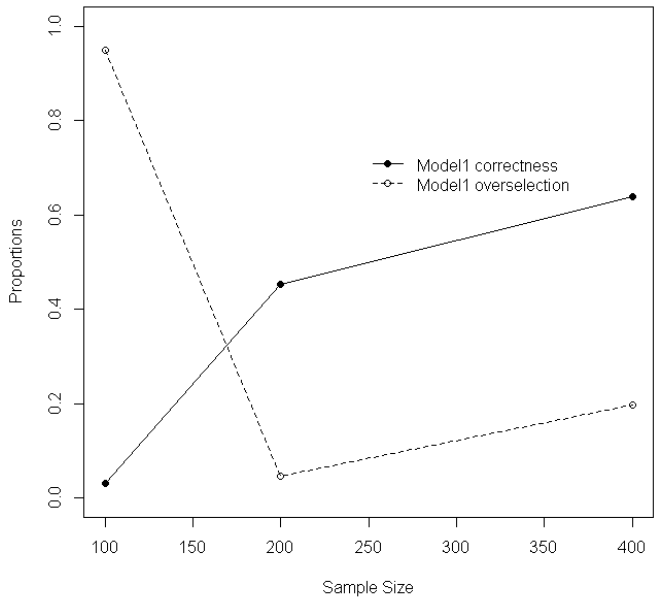

Figure 2.2: Patterns of correctness and over-fitting proportions with respect to sample sizes. Model 1, $E(y_i) = x_i + 3\log(z_{i1})\cos(z_{i2}) + 2\exp(z_{i3})$.

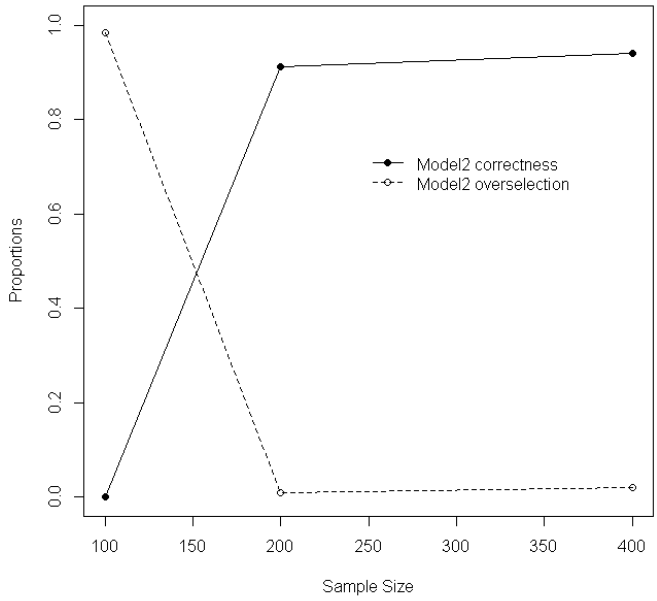

Figure 2.3: Patterns of correctness and over-fitting proportions with respect to sample sizes. Model 2, $E(y_i) = x_i + 3(z_{i1} - z_{i2})^2 + 2z_{i3}$.

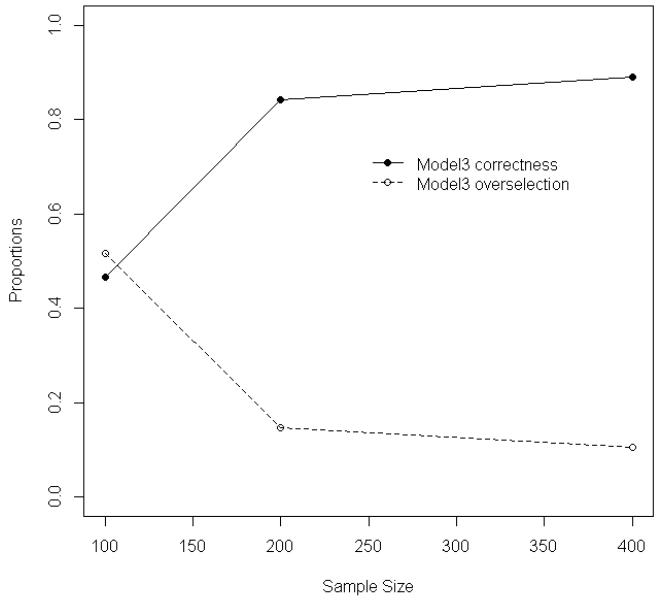

Figure 2.4: Patterns of correctness and over-fitting proportions with respect to sample sizes. Model 3, $E(y_i) = x_i + 2(z_{i1} - z_{i2}) + 3z_{i3}$.

are in chitinase 3-like 2 gene ("CHI3L2") (Table 2.4). Both genes are located on chromosome 1. Variable height is included as an adjusting factor due to its direct relationship with lung function (Walker et al. [1989]). The second data set is cotinine levels of 114 subjects and DNA methylation of 38 CpG sites of these 114 subjects. Cotinine is an alkaloid detected in tobacco and has been used as a biomarker of smoke exposure (Benowitz [1996]). These 38 CpG sites are selected due to their possible association with maternal smoking identified in our preliminary study and some of them are consistent with findings from a closely related study (Joubert et al. [2012]). There is evidence that maternal smoking may be linked to early onset of offspring smoking (Huizink and Mulder [2006]). Thus this selection process is to identify, among the CpG sites potentially associated with maternal smoking, CpG sites that also show relationship with active smoking. Information on some of the CpG sites including their reference genes and corresponding chromosomes is listed in Table 2.5.

We set $\alpha = 0.05$ for both selections. In the first example, we start from identifying

Table 2.4: Information of the 13 SNPs and the selection results. *Selected SNPs are marked with "✓". No SNPs are selected by using BIC, LASSO, and adaptive LASSO.*

Reference gene	SNP	Initial selection	Refinement	AIC	RF-based
CHI3L1	rs880633	✓	–	–	✓
	rs4950928	–	–	–	–
	rs4950929	–	–	–	–
	rs6698204	–	–	–	✓
	rs10399805	–	–	–	–
	rs10399931	–	–	–	–
CHI3L2	rs8535	–	–	–	–
	rs2147789	–	–	✓	–
	rs2255089	✓	✓	–	–
	rs3889380	✓	✓	✓	–
	rs3934922	–	–	✓	–
	rs3934923	✓	✓	✓	–
	rs17014713	✓	✓	–	–

a SNP set showing a significant set effect on lung function. To achieve this goal, we use the initial selection procedure discussed in the Method section, and use the IBS kernel to describe $h(\cdot)$. The initial selection procedure identifies a set of 5 SNPs that as a whole unit has a significant set effect on FVC (Column 3 in Table 2.4). Note that this identified set effect represents the overall effect of these 5 SNPs. It is possible that one or more SNPs among these 5 do not play any role in the determination of the overall effect and thus are redundant. Our further application of the refinement procedure indicates that including rs880633 on the "CHI3L1" gene actually decreases the statistical significance of set effect and thus should be excluded, resulting in 4 SNPs in the final parsimonious SNP set (Column 4 in Table 2.4). These four SNPs are all in the "CHI3L2" gene and no SNPs are from "CHI3L1". "CHI3L1" and "CHI3L2" are closely related chitinase-like family genes (Areshkov et al. [2012]) and there is

evidence that each of them is individually linked to lung function (Ober et al. [2008], Areshkov and Kavsan [2010]). However, when evaluating these two genes together with their complex interaction effects accounted, the effect of "CHI3L1" is negligible, which may deserve a closer investigation in the direction of gene network. The AIC, BIC, LASSO, adaptive LASSO, and random forest-based (RF-based) method are also applied to choose the SNPs. The AIC approach selects 4 SNPs (column 5 in Table 2.4), two of which agree with the results from the proposed method. The RF-based approach selected two SNPs (column 6 in Table 2.4) that are both different from the selections by using the proposed method and the AIC approach. The other three methods, BIC, LASSO, and adaptive LASSO, do not select any SNPs. Note that these three parametric methods assumed an additive model that includes main effects only. Connecting this to the findings in simulations, we postulate that the SNPs identified using the proposed method may interact with each other in a complex way, especially for SNPs rs2255089 and rs17014713, which are not selected by the AIC method.

For the second example, the set effect of the 38 CpG sites is already significant at $\alpha = 0.05$ (p-value=0.0059) with regard to cotinine levels. However, it is possible that there exist some redundant CpG sites not contributing to the overall effect. By applying the selection refinement procedure, 25 CpG sites are identified and included in the parsimonious CpG set (Table 2.5). This selection result indicates that, among the 38 CpG sites potentially associated with maternal smoking, about 70% of them also associated with active smoking (indicated by cotinine levels). This finding could be due to the association between maternal smoking and active smoking and may require a further cause-effect study. As for the competing methods, the AIC method selected 10 CpG sites with 6 overlapping with the selection from the proposed method, the BIC and LASSO methods selected one CpG site cg05575921, adaptive LASSO selected sites cg20418529 and cg17924476, and RF-based method selected 4 CpG sites and 3 of them also selected by the proposed method (Table 2.5). Based on these

findings, the 19 CpG sites identified by the proposed method but not by any of the parametric approaches are likely being selected due to complex interactions.

As seen in the simulations, BIC, LASSO, adaptive LASSO, and RF-based methods severely under select variables when the association is non-linear (model 1). The results from these methods in simulations under model 1 and in these two applications indicate a possible non-linear association that may include complex interactions between the genetic (SNPs) or epigenetic (methylation of CpG sites) variants. Furthermore, as demonstrated in our simulations, in all three simulation scenarios, AIC tends to partially select important variables, while the proposed methods correctly select truly important variables more often. This implies that the 4 SNPs and the 25 CpG sites selected from the refinement steps are more likely to be true contributors to the overall set effect.

2.7 Conclusion and Discussion

We proposed variable selection procedures easy to implement to select variables such as GEVs in semi-parametric models describing associations of candidate variables with a response variable. The association is described using reproducing kernels, which allows linear or non-linear effects of any form. The selection procedure is built upon a statistical testing in a set analysis and the variables are selected using the backward selection scheme. We proposed two selection scenarios: the initial selection emphasizing on detecting significant sets of variables and the refinement step focusing on identifying a parsimonious set of important variables with redundant variables removed.

The methods are demonstrated and evaluated through simulations. The simulation results show that the proposed methods can effectively identify the correct variables regardless of the feature of the association, linear or non-linear. We compare the methods with the standard AIC and BIC selection procedures, the LASSO

and adaptive LASSO methods, as well as the random forest-based approach. The AIC method tends to partially select while BIC usually under selects important variables. In the simulations, we assumed that the variables are not correlated. Thus, in a linear regression model (model 3), the LASSO and adaptive LASSO give similar results to those from the proposed methods, but they are inferior to the reproducing kernel-based approach when the variable effects are non-linear (models 1 and 2). The random forest-based approach severely under selects important variables in all situations. We applied the methods to two real data sets to select SNPs associated with lung function and CpG sites associated with smoke exposure. Based on the patterns observed in simulations, we postulate the existence of non-linear associations that involve complex interaction effects between genetic or epigenetic variants.

The proposed methods are ready to extend to choose variables in other types of statistical models including log-linear models and models applied to survival data analysis. On the other hand, the methods have some limitations that warrant a discussion. The variables are selected based on the strength of their joint effect. The procedure is able to exclude redundant variables via the refinement procedure, but the exclusion is based on an evaluation of overall effect and its significance. The amount of contribution of each individual variable is not estimable in the current framework. In some situations, it may be desirable to evaluate the effect of each selected variable, besides their joint effect. Furthermore, the proposed method assumes no missing values. Accounting for missing values in the kernels surely will extend the flexibility of the proposed selection procedure and is our on-going work.

Table 2.5: Information of the CpG sites selected by at least one method, ordered by gene names and CpG ID. *Symbol* "&" *indicates the CpG site is between two genes; symbol ";" indicates that the CpG site is on both genes.*

CpG ID	Reference	Chromsome	Refinement	AIC	BIC	LASSO	ALASSO	RF-based
cg05575921	AHRR	5	✓	✓	✓	✓	–	✓
cg17924476	AHRR	5	–	–	–	–	✓	–
cg21161138	AHRR	5	✓	–	–	–	–	–
cg23067299	AHRR	5	–	–	–	–	–	✓
cg01186919	ALG9	11	✓	–	–	–	–	–
cg03668078	C6orf103; LOC729176	5	✓	–	–	–	–	–
cg07442409	C14orf39	14	✓	–	–	–	–	–
cg02093176	COLEC11	6	–	✓	–	–	–	–

Continued on next page

Table 2.5 – continued from previous page

CpG ID	Reference	Chromsome	Refinement	AIC	BIC	LASSO	ALASSO	RF-based
cg11395306	CNTN5	11	✓	–	–	–	–	–
cg11207515	CNTNAP2	7	–	✓	–	–	–	–
cg05549655	CYP1A1	15	✓	✓	–	–	–	✓
cg11924019	CYP1A1	15	✓	–	–	–	–	–
cg17852385	CYP1A1	15	–	✓	–	–	–	–
cg18092474	CYP1A1	15	✓	–	–	–	–	–
ch_18_9250	DYM & ACAA2	18	✓	–	–	–	–	–
cg16116321	FAM124B	2	✓	–	–	–	–	–
cg18493761	FEZ1 & EI24	11	✓	–	–	–	–	–

Continued on next page

Table 2.5 – continued from previous page

CpG ID	Reference	Chromsome	Refinement	AIC	BIC	LASSO	ALASSO	RF-based
cg24874277	FMN1	15	✓	–	–	–	–	–
cg14179389	GFI1	1	✓	–	–	–	–	–
cg14282137	LIMS2	2	✓	–	–	–	–	–
cg08126560	LOC100129066	9	✓	–	–	–	–	–
cg04180046	MYO1G	7	–	✓	–	–	–	–
cg19089201	MYO1G	7	✓	–	–	–	–	–
cg00295418	MYOM2	8	✓	✓	–	–	–	–
cg19273101	NAB1 & GLS	2	✓	–	–	–	–	–
cg11881038	OPRM1	6	✓	✓	–	–	–	–

Continued on next page

Table 2.5 – continued from previous page

CpG ID	Reference	Chromsome	Refinement	AIC	BIC	LASSO	ALASSO	RF-based
cg00794911	RP11-252P19.3	6	✓	✓	–	–	–	–
cg18132363	RP11-252P19.3	6	✓	✓	–	–	–	–
cg20418529	RP11-252P19.3	6	✓	–	–	–	✓	–
cg21015808	RP11-545G3.1 &ACTR3C	7	✓	–	–	–	–	✓
cg14075934	SATB2	2	✓	–	–	–	–	–

Chapter 3

Joint Clustering

The work in this chapter is motivated by a collaborative project aiming to group dependent variables based upon the correlation of dependent variables and association of dependent variables and covariates of interest, along with the different response of individual subjects.

3.1 Introduction

With the development of technology, rich genetic and epigenetic information for each individual subject is available. To analyze this type of data efficiently, it is necessary to reduce its dimension. One way of dimension reduction is to perform cluster analysis. Jointly clustering individuals along with the genetic and epigenetic information (biclustering) was first introduced about a decade ago (Cheng and Church [2000]). The biclustering focuses on simultaneously clustering two-dimensional gene expression data and tries to optimize a pre-specified objective function (Freitas et al. [2012]). There are two main classes of biclustering algorithms: systematic search algorithms and stochastic search algorithms, and each class of algorithm has several different approaches (Freitas et al. [2012]). Various biclustering tools built upon these existing methods are available, for example biluster analysis in R (Kaiser and Leisch [2008]), BiVisu (Cheng et al. [2007]), GEMS (Wu and Kasif [2005]), Bayesian BiClustering model (BBC) (Gu and Liu [2008]), BicOverlapper (Santamaría et al. [2008]), and e-CCC-Biclustering (Madeira and Oliveira [2009]). These current bi-clustering approaches allow identifying sets of genes sharing compatible expression patterns

across subsets of samples, and have been demonstrated to be useful in varies of gene expression/microarray data in terms of dimension reduction for feature identification and easy interpretation. The bi-clustering concept considers the coherence of rows and columns in the data, and is a non-model based clustering technique. It is mainly restricted to the data only and external variables do not have any contribution to the evaluation of similarity between different clustering variables. Furthermore, some bi-clustering methods perform cluster analyses on the rows and columns separately, and do not consider the interrelationship between the rows and columns. Most importantly, existing methods overlook the correlations between the clustering dependent variables (DPVs), which can potentially cause mis-clustering.

In this article, we propose a clustering method, denoted as joint clustering, which takes into account the correlations between DPVs and the interrelationship between variables and subjects. The clusters are formed by consistent associations between a DPV and covariates of interests among a subset of subjects for a certain number of DPVs. Each joint cluster is composed a certain numbers of DPVs and a subset of subjects. To evaluate possibly non-linear associations between DPVs and covariates, a semi-parametric model via penalized splines (Eilers and Marx [1996]) is used. To cluster DPVs, an indicator variable is introduced for cluster assignment. To cluster subjects, a Dirichlet process mixture model is applied. The proposed joint clustering method has the ability to produce homogeneous clusters composed of a certain number of subjects sharing common features on the relationship between some DPVs and covariates.

The remainder of the article is organized as follows. Section 2 introduces the model of joint clustering under Bayesian framework and settings for the priors. The full conditional posteriors, detailed procedure and approach of joint clustering is also described in this section. We demonstrate and evaluate the performance of the proposed joint clustering in Section 3 through simulations. The proposed approach

is then applied to cluster methylation CpG sites and subject based on the association of methylation with cotinine levels. This is discussed in Section 4. We summarize our methods and findings in the Section 5.

3.2 The Method

We consider the following joint (two-dimensional) cluster which is illustrated in the Figure 3.1. For the ease of presentation, we dissect the unified clustering process into two parts. In part1, dependent variables (DPVs) are clustered; and in part 2, subjects within each DPV clusters are further clustered to form refined clusters, where the correlations between the DPVs will be taken into account.

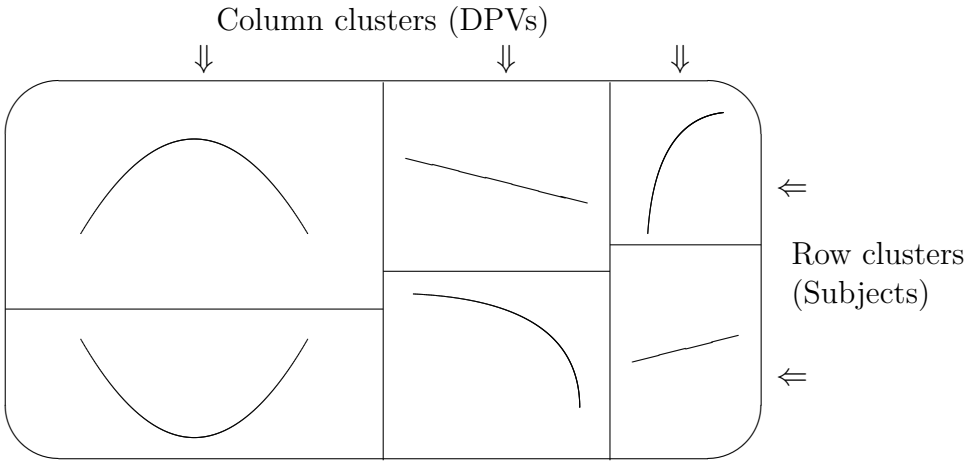

Figure 3.1: Illustration of joint clusters (DPVs first and then subjects).

3.3 Clustering the Variables

We cluster DPV based on agreement of relationships between DPV measures and covariates of interest in this part. Assume there are n subjects in the sample and in total K DPV variables are under consideration for clustering. For subject i, let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iK})$ denote the measures of K DPV components. Let M denote the number of clusters formed by DPV variables ($M \leq K$), and D be an $M \times K$ 0-1

matrix such that the kth column contains $M-1$ zeros and one element with the number 1 indicating which cluster the kth DPV ($k = 1, 2, \ldots, K$) variable belongs to. Then, the mth row indicates which DPV variables are in cluster m. We formulate the clustering procedure into the following:

$$\boldsymbol{y}_{i,m}|D_{m.} = \boldsymbol{Q}(\boldsymbol{x}_i, \boldsymbol{\beta}_m) + \boldsymbol{\varepsilon}_{i,m}^T. \tag{3.1}$$

Denote the k_m DPV variables $\boldsymbol{y}_{i,m} = (y_{i,(1)}, \ldots, y_{i,(k_m)})'$ as the mth cluster. $\boldsymbol{\varepsilon}_{i,m}^T$ follows a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix Σ_m, which informs the strength of correlations between the dependent variables in cluster m. Due to the property of homogeneity in a cluster (assuming variables are properly transformed when necessary), the variances of the variables in one cluster are assumed to be the same, i.e. the diagonal elements are the same. Also, we assume that variables from different DPV clusters are independent. Function $\boldsymbol{Q}(\cdot)$ is a vector function which describes the relationship between DPVs and covariates of interest x_i. We use semi-parametric models to model this relationship. Specifically, penalized splines will be applied (Eilers and Marx [1996]) due to its use of low rank bases. $\boldsymbol{\beta}_m$ is a vector of coefficients for the P-Splines for the mth cluster. In order to achieve the smoothness, quadratic splines will be used

$$Q(x, \boldsymbol{\beta}_m) = a_{m,1}x + a_{m,2}x^2 + \sum_{l=1}^{g} b_{m,l}(x - z_l)_+^2,$$

where g is the number of knots; $\boldsymbol{\beta}_m = (a_{m,1}, a_{m,2}, b_{m,1}, \ldots, b_{m,g})'$ of length $(g+2)$; z_l's are the spline knots; and

$$(x - z_l)_+ = \begin{cases} 0, & \text{if } x \leqslant z_l, \\ x - z_l, & \text{if } x > z_l. \end{cases}$$

We write $X = (x, x^2, (x - z_1)^2, \ldots, (x - z_g)^2)'$, and $Q(x, \boldsymbol{\beta}_m) = X'\boldsymbol{\beta}_m$. Since the dependent variables are multivariate, we write $\boldsymbol{X} = X \otimes \mathbf{1}_{k_m}$. Where $\otimes$ is the Kronecker product; $\mathbf{1}_{k_m}$ is a row vector of dimension k_m composed of 1's.

We use a fully Bayesian approach to infer the variable clusters. Following lists the prior distributions of the parameters $\Theta = (\boldsymbol{\beta}_m, D_{\cdot k}, \Sigma_m)$ and corresponding higher prior parameters, where $D_{\cdot k}$ denotes the kth column of D representing which cluster the kth DPV belongs to.

$$\begin{aligned}
\boldsymbol{\beta}_m|(\sigma^2, \sigma_m^2, D) &\sim \boldsymbol{N}\Big(\mathbf{0}, V(\sigma^2, \sigma_m^2)\Big), \\
\sigma_m^2|(a_1, c_1) &\sim \text{InvGamma}(a_1, c_1), \\
D_{\cdot k}|\boldsymbol{\pi} &\sim \text{Multinomial}(1, \boldsymbol{\pi}), \\
\boldsymbol{\pi}|\zeta &\sim \text{Dirichlet}(\zeta \mathbf{1}_M), \\
\zeta &\sim p(\zeta) = \frac{1}{2} \text{ if } 0 < \zeta \leq 1, \text{ and } \frac{1}{2}\zeta^{-2} \text{ if } \zeta > 1, \\
\Sigma_m|(S, \nu) &\sim \text{InvWishart}(S, \nu),
\end{aligned}$$

where σ^2 is the variance of $a_{m,1}$ and $a_{m,2}$, σ_m^2 is the variance of b_l's, $V(\sigma^2, \sigma_m^2)$ is a diagonal matrix with entries σ^2, and σ_m^2 corresponding to the order in $\boldsymbol{\beta}_m$, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_M)$ gives the probability that the kth DPV variable is in each of the M clusters, Σ_m is the variance-covariance matrix of $\boldsymbol{y}_{i,m}$, and ζ (Good [1965]), and ν are hyper-parameters, σ^2, a_1, c_1, and S are assumed to be known and selected to achieve vague priors.

3.4 Clustering the Subjects

The subjects within each DPV cluster are further grouped such that each group reflects a different relationship between DPV variables and covariates of interest. For this purpose, we relax $\boldsymbol{\beta}_m$ in (3.1) to let it be random across subjects, denoted as $\boldsymbol{\beta}_{i,m}$, and then group $\boldsymbol{\beta}_{i,m}$ through Dirichlet process. To cluster $\boldsymbol{\beta}_{i,m}$, we utilize the Dirichlet process due to its automatic detection of clusters and its ability to describe

non-standard distributions. Specifically, we have

$$\begin{aligned}
\boldsymbol{y}_{i,m}|(\boldsymbol{\beta}_{i,m}, D, \Sigma_m) &\sim \boldsymbol{N}\Big(\boldsymbol{X}_i'\boldsymbol{\beta}_{i,m}, \Sigma_m\Big), \\
\boldsymbol{\beta}_{i,m}|G &\sim G, \\
G &\sim \mathrm{DP}(G_0, \lambda), \\
G_0|\sigma_0^2 &\sim \boldsymbol{N}(\mathbf{0}, \sigma_0^2 I), \\
\sigma_0^2|(a_2, c_2) &\sim \mathrm{InvGamma}(a_2, c_2),
\end{aligned}$$

where $\boldsymbol{\beta}_{i,m}|G$ are independent given G, and $\mathrm{DP}(G_0, \lambda)$ represents the Dirichlet process with a measure having concentration λ and proportional to the base distribution $G_0 \sim \boldsymbol{N}(\mathbf{0}, \Sigma_0)$ with $\Sigma_0 = \sigma_0 I$. The conditional prior of $\boldsymbol{\beta}_{i,m}$ conditional on $\boldsymbol{\beta}_{-i,m}$ is the mixture distribution

$$\boldsymbol{\beta}_{i,m}\Big|\boldsymbol{\beta}_{-i,m} \sim \frac{1}{n-1+\lambda}\sum_{j\neq i}\delta_{\beta_{i,m}}(\boldsymbol{\beta}_{j,m}) + \frac{\lambda}{n-1+\lambda}G_0,$$

where $\delta_{\beta_{i,m}}(\boldsymbol{\beta}_{j,m})$ is a point mass concentrated at a single point where $\boldsymbol{\beta}_{i,m} = \boldsymbol{\beta}_{j,m}$.

The concentration parameter λ plays an important role in Dirichlet process and it controls the distribution over the number and sizes of the clusters. If λ is relatively large, the prior assigns distributions that are close to the baseline distribution, while it was pointed out by Antoniak [1974] to be careful of choosing small value for λ. Neal [2000], Bush and MacEachern [1996], Escobar [1994], Escobar and West [1995] proposed to use fixed concentration parameter. Later, both McAuliffe et al. [2006] and Dorazio et al. [2008] adopted a numerical approach based on the work of Liu [1996] to estimate λ. Our simulation indicates that this approach has the potential to under estimate λ. In this article, we propose to take λ fixed.

3.5 Posteriors Computing

The method we propose relies on the Gibbs sampling technology (Gelfand and Smith [1990]) which is to simulate successively observations of each parameter from its full

conditional posterior distribution. Most of the conditional posterior distributions listed below are obtained in a straightforward fashion.

The conditional posterior distributions related to the clustering of DPVs are:

$$\begin{aligned}
\boldsymbol{\pi}\Big|(D,\zeta) &\sim \text{Dirichlet}(\zeta\mathbf{1}+D_{\cdot k}),\\
D_{\cdot k}\Big|\boldsymbol{\pi} &\sim \text{Multinomial}(1,\boldsymbol{\pi}),\\
\boldsymbol{\beta}_m\Big|(Y,\sigma^2,\sigma_m^2,\Sigma_m,D) &\sim \boldsymbol{N}\Big((\sum_{i=1}^{n}\boldsymbol{X}_i'\Sigma_m^{-1}\boldsymbol{X}_i+V^{-1})^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i'\Sigma_m^{-1}\boldsymbol{y}_{i,m},\\
&\qquad (\sum_{i=1}^{n}\boldsymbol{X}_i'\Sigma_m^{-1}\boldsymbol{X}_i+V^{-1})^{-1}\Big),\\
\sigma_m^2\Big|\boldsymbol{\beta}_m &\sim \text{InvGamma}(a_1+\frac{g}{2},c_1+\frac{1}{2}\sum_{l=1}^{g}b_{m,l}^2).
\end{aligned}$$

The above conditional posterior distributions are all standard except for ζ. To sample ζ, we apply the Metropolis-Hastings algorithm (Metropolis et al. [1953], Hastings [1970]) and take the log-normal distribution as the proposal distribution.

The conditional posterior distributions in the procedure of further clustering subjects within each DPV cluster include the conditional posterior distribution of $\boldsymbol{\beta}_{i,m}$. Assuming the data are exchangeable (Neal [2000]), we have:

$$\boldsymbol{\beta}_{i,m}\Big|(\boldsymbol{\beta}_{-i,m},\boldsymbol{y}_{i,m})\sim\sum_{j\neq i}q_{i,j}\delta_{\boldsymbol{\beta}_{i,m}}(\boldsymbol{\beta}_{j,m})+r_iH_i,$$

where

$$\begin{aligned}
q_{i,j} &= b\frac{1}{n-1+\lambda}(2\pi)^{-\frac{k_m}{2}}|\Sigma_m|^{-\frac{1}{2}}\exp\Big[-\frac{1}{2}(\boldsymbol{y}_{i,m}-\boldsymbol{X}_i\boldsymbol{\beta}_{j,m})'\Sigma_m^{-1}(\boldsymbol{y}_{i,m}-\boldsymbol{X}_i\boldsymbol{\beta}_{j,m})\Big],\\
r_i &= b\frac{\lambda}{n-1+\lambda}(2\pi)^{-\frac{k_m}{2}}|\Sigma_m|^{-\frac{1}{2}}|\Sigma_0|^{-\frac{1}{2}}|\Sigma_{\beta_{i,m}}|^{\frac{1}{2}}\\
&\quad \exp\Big[-\frac{1}{2}\boldsymbol{y}_{i,m}'\Sigma_m^{-1}\boldsymbol{y}_{i,m}+\frac{1}{2}\boldsymbol{y}_{i,m}'\Sigma_m^{-1}\boldsymbol{X}_i\Sigma_{\beta_{i,m}}\boldsymbol{X}_i'\Sigma_m^{-1}\boldsymbol{y}_{i,m}\Big],\\
\Sigma_{\beta_{i,m}} &= (\boldsymbol{X}_i'\Sigma_m^{-1}\boldsymbol{X}_i+\Sigma_0^{-1})^{-1},\\
H_i &\sim \boldsymbol{N}\Big(\Sigma_{\beta_{i,m}}(\boldsymbol{X}_i'\Sigma_m^{-1}\boldsymbol{y}_{i,m}),\Sigma_{\beta_{i,m}}\Big),
\end{aligned}$$

and b is a normalizing constant. Other conditional posterior distributions involved in the clustering procedure are for Σ_m and Σ_0.

$$\begin{aligned}
\Sigma_m \Big| (Y, \boldsymbol{\beta}_{i,m}) &\sim \text{InvWishart}\Big(\sum_{i=1}^{n} (\boldsymbol{y}_{i,m} - \boldsymbol{X}_i' \boldsymbol{\beta}_{i,m})(\boldsymbol{y}_{i,m} - \boldsymbol{X}_i' \boldsymbol{\beta}_{i,m})' + S, n + \nu \Big) \\
\Sigma_0[j] \Big| (Y, \boldsymbol{\beta}_{i,m}) &\sim \text{InvGamma}\Big(a_2 + \frac{n(2+g)}{2}, c_2 + \sum_{i=1}^{n} \boldsymbol{\beta}_{i,m}[j] \Big),
\end{aligned}$$

where $\Sigma_0[j]$ denotes the jth diagonal element in Σ_0, and $\boldsymbol{\beta}_{i,m}[j]$ is the jth component of $\boldsymbol{\beta}_{i,m}$.

By applying the above mentioned Gibbs sampler, we have to draw $\boldsymbol{\beta}_{i,m}$ for each subjects, which is not efficient. The Algorithm 2 summarized in Neal [2000] involved a latent variable c_i to determine the subjects clustering (the method was initially used by Bush and MacEachern [1996]), in which all subjects are assigned values from $1, \ldots, n$, and subjects with the same value are from one cluster. Therefore, the Gibbs sampling process could be more efficient by drawing only those $\boldsymbol{\beta}_{c,m}$ that are currently associated with some subjects. The conditional posterior of c_i is

$$\begin{cases}
\text{if } c = c_j \text{ for some } j \neq i : P(c_i = c \Big| c_{-i}, Y, \boldsymbol{\beta}) \\
\quad = b \dfrac{n_{-i,c}}{n - 1 + \lambda} (2\pi)^{-\frac{k_m}{2}} |\Sigma_m|^{-\frac{1}{2}} \exp\Big[-\tfrac{1}{2} (\boldsymbol{y}_{i,m} - \boldsymbol{X}_i \boldsymbol{\beta}_{j,m})' \Sigma_m^{-1} (\boldsymbol{y}_{i,m} - \boldsymbol{X}_i \boldsymbol{\beta}_{j,m}) \Big], \\
P(c_i \neq c_j \text{ for all } j \neq i \Big| c_{-i}, Y, \boldsymbol{\beta}) \\
\quad = b \dfrac{\lambda}{n - 1 + \lambda} (2\pi)^{-\frac{k_m}{2}} |\Sigma_m|^{-\frac{1}{2}} |\Sigma_0|^{-\frac{1}{2}} |\Sigma_{\beta_{i,m}}|^{\frac{1}{2}} \\
\qquad \exp\Big[-\tfrac{1}{2} \boldsymbol{y}_{i,m}' \Sigma_m^{-1} \boldsymbol{y}_{i,m} + \tfrac{1}{2} \boldsymbol{y}_{i,m}' \Sigma_m^{-1} \boldsymbol{X}_i \Sigma_{\beta_{i,m}} \boldsymbol{X}_i' \Sigma_m^{-1} \boldsymbol{y}_{i,m} \Big],
\end{cases} \tag{3.2}$$

where b is an normalizing constant; c_{-i} denotes all c_j for $j \neq i$; $n_{-i,c}$ is the number of c_j for $j \neq i$ that are equal to c; $\boldsymbol{\beta}$ represents the set of $\boldsymbol{\beta}_{c,m}$ currently associated with at least one observation. The posterior for $\boldsymbol{\beta}_{c,m}$ is

$$\boldsymbol{\beta}_{c,m} \Big| (Y, \Sigma_m, D, c) \sim \boldsymbol{N}\Big((\sum_{c_i = c} \boldsymbol{X}_i' \Sigma_m^{-1} \boldsymbol{X}_i + \Sigma_0^{-1})^{-1} \sum_{c_i = c} \boldsymbol{X}_i' \Sigma_m^{-1} \boldsymbol{y}_{i,m}, (\sum_{c_i = c} \boldsymbol{X}_i' \Sigma_m^{-1} \boldsymbol{X}_i + \Sigma_0^{-1})^{-1} \Big)$$

3.6 Sampling Procedure

We now summarize the process of joint clustering as follows. At iteration t,

1. For the kth dependent variable, draw $\boldsymbol{\pi}$ from the distribution $\boldsymbol{\pi}\Big|(D_{.k}^{(t-1)}, \zeta^{(t-1)})$.
2. Draw $\zeta^{(t)}$ by the Metropolis-Hastings algorithm.
3. Draw $D_{.k}^{(t)}$ from the distribution of $D_{.k}\Big|(Y, \boldsymbol{\beta}_m^{(t-1)}, \Sigma^{(t-1)}, \boldsymbol{\pi}^{(t)})$, where Σ is the block-diagonal matrix of Σ_m's. This determines dependent variables clustering pattern.

The following steps are for each of the DPV clusters.

4. Draw $\boldsymbol{\beta}_m^{(t)}$ from the distribution of $\boldsymbol{\beta}_m\Big|(Y, \Sigma_m^{(t-1)})$.
5. Draw $\sigma_m^{(t)}$ from the distribution of $\sigma_m\Big|\boldsymbol{\beta}_m^{(t)}$.

Applying Dirichlet Process to cluster subjects. Draw for subjects $i = 1 \dots n$.

6. Draw $c_i^{(t)}$ by distribution given by Eq (3.2), where the state of c is $\{c_1^{(t)}, \dots, c_{i-1}^{(t)}, c_{i+1}^{(t-1)}, \dots, c_n^{(t-1)}\}$
7. Draw Σ_m from $\Sigma_m\Big|(Y, \boldsymbol{\beta}_{i,m}^{(t-1)})$ if $c_i^{(t)} \neq c_i^{(t-1)}$, where $\boldsymbol{\beta}_{i,m}^{(t-1)} = \boldsymbol{\beta}_m^{(t-1)}$ if cluster m is unique upto iteration $t-1$, $\boldsymbol{\beta}_{i,m}^{(t-1)} = \boldsymbol{\beta}_{i,m}^{(t'-1)}$ with $\boldsymbol{\beta}_{i,m}^{(t'-1)}$ being the coefficients at the latest iteration $t'-1$ such that DPVs form the same cluster m.

Draw $\boldsymbol{\beta}_{c,m}$ currently associated with at least one subject ($\boldsymbol{\beta}_{i,m} = \boldsymbol{\beta}_{c,m}$ for all $c_i = c$).

8. Draw a new value for $\boldsymbol{\beta}_{c,m}$ from the posterior distribution based on the prior G_0 and all observations currently associated with latent class c ($\boldsymbol{\beta}_m$ gives the initial value for $\boldsymbol{\beta}_{c,m}$).
9. For each component of Σ_0, draw from $\Sigma_0[j]\Big|(Y, \boldsymbol{\beta}_{i,m}^{(t)})$.

The above joint clustering process could dynamically determine the number of subject clusters within each dependent variable cluster. However, the number of DPV clusters M need to be specified in advance. To achieve the optimal results, we will let M varies in the process of DPVs clustering, and use the deviance information criterion (DIC) to determine the best value of M (Spiegelhalter et al. [2002]). Once the best number of DPVs clusters obtained, we will proceed for clustering subjects within each DPVs cluster.

In order to determine the final clusters, we consider the following procedure adapted from Dahl [2006], a procedure based on the method of "least-squares clustering":

1. After the MCMC burn-in, continue the MCMC simulations for an additional B iterations. Let A denote an $n \times n \times K$ matrix. The (i, j, k)th entry of A is the proportion of iterations such that the kth DPV of subjects i and j $(i, j = 1, ..., n)$ are in the same cluster. The matrix A is referred to as an averaged clustering matrix.

2. Continue to run an additional D iterations of the MCMC simulations. For each iteration,

    a) Form an $n \times n \times K$ matrix composed of indicators of clustering for that particular iteration. For instance, if the kth DPV of subjects i and j are in one cluster, then the (i, j, k)th entry is 1; otherwise, it is zero.

    b) Calculate the Euclidean distance between the matrix formed above and the averaged clustering matrix A.

3. Sort the Euclidean distances obtained from the D iterations, and the final selection on the number of clusters is in favor of simpler clusters and relatively small Euclidean distances.

3.7 Simulation Study: Settings

To demonstrate the methods and compare them with existing ones, we use simulations. All the programs are written in R (R Core Team [2012]).

We generate 500 data sets, with each of sample size 400 and having 10 DPVs, and one covariate x_i generated from a uniform distribution between 1 and 6. The 10 DPVs are grouped into 3 clusters and within each DPV cluster, the subjects are further clustered. Following is the setting of the clusters and the associations defined for each cluster:

- Cluster 1, $E(y_{ij}) = 6 + 5\sin(0.2\pi(x_i - 1))$ for $i = 1, \ldots, 250$ and $j = 1, \ldots, 5$
- Cluster 2, $E(y_{ij}) = -5 - 5\cos(0.2\pi(x_i - 3.5))$ for $i = 251, \ldots, 400$ and $j = 1, \ldots, 5$
- Cluster 3, $E(y_{ij}) = 10 - 0.8x_i$ for $i = 1, \ldots, 200$ and $j = 6, 7, 8$
- Cluster 4, $E(y_{ij}) = -5 - 3\exp\Big(0.4(x_i - 1)\Big)$ for $i = 201, \ldots, 400$ and $j = 6, 7, 8$
- Cluster 5, $E(y_{ij}) = 15 + 4\log\Big(0.4(x_i - 0.8)\Big)$ for $i = 1, \ldots, 180$ and $j = 9, 10$
- Cluster 6, $E(y_{ij}) = -2 + 0.1x_i$ for $i = 181, \ldots, 400$ and $j = 9, 10$

In total, we have 6 joint clusters with each cluster having a specific association between $\boldsymbol{y}$ and x. The random errors are assumed to be multivariate normal distribution with mean $\mathbf{0}$ and the following variance-covariance matrices for the three DPV clus-

ters

$$\Sigma_1 = \begin{bmatrix} 1 & -0.25 & 0 & 0 & 0 \\ -0.25 & 1 & -0.5 & 0 & 0 \\ 0 & -0.5 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0.3 \\ 0 & 0 & 0 & 0.3 & 1 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 1 & 0.1 & -0.4 \\ 0.1 & 1 & -0.1 \\ -0.4 & -0.1 & 1 \end{bmatrix}, \Sigma_3 = \begin{bmatrix} 0.6 & -0.25 \\ -0.25 & 0.6 \end{bmatrix}.$$

The patterns of the 6 clusters are displayed in Figure 1. We choose 10 knots evenly for the splines, and the following parameter in the Bayesian framework: $\nu = 12, a_1 = a_2 = c_1 = c_2 = 0.5, \sigma_m^2 = 100, \zeta = 0.75, v = 1.8, S = \frac{1}{2}\boldsymbol{I}$. For the concentration parameter λ, we take it as 1.5. Ten equally-spaced points between the range of x are chosen as the splines knots.

We further assume M is between 2 and 5. To draw inferences on the number of clusters and the parameters for each M, we run two MCMC chains with 1000 iterations each chain for each data set with 300 iterations as burn-in, the next 300 for the determination of the average clustering matrix, and the last 400 iterations for inferences. Our simulation s have shown a fast convergence of the MCMC chains.

To assess the quality of clustering, we record the number of joint clusters, accurate rate calculated based on pairwise agreement of clustering (i.e. where [subject, DPV] pairs $[i, j]$ and $[i', j']$ are in one cluster), sensitivity (Se=TP/(TP+FN)), and specificity (Sp=TN/(TN+FP)) with respect to a specific cluster, where "TP" denoting true positives (correct cluster identification), "FN" false negatives, "TN" true negatives, and "FP" false positives.

3.8 Simulation Study: Results

Among the 500 simulated data sets, three clusters are all preferred based on DIC. All subjects clustering summarized here are based on this preferred number of DPVs cluster. In total, joint clusters are correctly identified in 465 data sets. The average pairwise accurate rate is 0.997. The numbers of joint clusters are listed in Table 3.1 with the median be 6 and a 95% empirical interval is $(6, 9)$. The sensitivity and specificity based on the pairwise agreement are listed in Tables 3.2 and 3.3, which indicates the effectiveness of the method.

Table 3.1: List of the occurrence for the number of joint clusters

Number of clusters	5	6	7	8	9	10	11	12	13	15
Occurrence	2	466	17	2	2	3	1	4	2	1

Table 3.2: The average sensitivity for the pre-specified 6 joint clusters

	Sensitivity		
Cluster	Mean	Median	95% Empirical Interval
1	0.972	1	(0.936,1)
2	0.997	1	(0.990,1)
3	0.996	1	(1,1)
4	0.996	1	(1,1)
5	0.986	1	(1,1)
6	0.995	1	(0.965,1)

To illustrate the fitting performance of the proposed joint clustering method, we randomly choose one data set and the fitted curves vs true curves for all clusters are shown in Figures 3.2, 3.3, and 3.4. The proposed method could correctly identify the joint clustering and estimate the association between the dependent variables and the covariates of interest. The fitted curves are very close to the true curves.

Table 3.3: The average specificity for the pre-specified 6 joint clusters

	Specificity		
Cluster	Mean	Median	95% Empirical Interval
1	0.999	1	(1,1)
2	0.989	1	(0.649,1)
3	1	1	(1,1)
4	1	1	(1,1)
5	1	1	(1,1)
6	0.999	1	(1,1)

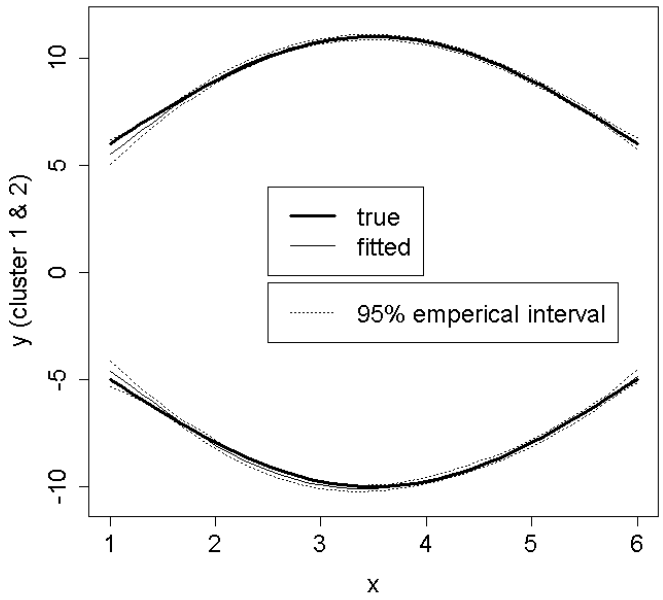

Figure 3.2: The fitted curves vs true curves for the first DPVs cluster.

3.9 Simulation Study: Comparisons with Existing Bicluster Methods

Existing biclustering methods focus on clustering data itself and ignore the contribution of external variables. In order to compare with those biclustering methods and demonstrate the effectiveness of the proposed method on data sets without covarites of interest, we generate 100 data sets with each of sample size 400 and having

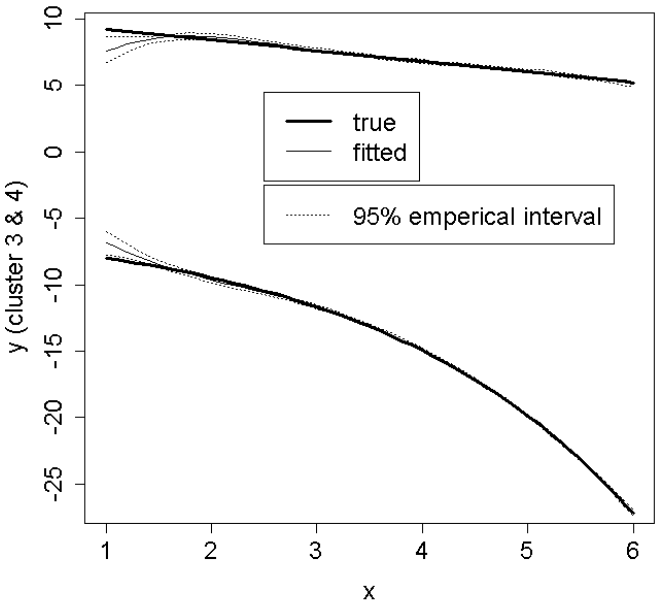

Figure 3.3: The fitted curves vs true curves for the second DPVs cluster.

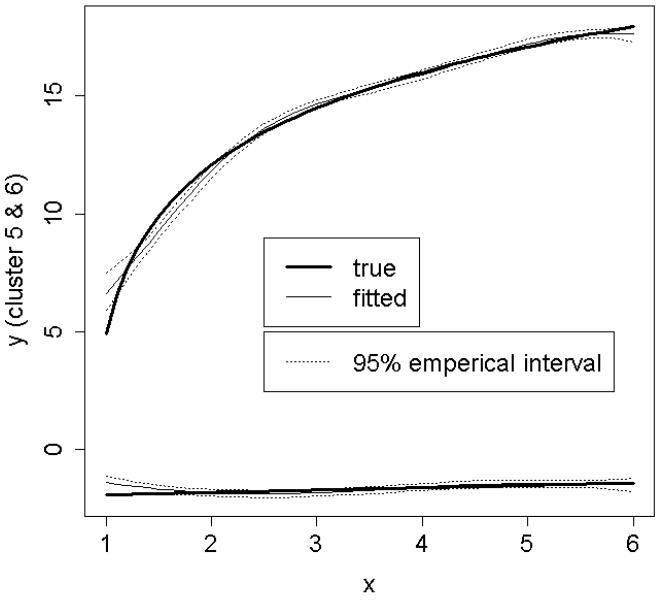

Figure 3.4: The fitted curves vs true curves for the third DPVs cluster.

10 DPVs. The 10 DPVs are grouped into 3 clusters and within each DPV cluster, the subjects are further clustered. Following is the setting of the clusters and the

associations defined for each cluster:

- Cluster 1, $E(y_{ij}) = 6$ for $i = 1, \ldots, 250$ and $j = 1, \ldots, 5$
- Cluster 2, $E(y_{ij}) = -5$ for $i = 251, \ldots, 400$ and $j = 1, \ldots, 5$
- Cluster 3, $E(y_{ij}) = 10$ for $i = 1, \ldots, 200$ and $j = 6, 7, 8$
- Cluster 4, $E(y_{ij}) = -8$ for $i = 201, \ldots, 400$ and $j = 6, 7, 8$
- Cluster 5, $E(y_{ij}) = 15$ for $i = 1, \ldots, 180$ and $j = 9, 10$
- Cluster 6, $E(y_{ij}) = -2$ for $i = 181, \ldots, 400$ and $j = 9, 10$

In total, we have 6 joint clusters. The random errors are assumed to be multivariate normal distribution with mean $\mathbf{0}$ and the following variance-covariance matrices for the three DPV clusters

$$\Sigma_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \Sigma_3 = \begin{bmatrix} 0.6 & 0 \\ 0 & 0.6 \end{bmatrix}.$$

Existing biclustering methods allow data points to be in more than one biclusters, while our proposed joint clustering method is mutually exclusive for the cluster assignment. We summarize the sensitivities and specificities for the 100 comparison data sets of the proposed method and bicluster methods by Cheng and Church [2000] (BCCC), and Prelić et al. [2006] (BCBimax). The results are listed in Table 3.4.

Table 3.4: Comparison of the average sensitivity for the pre-specified 6 joint clusters of proposed method, BCCC and BCBimax

	Sensitivity								
	Proposed			BCCC			BCBimax		
Cluster	Mean	Median	95% EI	Mean	Median	95% EI	Mean	Median	95% EI
1	1	1	(1,1)	0.713	0.712	(0.702,0.72)	0.990	1	(1,1)
2	1	1	(1,1)	0.997	1	(0.957,1)	0.990	1	(1,1)
3	1	1	(1,1)	0.891	0.890	(0.877,0.900)	0.990	1	(1,1)
4	1	1	(1,1)	0.748	0.750	(0.717,0.755)	0.990	1	(1,1)
5	1	1	(1,1)	0.990	0.989	(0.975,1)	0.950	1	(0.5,1)
6	1	1	(1,1)	0.680	0.682	(0.652,0.686)	0.765	0.8	(0.447,0.857)
	Specificity								
1	1	1	(1,1)	0.674	0.675	(0.655,0.681)	0.288	0.291	(0.291,0.291)
2	1	1	(1,1)	0.770	0.769	(0.766,0.779)	0.244	0.246	(0.246,0.246)
3	1	1	(1,1)	0.632	0.631	(0.617,0.639)	0.234	0.235	(0.235,0.235)
4	1	1	(1,1)	0.692	0.691	(0.689,0.705)	0.234	0.235	(0.235,0.235)
5	1	1	(1,1)	0.607	0.607	(0.594,0.614)	0.117	0.118	(0.082,0.153)
6	1	1	(1,1)	0.664	0.663	(0.661,0.678)	0.096	0.099	(0.067,0.125)

3.10 Real Data Analysis

We apply the proposed methods to a data set containing cotinine levels of 114 subjects and DNA methylation of 38 CpG sites of 114 subjects. As mentioned in Chapter 2, cotinine is an alkaloid detected in tobacco and has been used as a biomarker of smoke exposure (Benowitz [1996]). After the preliminary study, we choose the cotinine level as the covariate, and the following 10 CpG sites as the dependent variables for joint clustering: cg07442409, cg21015808, cg00295418, cg24874277, cg16116321, cg14179389, cg05575921, cg11207515, ch_18_9250, and cg18092474.

In order to fit the model, we take the log transformation of cotinine levels as the independent variable, and take the logit transformation for the methylation measures. After removing subjects with missing cotinine level and methylation measures, there are 114 subjects. Since there are fewer subjects compared to the simulation studies, we choose a big $alpha = 8$ as the concentration parameter. We propose to use 10 equally-spaced quantiles as the splines knots. All other model parameters are the same as in the simulation study. We use the DIC to determine the number of DPV clusters, and the following are the three most popular DPV clusters assignments in each there are 5 clusters (ordered by the number of occurrence in the MCMC sampling):

1. DPV clusters assignment 1 (occurred 759 times)
    - DPV cluster 1: cg14179389
    - DPV cluster 2: cg24874277, cg05575921
    - DPV cluster 3: cg11207515
    - DPV cluster 4: cg07442409, ch_18_9250
    - DPV cluster 5: cg21015808, cg00295418, cg16116321, cg18092474
2. DPV clusters assignment 2 (occurred 689 times)

- DPV cluster 1: cg07442409
- DPV cluster 2: cg24874277, cg05575921
- DPV cluster 3: cg11207515
- DPV cluster 4: cg14179389, ch_18_9250
- DPV cluster 5: cg21015808, cg00295418, cg16116321, cg18092474

3. DPV clusters assignment 3 (occured 421 times)

- DPV cluster 1: cg14179389
- DPV cluster 2: cg24874277, cg05575921
- DPV cluster 3: cg21015808, cg11207515
- DPV cluster 4: cg07442409, ch_18_9250
- DPV cluster 5: cg00295418, cg16116321, cg18092474

The difference of the DPV clusters assignments 1 & 2 is the exchanging positions of CpG sites cg14179389 and cg07442409. The difference of the DPV clusters assignments 1 & 3 is that the CpG site cg21015808 is assigned to a different CPV cluster. The proposed is pretty consistent to assign the vertical clustering.

We run two MCMC chains with 5000 iterations each chain with 1500 iterations as burn-in, the next 1500 for the determination of the average clustering matrix, and the last 2000 iterations for inferences. The subjects clustering within each DPV clusters is determined by the distance of each iterations to the average clustering matrix. We record the number of joint clusters and the minimum distance to the average clustering matrix. The result are depicted in Figure 3.5. From the figure, we can see that with the increasing of the number of joint clusters, the distance to the average clustering matrix increases, which indicates the model preferred small number of joint clusters. We notice that the bottom left of the figure contains most of the iterations

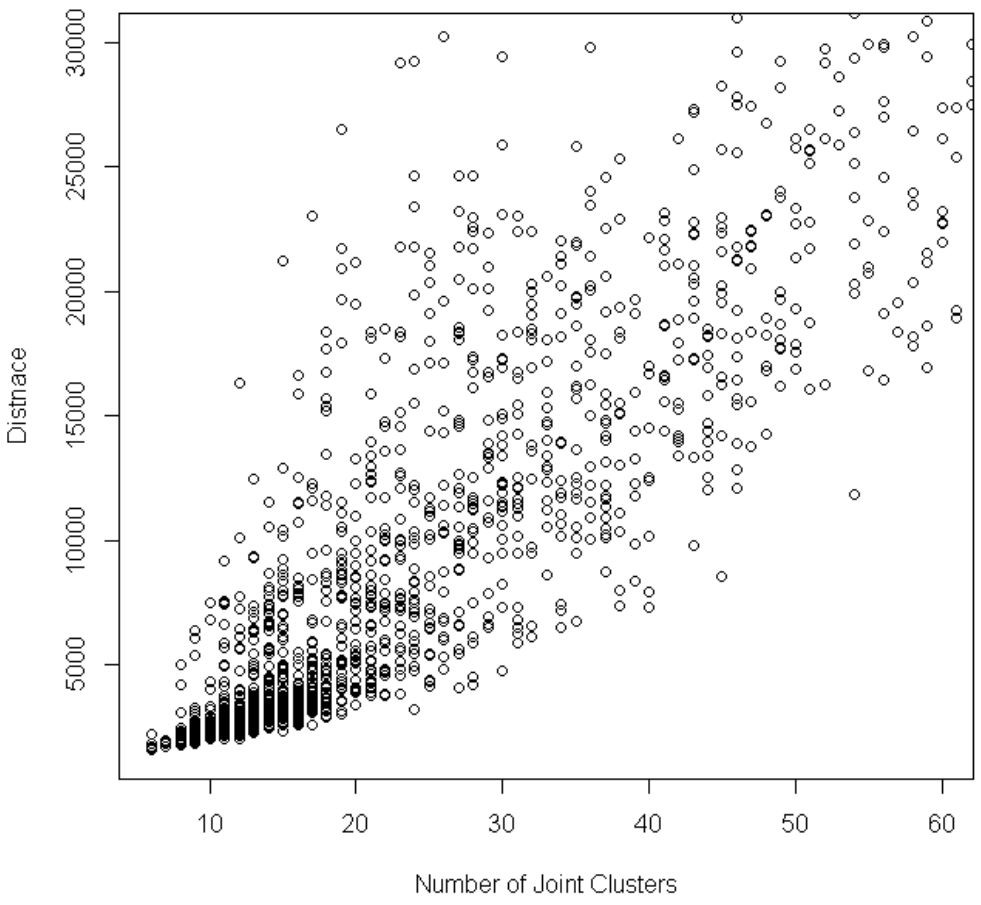

Figure 3.5: The relation of the number of joint clusters and the minimum distance to the average clustering matrix.

which corresponds to the number of joint clusters less than 20. Comparing to the 1140 pairs of data, the process of joint clustering reduces the dimension.

When we look at the details of joint clustering as the number of joint clusters varying, most of the DPV clusters have majority subjects grouping together and a few other subjects forming a small cluster. Below we listed the joint clustering assignment as the number of joint clusters varies:

Case 1: 6 joint clusters

The DPV cluster 1 has two nested subject clusters while subject ID 208 is a

cluster and all other subjects within that DPV cluster are in another cluster. All other DPV clusters have only one nested subject cluster.

Case 2: 7 joint clusters

The DPV cluster 1 has nested subject clustering as in Case 1. The DPV cluster 5 has two nested subject clusters while subject ID 946 is a cluster and all other subjects within that DPV cluster form another cluster. Other three DPV clusters have one nested subject cluster.

Case 3: 8 joint clusters

The DPV clusters 1 & 5 have nested subject clustering as in Case 2. The DPV cluster 3 has two subject clusters while subject IDs 53, 259 & 1504 form a cluster and all other subjects form another cluster. The other two DPV clusters have one nested cluster.

Case 4: 9 joint clusters

The DPV clusters 1, 3 & 5 have nested subject clustering as in Case 3. The DPV cluster 4 has two subject clusters while subject ID 191 is a cluster and all other subjects are in another cluster. The DPV cluster 2 has one nested subject cluster.

Regardless of the number of joint clusters taken based on relatively small distances, the message conveyed by the composition of each joint cluster indicates that the data might not have enough information to separate the (subject, DPV) pairs. This is likely due to the small sample size and, as indicated by our preliminary evaluations, possibly weak associations between cotinine levels and DNA methylation overall.

3.11 Conclusion

We proposed a joint clustering method to identify different associations between the response variables, subjects and covariates of interest in semi-parametric model. The

association is described using the penalized splines due to it low rank bases and ability to catch linear or non-linear effects. The joint clustering method is built upon Bayesian approach. We proposed two-step clustering: the column clustering emphasizing on detecting the different correlation among response variables, and the association of response variables and covariates of interest; the row clustering focusing on separating the different responses of individual subjects through the Dirichlet process mixture model.

The methods are demonstrated and evaluated through simulations. The simulation results show that the proposed methods can effectively identify different correlations/associations for the response variables and covariates, and also separate the different responses for the subjects. It has high sensitivity and specificity for each of the joint cluster.

The proposed methods are ready to extend to cluster other types of statistical models with multiple response variables which have different association with the same covariates of interest. On the other hand, the methods have some limitations that warrant a discussion. The two-step clustering requires grouping the response variables firstly and the number of column clusters need to be pre-specified. Then subjects clustering is nested within each column clusters. It may be desirable to relax the nesting condition and is our on-going work.

Chapter 4

Conclusion and Future Works

In this dissertation, motivated by the challenge of analyzing high dimensional epigenetic data sets, we investigated a variable selection method and a joint clustering through Bayesian approach built in semi-parametric models.

One way to reduce the dimension for the usually high dimensional epigenetic data is the variable selection, which could identify the important variants that are associated with a health outcome of interest. The existing variable selection methods applied in linear or non-linear models may not be applicable to the selection of GEVs due to the possibly complex and usually unknown form of association between GEVs and an outcome. We propose a simple method to select variables through set analyses that utilizes reproducing kernels to evaluate the relationship that is possibly non-linear and complex between the independent variables or predictors and the dependent variable. The proposed selection procedure can be used to select categorical variables, such as SNPs, and continuous variables, such as methylated CpG sites that are potentially associated with a disease. The selection procedure is built upon a statistical testing in a set analysis and the variables are selected using the backward selection scheme. We proposed two selection scenarios: the initial selection emphasizing on detecting significant sets of variables and the refinement step focusing on identifying a parsimonious set of important variables with redundant variables removed. Through simulation studies, we demonstrated that the proposed methods can effectively identify the correct variables regardless of the feature of the association, linear or non-linear. We compare the methods with the standard AIC

and BIC selection procedures, the LASSO and adaptive LASSO methods, as well as the random forest-based approach. The proposed methods are ready to extend to choose variables in other types of statistical models including log-linear models and models applied to survival data analysis. On the other hand, the methods have some limitations that warrant a discussion. The variables are selected based on the strength of their joint effect. The procedure is able to exclude redundant variables via the refinement procedure, but the exclusion is based on an evaluation of overall effect and its significance. The amount of contribution of each individual variable is not estimable in the current framework. In some situations, it may be desirable to evaluate the effect of each selected variable, besides their joint effect. Furthermore, the proposed method assumes no missing values. Accounting for missing values in the kernels surely will extend the flexibility of the proposed selection procedure and is our on-going work.

Clustering individuals along with the genetic and epigenetic information seems another good approach for dimension reduction. Traditional approaches focus on the clustering of either subjects or (response) variables. However, clusters formed through these approaches are possibly lack of homogeneity. To overcome this, biclustering was introduced [Cheng and Church, 2000], which focuses on simultaneously clustering two-dimensional gene expression data and tries to optimize a pre-specified objective function. The current bi-clustering approaches allow identifying sets of genes sharing compatible expression patterns across subsets of samples, and have been demonstrated to be useful in varies of gene expression/microarray data in terms of dimension reduction for feature identification and easy interpretation. The bi-clustering concept considers the coherence of rows and columns in the data, and is a non-model based clustering technique. It is mainly restricted to the data only and external variables do not have any contribution to the evaluation of similarity between different clustering variables. Furthermore, some bi-clustering methods per-

form cluster analyses on the rows and columns separately, and do not consider the interrelationship between the rows and columns. Most importantly, existing methods overlook the correlations between the clustering dependent variables (DPVs), which can potentially cause mis-clustering. In the dissertation, we proposed a joint clustering through Bayesian approach which considering the correlations between DPVs and the interrelationship between variables and subjects. To cluster the DPVs, a semi-parametric model with adoption of penalized splines is used to evaluate relationship between variables and covariates of interest. A Dirichlet process mixture model is applied in the process of the subjects clustering. The proposed joint clustering method has the ability to produce homogeneous clusters composed of a certain number of subjects sharing common features on the relationship between some (response) variables and covariates. On the other hand, the proposed joint clustering method assign all the variables and subjects to some cluster and does not take into account the background noise, i.e. some variables or subjects do not belong to any clusters. The proposed method considers the situation that subject clusters are nested within DPV clusters. It is worth to consider the model which can handle that DPV clusters are nested within subject clusters, or clustering process is not nested. Those are interesting future projects.

Bibliography

Anestis Antoniadis and Jianqing Fan. Regularization of wavelet approximations. *Journal of the American Statistical Association*, 96(455), 2001.

Charles E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The annals of statistics*, pages 1152–1174, 1974.

Pavlo A. Areshkov and Vadym M. Kavsan. Chitinase 3-like protein 2 (CHI3L2, YKL-39) activates phosphorylation of extracellular signal-regulated kinases ERK1/ERK2 in human embryonic kidney (HEK293) and human glioblastoma (U87 MG) cells. *Cytology and Genetics*, 44(1):1–6, 2010.

Pavlo A. Areshkov, Stanislav S Avdieiev, Olena V Balynska, Derek LeRoith, and Vadym M. Kavsan. Two closely related human members of chitinase-like family, CHI3L1 and CHI3L2, activate ERK1/2 in 293 and U373 cells but have the different influence on cell proliferation. *International Journal of Biological Sciences*, 8 (1):39, 2012.

Francis R. Bach. Consistency of the group LASSO and multiple kernel learning. *The Journal of Machine Learning Research*, 9:1179–1225, 2008.

Jeffrey D. Banfield and Adrian E. Raftery. Ice floe identification in satellite images using mathematical morphology and clustering about principal curves. *Journal of the American Statistical Association*, 87(417):7–16, 1992.

Jeffrey D. Banfield and Adrian E. Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, pages 803–821, 1993.

Neal L Benowitz. Cotinine as a biomarker of environmental tobacco smoke exposure. *Epidemiologic Reviews*, 18(2):188–204, 1996.

Christophe Biernacki and Gérard Govaert. Choosing models in model-based clustering and discriminant analysis. *Journal of Statistical Computation and Simulation*, 64(1):49–71, 1999.

David A. Binder. Bayesian cluster analysis. *Biometrika*, 65(1):31–38, 1978.

David Blackwell and James B. MacQueen. Ferguson distributions via pólya urn schemes. *The Annals of Statistics*, pages 353–355, 1973.

Hans H. Bock. Probabilistic models in cluster analysis. *Computational Statistics & Data Analysis*, 23(1):5–28, 1996.

Hans H. Bock et al. Probabilistic aspects in classification. *Data Science, Classification and Related Methods*, pages 3–21, 1998.

Charles S. Bos. Markov chain Monte Carlo methods: implementation and comparison. *WORK*, 2004.

George E. P. Box, Gwilym M. Jenkins, and Gregory C. Reinsel. *Time series analysis: forecasting and control*, volume 734. Wiley, 2011.

Russell A. Boyles. On the convergence of the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodological)*, pages 47–50, 1983.

Paul S. Bradley and Olvi L. Mangasarian. Feature selection via concave minimization and support vector machines. In *Machine Learning Proceedings of the Fifteenth International Conference (ICMLŠ98)*, pages 82–90, 1998.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Stephen P. Brooks and Andrew Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4): 434–455, 1998.

Philip J. Brown, Marina Vannucci, and Tom Fearn. Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(3):627–641, 1998.

Christopher A. Bush and Steven N. MacEachern. A semiparametric Bayesian model for randomised block designs. *Biometrika*, 83(2):275–285, 1996.

Bruce Carlson. SNPs — a shortcut to personalized medicine. *Genetic Engineering & Biotechnology News (Mary Ann Liebert, Inc.)*, 12, 2008.

George Casella and Edward I. George. Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

Jennifer L. Castle and David F. Hendry. *Automatic selection for non-linear models*. Springer, 2012.

Gilles Celeux and Gérard Govaert. A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis*, 14(3):315–332, 1992.

Zehua Chen. Fitting multivariate regression functions by interaction spline models. *Journal of the Royal Statistical Society: Series B (Statistical Methodological)*, pages 473–491, 1993.

Kin-On Cheng, Ngai-Fong Law, Wan-Chi Siu, and T.H. Lau. BiVisu: software tool for bicluster detection and visualization. *Bioinformatics*, 23(17):2342–2344, 2007.

Yizong Cheng and George M. Church. Biclustering of expression data. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, volume 8, pages 93–103, 2000.

Jukka Corander, Patrik Waldmann, and Mikko J. Sillanpää. Bayesian analysis of genetic differentiation between populations. *Genetics*, 163(1):367–374, 2003.

Jukka Corander, Patrik Waldmann, Pekka Marttinen, and Mikko J. Sillanpää. BAPS 2: enhanced possibilities for the analysis of genetic population structure. *Bioinformatics*, 20(15):2363–2369, 2004.

Jukka Corander, Pekka Marttinen, and Samu Mäntyniemi. A Bayesian method for identification of stock mixtures from molecular marker data. *Fishery Bulletin*, 104(4):550–558, 2006.

Nello Cristianini and John Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

David B. Dahl. Model-based clustering for expression data via a Dirichlet process mixture model. *Bayesian inference for gene expression and proteomics*, pages 201–218, 2006.

Kevin J. Dawson and Khalid Belkhir. A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genetical Research*, 78 (01):59–77, 2001.

Daniel Defays. An efficient algorithm for a complete link method. *The Computer Journal*, 20(4):364–366, 1977.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodological)*, pages 1–38, 1977.

Robert M. Dorazio, Bhramar Mukherjee, Li Zhang, Malay Ghosh, Howard L. Jelks, and Frank Jordan. Modeling unobserved sources of heterogeneity in animal abundance using a Dirichlet process prior. *Biometrics*, 64(2):635–644, 2008.

David Draper. Bayesian hierarchical modeling. *Department of Mathematical Sciences, University of Bath, UK*, 2000.

Richard O. Duda, Peter E. Hart, et al. *Pattern classification and scene analysis*, volume 3. Wiley New York, 1973.

Douglas F. Easton, Karen A. Pooley, Alison M. Dunning, Paul D.P. Pharoah, Deborah Thompson, Dennis G. Ballinger, Jeffery P. Struewing, Jonathan Morrison, Helen Field, Robert Luben, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, 447(7148):1087–1093, 2007.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

Paul H.C. Eilers and Brian D. Marx. Flexible smoothing with B-splines and penalties. *Statistical Science*, pages 89–102, 1996.

Michael D. Escobar. Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association*, 89(425):268–277, 1994.

Michael D. Escobar and Mike West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995.

Manel Esteller. Cancer epigenomics: DNA methylomes and histone-modification maps. *Nature Reviews Genetics*, 8(4):286–298, 2007.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pages 226–231, 1996.

Daniel Falush, Matthew Stephens, and Jonathan K. Pritchard. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4):1567–1587, 2003.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96 (456):1348–1360, 2001.

Jianqing Fan and Runze Li. New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association*, 99(467):710–723, 2004.

Jianqing Fan, Chunming Zhang, and Jian Zhang. Generalized likelihood ratio statistics and wilks phenomenon. *The Annals of Statistics*, 29(1):153–193, 2001.

Thomas S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, pages 209–230, 1973.

Mário A. T. Figueiredo. Adaptive sparseness for supervised learning. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(9):1150–1159, 2003.

Roger Fletcher. *Practical methods of optimization*. John Wiley & Sons, 2013.

LLdiko E. Frank and Jerome H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, pages 109–135, 1993.

Adelaide Freitas, Wassim Ayadi, Mourad Elloumi, José Luıs, and Jin-Kao Hao Oliveira. A survey on biclustering of gene expression data. 2012.

Herman P. Friedman and Jerrold Rubin. On some invariant criteria for grouping data. *Journal of the American Statistical Association*, 62(320):1159–1178, 1967.

Alan E. Gelfand and Adrian F.M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409, 1990.

Andrew Gelman and Donald B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, pages 457–472, 1992.

Andrew Gelman and Donald B. Rubin. Markov chain Monte Carlo methods in biostatistics. *Statistical Methods in Medical Research*, 5(4):339–355, 1996.

Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741, 1984.

Edward I. George and Robert E. McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, pages 881–889, 1993.

Edward I. George and Robert E. McCulloch. Approaches for Bayesian variable selection. *Statistica Sinica*, 7(2):339–373, 1997.

John Geweke et al. *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*. Federal Reserve Bank of Minneapolis, Research Department, 1991.

Charles J. Geyer. Practical Markov chain Monte Carlo. *Statistical Science*, 7(4): 473–483, 1992.

Walter R. Gilks, David G. Clayton, David J. Spiegelhalter, Nicky G. Best, and Alexander J. McNeil. Modelling complexity: applications of Gibbs sampling in medicine. *Journal of the Royal Statistical Society: Series B (Statistical Methodological)*, pages 39–52, 1993.

Oscar González-Recio, Daniel Gianola, Nanye Long, Kent A. Weigel, Guilherme J. M. Rosa, and Santiago Avendano. Nonparametric methods for incorporating genomic information into genetic evaluations: An application to mortality in broilers. *Genetics*, 178(4):2305–2313, 2008. doi: 10.1534/genetics.107.084293.

Irving John Good. *The estimation of probabilities: An essay on modern Bayesian methods*, volume 30. MIT press Cambridge, MA, 1965.

Peter J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

Chong Gu. Diagnostics for nonparametric regression models with additive terms. *Journal of the American Statistical Association*, 87(420):1051–1058, 1992.

Jiajun Gu and Jun S. Liu. Bayesian biclustering of gene expression data. *BMC Genomics*, 9(Suppl 1):S4, 2008.

Julius Gudmundsson, Patrick Sulem, Andrei Manolescu, Laufey T. Amundadottir, Daniel Gudbjartsson, Agnar Helgason, Thorunn Rafnar, Jon T. Bergthorsson, Bjarni A. Agnarsson, Adam Baker, et al. Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nature genetics*, 39 (5):631–637, 2007.

Steve R. Gunn and Jaz S. Kandola. Structural modeling with sparse kernels. *Machine Learning*, 48:137–163, 2002.

Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, 2002.

John A. Hartigan. Direct clustering of a data matrix. *Journal of the American Statistical Association*, 67(337):123–129, 1972.

John A. Hartigan and Manchek A. Wong. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):100–108, 1979.

Keith W. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

Hong He, Hongmei Zhang, Maity A., Yubo Zou, Hussey James, and Wilfred Karmaus. Investigating a reproducing kernel-based method for testing the combined e ect of a set of single-nucleotide polymorphisms. *Genetica*, 2012. doi: 10.1007/s10709-012-9690-5.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problesms. *Technometrics*, 12(1):55–67, 1970.

Joel L. Horowitz and Jian Huang. The adaptive LASSO under a generalized sparsity condition. *Manuscript, Northwestern University*, 2010.

Anja C. Huizink and Eduard J. H. Mulder. Maternal smoking, drinking or cannabis use during pregnancy and neurobehavioral and cognitive functioning in human offspring. *Neuroscience and Biobehavioral Reviews*, 2006.

David J. Hunter, Peter Kraft, Kevin B. Jacobs, David G. Cox, Meredith Yeager, Susan E. Hankinson, Sholom Wacholder, Zhaoming Wang, Robert Welch, Amy Hutchinson, et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nature Genetics*, 39(7):870–874, 2007.

Alejandro Jara. Applied Bayesian non- and semi-parametric inference using DPpackage. *R News*, 7(3):17–26, 2007. URL `http://CRAN.R-project.org/doc/Rnews/`.

Alejandro Jara, Timothy Hanson, Fernando Quintana, Peter Mueller, and Gary Rosner. DPpackage: Bayesian semi- and nonparametric modeling in R. *Journal of Statistical Software*, 40(5):1–30, 2011. URL `http://www.jstatsoft.org/v40/i05/`.

Karl G. Jöreskog. A general method for analysis of covariance structures. *Biometrika*, 57(2):239–251, 1970.

Bonnie R. Joubert, Siri E. Håberg, Roy M. Nilsen, Xuting Wang, Stein E. Vollset, Susan K. Murphy, Zhiqing Huang, Cathrine Hoyo, Øivind Midttun, Lea A. Cupul-Uicab, et al. 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. *Environ Health Perspect*, 2012. URL `http://dx.doi.org/10.1289/ehp.1205412`.

Andre G. Journel and Charles J. Huijbregts. *Mining geostatistics*, volume 600. Academic press London, 1978.

Sebastian Kaiser and Friedrich Leisch. A toolbox for bicluster analysis in R, compstat. *Proceedings in Computational Statistics*, 2008.

George S. Kimeldorf and Grace Wahba. Spline functions and stochastic processes. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 173–180, 1970.

Yuval Kluger, Ronen Basri, Joseph T. Chang, and Mark Gerstein. Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Research*, 13(4):703–716, 2003.

Keith Knight and Wenjiang Fu. Asymptotics for LASSO-type estimators. *Annals of Statistics*, pages 1356–1378, 2000.

Lynn Kuo and Bani Mallick. Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 65–81, 1998.

Emily K. Latch, Guha Dharmarajan, Jeffrey C. Glaubitz, and Olin E. Rhodes Jr. Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. *Conservation Genetics*, 7(2):295–302, 2006.

Laura Lazzeroni and Art Owen. Plaid models for gene expression data. *Statistica Sinica*, 12(1):61–86, 2002.

Runze Li and Hua Liang. Variable selection in semiparametric regression modeling. *Annals of Statistics*, 36(1):261, 2008.

Feng Liang, Rui Paulo, German Molina, Merlise A. Clyde, and Jim O. Berger. Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103(481), 2008.

Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002. URL `http://CRAN.R-project.org/doc/Rnews/`.

Yi Lin and Hao Helen Zhang. Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics*, 34(5):2272–2297, 2006.

Dawei Liu, Xihong Lin, and Debashis Ghosh. Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. *Biometrics*, 63(4):1079–1088, 2007. doi: 10.1111/j.1541-0420.2007.00799.x.

Jun S. Liu. Nonparametric hierarchical Bayes via sequential imputations. *The Annals of Statistics*, 24(3):911–930, 1996.

Yufeng Liu and Yichao Wu. Variable selection via a combination of the L_0 and L_1 penalties. *Journal of Computational and Graphical Statistics*, 16(4):782–798, 2007.

Stuart Lloyd. Least squares quantization in PCM. *Information Theory, IEEE Transactions on*, 28(2):129–137, 1982.

Amaia Lujambio, George A. Calin, Alberto Villanueva, Santiago Ropero, Montserrat Sánchez-Céspedes, David Blanco, Luis M. Montuenga, Simona Rossi, Milena S. Nicoloso, William J. Faller, et al. A microRNA DNA methylation signature for human cancer metastasis. *Proceedings of the National Academy of Sciences*, 105 (36):13556–13561, 2008.

David J. Lunn, John C. Whittaker, and Nicky Best. A Bayesian toolkit for genetic association studies. *Genetic Epidemiology*, 30(3):231–247, 2006.

Steven N. MacEachern. Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics-Simulation and Computation*, 23(3): 727–741, 1994.

Steven N. MacEachern and Peter Müller. Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2):223–238, 1998.

James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. California, USA, 1967.

Sara C. Madeira and Arlindo L. Oliveira. Biclustering algorithms for biological data analysis: a survey. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 1(1):24–45, 2004.

Sara C. Madeira and Arlindo L. Oliveira. A polynomial time biclustering algorithm for finding approximate expression patterns in gene expression time series. *Algorithms for Molecular Biology*, 4(1):8, 2009.

Robert J. May, Holger R. Maier, Graeme C. Dandy, and T. M. K. Fernando. Non-linear variable selection for artificial neural networks using partial mutual information. *Environmental Modelling & Software*, 23(10):1312–1326, 2008.

Jon D. McAuliffe, David M. Blei, and Michael I. Jordan. Nonparametric empirical Bayes for the Dirichlet process mixture model. *Statistics and Computing*, 16(1): 5–14, 2006.

Geoffrey J. McLachlan and Kaye E. Basford. Mixture models. inference and applications to clustering. *Statistics: Textbooks and Monographs, New York: Dekker, 1988*, 1, 1988.

Geoffrey J. McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*, volume 382. Wiley-Interscience, 2007.

Lukas Meier, Sara Van de Geer, and Peter Bühlmann. High-dimensional additive modeling. *The Annals of Statistics*, 37(6B):3779–3821, 2009.

James Mercer. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, 209:415–446, 1909.

Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21:1087, 1953.

Theo H. E. Meuwissen, B. J. Hayes, and M. E. Goddard. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819–1829, 2001.

Joanna L. Mountain and Luca L. Cavalli-Sforza. Multilocus genotypes, a tree of individuals, and human evolutionary history. *The American Journal of Human Genetics*, 61(3):705–718, 1997.

T. M. Murali and Simon Kasif. Extracting conserved gene expression motifs from gene expression data. In *Proc. Pacific Symp. Biocomputing*, volume 3, pages 77–88, 2003.

Fionn Murtagh and Adrian E. Raftery. Fitting straight lines to point patterns. *Pattern Recognition*, 17(5):479–483, 1984.

Radford M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, pages 249–265, 2000.

Ioannis Ntzoufras, Jonathan J. Forster, and Petros Dellaportas. Stochastic search variable selection for log-linear models. *Journal of Statistical Computation and Simulation*, 68(1):23–37, 2000.

Carole Ober, Zheng Tan, Ying Sun, Jennifer D. Possick, Lin Pan, Raluca Nicolae, Sadie Radford, Rodney R. Parry, Andrea Heinzmann, Klaus A. Deichmann, et al. Effect of variation in CHI3L1 on serum YKL-40 level, risk of asthma, and lung function. *New England Journal of Medicine*, 358(16):1682–1691, 2008.

Robert B. O'Hara and Mikko J. Sillanpää. A review of Bayesian variable selection methods: what, how and which. *Technometrics*, 12:55–67, 1970.

Finbarr O'Sullivan, Brian S. Yandell, and William J. Raynor Jr. Automatic smoothing of regression functions in generalized linear models. *Journal of the American Statistical Association*, 81(393):96–103, 1986.

Trevor Park and George Casella. The Bayesian LASSO. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

PPeter N. Peduzzi, R. J. Hardy, and Theodore R. Holford. A stepwise variable selection procedure for nonlinear regression models. *Biometrics*, pages 511–516, 1980.

Amela Prelić, Stefan Bleuler, Philip Zimmermann, Anja Wille, Peter Bühlmann, Wilhelm Gruissem, Lars Hennig, Lothar Thiele, and Eckart Zitzler. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22(9):1122–1129, 2006.

Jonathan K. Pritchard and William Wen. Documentation for STRUCTURE software: Version 2. 2003. *Department of Human Genetics, University of Chicago, Chicago*, 2003.

Jonathan K. Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

Peter Radchenko and Gareth M. James. Variable selection using adaptive nonlinear interaction structures in high dimensions. *Journal of the American Statistical Association*, 105(492):1541–1553, 2010.

Adrian E. Raftery and Steven Lewis. How many iterations in the Gibbs sampler. *Bayesian Statistics*, 4(2):763–773, 1992.

Pradeep Ravikumar, John Lafferty, Han Liu, and Larry Wasserman. Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030, 2009.

Gianluigi Rech, Timo Teräsvirta, and Rolf Tschernig. A simple variable selection technique for nonlinear models. *Communications in Statistics-Theory and Methods*, 30(6):1227–1241, 2001.

Christian P. Robert. *Bayesian computational methods*. Springer, 2012.

Christian P Robert and Kerrie L Mengersen. Reparametrization issues in mixture estimation and their bearings on the Gibbs sampler. *Computational Statistics and Data Analysis*, 29:325–343, 1999.

Lorenzo Rosasco, Matteo Santoro, Sofia Mosci, Alessandro Verri, and Silvia Villa. A regularization approach to nonlinear variable selection. In *Proceedings of the 13 International Conference on Artificial Intelligence and Statistics*, 2010.

Paul D. Sampson and Peter Guttorp. Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, 87 (417):108–119, 1992.

Rodrigo Santamaría, Roberto Therón, and Luis Quintales. A visual analytics approach for understanding biclustering results from microarray data. *BMC Bioinformatics*, 9(1):247, 2008.

A. J. Scott and Michael J. Symons. Clustering methods based on likelihood ratio criteria. *Biometrics*, pages 387–397, 1971.

Laura J. Scott, Karen L Mohlke, Lori L. Bonnycastle, Cristen J. Willer, Yun Li, William L. Duren, Michael R. Erdos, Heather M. Stringham, Peter S. Chines, Anne U. Jackson, et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science*, 316(5829):1341–1345, 2007.

John Shawe-Taylor and Nello Cristianini. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

Robin Sibson. SLINK: an optimally efficient algorithm for the single-link cluster method. *The Computer Journal*, 16(1):30–34, 1973.

Mikko J. Sillanpää and Madhuchhanda Bhattacharjee. Association mapping of complex trait loci with context-dependent effects and unknown context variable. *Genetics*, 174(3):1597–1611, 2006.

Robert Sladek, Ghislain Rocheleau, Johan Rung, Christian Dina, Lishuang Shen, David Serre, Philippe Boutin, Daniel Vincent, Alexandre Belisle, Samy Hadjadj, et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445(7130):881–885, 2007.

Adrian F. M. Smith and Alan E. Gelfand. Bayesian statistics without tears: a sampling–resampling perspective. *The American Statistician*, 46(2):84–88, 1992.

Michael Smith and Robert Kohn. Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, 75(2):317–343, 1996.

David J. Spiegelhalter, Nicola G. Best, Bradley P. Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002.

H. E. Theo and E. G. Mike. Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. *Genet. Sel. Evol*, 36:261–279, 2004.

Robert Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B (Statistical Methodological)*, pages 267–288, 1996.

Luke Tierney, Robert E. Kass, and Joseph B. Kadane. *Interactive Bayesian analysis using accurate asymptotic approximations*. University of Minnesota, School of Statistics, 1987.

Heather Turner, Trevor Bailey, and Wojtek Krzanowski. Improved biclustering of microarray data demonstrated through systematic performance tests. *Computational Statistics & Data ANalysis*, 48(2):235–254, 2005.

Vladimir Vapnik. *The nature of statistical learning theory*. springer, 1999.

Mary Walker, A. Gerald Shaper, Andrew N. Phillips, and Derek G. Cook. Short stature, lung function and risk of a heart attack. *International Journal of Epidemiology*, 18(3):602–606, 1989.

Stephen Walker and Paul Damien. Sampling methods for Bayesian nonparametric inference involving stochastic processes. In *Practical Nonparametric and Semiparametric Bayesian Statistics*, pages 243–254. Springer, 1998.

Hansheng Wang and Yingcun Xia. Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association*, 104(486), 2009.

Lifeng Wang, Guang Chen, and Hongzhe Li. Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics*, 23(12):1486–1494, 2007.

Joe H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

Jennifer Wessel and Nicholas J. Schork. Generalized genomic distance-based regression methodology for multilocus association analysis. *The American Journal of Human Genetics*, 79(5):792–806, 2006.

Mike West and Michael D. Escobar. *Hierarchical priors and mixture models, with application in regression and density estimation*. Institute of Statistics and Decision Sciences, Duke University, 1993.

Chang-Jiun Wu and Simon Kasif. Gems: a web server for biclustering analysis of expression data. *Nucleic Acids Research*, 33(suppl 2):W596–W599, 2005.

Jeff C. F. Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.

Michael C. Wu, Peter Kraft, Michael P. Epstein, Deanne M. Taylor, Stephen J. Chanock, David J. Hunter, and Xihong Lin. Powerful SNP-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics*, 86(6):929–942, 2010. doi: 10.1016/j.ajhg.2010.05.002.

Shizhong Xu. Estimating polygenic effects using markers of the entire genome. *Genetics*, 163(2):789–801, 2003.

Meredith Yeager, Nick Orr, Richard B. Hayes, Kevin B. Jacobs, Peter Kraft, Sholom Wacholder, Mark J. Minichiello, Paul Fearnhead, Kai Yu, Nilanjan Chatterjee, et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nature Genetics*, 39(5):645–649, 2007.

Nengjun Yi and Shizhong Xu. Bayesian LASSO for quantitative trait loci mapping. *Genetics*, 179(2):1045–1055, 2008.

Nengjun Yi, Varghese George, and David B. Allison. Stochastic search variable selection for identifying multiple quantitative trait loci. *Genetics*, 164(3):1129–1138, 2003.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

Arnold Zellner. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno De Finetti*, 6:233–243, 1986.

Daowen Zhang and Xihong Lin. Hypothesis testing in semiparametric additive mixed models. *Biostatistics*, 4(1):57–74, 2003.

Hao Helen Zhang, Grace Wahba, Yi Lin, Meta Voelker, Michael Ferris, Ronald Klein, and Barbara Klein. Variable selection and model building via likelihood basis pursuit. *Journal of the American Statistical Association*, 99(467), 2004.

Hongmei Zhang and Jianjun Gan. A reproducing kernel-based spatial model in poisson regressions. *International Journal of Biostatistics*, In Press, 2012.

Yuan-Ming Zhang and Shizhong Xu. A penalized maximum likelihood method for estimating epistatic effects of QTL. *Heredity*, 95(1):96–104, 2005.

Ji Zhu, Saharon Rosset, Trevor Hastie, and Rob Tibshirani. 1-norm support vector machines. *Advances in Neural Information Processing Systems*, 16(1):49–56, 2004.

Hui Zou. The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

Hui Zou and Runze Li. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36(4):1509, 2008.

Grietje Zuur, Paul H Garthwaite, and Rob J Fryer. Practical use of MCMC methods: lessons from a case study. *Biometrical Journal*, 44(4):433–455, 2002.