Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

Zhuxuan Jin	Date

Statistical Methods for Omics Data Integration

By

Zhuxuan Jin

Doctor of Philosophy

Biostatistics

Jian Kang, Ph.D.
Advisor

Tianwei Yu, Ph.D.
Advisor

Howard Chang, Ph.D.
Committee Member

Zhaohui Steve Qin, Ph.D.
Committee Member

Jing Zhang, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Statistical Methods for Omics Data Integration

By

Zhuxuan Jin
B.S., Peking University, 2013

Advisors: Jian Kang, Ph.D. and Tianwei Yu, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2017

Abstract

Statistical Methods for Omics Data Integration

By

Zhuxuan Jin

In this dissertation, we are interested in developing novel statistical methods for omics data integration with an application in various biostatistics problems.

In the first topic, we propose a statistical model to integrate gene expression profiles with gene network for feature classification. Existing methods do not allow flexible modeling of sub-types of genes and they ignore nodes without observed expressions. To address these limitations, we propose a Bayesian nonparametric method for gene classification. A new prior is developed for the class indicators incorporating the network dependencies. Missing gene nodes are handled by imputation. Our method can achieve increased classification accuracy in simulations. We illustrate our method on a survival analysis of the cutaneous melanoma dataset from the Cancer Genome Atlas and obtain some meaningful results.

In the second topic, we propose a computational method for integrating the LC-MS metabolomics data with the metabolic network and adduct ion relations for missing value imputation. Existing methods are mostly borrowed from microarray studies without considering feature relations or network information. Our algorithm incorporates the metabolic network, adduct ion relations, linear and nonlinear associations between features to build a feature-level network. The proposed method resorts to support vector regression for imputation based on features in the neighborhood on the network. It can achieve a smaller normalized root mean squared error in real data-based simulations.

In the third topic, we propose a statistical model to integrate genotypes with brain imaging phenotypes for activation shape estimation and gene discovery for Alzheimers disease. There is lack of statistical methods to perform genetic dissection of brain activation phenotypes such as shape and intensity. We propose a Bayesian hierarchical model which consists of two levels of hierarchy. At level 1, a Bayesian non-parametric level set model is used for studying the activation shape. At level 2, a regression model is constructed to select genetic variants that are strongly associated with the activation intensity, where a spike-and-slab prior and a Gaussian process prior are chosen for feature selection. The advantages of the method are illustrated via simulations and analyses of imaging genetics data from the Alzheimers disease neuroimaging initiative.

Statistical Methods for Omics Data Integration

By

Zhuxuan Jin
B.S., Peking University, 2013

Advisors: Jian Kang, Ph.D. and Tianwei Yu, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2017

Acknowledgement

I would like to thank my advisors, Dr. Tianwei Yu and Dr. Jian Kang, for their marvelous help and guidance towards becoming an outstanding researcher, for all the encouragement and support during my time at Emory. They have spent countless hours passing the torch of knowledge on to me. I thank them for believing in me, and in the process of doing so, helping me believe in myself. From them, I see the excellence, the truest love for research and the love for their students. Without their tremendous help, I would never have been able to finish my dissertation.

I would like to thank my dissertation committee members, Dr. Zhaohui Steve Qin, Dr. Howard Chang and Dr. Jing Zhang, for the rigorous and insightful academic advising, constructive comments and suggestions on this work. I thank them for their excellent guidance, caring and patience along the way.

Besides, I would like to thank Dr. Lance A. Waller and Dr. Howard Chang for their mentoring and consulting on another research project on which I was a research assistant. I thank them as they helped navigate my unique journey exploring another research field.

I also want to express my deepest gratitude to Melissa Sherrer, Mary Abosi and Bob Waggoner, for their care and support. Emory is a third hometown to me and my life is closely tied to the people here. I have met several batches of best friends: Yun Wei, Qingyang Xiao, Fengchao Liang, Ziyi Li, Na Bo. No matter where I am in the future, Atlanta is always the warmest home for me.

Last but most importantly, I would like to thank my father and mother for their longstanding love and care, for always encouraging me to follow my passions and to be my best self. I still remember the time when they told me that they would support whatever I would like to do, and wherever I am family would always be there for me. They deserve much credit for guiding me towards being a better person in life and

making me what I am today. I would also like to thank my sweetest boyfriend Jia, for all his love, support and company. He has always looked out for my future and been always standing by my side. I thank him for always seeing the best in me, bringing all the joy and happiness in my life. I feel extremely lucky to have you all in my life.

Thank again to everyone who made this dissertation and me possible.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

1.1 Omics Data

Omics data is originally proposed in the biology field, where it represents a large family of cellular molecules, including genes, proteins, metabolites, *etc.* We name them using the common suffix "omics": genomics, proteomics, metabolomics, *etc.* Motivated by this concept, omics data can also represent the quantitative features converted from tomographic images as well as the analysis of their correlations with genomic patterns, such as radiomics or radiogenomics (Gillies et al., 2015). Nowadays, the term "omics" is widely used in various biomedical research fields that generate high-dimensional and large-scale datasets from single objects or samples (Micheel et al., 2012). This complex data yields unprecedented opportunities as well as raises enormous challenges in biostatistics research.

Omics data in the biology field, as discussed by Joyce and Palsson (2006), can be presented at three different levels: omics data at the component level, omics data at the interaction level and omics data at the function-state level. Omics data at the component level refers to a particular type of molecule in a cell or in a biological system, such as genomics (the whole genome sequencing data), transcriptomics (the microarray-based/RNA-seq-based genome-wide expression profiles), proteomics (the protein mass spectrometry-based protein sequence and composition data) and metabolomics (metabolite quantities data). Omics data at the interaction level mainly includes the interactions among them, such as the protein-DNA interactions and the protein-protein interactions. The protein-DNA interactions are the interactions between transcription factors and their target promoters, which depict the genetic regulatory network. The protein-protein interactions are defined based on cellular functionality, which provide important information about the integrated cellular network. In biostatistics research, omics data at the function-states level is also referred to as phenotypes. They are collected as the physical, biochemical, clinical characteristics of the samples with/without changes in response to possible genetic mutation

or potential environmental influences.

Omics data in biomedical imaging filed represents the tremendous quantitative features extracted from tomographic images, including computed tomography (CT), magnetic resonance (MR) or positron emission tomography images (PET). Quantitative features can be presented at different levels: two-dimensional or three-dimensional image where there are millions of voxels; or derived measurable features from these imaging datasets such as size, shape, region, texture, *etc.*

Therefore, in order to draw a more comprehensive view of biological processes, to gain a deeper understanding of biological systems, it is suggested that omics data from different sources be integrated together in the data analysis. The integration can be accomplished at various levels. For instance, the integration of omics data within component levels, the integration of omics data from component levels with omics data from interaction levels or function-state levels, the integration of omics data from biology filed and from biomedical imaging filed.

However, the complexities of biological systems, the limitations of current technologies, the tremendous amount and the heterogeneity of omics data, raise great challenges in developing omics data integration methods. First, the dimensionality issue. Omics datasets are usually generated in high-throughput, resulting in datasets of high-dimension and a relatively small number of samples. Second, correlations between omics data are complicated. Although integrating omics data sounds appealing, the integration methods need to be deliberately designed without jeopardizing any of the uniqueness of omics data or introducing additional noise. And third, there are always unknown factors in the studies. However, as is known to all, great challenges come with great opportunities. In this dissertation, we are interested in developing statistical methods to integrate different omics data to address various biostatistical questions.

Omics data integration has been paid increasing attention in bioinformatics and

biomedical sciences, thus, tremendous work has been developed in this field. Researchers nowadays are mostly interested in integrating different omics datasets to examine a specific hypothesis, to answer biological questions or to better understand the biological processes. There are reviews on omics data integration but with different scopes. Selected literature includes: Joyce and Palsson (2006) reviewed the challenges, methods for studying biological systems and discussed the future research directions in this field. Zhang et al. (2010) reviewed the basic concepts, recent applications, and statistical methodologies in the scope of microbes. Berger et al. (2013a) reviewed current omics data integration methods in the mathematical aspects. Berger et al. (2013b) reviewed computation techniques and software tools. Gomez-Cabrero et al. (2014) focused on the integration methods in life science. Ritchie et al. (2015a) reviewed the approaches for genotype-phenotype interactions. Luo et al. (2016) discussed the data integration applications in biomedical and health-care informatics research. Buescher and Driggers (2016) focused on discussion within multi-omics studies in cancer research. Hasin et al. (2017) discussed the omics data integration methods regarding the research in human disease.

1.2 Applications

Much effort has been made in the omics data research. The objectives in omics data integration methods can be as specific as to annotate a gene of interest or as broad as to uncover the mystery of evolution. As it is impossible to cover every single question related to omics data integration methods in biostatistics, in this dissertation, we will only focus on some of the questions as an application.

In Chapter 2, we focus on data integration methods in feature classification. Feature classification, which is inherited from feature selection, aims to provide more insight into sub-category-related information for features. It is not a only "selecting"

process, but more of a "classifying" process. In Chapter 3, we focus on the missing value imputation problem. High-throughput technologies, such as gene expression microarray or liquid chromatography-mass spectrometry (LC-MS) data suffer from missing values due to complicated technical or experimental reasons. The way to handle missing values will largely influence downstream analysis. Thus, it is also of great importance for us to investigate methods to handle missing values in omics data. In Chapter 4, we are interested in Alzheimer's disease. Alzheimer's disease, one of the most common forms of dementia, causes inevitable losses of memory, cognitive and physical abilities. However, the etiology of Alzheimer's disease is still mostly unknown and there is no cure for Alzheimer's disease currently. To make it worse, Alzheimer's disease is usually mistakenly recognized as a form of normal aging, causing a large delay in treatment. Thus, in this topic, we are interested in omics data integration methods with an application to Alzheimer's disease, aiming to provide some insight into the genetic dissection of brain activation phenotypes that are potentially helpful in early disease diagnostics.

1.2.1 Feature Classification

Based on our knowledge, there are many fewer statistical methods developed so far in feature classification, compared to the tremendous work in feature selection. Although feature classification can be achieved via a selection process followed by a classification process, this intuitively appealing two-stage process will suffer from reduced performance in classification due to the complete separation of the two individual processes without information exchange. Thus, developing a multi-class classification model will be of great importance by filling the gap.

There is some literature designed for feature selection with or without disease (sample) classification in genetics studies, which is different from the question in our scope, where we aim to conduct classifications on the features themselves. Feature

selection is usually accomplished by selecting some genes from thousands of candidate genes that are believed to be essential or influential. Roughly, there are mainly two categories of feature selection methods that are popular in this field: the filter methods and the wrapper methods (Saeys et al., 2007a). Filter methods aim to score each gene by some metrics or statistics and then to select genes with scores within a pre-defined threshold, such as Dudoit et al. (2002); Jafari and Azuaje (2006); Sartor et al. (2006), especially the methods built under multiple hypothesis testing framework: Efron, Tibshirani, Storey and Tusher (2001); Tusher et al. (2001); Pan (2003); Kendziorski et al. (2003); Dudoit et al. (2003); Newton and Kendziorski (2003); Dudoit et al. (2004); Pollard et al. (2005); Efron (2005); Ploner et al. (2006), *etc.*, where their work inspired us to develop a feature classification method that should also better control the false discovery rate. Other works built within the context of network were developed later on, such as Winter et al. (2012); Cun and Fröhlich (2013); Ding and Peng (2005); Zou et al. (2016); Mohammadi et al. (2016), they inspired us to incorporate the network information into feature classification. Compared to the filter methods, wrapper methods work in an interactive way. A subset of candidate features are selected first and scored by the overall classification accuracies or clustering goodness *etc*, then the subset is adjusted accordingly until the optimal / sub-optimal feature subset is found. Works include: Kohavi and John (1997); Xiong et al. (2001); Ghosh and Chinnaiyan (2005); Jirapech-Umpai and Aitken (2005); Ruiz et al. (2006); Ma and Huang (2008); Solorio-Fernández et al. (2016). But as it is an NP-hard problem to search for all possible feature subsets, one major issue with wrapper methods is their high computational complexity. Besides filter and wrapper, there are some methods that combine both filter and wrapper, such as Soufan et al. (2015); Apolloni et al. (2016a).

In Chapter 2, we are interested in feature classification methods inspired by some of the work in feature selection. We aim to provide a classification algorithm that

can select and classify features into different subtypes interactively in the model. Different from the supervised methods in the field of text mining or computer vision, our method is built under an unsupervised framework and to make the method flexible enough, we aim to provide a solution nonparametrically. As some of the genetics datasets do not have an outcome variable attached, such as the Spellman yeast cell cycle microarray dataset (Spellman et al., 1998), our method is built on the test-statistics from a generalization perspective. Additionally, as test statistics are generated from thousands of hypothesis tests applied on each of the genes, our methods need to better control the false discovery rate. Lastly, existing network information including biological pathways and molecular interactions have been found to be helpful in depicting the feature relations, thus, a network-based feature classification is of great interest for us.

1.2.2 Missing Values in Omics Data Research

Although various platforms for generating omics data have been developed significantly during the past years, technologies are still rather prone to errors, resulting in datasets with different levels of missingness. In our motivating dataset used in Chapter 2, among the genes with at least one connection in the network, around 6% are not linked with a gene expression profile and another 14% are observed genes but of low and unreliable expressions for statistical testing. However, most of the current methods remove missing gene nodes and use a reduced gene network in data analysis, which will potentially jeopardize the performance of the methods or their downstream analysis. To address these limitations, another objective in the first topic is to develop methods that can appropriately handle the missing gene nodes in the network. After that, inspired by the imputation techniques we developed in the first topic, in the second topic in Chapter 3, we focus on the imputation methods in the LC-MS metabolomics datasets.

Little and Rubin (2014) have laid the foundation of missing mechanisms in data matrices and provided fundamental ideas on how to handle missingness. For example, when features in some samples are missing due to some technical problems that are not related to the values, missing or observed, we refer to it as "missing completely at random" (MCAR). Usually it is the simplest to handle. When missingness in features is only related to the observed values but not the missing values, we refer to it as "missing at random" (MAR). When the missing is related to the missing values itself, we refer to it as "missing not at random" (MNAR). Usually it is hard to deal with. Ideally, methods to handle missingness in datasets should rely on the mechanism that caused the values to be missing. However, it is very hard to justify the missing mechanisms in real data applications sometimes. For instance, the missing mechanism in metabolomics studies can result from a failure in computational detection or a low signal that is hard to detect. Based on the analysis of DI FT- ICR MS metabolomics datasets, Hrydziuszko and Viant (2012) argued that missing values in metabolomics data do not occur randomly but may be closely related to missing features' signal intensities and mass-to-charge ratios. As a result, it will be risky to assume one particular type of missing mechanisms (MCAR or MAR) in the study. Missing value problems need to be carefully handled in omics data research.

In general, there are two types of methods for handling missingness in the existing omics data research, perhaps largely overlapping with each other: one is the generic method and the other is the application-specific method. For methods built for a generic purpose, they can be roughly categorized into different subcategories following a similar taxology in the field of pattern recognition:

- Remove missing values and continue data analysis completely based on the available observations.

- Impute missing values and use the imputed values in the data analysis.

- Utilize multiple imputation techniques where several estimated matrices are evaluated in the downstream analysis and then final imputation will be pooled together.

- Model missing values in a probabilistic model with one pre-specified missing mechanism under either the parametric or nonparametric framework, continue data analysis using the expected values or similar. Methods include various expectation-maximization (EM) based algorithms or posterior samples in the Bayesian analysis.

- Integrate with external sources of information, such as other experimental data or results from other similar studies.

In Chapter 2, the issue of missing nodes in the network is not a typical missing value issue. As our proposed method is built under a Bayesian framework, we can impute the missing test statistics by its posterior samples. We also propose a fast version by inserting the values calculated based on the nodes' nearest neighbors. In Chapter 3, missing value imputation methods in metabolomics studies are mostly borrowed from the microarray studies, such as imputation by inserting a single value, like mean, median values, half of the minimum values, *etc.*; imputation based on global similarity, like the Bayesian principal component analysis (Oba et al., 2003), singular value decomposition (Troyanskaya et al., 2001), and imputation based on local similarity, like the K-nearest-neighbors (Hastie et al., 1999), *etc.* However, metabolomics data have their own uniqueness that should draw great attention, for instance, the existing metabolic network, adduct ion relations even for unknown compounds, as well as linear and nonlinear associations between feature intensities. Compared to the methods developed more from a generic point of view, in Chapter 3, we are more interested in the application-specific methods that are specially designed for LC-MS metabolomics data.

1.2.3 Alzheimer's Disease

Alzheimer's disease is the most prevalent form of dementia. It is a chronic and progressive neurodegenerative disease, resulting in large deficits in the brain cells, leading to a great loss in memory, cognitive skills, and physical activity abilities that disrupt daily life. In America, there are so far more than 5 million patients living with it and the treatment costs around $236 billion, the patient number will rise to 16 million and the estimated cost will reach more than $1 trillion by 2050 (http://www.alz.org/).

Even though it has been more than 100 years since its discovery, we still do not know the exact causes and there is still no cure for Alzheimers disease. One difficulty is the detection and early diagnosis of Alzheimer's disease, because it is largely neglected as a form of normal aging. In fact, patients will benefit from early diagnosis of Alzheimer's, such as access to available treatment with cholinesterase inhibitors, access to a care system from health providers and supportive services, help in developing clinical trials. Thus, methods developed for early detection and diagnosis will be beneficial to society.

Several studies have already unveiled the fact that the activation shape formation, such as the activation shapes in the hippocampus region, is known to be atrophied in Alzheimer's patients (Scheltens et al., 1992; Braak and Braak, 1995; Seab et al., 1988; Dickerson et al., 2001; Frisoni et al., 2010a). Shape analysis for local brain atrophy based on imaging data has drawn increasing attention to study Alzheimer's disease. Rathore et al. (2017) gave a comprehensive review of neuroimaging-based classification studies based on publications from PubMed and Google Scholar from January 1985 to June 2016. From their studies, methods can be roughly categorized into two approaches: the region-of-interest (ROI) based approach, such as Lerch et al. (2008); Cuingnet et al. (2011a), which requires prior knowledge for region selection; and the machine learning based methods, such as Chiang and Pao (2016); Habes

et al. (2016), where manually labeled data are used for training and testing purposes. For studying the regional shape abnormalities, Li et al. (2007) built a classification method based on surface-based mesh modeling techniques; Cuingnet et al. (2011a) modeled the activation shape of the hippocampus based on spherical harmonics; Shen et al. (2012) detected global and local shape changes using statistical shape models. Lao et al. (2004); Wang et al. (2007); Tang et al. (2016) focused on the application of dimension reduction methods in classification. Sajda (2006); Chu et al. (2011); Mller et al. (2015) utilized advanced machine learning algorithms for differentiating Alzheimer's disease patients.

Inspired by the recent development of shape analysis, in Chapter 4, we aim to develop a novel statistical method to detect potential atrophy in brain activation shapes. Moreover, compared to the research in the imaging field, the Alzheimer's Disease Neuroimaging Initiative (ADNI, http://adni.loni.usc.edu/) provides a new regimen for brain shape activation analysis by integrating external data sources, such as genetics data, into data analysis. Thus, in this topic, our study of Alzheimer's disease mainly focuses on jointly analyzing imaging data and genetics data. Our objective is to detect genes that are associated with brain activation shape atrophy in Alzheimer's disease.

1.3 Omics Data Sources

There is an increasing number of open data sources for omics data research. The popular ones include: The Human Genome Project (Lander et al., 2001; Venter et al., 2001), Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005), 1000 Genomes Project (Consortium et al., 2010, 2012), The Encyclopedia of DNA Elements Project (ENCODE) (Ecker et al., 2012), Immunological Genome Project (ImmGen) (Shay and Kang, 2013), The Gene Expression Omnibus (GEO) (Edgar

et al., 2002; Barrett et al., 2013), *etc.* I will only introduce the databases used in this dissertation in the following:

- The Cancer Genome Atlas Project (TGCA). It aims to provide insights into the heterogeneity of different cancer subtypes by creating a map of molecular alterations for every type of cancer. For instance, in chapter 2, the skin cutaneous melanoma dataset has been characterized by mRNA/ miRNA expression, protein expression, pathology review, whole genome, copy number variations, DNA methylation profiling, *etc.* (http://cancergenome.nih.gov/)

- The Gene Ontology Consortium (Gene Ontology). It aims to produce a dynamic, controlled vocabulary that can be applied to all eukaryotes, even as a dictionary of gene and protein roles in cells (Ashburner et al., 2000).

- Kyoto Encyclopedia of Genes and Genomes (KEGG). It aims to provide the reference knowledge base that integrates current knowledge on molecular interaction networks such as pathways and complexes, the information about genes and proteins generated by genome projects (GENES/SSDB/KO databases) and information about biochemical compounds and reactions are extracted from COMPOUND/GLYCAN/REACTION databases (Kanehisa et al., 2004).

- Alzheimers Disease Neuroimaging Initiative (ADNI). It is an ongoing project initiated in 2003 by NIA, NIBIB, FDA, *etc.* It aims to uncover the etiology of Alzheimer's disease. ADNI collects various types of data, including serial magnetic resonance imaging (MRI), positron emission tomography (PET), genetic factors such as single nucleotide polymorphisms (SNPs), other biological markers and clinical and neuropsychological assessments. With the help of ADNI, researchers can benefit from a whole genome and whole brain data to study the genotype-phenotype associations for studying Alzheimer's disease (Jack et al., 2008).

1.4 Outline

Following this introduction, there are three chapters. In Chapter 2, we propose a Bayesian nonparametric modeling approach incorporating network information and imputing missing gene nodes in the network for feature classification. In Chapter 3, we continue exploring the imputation techniques incorporating network information, adduct ion relations, various linear or nonlinear correlations among the features in the metabolite studies. Lastly, in Chapter 4, we investigate the integration methods in genotypes with brain imaging phenotypes for discovering the gene influence with activation shape atrophy looking for neurodegeneration evidence in Alzheimer's disease.

Chapter 2

Integrate Gene Expression Profiles with Gene Network for Feature Classification

2.1 Introduction

Feature selection is a fundamental problem in high-dimensional data analysis, especially in the field of genomics, where researchers are interested in classifying features in different categories according to their biological characteristics. We refer to this procedure as feature classification. Traditional differential expression framework calculates false discovery rates, i.e. posterior probabilities of differential expression using parametric or nonparametric density estimations, without considering biological relations between features (Efron and Tibshirani, 2002; Do et al., 2005). However, existing biological networks including biological pathways and molecular interactions have been found to be helpful for depicting the biological relationship between the features. Thus researchers in statistics and bioinformatics have paid increasing attention to developing network-based approaches.

Some filtering algorithms were developed in the machine learning and bioinformatics fields, without much consideration of statistical inference (Cun and Fröhlich, 2012; Cun and Fröhlich, 2013; Apolloni et al., 2016a). In the statistics field, the main approach for network-based feature selection is built under the parametric/ regression framework, such as Wei and Li (2007); Li and Li (2008); Pan et al. (2010); Li and Zhang (2012); Stingo and Vannucci (2011); Stingo et al. (2011); Ročková and George (2014), where model structures are developed to capture the dependency of genes by using various penalties that smooth the regression coefficients of the features over the network, or applying different priors utilizing the structure of the network. We have previously proposed a Bayesian nonparametric feature selection approach incorporating the network information (Zhao et al., 2014). Based on the test statistics on a per-feature basis, the method is flexible enough to allow any type of association between features and outcome variables or even testing behavior of the features without an outcome variable. However, it has some limitations. It assumes the test statistics of the null genes follow a symmetric distribution. Also, it lumps up-regulated and

down-regulated genes into a single group and assumes they behave symmetrically. In this work, we address these issues by developing a more flexible framework. To the best of our knowledge, we are the first to develop a network-based feature classification method that allows asymmetric null distribution, as well as different levels of deviation from the null for down-regulated and up-regulated genes.

Another important issue in network-based analysis is that some nodes may not be observed in the expression data. Little and Rubin (2014) laid the foundation of missing mechanisms and provided ideas on how to handle missingness. However such approaches do not handle the missing of entire rows in the data well. In our motivating dataset, among the genes with at least one connection, around 6% are not measured, and another 14% are observed but of a low and unreliable expression level for statistical testing. We treat them as nodes with missing observations in the network. As a result, the occurrence of missingness makes our problem even more challenging. Ignoring such nodes and their edges, as all existing methods do, causes severe loss of network structure and biases the results. Thus handling missing nodes in the network is of great importance to our problem, due to the fact that missing nodes are possible to be either down-regulated or up-regulated genes, and/or serve to communicate information via their edges with observed genes (See Figure 2.1).

Our method is based on Bayesian nonparametric posterior inference on the class indicators. We do not impose any parametric assumptions on the distributions of the test statistics for each class. Instead, we use the Dirichlet process mixture (DPM) model. DPM is widely used and extensively studied from the literature (See Neal (2012) for an overview of DPM). Other than that, in our model, we assign a weighted Potts prior to the class indicators in order to capture the dependence among genes. The weighted Potts prior is a generalization of the Ising prior from two categories to multi-categories that can satisfy the three-class feature classification problem. We then perform a fully Bayesian inference on the proposed model via Metropolis-

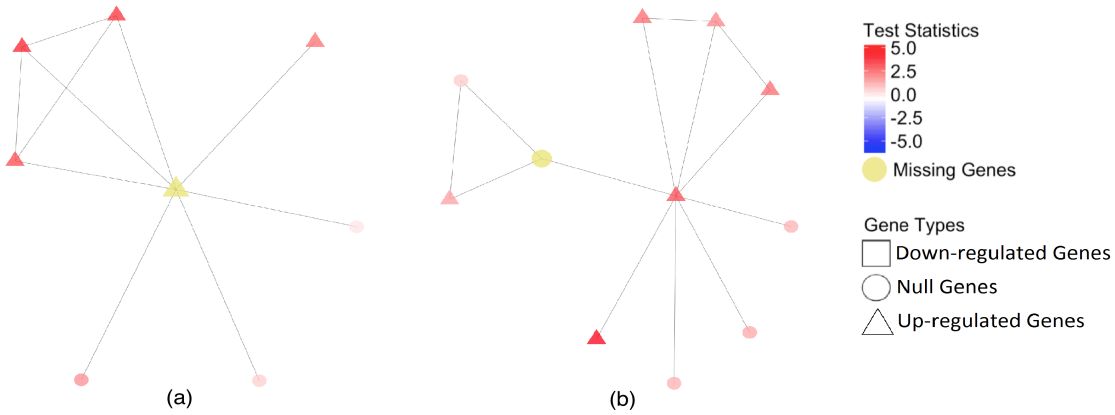


Figure 2.1: The impact of missing genes in the network. (a) the missing gene serves as a "bridge" for information exchange. If it is simply removed, the light red node located on the top right side would not be able to be recalled as up-regulated gene; (b) the missing gene is itself an up-regulated gene, it would be excluded if missing genes are removed from data analysis.

Hastings within Gibbs sampling. The proposed method has been implemented in the BANFF package by Lan et al. (2016) using the function BANFF^2. It is worth noting that since we study thousands of nodes simultaneously, our method utilizes the local false discovery rate control rule proposed by Efron, Storey and Tibshirani (2001).

The remainder of the manuscript is organized as follows. In Section 2.2, we describe the proposed model and the prior specifications. In Section 2.3, we present the posterior computation algorithms. In Section 2.4, we compare the performance of the proposed method with the traditional methods via extensive simulation studies. In Section 4.4, we analyze the cutaneous melanoma dataset and provide biologically meaningful results.

2.2 Bayesian Nonparametric Feature Classification

2.2.1 The Model

Let n be the total number of genes in the gene network. Let i index gene for $i = 1, 2, \cdots, n$. For each gene i, we can obtain a test statistic r_i, which can be considered as a function of the gene expression profile, the collected phenotypes and the clinical outcome. To perform the feature classification, we introduce a latent class indicator z_i which takes values $-1, 0$ and 1 representing gene behaviors: down-regulated, not differentiated expressed (null) and up-regulated, respectively. Given $z_i = k$, we further introduce a cluster index g_i, which represents the cluster index indicating which component in the mixture model that r_i is associated with. In particular, r_i given g_i is assumed to be normally distributed with mean $\widetilde{\mu}_{g_i}$ and variance $\widetilde{\sigma}^2_{g_i}$, denoted by $N(\widetilde{\mu}_{g_i}, \widetilde{\sigma}^2_{g_i})$. We write $\widetilde{\boldsymbol{\theta}}_g = (\widetilde{\mu}_{g_i}, \widetilde{\sigma}^2_{g_i})$ and assume they are independently drawn from a base measure called G_{0k}. The $\widetilde{\boldsymbol{\theta}}$ denotes all the $\widetilde{\boldsymbol{\theta}}_g$s for simplicity. Given $z_i = k$, g_i follows a discrete distribution with parameter $\mathbf{a}_k, \mathbf{q}_k$, which means g_i can take values in $\boldsymbol{a_k} = (a^k_1, a^k_2, \cdots, a^k_{L_k})$ with probability $\mathbf{q}_k = (q^k_1, q^k_2, \cdots, q^k_{L_k})$, denoted Discrete$(\boldsymbol{a}_k, \boldsymbol{q}_k)$. In fact, the actual values of g_i given $z_i = k$ is arbitrary, thus we can assume $\mathbf{a}_k = (1, 2, \cdots, L_k)$ without loss of generality. The probability $\mathbf{q}_k$ follows a Dirichlet distribution with parameters $(\tau_k/L_k, \tau_k/L_k, \cdots, \tau_k/L_k)$. Note that the total number of components L_k for all $k = -1, 0, 1$ are also unknown, thus this extended DPM model is nonparametric in nature. In summary, we have the following Bayesian hierarchical model:

$$
\begin{aligned}
& r_i | g_i, \widetilde{\boldsymbol{\theta}} \sim N(\widetilde{\mu}_{g_i}, \widetilde{\sigma}^2_{g_i}), \\
& g_i | z_i = k, \boldsymbol{q}_k \sim \text{Discrete}(\boldsymbol{a}_k, \boldsymbol{q}_k), \\
& \widetilde{\boldsymbol{\theta}}_g \sim G_{0k} \quad \text{for } g \in \boldsymbol{a}_k, \\
& \boldsymbol{q}_k \sim \text{Dirichlet}(\tau_k \mathbf{1}_{L_k}/L_k).
\end{aligned}
\tag{2.1}
$$

2.2.2 Prior Specifications

Suppose the gene network is provided. Let $\mathbf{C} = \{c_{ij}\}$ be the adjacency matrix characterizing the gene network configuration, where $c_{ij} = 1$ if genes i and j are biologically connected and $c_{ij} = 0$ otherwise. To incorporate this topology structure, we assign a weighted Potts prior to $\mathbf{z} = (z_1, \ldots, z_n)$, denoted by wPotts$(\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{w}, \mathbf{C})$, where $\boldsymbol{\pi} = (\pi_{-1}, \pi_0, \pi_1)$ with $\pi_k > 0$, $\boldsymbol{\rho} = (\rho_{-1}, \rho_0, \rho_1)$ with $\rho_k \geq$ and $\boldsymbol{w} = (w_1, \ldots, w_n)$ with $w_i \geq 0$. Then the probability mass function is proportional to

$$\exp\left[\sum_{i=1}^{n}(\widetilde{\omega}_i \log(\pi_{z_i}) + \rho_{z_i}\sum_{i\neq j}\omega_j c_{ij} I[z_i = z_j])\right]. \tag{2.2}$$

The parameter $\boldsymbol{\pi}$ contains prior knowledge about the class indicator $\mathbf{z}$. We assume that $\pi_1 + \pi_{-1} < \pi_0$ implying that signals are sparse. Similar to the Ising model, parameter ρ_k controls the global strength of the neighborhood similarity. When $\rho_k = 0$, z_i is independent with z_j for j in the neighborhood of i. However, when $\rho_k > 0$, z_i has a larger probability to take the value of k when $z_j = k$ for j in the neighborhood of i. Across the whole gene network, the larger the ρ_k is, the stronger the tendency to share the same memberships with neighbors. Weight w_i can be elicited from the prior biological knowledge. A larger weight w_i implies a stronger prior belief of the similarity between gene i and its neighbors locally. The neighbor weight $\widetilde{w}_i = \sum_{j=1}^{n} c_{ij}w_j / \sum_{i=1}^{n} c_{ij}$ represents an average of weights from neighbors for gene i. For each regulation type k, we assume the base measure $G_{0k} = P(\mu, \Sigma)$ is formed as a conjugate Normal-inverse-Wishart distribution with parameters $(\mu_{0k}, c_{0k}, S_{0k}, \psi_{0k})$ and the scale parameter c_{0k} in the normal part of the base measure follows a gamma distribution with parameters (a_{0k}, b_{0k}), thus we denote the the distribution G_{0k} as $NIWG(\mu_{0k}, S_{0k}, \psi_{0k}, a_{0k}, b_{0k})$.

2.2.3 Missing Data Imputation

Suppose the test statistics $\mathbf{r}$ is only partially observed and can be further partitioned as $\mathbf{r} = (\mathbf{r}_{mis}, \mathbf{r}_{obs})$ where the $\mathbf{r}_{obs}$ represents the observed part and the $\mathbf{r}_{mis}$ denotes the missing part. Similarly, we can also partition the cluster indexes into the observed part and the missing part as $\mathbf{g} = (\mathbf{g}_{mis}, \mathbf{g}_{obs})$. The element-wise representation of the missing part of the test statistics is rewritten as $\mathbf{r}_{mis} = (r_{mis,1}, \cdots, r_{mis,m})$ where m is the number of missing nodes in the network. We have $\mathbf{g}_{mis} = (g_{mis,1}, \cdots, g_{mis,m})$.

Under the fully Bayesian inference framework, we handle the missingness by making posterior inference on the joint distribution of $\mathbf{r}_{mis}$ and all other latent quantities in the model (2.1), where the test statistics are conditionally independent given the cluster indexes and the density specifications. Thus, the conditional distribution for $\mathbf{r}_{mis}$ given $\mathbf{r}_{obs}, \mathbf{g}, \mathbf{z}, \widetilde{\boldsymbol{\theta}}$ only depends on $\mathbf{g}_{mis}, \widetilde{\boldsymbol{\theta}}$, that is

$$P(\mathbf{r}_{mis}|\mathbf{r}_{obs}, \mathbf{g}_{obs}, \mathbf{g}_{mis}, \mathbf{z}_{obs}, \mathbf{z}_{mis}, \widetilde{\boldsymbol{\theta}}) = P(\mathbf{r}_{mis}|\mathbf{g}_{mis}, \widetilde{\boldsymbol{\theta}}) = \Pi_{i=1}^{m} P(r_{mis,i}|g_{mis,i}, \widetilde{\boldsymbol{\theta}}).$$

This further implies that in the posterior computation algorithm for complete data analysis (See Section 2.3), we only need to introduce one more step to impute the missing test statistics $r_{mis,i}, i = 1, \cdots, m$. Assume the superscript represents the results from the previous iteration t, then, for the $(t+1)$th iteration, we draw a imputed value for $r_{mis,i}^{(t+1)}$ from $N(\widetilde{\mu}_{g_{mis,i}}^{(t)}, \widetilde{\sigma}^{2(t)}_{g_{mis,i}})$.

We also propose a fast imputation approach by approximating the fully Bayesian inference in light of the assumption that neighboring genes are more likely to function together and share the same functionality due to the network dependencies. Suppose we integrate out all the latent quantities in the model and impute $r_{mis,i}$ directly using $\mathbf{r}_{obs}$ based on the conditional expectation which is given by

$$E(r_{mis,i} \mid \mathbf{r}_{obs}) = \int r_{mis,i} P(r_{mis,i} \mid \mathbf{r}_{obs}) dr_{mis,i}, \text{ with}$$

$$P(r_{mis,i} \mid \mathbf{r}_{obs}) = \int \int P(r_{mis,i} \mid z_{mis,i}, \widetilde{\boldsymbol{\theta}}) P(z_{mis,i}, \widetilde{\boldsymbol{\theta}} \mid \mathbf{r}_{obs}) dz_{mis,i} d\widetilde{\boldsymbol{\theta}}.$$

Thus, if we have N samples of $(z_{mis,i}, \widetilde{\boldsymbol{\theta}})$ from the posterior distribution given $\mathbf{r}_{obs}$,

denoted as $(z_{mis,i}^{(1)}, \widetilde{\boldsymbol{\theta}}^{(1)}), \ldots, (z_{mis,i}^{(N)}, \widetilde{\boldsymbol{\theta}}^{(N)})$, then $P(r_{mis,i} \mid \mathbf{r}_{obs})$ can be approximated by $\frac{1}{N}\sum_{n=1}^{N} P(r_{mis,i} \mid z_{mis,i}^{(n)}, \widetilde{\boldsymbol{\theta}}^{(n)})$. As indicated by model (2.2), when $\boldsymbol{\rho} > 0$, $z_{mis,i}$ has a larger probability to take the value of k when $z_j = k$ for j in the neighborhood of i. From our experience, we can well approximate $P(r_{mis,i} \mid \mathbf{r}_{obs})$ by a discrete distribution $P(r_{mis,i} = r_j \mid \mathbf{r}_{obs}) = 1/|nbr(i)|$ for $j \in nbr(i)$, where $nbr(i)$ represents the neighborhood of i with r_j observed. Then $E(r_{mis,i} \mid \mathbf{r}_{obs})$ can be approximated by $\sum_{j \in nbr(i)} r_j/|nbr(i)|$. We refer to this approach as the nearest-neighbor imputation method.

2.3 Posterior Computation

The posterior computation algorithm has three major steps in each iteration: 1) Impute missing test statistics $\mathbf{r}_{mis}$ (if any) either by conditional sampling (fully Bayesian inference) or by the nearest-neighbor imputation method; 2) Update class indicators $\mathbf{z}$ by the Swendsen-Wang algorithm, and 3) Update $\widetilde{\boldsymbol{\theta}}$ by refitting a DPM to estimate densities for each regulation type. Others including L_k, g_k are omitted temporarily for simplicity. For updating the hyperparameters in the Potts model for $\mathbf{z}$, we adopt the method of Double Metropolis-Hastings (DMH) sampler proposed by Liang (2010a).

Swendsen-Wang algorithm: it has been widely used in the Potts model. It works by introducing another set of auxiliary variables denoted as $\mathbf{W} = \{W_{ij}, i \sim j\}$. W_{ij} is defined only when gene pairs i and j are connected. Given z_i, z_j, W_{ij} is uniformly distributed between 0 and $\exp(\rho_{z_i}\omega_j c_{ij} I[z_i = z_j])$. Then the full conditional distribution for $\mathbf{z}$ given $\mathbf{W}$ can be simplified as proportional to $P(\mathbf{r}|\mathbf{z}, \widetilde{\boldsymbol{\theta}}) \exp\left[\sum_{i=1}^{n} \widetilde{\omega}_i \log(\pi_{z_i})\right]$. See Appendix Equation (A.1) for more details.

The posterior sampling scheme has two steps: the network partitioning step (sample $\mathbf{W}$ given $\mathbf{z}$) and the network relabeling step (sample $\mathbf{z}$ given $\mathbf{W}$). The objective for network partitioning is to cut the network into smaller connected subnetworks

so that the genes located within the same subnetwork share the same class indicators. Then in the network relabeling step, the class indicators of all the genes located within the same subnetwork can be flipped simultaneously. Comparing to the Gibbs sampler when it updates the genes each one at a time, the Swendsen-Wang algorithm advantages itself by a more efficient group level updating scheme and a better convergence.

DPM Density Updating Conditional on the class indicators, we update g_i and $\widetilde{\boldsymbol{\theta}}_i$ given $g_1, \ldots, g_{i-1}, \widetilde{\boldsymbol{\theta}}_1, \ldots, \widetilde{\boldsymbol{\theta}}_{i-1}$. Utilizing Algorithm 8 in Neal (2000), we firstly summarize the frequency for each of the total l unique g values ever appeared in set $(g_1, \ldots, g_{i-1})$, denoted as $(1, 2, \ldots, l)$ with cluster parameters $(\widetilde{\boldsymbol{\theta}}_1, \ldots, \widetilde{\boldsymbol{\theta}}_l)$. It is $n_{i,g} = \sum_{j=1}^{i-1} I[g_j = g], g = 1, 2, \cdots, l$. Then the prior probability of g_i equals to any of the ever-appeared cluster index g is given by $n_{i,g}/(i - 1 + \tau_k), g \in (1, 2, \ldots, l)$ (See Appendix Equation (A.2)), if the sampled g_i equals to any appeared cluster index g, then we set $\widetilde{\boldsymbol{\theta}}_i = \widetilde{\boldsymbol{\theta}}_g$; on the other hand, the prior probability of g_i being a new index is given by $\tau_k/(i - 1 + \tau_k), g \notin (1, 2, \cdots, l)$ (See Appendix Equation (A.2)), if the sampled g_i is a new index, then we sample a new set of parameter $\widetilde{\boldsymbol{\theta}}_g$ from base measure G_{0k}. Given the cluster index g, r_i follows a normal distribution with parameter $\widetilde{\boldsymbol{\theta}}_g$.

Choice of Initial Values In order to speed up the convergence in Markov Chain Monte Carlo, we specify the initial values for $G_{0k}, (k = -1, 0, 1), \mathbf{z}, \mathbf{g}, \widetilde{\boldsymbol{\theta}}$ and $\mathbf{L}$ based on the DPM density fitting of the test statistics $\boldsymbol{r}$ without the network information, we develop the Kullback-Leibler-divergence-based hierarchical ordered density clustering algorithm (KL-HODC). In the beginning, we order all the small cluster density parameters $\widetilde{\boldsymbol{\theta}}_g, \widetilde{\boldsymbol{\theta}}_g = (\widetilde{\mu}_g, \widetilde{\sigma}_g^2)$ based on their mean value $\widetilde{\mu}_g$ locations. Each time, we pick several clusters to form a proposed null. We calculate the KL distance between this proposed null and a prior null which is pre-determined by biological knowledge. The combination of the clusters with the smallest KL distance is selected and added

as the initial value for the null densities. Once all the clusters are assigned to three classes, $\mathbf{z}, \mathbf{g}, \widetilde{\boldsymbol{\theta}}, \mathbf{L}$ can be determined as well. When the biology knowledge is not available for the prior null, it can be estimated by a truncated bi-Gaussian distribution using the central part of the test statistics such as statistics within 15% and 75% quantiles.

KL-HODC is a hierarchical density clustering algorithm that extends the HODC proposed by Zhao et al. (2014). It incorporates the prior biological knowledge used as a prior null density and it handles the multi-class feature classification problem, while HODC can only be used for selecting features not further differentiating their subtypes.

2.4 Simulation Studies

We conduct extensive simulation studies to evaluate the performance of the proposed methods for the complete data case and the missing data case.

Settings The network used in the simulation studies is a subnetwork of the real biological network used in real data analysis downloaded from the High-quality INTeractomes (HINT) database (Das and Yu, 2012). It is formed by a total of 776 nodes with a median degree of 3, a mean degree of 5.2 and a maximum degree of 30. The underlying true gene regulation types are assigned based on the merged communities by the fast greedy modularity optimization algorithm (Clauset et al., 2004). We assign the genes located in the largest community as the null class and then we randomly assign the down-regulated or the up-regulated class to the other two. For the null genes, their test statistics are independently drawn from a normal distribution, and for the up-regulated or the down-regulated genes, their test statistics are independently drawn from one of the following three distributions: a normal, a gamma or a lognormal. (See Figure A.1 for an illustration of one simulated dataset; See Table 2.1 for the designs of the simulation settings). The missing locations are randomly

selected among the genes with network degrees less than 6, which is the 66% quantile of the degrees of the nodes in the network. We simulate 20% missingness since it is the missing rate in the real dataset.

Table 2.1: Simulation settings.

	down-regulated class	null class	up-regulated class
Gaussian	N(-0.6, 0.2)	N(0, 0.2)	N(0.6, 0.2)
Gamma	Gamma(shape=2, scale=0.5) truncated within $(-\infty, 2]$, shifted -1.9	N(0, 0.4)	Gamma(shape=2, scale=0.3) truncated within $(-\infty, 1.8]$, shifted +1.7
log-normal	log N(0, 1) truncated within $(-\infty, 2]$, shifted -1.9	N(0, 0.4)	log N(0, 1)+2.2 truncated within $(-\infty, 2.3]$, shifted +2.2

Evaluation Criteria For each simulation setting, we simulate 50 datasets. We define a "rate" of the true class $z_i = a$ is classified as b for $a, b = -1, 0, 1$ as the $\sum_{s=1}^{50} I[\widehat{z}_i^{(s)} = b, z_i = a]/50$ where $\widehat{z}_i^{(s)}$ is the estimate of z_i in the simulated dataset s. Denote TP-down, TP-up and TN the true positive rate averaged across all simulations for the down-regulated ($a = b = -1$), up-regulated ($a = b = 1$) and null genes ($a = b = 0$) respectively. Denote FN-down and FN-up the averaged false negative rates for the down-regulated and up-regulated genes. Additionally, FP-down and FP-up are the averaged false positive rates. And finally, FDR is the false discovery rate defined as the proportion of false discoveries among all the discoveries on average.

Hyperprior Specifications As for the Potts prior model (2.2), set weights as $\omega_j = 1, j = 1, 2, \cdots, n$ so that $\widetilde{w}_i = 1, i = 1, 2, \cdots, n$. Set $\boldsymbol{\rho} = (1.001, 0.497, 0.998)$, $\boldsymbol{\pi} = (0.15, 0.70, 0.15)$ as an output from DMH of a 10000-iteration run with 5000 burn-ins. The proposal used in DMH for $(\boldsymbol{\pi}, \boldsymbol{\rho})$ is an independent random walk proposal for $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$: for each element of $\boldsymbol{\rho}$, it is a truncated Gaussian distribution with a mean of 0, a standard deviation of 0.03, a lower-bound of 0 and a upper-bound of 1.5; for $\boldsymbol{\pi}$, since it must satisfy $\pi_2 = 1 - \pi_1 - \pi_3$ and $\pi_1 + \pi_3 < 0.5$, thus we assume π_1 and π_3 follow the truncated Gaussian distribution with a mean of 0, a standard deviation of 0.03, a lower-bound of 0 and a upper-bound of 0.5. As for the

hyperparameters for the DPM model fitting, the base measure G_{0k}s are NIWG with parameters $\mu_{0k}, S_{0k}, \phi_{0k}, a_{0k}, b_{0k}$. For this prior model, we firstly apply the normal mixture modeling for model-based clustering method (Mclust by Fraley and Raftery (1999)) where the parameter indicating the total number of groups is set to be 3. Then we use the estimated mean and variance from each group k as μ_{0k} and S_{0k}. And we set $\phi_{0k} = 3, a_{0k} = 1, b_{0k} = 100, \tau_k = 3$.

2.4.1 Complete Data Cases

We first consider the cases when all the test statistics are fully observed. For each of these simulation settings, we compare our method (BANFF^2) with the Bayesian nonparametric mixture model for selecting genes (BANFF) by Zhao et al. (2014) and the false discovery rate controlling procedures for identifying differentially expressed genes (locfdr) by Efron and Tibshirani (2002). The locfdr method does not consider the network structure and only use the gene expression data matrix.

BANFF is a Bayesian nonparametric gene and gene-network selection method, it can also utilize the network information but it is mainly for selecting the activated-state features from the null-state features. In order to modify the BANFF for this feature classification problem, we firstly classify genes into three by Mclust. Then we flip the sign of the test statistics of the genes assigned to the down-regulated class so that ideally those genes combined with the up-regulated class should be of the active state. Then the finalized class indicators are assigned based on the results from BANFF being flipped back. For the locfdr, it is a kernel density-based non-parametric method for selecting differentially expressed genes without considering the network. To be specific, we applied the central matching for estimating the null densities and then calculated the estimated local false discovery rate for each gene. We adopt a commonly used cutoff of 0.2 so that the genes with the posterior probability of being in the null class below 0.2 will be identified as differentially expressed, and the null class otherwise. Then the differentially expressed genes can be further classified by

comparing the relative locations of their test statistics with 0.

Table 2.2 indicates that for Gaussian simulations, under each regulation type, BANFF^2 performs better than the BANFF and locfdr. Our method achieves classification accuracies as high as 0.87 for the down-regulated genes, 0.91 for the up-regulated genes, and 0.97 for the null genes. At the same time, BANFF^2 achieves the false positive rates as lower as 0 for the down-regulated genes, 0.03 for the up-regulated genes, 0.12 for the null genes to be classified as down-regulated genes, 0.09 for the null genes to be classified as up-regulated genes. Overall, our method can achieve higher accuracies and lower false positive and false negative rates. BANFF performs worse in the true negative rates and false positive rates. locfdr performs well at selecting the null genes, with a true negative rate of 1. However, it gives a false negative rate as high as 0.49 for the down-regulated genes and 0.5 for the up-regulated genes, indicating the procedure is overly conservative.

Comparing the classification accuracies for Gamma and log-normal settings, our approach outperforms all the others in all the measures. The BANFF performs worse than the BANFF^2. It is because the proposed method can flexibly model the gene subtypes so that it can allow for different levels of deviation from the null for down-regulated and up-regulated genes. The worse performance of locfdr compared to BANFF^2 indicates that by utilizing network information, better classification accuracies can be obtained.

Table 2.2: Algorithm performance for complete data cases.

Generative model	Methods	TP-down	TP-up	TN	FP-down	FP-up	FN-down	FN-up	FDR
Gaussian	BANFF^2	0.87	0.91	0.97	0	0.03	0.12	0.09	0.03
	BANFF	0.75	0.87	0.62	0.03	0.36	0.2	0.13	0.3
	locfdr	0.5	0.51	1	0	0	0.49	0.5	0.01
Gamma	BANFF^2	0.92	0.96	0.99	0	0.01	0.08	0.04	0.01
	BANFF	0.5	0.89	0.69	0	0.31	0.38	0.11	0.2
	locfdr	0.57	0.71	0.98	0.01	0.01	0.43	0.29	0.03
log-normal	BANFF^2	0.9	0.96	0.99	0	0.01	0.1	0.04	0.01
	BANFF	0.73	0.92	0.55	0.05	0.04	0.17	0.08	0.31
	locfdr	0.59	0.72	0.99	0.01	0.01	0.41	0.28	0.03

2.4.2 Missing Data Cases

We further compare our proposed method with the others when there are missing node observations in the network. We only focus on the symmetric cases as described in Table 2.1, and compare five methods to perform feature classification and to handle missingness simultaneously: 1) BANFF^2+Bayes: we apply the BANFF^2 for feature classification and the conditional sampling for fully Bayesian inference to impute the missing test statistics. 2) BANFF^2+NN: we apply the BANFF^2 for feature classification combined with the nearest neighbor imputation method to impute the missing test statistics. 3) BANFF^2+NArm: we firstly remove all the missing nodes and their edges in the network and then use BANFF^2 for feature classification. In this case, only the estimated class indicators for gene nodes with observed test statistics can be obtained. 4) BANFF+NN: we utilize the BANFF for feature classification and use the nearest neighbor imputation method to impute the missing test statistics. 5) BANFF+NArm: we apply BANFF to the reduced network comprised of nodes with observed test statistics. Similar to BANFF^2+NArm, only replace the BANFF^2 with BANFF for feature classification.

To summarize the classification accuracies, we separate different types of nodes to calculate the averaged rates: 1. Missing: only average the rates among the genes

whose test statistics are missing. 2. Observed: only average the rates among all observed genes. 3. Total: average among all the genes nodes.

From Table 2.3, we observe that BANFF^2+NN performs the best in general. The overall classification accuracies for the all the down-regulated, the up-regulated and the null genes to be correctly classified are 0.87, 0.87, 0.89. The averaged false positive rates for the null genes being classified as down-regulated or up-regulated are 0 and 0.01. The averaged false negative rates for the down-regulated or the up-regulated genes are 0.12 and 0.13, respectively. The estimated false discovery rate is 0.12. This performance keeps consistent among missing genes and the observed genes. Compared to BANFF^2+Bayes, BANFF^2+NN is slightly better. It is because the nearest-neighbor imputation scheme is more flexible than the model-based Bayesian posterior inferences since Bayesian posterior sampling needs to specify a proper prior. The Bayesian model we are utilizing might not characterize very well the predictive distribution of the missing test statistics given the observed test statistics across the network while utilizing the information from the nearest neighbors might help to improve.

The accuracies will drop if we use the BANFF for feature classification regardless of which schemes are used for handling the missingness. It indicates that our proposed algorithm outperforms BANFF when there are missing observations in the network, which is consistent with the simulation results in fully observed cases. Moreover, regardless of which feature classification algorithms we utilize, either BANFF^2 or BANFF, compare the imputations methods Bayes or NN with NArm among the observed gene nodes, we observe that the classification accuracies drop and the false positive/ false negative rates increase, so as the averaged false discovered rates. Thus, imputation methods are recommended for feature classification problem with missing gene observations.

Table 2.3: Algorithm performance for missing data cases.

Algorithm	Gene nodes type	TP-down	TP-up	TN	FP-down	FP-up	FN-down	FN-up	FDR
BANFF^2+Bayes	Missing	0.73	0.75	0.77	0.01	0.22	0.26	0.25	0.27
	Observed	0.92	0.9	0.78	0	0.22	0.05	0.1	0.2
	Total	0.88	0.88	0.78	0.01	0.22	0.1	0.12	0.21
BANFF^2+NN	Missing	0.83	0.81	0.88	0.01	0.11	0.17	0.19	0.15
	Observed	0.88	0.88	0.89	0	0.01	0.11	0.12	0.11
	Total	0.87	0.87	0.89	0	0.01	0.12	0.13	0.12
BANFF^2+NArm	Observed	0.87	0.88	0.66	0.01	0.33	0.07	0.12	0.3
BANFF+NN	Missing	0.6	0.79	0.48	0.04	0.48	0.26	0.21	0.47
	Observed	0.7	0.86	0.41	0.05	0.54	0.19	0.14	0.46
	Total	0.68	0.85	0.42	0.05	0.53	0.21	0.15	0.46
BANFF+NArm	Observed	0.67	0.82	0.5	0.05	0.45	0.23	0.18	0.43

2.5 Survival Analysis of Cutaneous Melanoma

We analyze the cutaneous melanoma dataset from The Cancer Genome Atlas (TCGA) Network (2015), downloaded from the cBio Cancer Genomics Portal (Cerami et al., 2012). There are 478 patient records by the time we downloaded. After removing six patient records that lack gene expression profiles, one patient record that is recorded a negative survival month due to possible errors, one patient record that is missing survival status, and one patient record that is missing the sample type which is one of the covariates we are interested in, we use the remaining 469 patient records in a Cox proportional hazard model to assess the association between the expression levels of individual genes and survival time. In our model, we control for three covariates: age at initial pathologic diagnosis (minimum 15, median 58, mean 58.08, max 90, and 8 are missing), gender (180 females and 290 males), and sample type (366 of metastatic, 102 of primary tumor and 1 of additional metastatic).

We downloaded the protein-protein interactions in Homo Sapiens from the High-quality INTeractomes (HINT) database by Das and Yu (2012). After data cleaning, there are a total of 11,662 genes and 87,482 edges. Then we apply the community

detection algorithm by Clauset et al. (2004) to extract the largest connected subnetwork as our network input. To be specific, the largest connected component contains 10,484 genes while the remaining genes form 1097 tiny islands (1 island is of five genes, 2 islands are of four genes, 5 are of three genes, 61 of two genes and 1028 are formed by a single gene node). By excluding these tiny islands, the network contains a total of 10484 nodes, with a degree distribution of a minimum of 1, a median of 3, a mean of 8.328, and a maximum of 400.

For the gene expression profile, we firstly map all 20530 unique gene names to 18978 Entrez IDs. Among the 10484 genes in the network, 9833 can be mapped to an expression profile. There are 651 (6.21%) genes that do not have any expression profile and another 1433 (13.67%) genes that are considered unreliably measured based on their low maximum expression level across the samples. Removing such genes leads to a total missing rate of 19.88% in our real data analysis.

Similar to the simulations, for the Potts prior model, the hyperparameters in Equation (2.2), we prefix the $\omega_j = 1, j = 1, 2, \cdots, n$ so that the $\widetilde{w}_i = 1, i = 1, 2, \cdots, n$. Set $\boldsymbol{\rho} = (1.003, 0.479, 0.988)$ and $\boldsymbol{\pi} = (0.15, 0.70, 0.15)$ as an output from the DMH of 10000 iterations with 5000 burn-in. Other hyperparameters settings are the same with the settings used for simulations. In the following discussion, we refer to genes that significantly increase the risk of death as high-risk genes and genes that significantly decrease the risk of death as low-risk genes.

Our method finds 144 high-risk genes and 263 low-risk genes. Compared to ours, the locfdr method finds 217 low-risk genes by central matching estimation for a symmtric null while it does not identify any differentially expressed genes by applying a split normal version of central matching estimation for an asymmetric null. Thus, for the following discussion, we will focus on the comparison between the proposed method and the locfdr utilizing central matching estimation (See Figure 2.2a) even though the null density is asymmetric and the mode of the distribution is away

from zero for the motivating dataset.

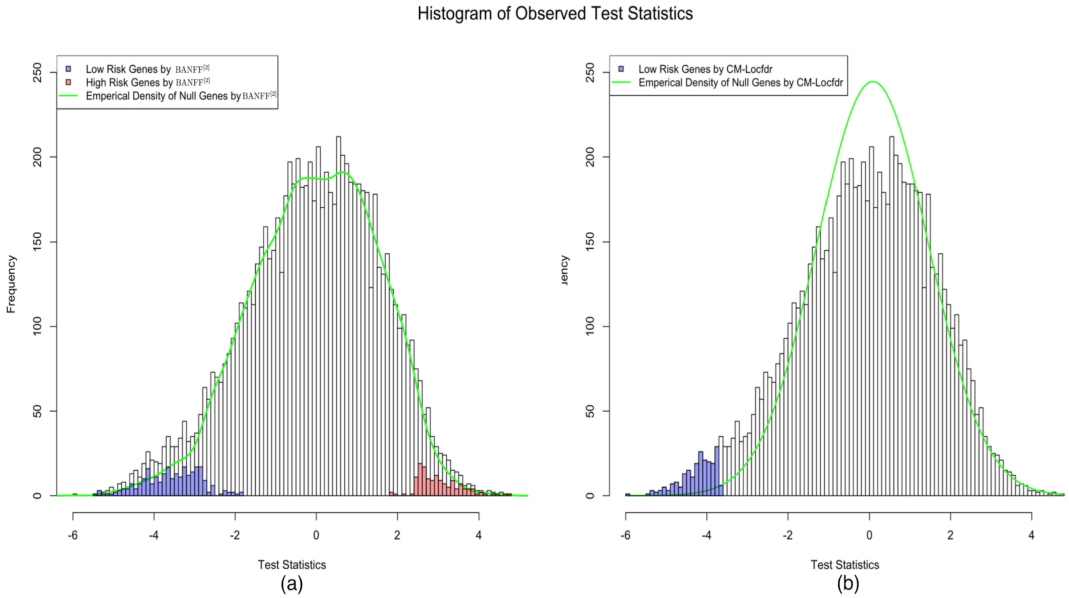


Figure 2.2: Histogram of the test statistics, with estimated null density and frequencies of the selected genes. (a) Results by BANFF^2; (b) Results by locfdr with center matching estimation for a symmetric null. Local false discovery rate is controlled at 0.2 for both methods. Blue: low-risk genes; red: high-risk genes.

Using the test statistics alone, combined with the common assumption of symmetric null distribution, locfdr identifies significant genes only on the low-risk side (See Figure 2.2b). On the other hand, when the existing network is utilized, the proposed method can detect both high-risk and low-risk genes.

To facilitate interpretation, we further find modules by applying the fast greedy community detection algorithm among the selected nodes and their one-step neighbors (Clauset et al., 2004). There is a total of 56 modules selected, 16 of which contain more than 10 selected genes.

Here we present some example modules and discuss their biological functions in relation to the clinical outcome. The module shown in Figure 2.3a contains 48 selected genes. There are 39 low-risk genes and 9 high-risk genes in this module. Analyzing the biological functions of the selected genes using GOstats (Falcon and Gentleman, 2007), we find the biological function of the low-risk genes are focused in the area

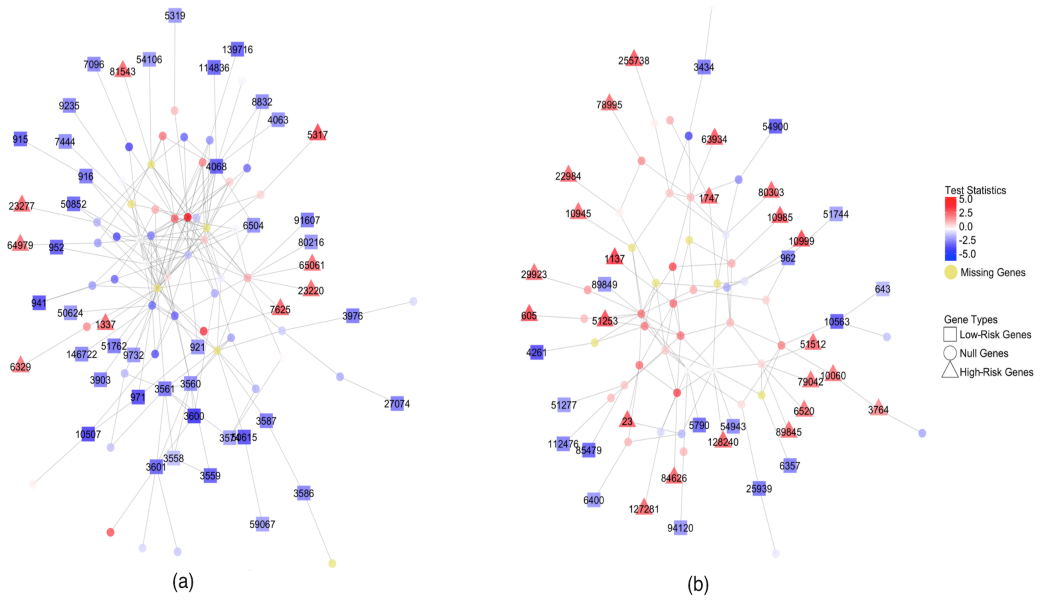


Figure 2.3: Two example modules for discussion about biological functions in relation to the clinical outcome.

of immune responses, with 18 of the 39 genes falling into the biological process of "regulation of immune response", and various related functions. The prognosis of melanoma is closely related to tumor-infiltrating lymphocytes (Taylor et al., 2007). A cross-platform meta-analysis has shown that the increased expression of immune function-related genes in melanoma is associated with longer patient survival, and B and T cells are enriched in melanoma biopsies from patients with favorable outcome (Lardone et al., 2016).

The module shown in Figure 2.3b contains 23 high-risk genes and 17 low-risk genes. An interesting finding is that the top gene ontology biological process being over-represented by the high-risk genes is transmembrane transport, with eight of the 23 genes falling into this category. Six of the high-risk genes are involved in ion transport. Although transmembrane transporters haven't been systematically studied in melanoma progression, recent developments in other cancer have indicated their role in cancer prognosis (Elsnerova et al., 2016). For example, gene 3764 (KCNJ8) encodes a potassium channel. It is found to be over-expressed in nasopharyngeal carcinoma

(NPC) tissues as well as in esophageal cancer (Zhou et al., 2007; Warnecke-Eberz et al., 2016). The gene 6520 (SLC3A2) encodes the heavy chain of the transmembrane protein CD98 that regulates intracellular calcium levels and transports L-type amino acids. It has been linked to Ras-driven skin carcinogenesis and prognosis of lung cancer (Guo et al., 2015; Estrach et al., 2014). Gene 11660 (ABCC9) is a member of the ATP-binding cassette transporter (ABC transporter) family. Recently the down-regulation of ABC transporters, including ABCC9, has been observed in prostate cancer (Demidenko et al., 2015). Gene 255738 (PCSK9) is involved in peptide precursors trafficking. It has been shown that tumor development influences the host lipid metabolism through PCSK9-mediated degradation of hepatic LDLR, and PCSK9 is suppressed in hepatocellular carcinoma (Bhat et al., 2015; Huang et al., 2016). Combined with these evidence in other types of cancer, our results indicate a link between transmembrane transporters and the prognosis of melanoma.

Six of the 17 low-risk genes belong to cytokine-mediated signaling pathways, which are critical in leukocyte trafficking and immune functions (Zlotnik and Yoshie, 2012). Gene 643 (CXCR5), a member of the CXC chemokine receptor family, is expressed in mature B-cells and Burkitt's lymphoma. The loss of CXCR5 in naive T cells is linked to the metastatic dissemination of melanoma into lungs (Jacquelot et al., 2016). Gene 3434 (IFIT1) is an interferon-induced protein. Overexpression of IFIT1 has been shown to predict improved outcome in newly diagnosed glioblastoma (Zhang et al., 2016). Gene 4261 (CIITA) regulates class II major histocompatibility complex gene transcription. CIITA overexpression facilitates engulfment of the T-cell material by melanoma cells, which can blunt the anti-tumor response (Lloyd et al., 2015). Gene 10563 (CXCL13) is a cytokine that belongs to the CXC chemokine family.Its expression is correlated with the densities of tumor high endothelial venules (HEVs), which allows the recruitment of tumor-infiltrating lymphocytes (TILs) (Martinet et al., 2012). CXCL13 is also found to be one of a group of diagnostic markers

of melanoma (Liu et al., 2013). Gene 25939 (SAMHD1) is a deoxyribonucleoside triphosphate triphosphohydrolase that decreases dNTP pools, which in turn affects DNA replication fidelity. Although it hasn't been well studied in melanoma, SAMHD1 is found to be frequently mutated in colon cancers, resulting in decreased SAMHD1 activity and thereby facilitating cancer cell proliferation (Rentoft et al., 2016).

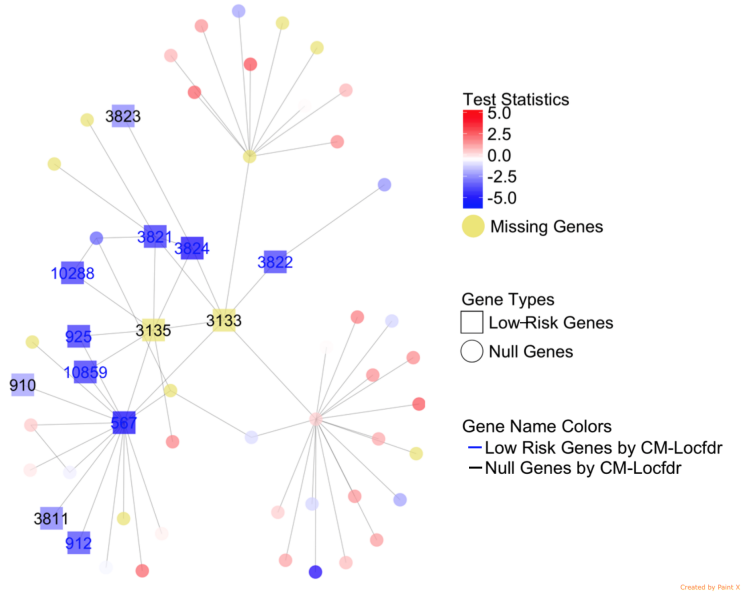


Figure 2.4: A module containing two nodes with missing observations being identified as low-risk genes by BANFF^2.

Figure 2.4 shows a module where two nodes with missing observations are identified as low-risk genes. These two genes, 3135 (HLA-G) and 3133 (HLA-E) have both been implicated in melanoma immunomodulation. HLA-G can inhibit the function of T cells, natural killer cells, and dendritic cells. It has been documented that HLA-G is inconsistently expressed in melanoma, and its expression can provide the malignant cells a mechanism of escaping immune surveillance (Paul et al., 1998; Yan et al., 2005). Similarly, HLA-E expression on the cell surface facilitates the melanoma cells' escape from CTL and NK cell surveillance (Derré et al., 2006). Among all the 13 genes in this module, 10 are annotated to the biological process of regulation of immune response, which is consistent with our earlier discussion about the associ-

ation of immune function-related genes with patient survival (Lardone et al., 2016; Taylor et al., 2007). The figure also shows that by test statistic alone, three of the 13 genes are not selected by locfdr. They are selected by BANFF^2 because their connections in the network offer extra evidence that they are related to the clinical outcome. These three genes are 910 (CD1B), 3811 (KIR3DL1), 3823 (KLRC3). It has been found that down-regulating CD1 molecules including CD1B on infiltrating dendritic cells by secreting IL-10 are associated with metastasis of melanoma (Gerlini et al., 2004). Both KIR3DL1 and KLRC3 are receptors expressed on natural killer (NK) cells, the induction of which shows the potential of suppressing solid melanoma tumors (Wennerberg et al., 2015).

Besides being biologically relevant, the selected modules each present a good predictive power on the clinical outcome. Here we compare concordance statistics which is commonly used in survival analysis to check on model validity. Concordance statistics (C-statistics) is defined as the probability of agreement between any two randomly chosen observations. If a model predicts a higher risk of death of one patient when it is observed with a shorter survival time compared to the other, then we define this pair as "agree", otherwise as "disagree". Since ties of the predicted and the observed survival time may occur, we refer to those pairs are "tie". Then, the C-statistics is defined as $P(agreement) = (agree + tied/2)/(agree + disagree + tied)$ for all possible comparable pairs (T., 2015). By saying "comparable", it is defined as the opposite to "uncomparable". The "uncomparable" pairs are the pairs when we lack the information of whether the predicted and the survival time agree or disagree with each other. For example, one patent record is censored at time 2 while the survival time we predict is 4. In general a C-statistic of 1 means perfect agreement; 0.6-0.7 is a common result for survival data while 0.5 is an agreement that is no better than the random guess.

We then calculate the C-statistics by the direct comparisons between the observed

survival time and the predicted survival time generated by the model fitting results of the Cox proportional hazard model for each selected module. Due to the lack of the ability to handle the nodes when their expression profiles are completely missing in the Cox proportional hazard model, thus, all the models are fitted using data except for those missing nodes. The modules with the number of genes larger than 20 are outputted in Table 2.4. From Table 2.4, we observe that our proposed method can successfully recall the high-risk genes when they cannot be discovered by locfdr method. The averaged C-statistics for these top 16 modules are 0.70 for our method while it is 0.66 for locfdr. This indicates a better predicting power using our method.

Table 2.4: Module group sizes and concordant scores.

Module ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Total number of genes nodes	116	101	86	61	57	40	40	37	28	27	26	26	24	21	21	20
high-risk genes by BANFF^2	24	9	23	7	9	10	7	1	3	2	4	8	6	4	2	0
low-risk genes by BANFF^2	29	39	17	24	15	12	11	16	11	10	8	8	8	5	8	11
low-risk genes by locfdr	13	16	7	8	11	5	6	8	7	7	4	3	1	2	0	7
low-risk genes by both methods	12	13	6	8	9	5	5	8	7	7	3	3	1	2	0	7
C-statistics by BANFF^2	0.7491	0.7363	0.7311	0.7113	0.7196	0.7146	0.7097	0.6846	0.6788	0.6902	0.688	0.6996	0.6928	0.7006	0.6695	0.6806
C-statistics by locfdr	0.6702	0.6504	0.6648	0.6899	0.6761	0.6635	0.6688	0.649	0.6497	0.661	0.6558	0.6679	0.6557	0.6612	NA	0.6667

2.6 Discussion

The feature classification problem utilizing the network information is a novel problem and it has been drawn an increasing attention recently. Based on our knowledge, we are the first to propose a non-parametric Bayesian framework not only to select features but also to differentiate the subtypes of the selected features over large-scale gene networks, and to handle the missing gene node observations simultaneously.

We have applied our method to the cutaneous melanoma dataset from the Cancer Genome Atlas and provided novel gene regulation evidence for unveiling the disease mechanism. In general, we recommend BANFF^2 for feature classification and if there are missing node observations in the network, we recommend nearest-neighbor imputation method to handle missingness.

It is noteworthy that in the application section, we do not consider genes that are

not part of the network because the main purpose of the subsequent analysis is to select subnetworks, which are defined as functionally coherent and easy to interpret since most of the tiny islands are formed by a single node.

Moreover, the KL-HODC algorithm we proposed for setting up the initial values for fast convergence can be further utilized in another fast version of our proposed algorithm based on density approximations, which can be implemented in our package. The fast algorithm works by fitting DPM densities for several iterations and then the densities are fixed, the algorithm continues to run but only update the class indicators given the densities until the Markov Chain reaches its equilibrium. For this fast version, it is of great importance to choose an initial value based on our experience. Thus, KL-HODC advantages itself by providing a better inference of the density specifications and class indicators since it can properly incorporate the prior biological knowledge.

Future work may be focused on the extension of our method to a multivariate statistics cases when combined information can provide more aspect of the information for classifying features, which can intuitively improve the classification accuracies.

Chapter 3

Integrate the LC-MS Metabolomics Data with Metabolic Network and Adduct Ion Relations for Missing Value Imputation

3.1 Introduction

Metabolomics aims to comprehensively identify and quantify all metabolites in a system and to study their changes in relation to diet, environment, disease status, genetic effects, pharmaceutical interventions, *etc* (Lindon et al., 2007). By profiling and analyzing metabolite abundance, it can be helpful for unveiling the etiology of diseases and providing a functional readout of the physiological state of the human body. Liquid chromatography-mass spectrometry (LC-MS) is a commonly used metabolomics platform due to its feasibility to measure complex samples, such as human plasma and urine (Jones et al., 2012).

The quality of the LC-MS data influences the downstream analysis, including metabolite quantitation, functional interpretation, pathway analysis for disease mechanisms. The datasets normally contain large portions of metabolites with missing observations in some samples. The underlying missingness mechanism is complex. As discussed by Gromski et al. (2014), the missingness can be the result of one or any combination of the following factors: 1. the failure in computational detection; 2. measurement error; 3. signals are of low intensity which cannot be distinguished from background noise; 4. imperfection of the detection algorithms used; 5. deconvolution that may result in false negatives. They also argued that imputation techniques should be favored over other methods of handling missingness in LC-MS metabolomics studies.

Various imputation techniques have been developed and applied in metabolomics studies (Armitage et al., 2015; Gromski et al., 2014; Hrydziuszko and Viant, 2012; Taylor et al., 2016), many of which were carried over from the field of microarray gene expression. They do not utilize two pieces of valuable information that are unique to metabolomics data. The first piece of information is the known metabolic network, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa and Goto, 2000). There are plenty of literature supporting the idea of

utilizing network information in data analysis procedures to improve variable selection and functional interpretation (Xia and Wishart, 2010; Kessler et al., 2013; Ravasz et al., 2002; Stelling et al., 2002; Aggio et al., 2010; Li et al., 2013; Barupal et al., 2012; Cai et al., 2017b). Given their close co-regulation, features matched to neighboring metabolites on the network could help predict each other's abundance in the sample. This may only be true for a subset of the metabolites, and the relation could be non-linear, creating a challenge in utilizing such information. However advanced machine learning techniques such as support vector regression (SVR) can utilize non-linear relations, as well as resist the impact of nuisance variables, i.e. those included in the model but have no true predictive power. With the help of such techniques, network information could contribute to missing value imputation.

The second piece of information that we try to utilize is the relationship between features that are likely derived from the same metabolite. Grouping and annotating features based on their mass-to-charge ratio (m/z) and retention time (RT) characteristics have been utilized in feature identification (Silva et al., 2014; Kuhl et al., 2012; Uppal et al., 2017). Potentially features derived from the same metabolite, even if the identity of the metabolite is unknown, can help the imputation of each other. For example, if the monoisotopic weight of a hypothetical molecule M is 100.000, then in data from positive ion mode with ESI ionization, the theoretical m/z values of two of its likely adduct ions are: $[M+H]^+, 101.007276$ and $[M+Na]^+, 122.989218$. Here "M" represents the metabolite, the element after the plus sign represents the adduct, and the "+" outside the bracket represents the charge state. The difference between the two m/z values does not change with the molecular weight of M. That is, even if a chemical is not in the database, its adduct ions still follow the same pattern in terms of the difference between their m/z values. For example, if we observe two m/z values in the data, and $|m/z_1 - m/z_2|$ is different from $22.989218 - 1.007276$ by no more than $m/z_2 \times 10^{-5}$, and the two features have close RT values, then we consider

they are highly likely to be derived from the same metabolite. We note that this relation is *likely* but not *definitive*. We will again rely on the SVR's capability to resist nuisance variables when a false relation is included in the imputation.

Combining the afore-mentioned information and traditional approaches, we propose a missing value imputation algorithm for LC-MS metabolomics data by applying the support vector regression (SVR) algorithm to a predictor network newly constructed among the features. To be specific, the predictor network is built by incorporating the metabolic network and adduct ion relations, together with linear and nonlinear associations between feature abundance levels calculated directly from the data (Figure 3.1 a). And then we impute each feature with missing values by fitting an SVR model on the dataset where the neighboring features on the predictor network are utilized (Figure 3.1b). An R package called MINMA (Missing data Imputation incorporating Network and adduct ion information in Metabolomics Analysis http://web1.sph.emory.edu/users/tyu8/MINMA) has been developed to implement the algorithm.

3.2 Methods

3.2.1 Building the predictor network

The predictor network was constructed on the feature level. The purpose of this network was to represent the feature relations. Essentially, every node on this network was a feature. If two features were considered as "potentially helpful in imputing each other's missing values", they were connected by an edge between them in the network. To define the "potentially helpful" features, we mainly considered the feature relations from three sources (Figure 3.1a):

- Metabolic Network

  The metabolic network we used in this paper was extracted from the KEGG

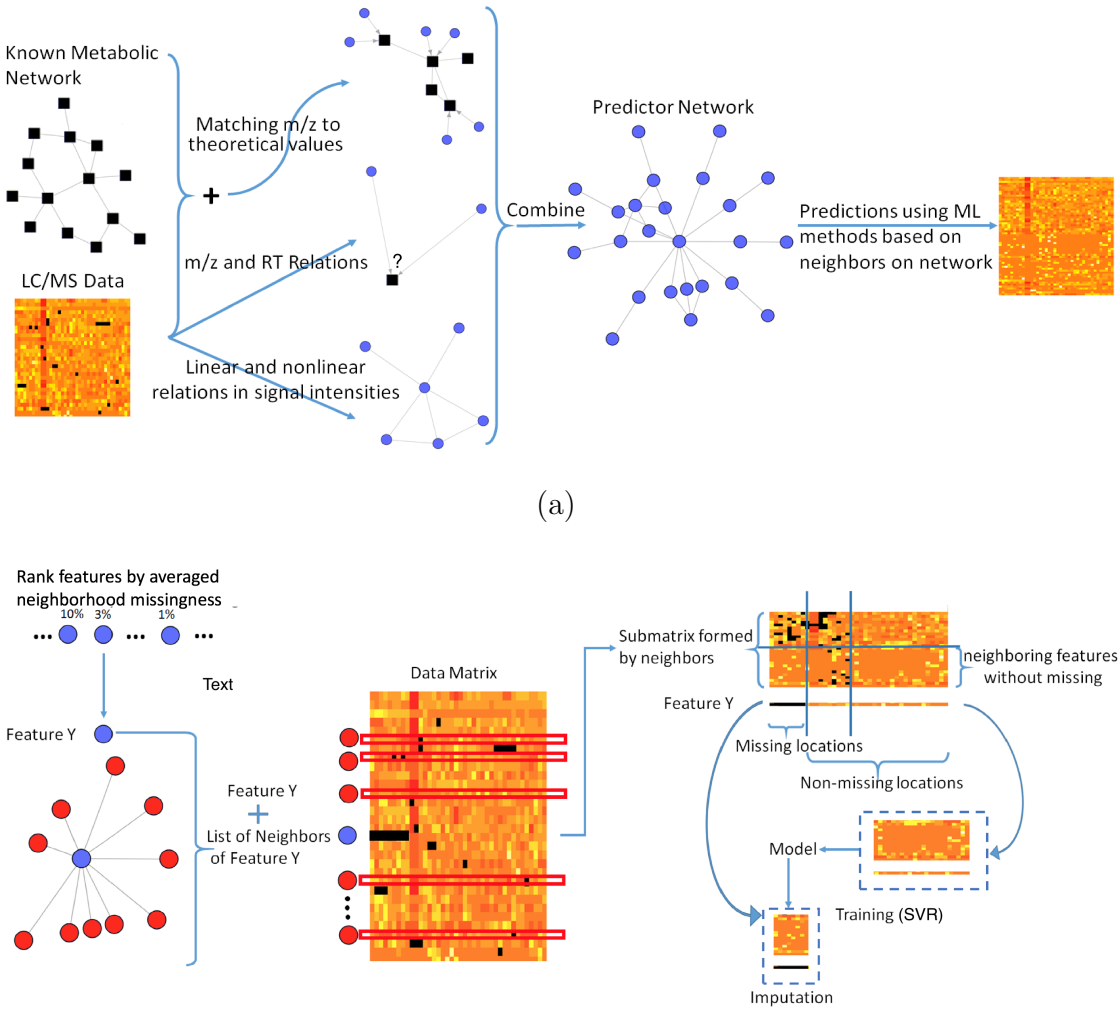


Figure 3.1: The workflow of the proposed method. (a) building the predictor network for imputation; (b) the imputation procedure given the predictor network.

database (Kanehisa and Goto, 2000). If two metabolites are involved in the same reaction, then they are linked in the metabolic network. Features matched to these two metabolites were considered connected. The matching of features to metabolites was based on matching the theoretical m/z of some common adduct ions of the metabolites to the observed m/z values of the features at a certain tolerance level (10*ppm* in this study). In this proof-of-concept study, as the data were generated from positive ion mode with electrospray ionization (ESI), we considered five adduct ions that are common in this type of data: $[M+H]^+$,

$[M + NH4]^+$, $[M + Na]^+$, $[M + K]^+$, $[M + 2Na - H]^+$. The specification of ion types can easily be done by user choice of the package.

- m/z value differences of common adduct ions

  First we determined what adduct ion forms were included. Then the m/z differences between the adduct ions of the same charge were calculated. Pairwise m/z differences were calculated for all features in the data. When the m/z difference between two features match closely with the theoretical difference between two adduct ions ($10ppm$ in this study), and their RT difference was less than a predefined threshold ($100seconds$ in this study), the two features were considered likely to be derived from the same metabolite. They were connected in the feature level network. The same set of five common adduct ions as mentioned above were used in this study.

- Correlation Inferred from Data Matrix

  We consider two features "neighbors" if they were highly correlated base on the following correlation measures:

  1. Linear correlation: consider "neighbors" based on n_1 largest pairwise Pearson correlations ($n_1 = 10$ in this study).
  2. DCOL correlation: consider "neighbors" based on n_2 largest pairwise nonlinear correlations defined by Distance based on Conditional Ordered List (DCOL) (Yu and Peng, 2013). ($n_2 = 10$ in this study)
  3. dCov dependency: consider "neighbors" based on n_3 largest pairwise general dependencies defined by Brownian distance covariance (Kosorok, 2009). ($n_3 = 10$ in this study)

  These three criteria might generate overlapping feature pairs.

By building the predictor network from multiple sources, it was guaranteed that each feature had at least k connections in the network.

3.2.2 The imputation procedure

The imputation was based on the predictor network. In the following discussions when network neighborhood is mentioned, we refer to the predictor network. In the imputation of every feature, only its connected features on the predictor network were used as predictors. We firstly introduce some mathematical notations here: $e_{(i,j)}$ represents the value at location (i, j) in data matrix $E = \{e_{(i,j)}, i = 1, \cdots, m, j = 1, \cdots, n\}$, i represents the ith feature (row) and j represents the jth sample (column). If the ith row, feature $e_i = \{e_{(i,j)}, j = 1, \cdots, n\}$ has missing locations, we denote: $e_{i,mis} = \{e_{(i,j)}, j = 1, \cdots, n, \text{where}, e_{(i,j)} = \text{NA}\}$, similarly, we denote $e_{i,obs} = \{e_{(i,j)}, j = 1, \cdots, n, \text{where}, e_{(i,j)} \neq \text{NA}\}$ as the observed locations in feature e_i, here the mis and obs only indicate locations instead of the values. All the neighboring features of feature i are indexed as $nbr(i)$.

For feature i, we selected the non-missing locations of feature i and used $e_{i,obs}$ as response vector and those neighboring features where they were fully observed in these observed locations denoted as $e_{nbr(i),obs}$ were formed as the predictor matrix. Then we trained the SVR model using $e_{i,obs} \sim e_{nbr(i),obs}$ and extracted the predicted value $\hat{e}_{i,mis}$ when $e_{nbr(i),mis}$ was used as the testing data for imputation.

Before imputation, the sequence for imputing the features with missing locations needs to be decided first or to be updated along the way. In this paper, we utilized a pre-fixed imputation sequence scheme for computation consideration. Specifically speaking, features were firstly ranked by a measure called averaged neighborhood missingness. The averaged neighborhood missingness of one feature was defined as the average number of missing locations of its neighboring features. Then the features with smaller averaged neighborhood missingness were imputed first. After imputing

the feature, the imputed values were filled in the original missing locations and were treated as non-missing locations in the following iterations (Figure 3.1b). However, the imputation sequence still stayed the same.

3.2.3 Performance Comparision

We compared the proposed imputation algorithm (denoted as Net_SVR) with other commonly used imputation algorithms in metabolomics studies, including the K-Nearest Neighbors (KNN) (Hastie et al., 1999; Troyanskaya et al., 2001), the Bayesian Principal Component Analysis (BPCA) (Oba et al., 2003), the imputation based on Simple Linear Regression (SLR), the imputation based on Singular Value Decomposition (SVD), the imputation by inserting Single Values (SVI: Min/2, Mean, Median). We briefly describe those methods:

- The K-Nearest Neighbors (KNN) (Hastie et al., 1999; Troyanskaya et al., 2001) finds the k nearest neighboring features $\{e_j, j = l_1, \cdots, l_k\}$ by a Euclidean metric calculated among those whose feature columns are not missing at location mis, and then takes the average values of non-missing locations $e_{j,mis}$ calculated as $\frac{1}{k}\sum_{j=l_1}^{l_k} e_{j,mis}$ for imputation.

- The Bayesian Principal Component Analysis (BPCA) (Oba et al., 2003) simultaneously estimates a probabilistic model for the data matrix and estimates some latent parameter sets within the framework of Bayesian inference, and then impute the missing values in the data matrix by the expectation with respect to the estimated posterior distribution.

- The imputation based on Simple Linear Regression (SLR) is conducted by first fitting a series of univariate simple linear regression models and collecting the predicted value from each SLR model, and then imputing $e_{i,mis}$ by a weighted

summation of all these predicted values, where the weights are decided by their pairwise Pearson correlation only using observed data.

- The imputation based on Singular Value Decomposition (SVD) (Troyanskaya et al., 2001) firstly initializes all missing values by their row means. Each time, given a complete observed matrix, it conducts a SVD procedure that obtains a set of mutually orthogonal expression patterns (eigen-features). And then it imputes the missing values by regressing the features with missing values against the nPC eigen-features ($nPCs$ need to be pre-specified). This imputation is repeated until the total change of two successive imputations is less than the tolerance value.

- The imputation by inserting a Single Value (SVI: Min/2, Mean, Median). These methods replace all missing values by a pre-calculated value. Common choices are: half of the minimum (Min/2), the mean (Mean) and the median (Median) calculated from all the observed values in the data matrix.

In order to evaluate the performance of each method, we calculated the normalized root mean squared error (NRMSE) of the imputed values. The NRMSE was calculated for all the simulated missing locations that were non-zero in the original data matrix. Suppose the total number of locations we use in calculation is K, the imputed values are $\hat{e} = \{\hat{e}_k, k = 1, \cdots, K\}$ and the ground-truth from the original observed data matrix are $e = \{e_k, k = 1, \cdots, K\}$. The NRMSE is defined as follows:

$$\text{NRMSE}(\hat{e}, e) = \sqrt{\frac{\sum_{k=1}^{K}(\hat{e}_k - e_k)^2/K}{Var(e)}}$$

The smaller NRMSE is, the lower the prediction errors and the better the imputation method. For better illustration, we further used a metric called "NRMSE Ratio" for algorithm comparison, such that the plot is on similar scale for all missing rates. For every missing imputation method (MI), it is defined as the ratio of its NRMSE

taken over the NRMSE of KNN. Due to the popularity of KNN in this field, we chose to use KNN in the denominator for calculation.

$$\text{NRMSE Ratio(MI)} = \text{NRMSE}(\hat{e}_{\text{MI}}, e)/\text{NRMSE}(\hat{e}_{\text{KNN}}, e)$$

Based on the definition, if we compare two methods, the smaller the NRMSE Ratio is, the better imputation performance. Pseudocodes for the proposed algorithm are listed the Appendix (see Algorithm 8, 9 and 10).

3.3 Results

3.3.1 Datasets and Simulation Setup

In this study, we used two metabolomics datasets denoted as CAD and CHD to assess the performance of different methods. The CAD dataset is from the Emory Cardiovascular Biobank, which consists of patients who have undergone coronary angiography to document the presence/absence of coronary artery disease. Demographic characteristics, medical histories, behavioral factors and fasting blood samples have been documented and details about risk factor definitions and coronary angiographic phenotyping have been described previously (Patel et al., 2012). Each sample was analyzed in triplicate with high-resolution liquid chromatography-mass spectrometry (LC-MS), using anion exchange column combined with the Thermo-Orbitrap-Velos (Thermo Fisher, San Diego, CA) mass spectrometer in positive ion mode, with a m/z range of 85 to 850.

The CHD dataset is a dataset from the Emory-Georgia Tech Predictive Health Initiative Cohort of the Center for Health Discovery and Well Being. This is a cohort of generally healthy university employees aged 18 and older (`http://predictivehealth.emory.edu`) (Brigham, 2010). The data was generated by C18 column combined with the Thermo-Orbitrap-Velos mass spectrometer in positive ion mode, with a m/z range of 85 to 850.

Both datasets were pre-processed using xMSAnalyzer (Uppal et al., 2013) in combination with apLCMS (Yu et al., 2009, 2013). Each sample was run in triplicates in the datasets. For each feature, there were three readings per subject. An average feature intensity value was calculated from the non-zero readings of the three. For filtering the data matrix, rows with more than 20% of zeros were removed. Finally, the data matrix was log-transformed by the function $y = log(1 + x)$. The CAD dataset contains 18434 features and 489 samples with 41.34% of the locations being zero. We removed rows with over 20% zeros, resulting in a data matrix of 7033 rows with an overall missing rate of 2%. The CHD dataset contains 8942 features and 415 samples with 43.54% zeros. We removed rows with over 20% zeros, resulting in a data matrix of 3187 rows with an overall missing rate of 7%. In the following simulation procedure, the non-zero values in these matrices served as ground truth. They were knocked out and then imputed, and the imputation accuracy was assessed by NRMSE over these non-zero ground truth values.

As described in the Methods section, we built the predictor network using: (1) linear correlation, (2) DCOL correlation, (3) dCov dependency, (4) difference in m/z (relative difference is less than 10 ppm) and RT values (difference less than 100) between any pair of features, indicating high likelihood of them being derived from the same metabolite, (5) m/z matching to neighboring metabolites on the KEGG metabolic network. For all KEGG metabolites, we first computed the theoretical m/z values of common adduct ions, and then computed the difference between these m/z values and the feature m/z values. A relative difference less than 10 ppm suggests a potential match. Two features matched to connected metabolites on the KEGG network are connected in the predictor network. For (4) and (5), five adduct ions were considered in this study: $[M + H]^+$, $[M + NH4]^+$, $[M + Na]^+$, $[M + K]^+$, $[M + 2Na - H]^+$. The MINMA package provides the option of using other adduct ions.

3.3.2 Computation

As indicated by Hrydziuszko and Viant (2012), the missingness may not occurr randomly in metabolomics data based on the analysis of DI FT-ICR MS metabolomics datasets. As a result, assuming a complete random missing mechanism may not be appropriate for imputation. Inspired by their work, we created the simulated datasets by knocking out a portion of locations from the ground-truth matrix by mimicking real missing patterns, and then evaluated each of the algorithms. To be specific, when we simulated a missing rate of r, each time we randomly selected one feature a from this ground-truth matrix, and one feature b from the original input matrix (before removing rows with $> 20\%$ missing). We knocked out the locations (encoded as NA) in feature a where there were observed zero values in the corresponding location in feature b, until the simulated dataset hit the missing rate of r. Similar approach has been taken in microarray missing value imputation study (Yu et al., 2011). In this way, without any assumptions of missing mechanism, imputation algorithms were all evaluated based on the real data missing pattern.

We simulated the datasets with various missing rates: 1%, 5%, 10%, 15%, 20%, 25%, 30%, 35% and 40%. For each of them, we generated 50 datasets and used the averaged NRMSE Ratio for evaluation. For each missing percentage, we tested various parameter settings for each method, i.e. $k = 5, 10, 15$ for KNN and $n_1, n_2, n_3 = 5, 10, 15$ for Net_SVR, and $nPCs = 5, 10, 15$ for BPCA and SVD, using 5 simulations, and then used the best parameter setting in the full simulation of 50 datasets.

All computations were run under R version 3.3.1. KNN was implemented using the function "impute.knn" from the package "impute"; BPCA was performed using the function "bpca" from the package "pcaMethods" (Stacklies et al., 2007); SVD was applied using function "impute.svd" in the package "bcv". For our mehtod Net_SVR, the SVR model was fitted using the function "svm" from the package "e1071" (Dimitriadou et al., 2009). The packages "impute" and "pcaMethods" are

Bioconductor packages.

3.3.3 Simulation Results

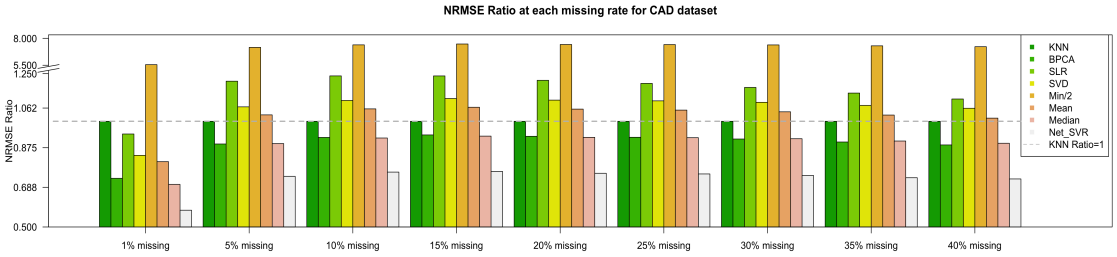
The simulation results are presented in Figure 3.2, where we applied all the candidate algorithms for imputation to two real datasets: CAD and CHD. For the simulation results of CAD dataset (Figure 3.2 a), at each missing rate ranging from 1% to 40%, BPCA, Median Imputation, and Net_SVR were below the dash line of 1, which means all three methods outperformed KNN (recall that NRMSE Ratio of KNN is always 1). The averaged NRMSE Ratio for them were 0.893, 0.890 and 0.727, respectively. SLR, SVD and Mean Imputation outperformed KNN only when missing rate was 1% and performed worse than KNN when missing rate was increased. Among all top three methods: BPCA, Median, and Net_SVR, when missing rate was as low as 1%, all three of them performed significantly better than KNN, as the missing rate increased, the gap compared to KNN shrank. Across all missing rates, our proposed algorithm Net_SVR performed the best as it obtained the smallest NRMSE Ratio compared to others with a minimum of 0.579 and maximum of 0.762.

The Net_SVR method also outperformed others when we applied all algorithms to the CHD dataset (Figure 3.1 b). It was the only algorithm that achieved an NRMSE Ratio below 1 across all missing rates. The averaged NRMSE Ratio of Net_SVR was 0.726 with a minimum of 0.518 and a maximum of 0.806. BPCA performed slightly better than KNN in most of the cases, but still yielded larger NRMSE Ratio at missing rate 10% (1.013), and was very close to KNN at missing rates 15% (NRMSE Ratio 1.000) and 20% (NRMSE Ratio 0.996). Median performed worse than KNN for the CHD dataset while it performed better in the CAD dataset, but the overall NRMSE Ratio of Median is around 1.

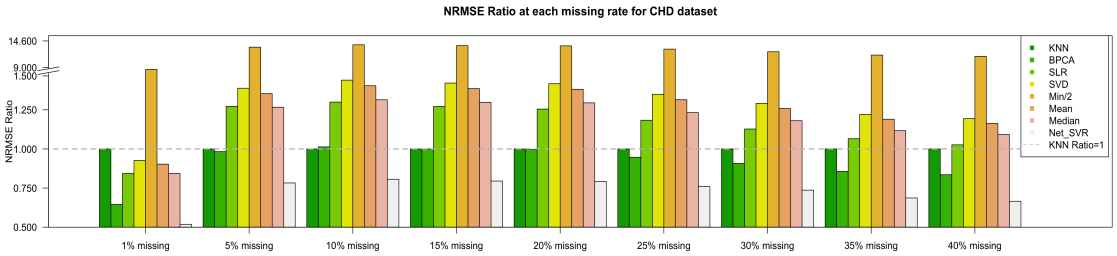
Additionally, of all the algorithms evaluated, imputing the missing locations by half of the minimum value yielded the largest NRMSE values, even though the data

was already log-transformed. It is because the Min/2 approach takes a different assumption than all the other methods. It assumes the unobserved values are missed only when the signal is below a detection threshold, which largely doesn't hold true in metabolomics data, thus, it is the worst among all the imputation algorithms. Our results are generally consistent with previous studies. The studies were somewhat diverse in terms of the data used, as well as the objectives used in judging the performance. Overall they showed a mixed performance between KNN, BPCA and SVD, while simple imputation methods such as Min/2 are in general unfavorable (Armitage et al., 2015; Gromski et al., 2014; Hrydziuszko and Viant, 2012; Taylor et al., 2016). Given the methods' performance may depend on the data type, sample size and missing mechanisms, it is most likely that no method is universally better. On the other hand, the knockout-impute simulation approach can be helpful. Given a specific dataset, a simulation similar to the current manuscript or those previously reported may be helpful in determining which imputation method best suites the data.

In metabolomics data, the underlying missing data pattern is unclear, and the assumptions needed for modeling missing mechanism is hard to justify. Thus in these two simulation studies, the missing locations were generated in a way to mimic the real data missing pattern, which is not in favor of any of the algorithms tested. The results indicated that Net_SVR may be a safer choice given it utilizes diverse information.



(a)



(b)

Figure 3.2: Simulation results. (a) CAD (AE) data; (b) CHD (C18) data.

The two datasets used both contained over 400 samples. However, some datasets in real-world applications may contain fewer samples. In order to evaluate how the methods perform under the situation of smaller sample sizes, we randomly subsampled the columns of the CHD dataset. The simulation result when we subsampled 100 columns is presented in Figure 3.3. All the algorithms performed similarly to the results in Figure 3.2 (b). With the sample size reduction, BPCA had better relative performance compared to itself but still worse than Net_SVR. Our proposed method still outperformed the others at most missing rates.

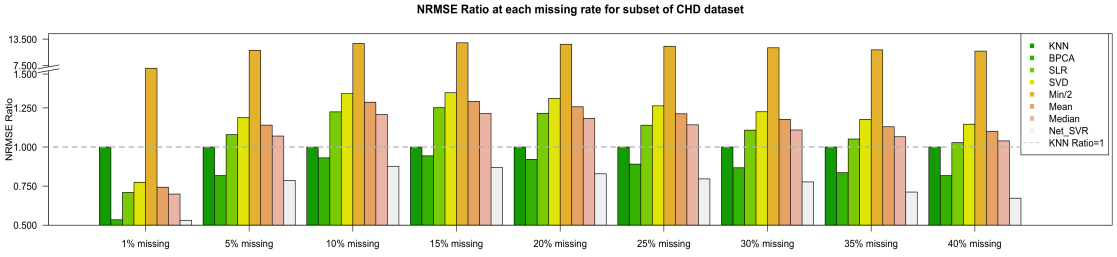


Figure 3.3: Simulation results from a subset of the CHD data with 100 columns.

3.4 Discussion and Conclusion

Imputation techniques are widely used for handling missing data in metabolomics studies. In this paper, we proposed a missing data imputation algorithm where a feature-level predictor network is constructed and then utilized for imputation. We incorporated different information for constructing the predictor network: the existing metabolic network structure, adduct ion relations among features, and various linear/ nonlinear pairwise correlations calculated from feature abundance levels. They are believed to be potentially helpful in depicting related features which may help in imputing each other's missing values. As this predictor network may include some false edges, hence noise in the imputation model, we applied the SVR model for reducing the influence of possible nuisance variables in the imputation process.

In real-world metabolomics studies, missing mechanism is hard to ascertain and the assumptions needed for modeling real data missing pattern is sometimes hard to justify. In order to better compare some of the widely-used algorithms in this field, we randomly sampled missing patterns from real features to mimic the real data missing pattern in the simulation studies. Simulation results showed that in high-resolution LC-MS data, the proposed algorithm Net_SVR outperforms the others at most missing rate settings.

In the application of the Net_SVR method, correctly specifying the types of adduct ions is important. Using too few adduct ion types causes the loss of valuable links that could contribute to imputation, while using adduct ions that are uncommon in the specific experimental platform may add many false edges in the predictor network. MINMA provides a function to match feature m/z values to 32 positive adduct ions, or 13 negative adduct ions. Alternatively, xMSannotator provides matching to more adduct ions (Uppal et al., 2017). Although m/z matching can always yield some false positives, nonetheless the frequency of adduct ion in the match can indicate which types of adduct ions are more common in the data, which can serve as the basis for

selecting adduct ions to use.

To summarize, by constructing a feature-level predictor network and then imputing missing values using a SVR model that uses neighborhood predictors on the network, the Net_SVR is an effective imputation method. The method can be extended in several directions: 1. other machine learning methods that are better resistant to nuisance variables can be used in place of the SVR; 2. when constructing the predictor network, different sources of information could be weighted differently based on the user's prior knowledge; 3. other feature relations can be incorporated; 4. if computationally feasible, the imputation sequence can be constantly updated along the way for better utilizing the network information.

Chapter 4

Integrate Genotypes with Imaging Phyenotypes for Shape Analysis and Gene Discovery for Alzheimer's Disease

4.1 Introduction

Imaging genetics is an emerging interdisciplinary field with a focus on assessing the impact of genetic variation on brain function and structure. It is a useful tool to uncover the etiologies complex neuropsychiatric diseases, such as Autism (Ameis and Szatmari, 2012) schizophrenia (Meyer-Lindenberg, 2010a) and Alzheimer's disease (Weiner et al., 2013). Traditional genetics studies have attempted to search genetic variants that are strongly associated with a behavior or related phenotypes; however, some findings were weak and inconsistent. There are considerable inter-subject differences in the behavioral measures, usually requiring large sample sizes to detect a signal. For neuropsychiatric disease, many genetic variants may not be directly associated with a clinical outcome or a behavior response but have a strong indirect effect which is mediated through molecular and cellular level information processing by neurons in the brain. We refer to this information processing procedure as brain activity. Functional neuroimaging, including functional magnetic resonance imaging (fMRI) and positron emission tomography (PET), is a set of powerful techniques to indirectly measure the brain activity at each location in the brain. Many current functional neuroimaging studies have focused on detecting the brain activation regions in association with particular cognitive and emotional tasks or at resting state.

Therefore, in imaging genetics studies, it is of great interest is to simultaneously select important genetic variants and detect brain activation regions where the genetic effects are strongly associated with brain activity. We refer to this procedure as genetic dissection of brain activation regions. However, to the best of our knowledge, none of the existing approaches can adequately address this question, although many of them have been adopted to detect the association between imaging biomarkers and genetic variants. The pioneer work includes voxelwise genome-wide association (vGWAS) study (Stein, Hua, Lee, Ho, Leow, Toga, Saykin, Shen, Foroud, Pankratz et al., 2010) where each voxel is considered as a phenotype and univariate regression

models were fitted for all the combinations of voxels and genetic variants. This approach enjoys the simplicity and fast computations but suffers from the difficulty of the multiple testing problem since the number of voxels often can be up to more than 10,000. To address those limitations, Huang et al. (2015) proposed a joint modeling approach with a well family-wise error control procedure and developed efficient computing tools for large-scale imaging genetics studies. Alternatively, Vounou, Nichols, Montana, Initiative et al. (2010) and Zhu et al. (2014a) proposed to use low rank regression to handle the high-dimensional neuroimaging phenotype, where a latent structure are imposed in the regression coefficients. Besides reduced rank approximation approaches, independent component analysis (ICA) (Liu et al., 2009) and canonical correlation analysis (CCA) (Chi et al., 2013) have been applied to discover the association between the imaging biomarkers and genetic variants with different latent structure assumptions. Different from all the existing methods, in this work, we propose a Bayesian hierarchical model for genetic dissection of brain activation regions. Our model consists of two levels of hierarchy.

At level 1, a Bayesian nonparametric level set model is developed for characterizing the shape of consistent brain activation regions across multiple subjects. The level set method has been widely used in image segmentation problems (e.g. Balafar et al., 2010; Li et al., 2011; Bergeest and Rohr, 2012), where contours (2D) or surfaces (3D) are represented as the zero-level set of a higher dimensional function, thus spatial voxels can be classified based on the function values: positive (inside the region) or negative (outside the region). We refer to this function as the level set function. The corresponding shape representation can characterize complex topological variations: the appearance of holes or tails, shapes that break down into smaller pieces, *etc.* The traditional level set based shape estimation problem can be solved by the numerical methods for partial differential equations. In our model, we propose to assign a Gaussian process prior to the level set function and make fully posterior inference on

the level set function as well as the shape of the activation regions, taking advantages of the good statistical properties of Gaussian processes.

At level 2, a regression model is adopted to select genetic variants that are strongly associated with the average brain activity within the region over multiple subjects, where a spike-and-slab prior and a Gaussian prior are chosen for feature selection. In particular, we model the average brain activation intensity within the region for each subject as the response variable; and we consider all the genetic variants as well as some clinical factors as predictors. We assign the Bayesian spike and slab prior on the regression coefficients for variable selection and thus to detect the important genetic variants of interest. The spike and slab prior was initially proposed by Mitchell and Beauchamp (1988); George and McCulloch (1993) and has been broadly adopted for various applications (Chipman et al., 2001; Ishwaran and Rao, 2005b,a). In the spike and slab prior specifications, the coefficients are mutually independent with a two-point mixture distribution made up of a 'uniform-like' flat distribution (called 'slab') and a 'degenerated-point-mass-at-zero-like' distribution (called 'spike'), leading to sparsity in the posterior inference.

Our motivating example is joint analysis of the fluorodeoxyglucose positron emission tomography (FDG-PET) data, single nucleotide polymorphisms (SNP) data and clinical data in the Alzheimer's Disease Neuroimaging Initiative (ADNI) study. Alzheimer's disease (AD) is one of the most common neurodegenerative disorders that impair mental functioning. It affects approximately eight percent of people who are 65 years of age or older. It has been shown that AD leads to nerve cell death and tissue loss in the brain (Bookheimer et al., 2000). As AD progresses, the brain shrinks dramatically; and abnormal changes in the brain worsen over time, eventually interfering with many aspects of brain function, such as memory loss, resulting in a decline in some intellectual abilities and changes in personality and behavior. New and potential treatments for AD focus on slowing the progression of the dis-

ease, making it important to identify at an early stage markers of future cognitive decline. Genetics studies showed that the presence of some of genes such as APOE and NEDD9 may be associated with cognitive decline in older persons (Wang, Nie, Huang, Kim, Nho, Risacher, Saykin and Shen, 2011). Structural magnetic resonance imaging (sMRI) studies (Bookheimer et al., 2000) identified that older persons with normal cognition may show medial temporal atrophy and thus indicate the possibility of future cognitive impairment. Many ADNI studies have focused on the joint analysis of sMRI and SNPs to discover the genetic effects on brain structure (Stein, Hua, Lee, Ho, Leow, Toga, Saykin, Shen, Foroud, Pankratz et al., 2010; Zhu et al., 2014a; Huang et al., 2015). Functional neuroimaging techniques can facilitate to discover more subtle alternations in brain function as AD progresses, thus analyses of PET or fMRI data in the ADNI studies have drawn much attention recently as well. For example, Huang et al. (2010) and Kundu and Kang (2016) developed statistical methods for leaning the genetic effects on the functional connectivity of AD. In this work, our goal is to study the genetic effects on functional brain activity for people at risk of AD, based on which we can identify the consistent brain activation regions across multiple subjects and quantify the changes of their shapes over times.

The remainder of the manuscript is organized as follows. In Section 4.2, we present the proposed model with prior specifications, and develop the posterior computation algorithms for fully Bayesian model. In Section 4.3, we evaluate the performance of the proposed method via extensive simulation studies. In Section 4.4, we illustrate the proposed method on analysis of the PET and SNP data from the ADNI study to detect influential SNPs and consistent activation regions across subjects. Finally, we conclude our paper by discussion in Section 4.5.

4.2 The Model

We propose a two-level Bayesian hierarchical model for fitting the brain activation regions that can simultaneously select important genetic variants. At Level 1, we focus on identifying the consistent activation regions across subjects, where the brain activation intensity may be different for different subjects. At Level 2, we are interested in identifying the important genetic variants (such as SNPs) that are strongly associated with brain activation intensities.

4.2.1 Two-Level Model

Suppose we collect brain images consisting of p voxels in a brain region $\mathcal{B} \subset \mathbb{R}^3$ and genetic variants of m SNPs from n subjects. Let $i(i = 1, \ldots, n)$ index the subject, $j(j = 1, \ldots, p)$ index the voxels and $k(k = 1, \ldots, m)$ index the SNPs. Denote by y_{ij} the observed imaging signal at voxel $\mathbf{v}_j \in \mathcal{B}$. Let S_{ik} be the genetic variant for SNP k.

At Level 1, we model the brain signal intensity within brain activation regions by assuming y_{ij} follow a normal mixture model:

$$(y_i(\mathbf{v}_j) \mid \phi, \mu_i, \sigma_i^2) \sim \mathrm{N}\left[\mu_i \delta\{\phi(\mathbf{v}_j)\}, \sigma_i^2\right], \tag{4.1}$$

where $\delta(x) = 1$ if $x > 0$ and $\delta(x) = 0$ if $x \leq 0$. The level set function $\phi(\mathbf{v}) : \mathcal{B} \to \mathbb{R}$ determines the brain activation regions. For any voxel $\mathbf{v}$ in the brain, if $\phi(\mathbf{v}) > 0$ implying that $\delta\{\phi(\mathbf{v}_j)\} = 1$, then it is located in a activation region and the brain signal y_{ij} has an average activation intensity μ_i. Otherwise, the voxel is located outside the brain activation regions with a mean intensity zero. The parameter σ_i^2 is the variance of the signal y_{ij} across all voxels j for subject i.

At Level 2, we link the activation intensity to the genetic variant by using a regression model

$$\mu_i \sim \mathrm{N}\left(\sum_{k=1}^{m} S_{ik}\eta_k, \tau_\mu^2\right). \tag{4.2}$$

where η_k is the genetic effects of SNP k on the brain activation intensity. The variance parameter τ_μ^2 characterizes the variability of the average activation intensity that are not from the genetic variants.

4.2.2 Prior Specifications

In this section, we discuss the prior specifications for models (4.1) and (4.2).

At Level 1, to guarantee the robustness and flexibility of modeling the activation regions shape, we assign a Gaussian process prior to the level set function $\phi(\mathbf{v})$ with mean zero and covariance kernel function, denoted as

$$\phi \sim \mathcal{GP}(0, \kappa),$$

where $\kappa(\mathbf{v}, \mathbf{v}') : \mathcal{B} \times \mathcal{B} \to \mathbb{R}$ is a symmetric positive definite kernel function.

At Level 2, we impose sparsity on η_k for identify the important SNP sets that are strongly associated with the brain activation intensity. We assign the spike-and-slab prior proposed by Ishwaran and Rao (2005c) to η_k. denoted this prior as spikeslab$(\cdot)$ with the following conjugate bimodal hyperparameter setting:

$$\begin{aligned} \left[\eta_k \mid \gamma_k, \tau_k^2\right] &\sim \mathrm{N}[0, \gamma_k \tau_k^2], \\ [\gamma_k \mid \nu_0, w] &\sim (1-w)\delta_{\nu_0} + w\delta_1, \\ w &\sim \mathrm{Uniform}[0, 1]. \end{aligned}$$

where δ_{ν_0} is the point mass at ν_0, δ_1 is the point mass at 1. w is the prior inclusion probability indicating how likely each feature to be selected. Pre-defined value ν_0 usually select very small so that the "spike" (δ_{ν_0}, *i.e.* $\mathrm{N}[0, \nu_0\tau_k^2]$) part and "slab" ($\delta_1$, *i.e.* $\mathrm{N}[0, \tau_k^2]$) part can be mostly differentiated from each other.

For all the variance parameters σ_i, τ_μ^2 and τ_k^2, we assume they are mutually indenpendent and follow conjugate priors:

$$\sigma_i^2 \sim \mathrm{IG}(a_1, a_2), \qquad \tau_\mu^2 \sim \mathrm{IG}(b_1, b_2), \qquad \tau_k^2 \sim \mathrm{IG}(c_1, c_2),$$

where $\mathrm{IG}(w_1, w_2)$ represents an inverse gamma prior with shape w_1 and rate w_2.

4.2.3 Model Representation

To implement posterior computation algorithm, we need to consider model approximations. First, we consider the basis expansion approximation $\phi(\mathbf{v}) = \sum_{l=1}^{L} \beta_l \psi_l(\mathbf{v})$ with $\beta_l \overset{\text{iid}}{\sim} \mathrm{N}(0, \Lambda)$, where $\{\psi_l(\cdot)\}$ and $\{\lambda_l\}$ are respectively eigen functions and eigenvalues for the kernel function $\kappa(\cdot, \cdot)$ that are shared cross all patient samples. Second, we introduce the function $H_\epsilon[x] = \frac{1}{2}[1 + \frac{2}{\pi}\arctan(\frac{x}{\epsilon})]$ with $H_\epsilon[x] \to \delta[x]$ as $\epsilon \to 0$. Note that its first derivative is $H_\epsilon^{(1)}[x] = \frac{1}{\pi}\frac{\epsilon}{\epsilon^2+x^2}$. Let $\mathbf{H}_\epsilon(\boldsymbol{\beta}) = (H_\epsilon(\boldsymbol{\psi}_1^{\mathrm{T}}\boldsymbol{\beta}), \cdots, H_\epsilon(\boldsymbol{\psi}_p^{\mathrm{T}}\boldsymbol{\beta}))^T = (H_1, \cdots, H_p)^T$

Write $\mathbf{y}_i = (y_{i1}, \ldots, y_{ip})^{\mathrm{T}}$, $\mathbf{S} = (S_{ik})$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^{\mathrm{T}}$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_L)^{\mathrm{T}}$ and $\boldsymbol{\Psi} = (\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_p)^{\mathrm{T}}$ with $\boldsymbol{\psi}_j = [\psi_{1,j}, \ldots, \psi_{L,j}]^{\mathrm{T}}$ and $\psi_{l,j} = \psi_l(\mathbf{v}_j)$. Then our Bayesian hierarchical model with prior specifications can be represented as

$$
\begin{aligned}
\mathbf{y}_i \mid \boldsymbol{\beta}, \mu_i, \sigma_i^2 &\sim \mathrm{N}_p\left[\mu_i \mathbf{H}_\epsilon(\boldsymbol{\beta}), \sigma_i^2 \mathbf{I}_p\right], \\
\boldsymbol{\beta} &\sim \mathrm{N}_L\left[\mathbf{0}, \boldsymbol{\Lambda}_L\right], \\
\boldsymbol{\mu} &\sim \mathrm{N}_n[\mathbf{S}^T\boldsymbol{\eta}, \tau_\mu^2 \mathbf{1}_n], \\
\boldsymbol{\eta} \mid \boldsymbol{\gamma}, \boldsymbol{\tau}^2 &\sim \mathrm{N}_m[\mathbf{0}, \boldsymbol{\Gamma}(\boldsymbol{\gamma}, \boldsymbol{\tau}^2)], \\
\gamma_k | w &\sim (1-w)\delta_{\nu_0} + w\delta_1 \\
\sigma_i^2 &\sim \mathrm{IG}[a_1, a_2], \\
\tau_\mu^2 &\sim \mathrm{IG}[b_1, b_2] \\
\tau_k^2 &\sim \mathrm{IG}[c_1, c_2] \\
w &\sim \mathrm{Uniform}[0, 1]
\end{aligned}
$$

where $\boldsymbol{\Gamma}$ is a diagonal matrix with (k, k) element being $\gamma_k \tau_k^2$, $\mathbf{Y}_{n\times p}$ is signal matrix and $\mathbf{y}_i$ as the signal vector for subject i, $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_n^2)$, $\boldsymbol{\tau}^2 = (\tau_1^2, \ldots, \tau_m^2)$.

4.2.4 Posterior Computation

We use the Metropolis adjusted Langevin algorithm MALA by Girolami and Calderhead (2011) and Stochastic Search Variable Selection SSVS by George and McCulloch (1997) within Gibbs sampling for posterior computation. The joint posterior distribution is given by

$$\begin{aligned}\pi(\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\sigma}^2,\boldsymbol{\eta},\boldsymbol{\gamma},\boldsymbol{\tau}^2,\tau^2,w \mid \mathbf{Y}) \;\propto\; & \pi(\mathbf{Y}|\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\sigma}^2)\pi(\boldsymbol{\beta})\pi(\boldsymbol{\sigma}^2)\pi(\boldsymbol{\mu}|\boldsymbol{\eta},\tau^2)\pi(\boldsymbol{\eta}|\boldsymbol{\gamma},\boldsymbol{\tau}^2) \\ & \pi(\boldsymbol{\gamma}|w)\pi(\boldsymbol{\tau}^2)\pi(\tau^2)\pi(w)\end{aligned}$$

The Gibbs sampler works as follows:

1. Update $\boldsymbol{\mu}$

   The full conditional of $\boldsymbol{\mu}$ is given by

$$\begin{aligned}\pi(\boldsymbol{\mu} \mid \bullet) \;&\propto\; \pi(\mathbf{Y}|\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\sigma}^2)\pi(\boldsymbol{\mu}|\boldsymbol{\eta},\tau_\mu^2) \\ &\propto\; \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{q}\frac{1}{\sigma_i^2}(\mathbf{y}_i-\mu_i\mathbf{H}_\epsilon(\boldsymbol{\beta}))^T(\mathbf{y}_i-\mu_i\mathbf{H}_\epsilon(\boldsymbol{\beta}))+\sum_{i=1}^{q}\frac{1}{\tau_\mu^2}(\mu_i-\mathbf{S}_i\boldsymbol{\eta})^2\right]\right\} \\ &\propto\; \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{q}\left(\frac{1}{\sigma_i^2}\sum_{j=1}^{p}H_j^2+\frac{1}{\tau_\mu^2}\right)\mu_i^2-2\left(\frac{1}{\sigma_i^2}\sum_{j=1}^{p}H_jy_{ij}+\frac{1}{\tau_\mu^2}\sum_{k=1}^{m}S_{ik}\eta_k\right)\mu_i\right]\right\}\end{aligned}$$

   This implies that we can update each μ_i by sampling from

$$[\mu_i \mid \bullet] \;\sim\; \mathrm{N}\left[\left(\frac{1}{\sigma_i^2}\sum_{j=1}^{p}H_j^2+\frac{1}{\tau_\mu^2}\right)^{-1}\left(\frac{1}{\sigma_i^2}\sum_{j=1}^{p}H_jy_{ij}+\frac{1}{\tau_\mu^2}\sum_{k=1}^{m}S_{ik}\eta_k\right),\left(\frac{1}{\sigma_i^2}\sum_{j=1}^{p}H_j^2+\frac{1}{\tau_\mu^2}\right)^{-1}\right]$$

2. Update $\boldsymbol{\sigma}^2$

   The full conditional of $\boldsymbol{\sigma}^2$ is given by

$$\begin{aligned}\pi(\boldsymbol{\sigma}^2 \mid \bullet) \;&\propto\; \pi(\mathbf{Y}|\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\sigma}^2)\pi(\boldsymbol{\sigma}^2) \\ &\propto\; \prod_{i=1}^{q}\frac{1}{\sigma_i^p}\exp\left\{-\frac{1}{2\sigma_i^2}\left(\mathbf{y}_i-\mu_i\mathbf{H}\right)^T\left(\mathbf{y}_i-\mu_i\mathbf{H}\right)\right\}\times\frac{1}{\sigma_i^{2a_1+2}}\exp\left\{-\frac{a_2}{\sigma_i^2}\right\} \\ &\propto\; \prod_{i=1}^{q}\frac{1}{\sigma_i^{2a_1+p+2}}\exp\left\{-\frac{a_2+\sum_{j=1}^{p}(y_{ij}-\mu_iH_j)^2/2}{\sigma_i^2}\right\}\end{aligned}$$

This implies that we can update each σ_i^2 by sampling from

$$[\sigma_i^2 \mid \bullet] \sim \text{IG}\left[a_1 + \frac{p}{2}, a_2 + \frac{1}{2}\sum_{j=1}^{p}(y_{ij} - \mu_i H_j)^2\right]$$

3. Update $\boldsymbol{\beta}$

The full conditional distribution of $\boldsymbol{\beta}$ is given by

$$\begin{aligned}\pi(\boldsymbol{\beta}|\bullet) &\propto \pi(\mathbf{Y}|\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2)\pi(\boldsymbol{\beta}) \\ &\propto \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{q}\frac{1}{\sigma_i^2}(\mathbf{y}_i - \mu_i\mathbf{H}_\epsilon(\boldsymbol{\beta}))^T(\mathbf{y}_i - \mu_i\mathbf{H}_\epsilon(\boldsymbol{\beta})) + \boldsymbol{\beta}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{\beta}\right]\right\}\end{aligned}$$

Then log full conditional distribution of $\boldsymbol{\beta}$ is given by

$$\mathcal{L}(\boldsymbol{\beta}) = \log[\pi(\boldsymbol{\beta} \mid \bullet)] = C - \frac{1}{2}\left[\sum_{i=1}^{q}\frac{1}{\sigma_i^2}\sum_{j=1}^{p}\left(y_{ij} - \mu_i H_\epsilon\left[\sum_{l=1}^{L}\beta_l\psi_{l,j}\right]\right)^2 + \sum_{l=1}^{L}\frac{\beta_l^2}{\lambda_l}\right],$$

and

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\beta_l} &= \sum_{i=1}^{q}\sum_{j=1}^{p}\frac{\mu_i}{\sigma_i^2}\psi_{l,j}H_\epsilon^{(1)}\left[\boldsymbol{\psi}_j^{\mathrm{T}}\boldsymbol{\beta}\right]\left(y_{ij} - \mu_i H_\epsilon\left[\boldsymbol{\psi}_j^{\mathrm{T}}\boldsymbol{\beta}\right]\right) - \frac{\beta_l}{\lambda_l} \\ \frac{\partial^2\mathcal{L}}{\partial\beta_l^2} &= \sum_{i=1}^{q}\sum_{j=1}^{p}\left\{\frac{\mu_i}{\sigma_i^2}\psi_{l,j}^2H_\epsilon^{(2)}\left[\boldsymbol{\psi}_j^{\mathrm{T}}\boldsymbol{\beta}\right]\left(y_{ij} - \mu_i H_\epsilon\left[\boldsymbol{\psi}_j^{\mathrm{T}}\boldsymbol{\beta}\right]\right) - \frac{\mu_i^2}{\sigma_i^2}\psi_{l,j}^2H_\epsilon^{2(1)}\left[\boldsymbol{\psi}_j^{\mathrm{T}}\boldsymbol{\beta}\right]\right\} - \frac{1}{\lambda_l} \\ \frac{\partial^2\mathcal{L}}{\partial\beta_l\partial\beta_k} &= \sum_{i=1}^{q}\sum_{j=1}^{p}\left\{\frac{\mu_i}{\sigma_i^2}\psi_{l,j}\psi_{k,j}\mu H_\epsilon^{(2)}\left[\boldsymbol{\psi}_j^{\mathrm{T}}\boldsymbol{\beta}\right]\left(y_{ij} - \mu_i H_\epsilon\left[\boldsymbol{\psi}_j^{\mathrm{T}}\boldsymbol{\beta}\right]\right) - \frac{\mu_i^2}{\sigma_i^2}\psi_{l,j}\psi_{k,j}H_\epsilon^{2(1)}\left[\boldsymbol{\psi}_j^{\mathrm{T}}\boldsymbol{\beta}\right]\right\},\end{aligned}$$

This further implies that

$$\begin{aligned}\nabla_{\boldsymbol{\beta}}\mathcal{L}(\boldsymbol{\beta}) &= \sum_{i=1}^{q}\frac{\mu_i}{\sigma_i^2}\sum_{j=1}^{p}(y_{ij} - \mu_i H_j)H_j^{(1)}\boldsymbol{\psi}_j - \boldsymbol{\Lambda}^{-1}\boldsymbol{\beta} \\ &= \left(\frac{\boldsymbol{\mu}}{\boldsymbol{\sigma}^2}\right)^T\left(Y_{q\times p} - \boldsymbol{\mu}_{q\times 1}H_\epsilon^T\left[\boldsymbol{\psi}_j^{\mathrm{T}}\boldsymbol{\beta}\right]_{1\times p}\right)diag(H_{p\times 1}^{(1)})_{p\times p}\boldsymbol{\psi}_j - \boldsymbol{\Lambda}^{-1}\boldsymbol{\beta}, \\ \mathbf{G}(\boldsymbol{\beta}) &= \{g_{l,k}(\boldsymbol{\beta})\}_{L\times L},\end{aligned}$$

where

$$\begin{aligned}g_{l,l}(\boldsymbol{\beta}) &= -\mathrm{E}[\frac{\partial^2 g}{\partial\beta_l^2}] = \sum_{i=1}^{q}\sum_{j=1}^{p}\frac{\mu_i^2}{\sigma_i^2}\psi_{l,j}^2H_\epsilon^{2(1)}\left[\boldsymbol{\psi}_j^{\mathrm{T}}\boldsymbol{\beta}\right] + \frac{1}{\lambda_l} \\ g_{l,k}(\boldsymbol{\beta}) &= -\mathrm{E}[\frac{\partial^2 g}{\partial\beta_l\partial\beta_k}] = \sum_{i=1}^{q}\sum_{j=1}^{p}\frac{\mu_i^2}{\sigma_i^2}\psi_{l,j}\psi_{k,j}H_\epsilon^{2(1)}\left[\boldsymbol{\psi}_j^{\mathrm{T}}\boldsymbol{\beta}\right].\end{aligned}$$

Thus, the proposal distribution for $\boldsymbol{\beta}$ of the MALA is given by

$$\boldsymbol{\beta}^* \sim \mathrm{N}\left[\boldsymbol{\beta} + \frac{\Delta^2}{2}\mathbf{G}^{-1}(\boldsymbol{\beta})\nabla_{\boldsymbol{\beta}}\mathcal{L}(\boldsymbol{\beta}), \Delta^2\mathbf{G}^{-1}(\boldsymbol{\beta})\right],$$

4. Update $\boldsymbol{\eta}$

The full conditional distribution of $\boldsymbol{\eta}$ is given by

$$\begin{aligned}\pi(\boldsymbol{\eta} \mid \bullet) &\propto \pi(\boldsymbol{\mu}|\boldsymbol{\eta}, \tau_\mu^2)\pi(\boldsymbol{\eta}|\boldsymbol{\gamma}, \boldsymbol{\tau}^2) \\ &\propto \exp\left\{-\frac{1}{2\tau_\mu^2}(\boldsymbol{\mu} - \mathbf{S}\boldsymbol{\eta})^T(\boldsymbol{\mu} - \mathbf{S}\boldsymbol{\eta}) - \frac{1}{2}\boldsymbol{\eta}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\eta}\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{\eta}^T\left(\frac{1}{\tau_\mu^2}\mathbf{S}^T\mathbf{S} + \boldsymbol{\Gamma}^{-1}\right)\boldsymbol{\eta} - 2\left(\frac{\boldsymbol{\mu}^T\mathbf{S}}{\tau_\mu^2}\right)\boldsymbol{\eta}\right]\right\}\end{aligned}$$

This implies that we can update $\boldsymbol{\eta}$ by sampling

$$[\boldsymbol{\eta} \mid \bullet] \sim \mathrm{N}\left[\left(\frac{1}{\tau_\mu^2}\mathbf{S}^T\mathbf{S} + \boldsymbol{\Gamma}^{-1}\right)^{-1}\frac{\boldsymbol{\mu}^T\mathbf{S}}{\tau_\mu^2}, \left(\frac{1}{\tau_\mu^2}\mathbf{S}^T\mathbf{S} + \boldsymbol{\Gamma}^{-1}\right)^{-1}\right]$$

5. Update $\boldsymbol{\gamma}$

The full conditional distribution of $\boldsymbol{\gamma}$ is given by

$$\begin{aligned}\pi(\boldsymbol{\gamma} \mid \bullet) &\propto \pi(\boldsymbol{\eta}|\boldsymbol{\gamma}, \boldsymbol{\tau}^2)\pi(\boldsymbol{\gamma}|w) \\ &\propto \prod_{k=1}^{m}\left[\frac{\omega_{0,k}}{\omega_{0,k} + \omega_{1,k}}\delta_{\nu_0} + \frac{\omega_{1,k}}{\omega_{0,k} + \omega_{1,k}}\delta_1\right]\end{aligned}$$

where $\omega_{0,k} = (1-w)\nu_0^{-1/2}\exp\left(-\frac{\eta_k^2}{2\nu_0\tau_k^2}\right)$ and $\omega_{1,k} = w\exp\left(-\frac{\eta_k^2}{2\tau_k^2}\right)$

We can update each γ_k by sampling from

$$[\gamma_k \mid \bullet] \sim \frac{\omega_{0,k}}{\omega_{0,k} + \omega_{1,k}}\delta_{\nu_0} + \frac{\omega_{1,k}}{\omega_{0,k} + \omega_{1,k}}\delta_1$$

6. Update $\boldsymbol{\tau}^2$

The full conditional of $\boldsymbol{\tau}^2$ is given by

$$\begin{aligned}\pi(\boldsymbol{\tau}^2 \mid \bullet) &\propto \pi(\boldsymbol{\eta}|\boldsymbol{\gamma}, \boldsymbol{\tau}^2)\pi(\boldsymbol{\tau}^2) \\ &\propto \prod_{k=1}^{m}\frac{1}{\tau_k}\exp\left\{-\frac{\eta_k^2}{2\tau_k^2\gamma_k}\right\} \times \frac{1}{\tau_k^{2c_1+2}}\exp\left\{-\frac{c_2}{\tau_k^2}\right\} \\ &\propto \prod_{k=1}^{m}\frac{1}{\tau_k^{2c_1+3}}\exp\left\{-\frac{c_2 + \eta_k^2/2\gamma_k}{\tau_k^2}\right\}\end{aligned}$$

This implies that we can update each τ_k^2 by sampling

$$[\tau_k^2 \mid \bullet] \sim \mathrm{IG}\left[c_1 + \frac{1}{2}, c_2 + \frac{\eta_k^2}{2\gamma_k}\right]$$

7. Update w

The full conditional of w is given by

$$\begin{aligned}\pi(w \mid \bullet) &\propto \pi(\boldsymbol{\gamma}|w)\pi(w) \\ &\propto \prod_{k=1}^{m}[(1-w)I[\gamma_k = \nu_0] + wI[\gamma_k = 1]] \\ &\propto w^{\sum_{k=1}^{m} I[\gamma_k=1]}(1-w)^{\sum_{k=1}^{m} I[\gamma_k=\nu_0]}\end{aligned}$$

We update from

$$[w \mid \bullet] \sim \text{Beta}\left[1 + \sum_{k=1}^{m} I[\gamma_k = 1], 1 + \sum_{k=1}^{m} I[\gamma_k = \nu_0]\right]$$

8. Update τ_μ^2

The full conditional of τ_μ^2 is given by

$$\begin{aligned}\pi(\tau_\mu \mid \bullet) &\propto \pi(\boldsymbol{\mu}|\boldsymbol{\eta}, \tau_\mu^2)\pi(\tau_\mu) \\ &\propto \frac{1}{\tau_\mu^q}\exp\left\{-\frac{1}{2\tau_\mu^2}\left(\boldsymbol{\mu} - \mathbf{S}\boldsymbol{\eta}\right)^T\left(\boldsymbol{\mu} - \mathbf{S}\boldsymbol{\eta}\right)\right\} \times \frac{1}{\tau_\mu^{2b_1+2}}\exp\left\{-\frac{b_2}{\tau_\mu^2}\right\} \\ &\propto \frac{1}{\tau_\mu^{2b_1+q+2}}\exp\left\{-\frac{b_2 + \left(\boldsymbol{\mu} - \mathbf{S}\boldsymbol{\eta}\right)^T\left(\boldsymbol{\mu} - \mathbf{S}\boldsymbol{\eta}\right)/2}{\tau_\mu^2}\right\}\end{aligned}$$

Thus we sample from

$$[\tau_\mu \mid \bullet] \sim \text{IG}\left[b_1 + \frac{q}{2}, b_2 + \frac{\sum_{i=1}^{q}\left(\mu_i - \mathbf{S}_i\boldsymbol{\eta}\right)^T\left(\mu_i - \mathbf{S}_i\boldsymbol{\eta}\right)}{2}\right]$$

4.2.5 Non-sparse Bayesian Variable Selection Model

For fast computation purpose, we also propose another non-sparse version of the proposed algorithm, where we impose a conjugate normal prior on $\boldsymbol{\eta}$. The model is

represented as

$$\begin{aligned}
\mathbf{y}_i \mid \boldsymbol{\beta}, \mu_i, \sigma_i^2 &\sim \mathrm{N}_p\left[\mu_i \mathbf{H}_\epsilon(\boldsymbol{\beta}), \sigma_i^2 \mathbf{I}_p\right], \\
\boldsymbol{\beta} &\sim \mathrm{N}_L\left[\mathbf{0}, \boldsymbol{\Lambda}_L\right], \\
\boldsymbol{\mu} &\sim \mathrm{N}_q[\mathbf{S}^T\boldsymbol{\eta}, \tau_\mu^2 \mathbf{I}_q], \\
\boldsymbol{\eta} &\sim \mathrm{N}_m[\mathbf{0}, \tau_\eta^2 \mathbf{I}_m] \\
\sigma_i^2 &\sim \mathrm{IG}[a_1, a_2], \\
\tau_\mu^2 &\sim \mathrm{IG}[b_1, b_2] \\
\tau_\eta^2 &\sim \mathrm{IG}[d_1, d_2]
\end{aligned}$$

For posterior computation,

1. Update $\boldsymbol{\eta}$

   The full conditional distribution of $\boldsymbol{\eta}$ is given by

$$\begin{aligned}
\pi(\boldsymbol{\eta} \mid \bullet) &\propto \pi(\boldsymbol{\mu}|\boldsymbol{\eta}, \tau_\mu^2)\pi(\boldsymbol{\eta}|\tau_\eta^2) \\
&\propto \exp\left\{-\frac{1}{2\tau_\mu^2}(\boldsymbol{\mu} - \mathbf{S}\boldsymbol{\eta})^T(\boldsymbol{\mu} - \mathbf{S}\boldsymbol{\eta}) - \frac{1}{2\tau_\eta^2}\boldsymbol{\eta}^T\boldsymbol{\eta}\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{\eta}^T\left(\frac{1}{\tau_\mu^2}\mathbf{S}^T\mathbf{S} + \frac{1}{\tau_\eta^2}\mathbf{I}_m\right)\boldsymbol{\eta} - 2\left(\frac{\boldsymbol{\mu}^T\mathbf{S}}{\tau_\mu^2}\right)\boldsymbol{\eta}\right]\right\}
\end{aligned}$$

   This implies that we can update $\boldsymbol{\eta}$ by sampling

$$[\boldsymbol{\eta} \mid \bullet] \sim \mathrm{N}\left[\left(\frac{1}{\tau_\mu^2}\mathbf{S}^T\mathbf{S} + \frac{1}{\tau_\eta^2}\mathbf{I}_m\right)^{-1}\frac{\boldsymbol{\mu}^T\mathbf{S}}{\tau_\mu^2}, \left(\frac{1}{\tau_\mu^2}\mathbf{S}^T\mathbf{S} + \frac{1}{\tau_\eta^2}\mathbf{I}_m\right)^{-1}\right]$$

2. Update τ_η^2

   The full conditional of τ_η^2 is given by

$$\begin{aligned}
\pi(\tau_\eta \mid \bullet) &\propto \pi(\boldsymbol{\eta}|\tau_\eta^2)\pi(\tau_\eta) \\
&\propto \frac{1}{\tau_\eta^m}\exp\left\{-\frac{1}{2\tau_\eta^2}\boldsymbol{\eta}^T\boldsymbol{\eta}\right\} \times \frac{1}{\tau_\eta^{2d_1+2}}\exp\left\{-\frac{d_2}{\tau_\eta^2}\right\} \\
&\propto \frac{1}{\tau_\eta^{2d_1+m+2}}\exp\left\{-\frac{d_2 + \boldsymbol{\eta}^T\boldsymbol{\eta}/2}{\tau_\eta^2}\right\}
\end{aligned}$$

Thus we sample from

$$[\tau_\eta \mid \bullet] \sim \text{IG}\left[d_1 + \frac{m}{2}, d_2 + \frac{\sum_{k=1}^{m} \eta_k^2}{2}\right]$$

For the variable selection, we apply an ad-hoc method based on posterior credible intervals. For correlating the variables (clinical or SNPs) with brain image intensity levels, we use the null hypothesis that SNP k is uncorrelated with the intensity level inside the activation region ($H_0 : \eta_k = 0$) and the alternative hypothesis that SNP k is not uncorrelated with the intensity level inside the activation region ($H_a : \eta_k \neq 0$). Based on the marginal posterior distribution for η_k, if 0 is included in the posterior 95% credible interval, we assign $\gamma_k = 1$, otherwise $\gamma_k = 0$ where γ_k is the same indicator variable introduced in SSVS. We approximate the posterior inclusion probability of SNP k: $E\gamma_k$ using the averaged values after burn-in $\overline{\gamma}_k$. Then the SNPs with posterior inclusion probability larger than 0.01 are selected as important.

4.3 Simulation Studies

We tested the performance for learning activation region shapes and selection influential variables using proposed method starting from the simplest scenario and then gradually extended to the most complicated scenario. For the simplest simulation setting, we simulated a single subject, 2D imaging data and zero predictor matrix, *i.e.* set $n = 1, d = 2, \mathbf{S} = \mathbf{0}$ thus no variable selection involved. For the most complicated simulation setting, we simulated multiple subjects, 3D imaging data and utilizing the predictors in real data analysis for selection.

4.3.0.1 Single Subject with 2D Image and no Variable Selection

In this simulation study, the objective is to test the Bayesian nonparametric level set method for random shape fitting. We simulated 2D images of size 150×150 on a square region $[-1, 1]^2$ ($d = 2$). We considered three activation region shapes: circles,

squares and random shapes. We simulated data by setting $\sigma^2 = 1$, and the signal intensities μ and the level set function were set as follows:

- Circle shapes: set the signal intensity $\mu = 1$ (weak) and the true level set function $\phi(\mathbf{v}) = \exp\{-0.5(v_1^2 + v_2^2)\} - 0.8$
- Square shapes: set the signal intensity $\mu = 3$ (strong) and the true level set function $\phi(\mathbf{v}) = \exp\{-0.5(|v_1| + |v_2|)\} - 0.8$
- Random shapes: set the signal intensity $\mu = 2$ (intermediate) and draw the true level set from a Gaussian process with mean zero and covariance kernel $\kappa(v_1, v_2) = \exp(-10(v_1 - v_2)^2)$

For the posterior computation, we set $\epsilon = 1 \times 10^{-3}$ and run 5000 iterations with 2000 burn-in. The shape estimation results were presented in Figure 4.1.

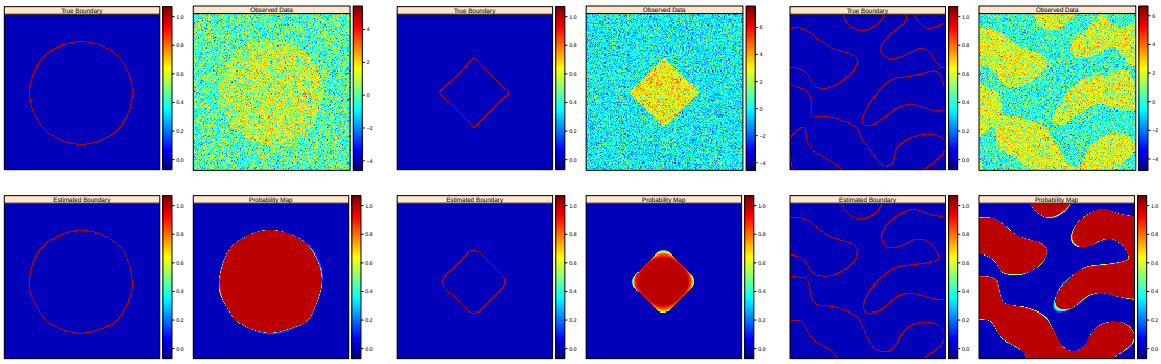


Figure 4.1: Single subject with 2D image and no variable selection: from top to bottom, left to right: simulated boundary in red, simulated intensity data, estimated boundary in red and inclusion probability map

4.3.0.2 Multi-subjects with 3D Image and no Variable Selection

We evaluated the proposed method on a total of $m = 50$ subjects with 3D images simulated for each of them. The 3D image grid was of $20 \times 20 \times 20$ ($p = 8000$) on a square region $[-1, 1]^2$ ($d = 3$). Again we set $\mathbf{S} = \mathbf{0}$ so that there is no variable

selection involved. We considered three different shapes of activation region: spheres, diamonds, and random shapes. We set $\sigma_i^2 = 1, i = 1, \ldots, n$. The signal intensities μ_i $(i = 1, \ldots, n)$ and the level set function $\phi(\mathbf{v})$ were set as follows:

- Sphere shapes: set the signal intensity $\mu_i \sim N(1, 1)$ and the true level set function $\phi(\mathbf{v}) = \exp\{-0.5(v_1^2 + v_2^2 + v_3^2)\} - 0.7$
- Diamond shapes: set the signal intensity $\mu_i \sim N(3, 1)$ and the true level set function $\phi(\mathbf{v}) = \exp\{-0.5(|v_1| + |v_2| + |v_3|)\} - 0.6$
- Random signal shapes: set the signal intensity $\mu_i \sim N(2, 1)$ and draw the true level set from a Gaussian process with mean zero and covariance kernel $\kappa(\mathbf{v_1}, \mathbf{v_2}) = \exp(-10(\mathbf{v_1} - \mathbf{v_2})^T(\mathbf{v_1} - \mathbf{v_2}))$

For the posterior computation, we set $\epsilon = 1 \times 10^{-3}$, $\alpha = 0.8$ as PCA percent and run 5000 iteration with 2000 burn-in. The shape segmentation results were respectively summarized in Figure 4.2.

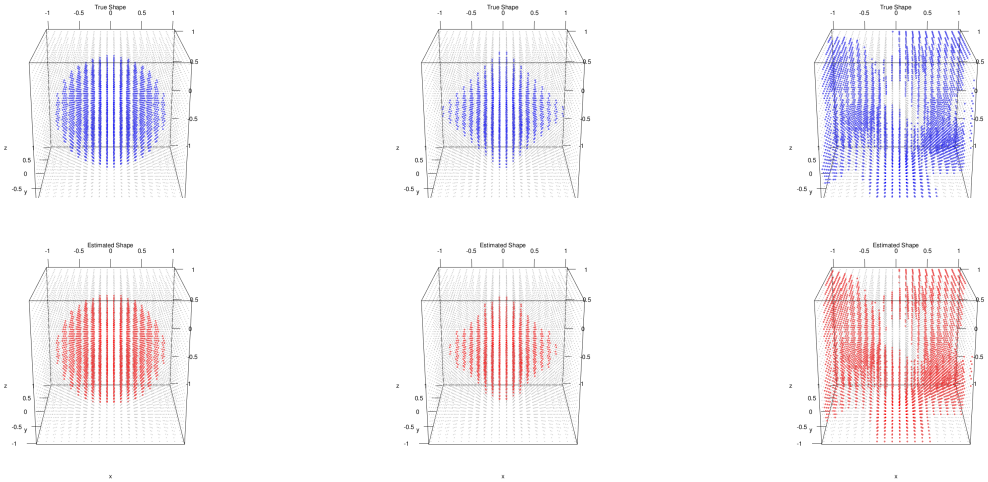


Figure 4.2: Multiple subjects with 3D image and no variable selection: top/ bottom: simulated/ estimated shapes; classification accuracies; left to right: MSE($\boldsymbol{\mu}$) are sphere 0.98, 0.000218; diamond 0.98, 0.000728; random 0.96, 0.000158

4.3.0.3 Multi-subjects with 3D Image and variable selection

In the simulation study, we evaluated the proposed method on the most complicated scenario where there is a total of $n = 235$ subjects with 3D images simulated for each of them. We only took the first 200 columns ($m = 200$) from the SNP matrix in real data analysis to form $\mathbf{S}$ in the simulations. We randomly selected 5 of them as signal (without loss of generality, set $\eta_k = 1, k = 1, \ldots, 5$) and the remaining 195 ($\eta_k = 0, k = 6, \ldots, 200$) as noise. Like previous simulation studies, we considered three different activation region shapes with different combination of abilities for shape estimation and variable selection quantified by signal-to-noise-ratio $SNR(\bullet)$.

$$\begin{aligned} SNR(\boldsymbol{\beta}) &= \frac{1}{q}\sum_{i=1}^{q} SNR(\boldsymbol{\beta}|\mathbf{y}_i) = \frac{1}{q}\sum_{i=1}^{q}\frac{|\mu_i|}{\sigma_i} \\ &\approx \frac{1}{q}\sum_{i=1}^{q}\frac{\sum_{iter=1}^{n}|\mu_i^{(iter)}|}{\sum_{iter=1}^{n}\sigma_i^{(iter)}} \\ SNR(\boldsymbol{\eta}) &= \frac{V(E\boldsymbol{\mu})}{E\epsilon^2} = \frac{V(\mathbf{S}\boldsymbol{\eta})}{E\epsilon^2} \\ &\approx \frac{\sum_{iter=1}^{n} V(\mathbf{S}\boldsymbol{\eta}^{(iter)})}{\sum_{iter=1}^{n}\epsilon^{2(iter)}} \end{aligned}$$

where $SNR(\boldsymbol{\beta})$ is the signal-to-noise ratio for activation shape estimation and $SNR(\boldsymbol{\eta})$ is the signal-to-noise ratio for variable selection. And in simulations, we simulated datasets of different combinations: $SNR(\boldsymbol{\beta}) = 8, 5, 2$ and $SNR(\boldsymbol{\eta}) = 8, 5, 2$.

For the posterior computation, we set $\epsilon = 1 \times 10^{-4}$, $\alpha = 0.75$ as PCA percent. We run 6000 iterations with 4000 burn-in and thin 2. For each of the simulation settings, we simulated 50 datasets in total and evaluated the algorithm performance based on some proposed metrics averaged across different datasets. The voxels inside activation regions were selected if their posterior inclusion probability is larger than 0.5. The variable are selected if their posterior inclusion probability is larger than 0.02 for SSVS and 0.01 when used non-sparse prior.

For activation shape estimation and variable selection, as there are only two possible values that voxels can take: "inside the region" or "outside the region", also two

possible values that variables can take: "selected" or "not-selected", we can summarize spatial voxels and variable selection results by their averaged accuracy, sensitivity, and specificity respectively. We also provided the averaged mean-squared-errors (MSE) for $\boldsymbol{\eta}$ and $\boldsymbol{\mu}$. The simulation results using SSVS are presented in Table 4.1 and results using non-sparse prior are presented in Table 4.2

Table 4.1: Different shapes with various signal-to-noise ratios using spike and slab prior

	Sphere			Diamond			Random		
	$SNR(\boldsymbol{\beta}) = 8$			$SNR(\boldsymbol{\beta}) = 5$			$SNR(\boldsymbol{\beta}) = 2$		
	$SNR(\boldsymbol{\eta}) = 8$	$SNR(\boldsymbol{\eta}) = 5$	$SNR(\boldsymbol{\eta}) = 2$	$SNR(\boldsymbol{\eta}) = 8$	$SNR(\boldsymbol{\eta}) = 5$	$SNR(\boldsymbol{\eta}) = 2$	$SNR(\boldsymbol{\eta}) = 8$	$SNR(\boldsymbol{\eta}) = 5$	$SNR(\boldsymbol{\eta}) = 2$
accuracy($\mathbf{z}$)	0.992	0.992	0.991	0.991	0.991	0.991	0.952	0.954	0.952
sensitivity($\mathbf{z}$)	0.972	0.969	0.965	0.949	0.949	0.948	0.950	0.952	0.950
specificity($\mathbf{z}$)	1.000	1.000	1.000	1.000	1.000	1.000	0.948	0.950	0.947
accuracy($\boldsymbol{\gamma}$)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
sensitivity($\boldsymbol{\gamma}$)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
sensitivity($\boldsymbol{\gamma}$)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
MSE($\boldsymbol{\eta}$)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
MSE($\boldsymbol{\mu}$)	0.000	0.000	0.000	0.004	0.003	0.003	0.179	0.162	0.153

Table 4.2: Different shapes with various signal-to-noise ratios using non-sparse prior

	Sphere			Diamond			Random		
	$SNR(\boldsymbol{\beta}) = 8$			$SNR(\boldsymbol{\beta}) = 5$			$SNR(\boldsymbol{\beta}) = 2$		
	$SNR(\boldsymbol{\eta}) = 8$	$SNR(\boldsymbol{\eta}) = 5$	$SNR(\boldsymbol{\eta}) = 2$	$SNR(\boldsymbol{\eta}) = 8$	$SNR(\boldsymbol{\eta}) = 5$	$SNR(\boldsymbol{\eta}) = 2$	$SNR(\boldsymbol{\eta}) = 8$	$SNR(\boldsymbol{\eta}) = 5$	$SNR(\boldsymbol{\eta}) = 2$
accuracy($\mathbf{z}$)	0.991	0.991	0.991	0.993	0.992	0.992	0.954	0.952	0.954
sensitivity($\mathbf{z}$)	0.969	0.969	0.969	0.956	0.950	0.955	0.952	0.950	0.954
specificity($\mathbf{z}$)	1.000	1.000	1.000	1.000	1.000	1.000	0.952	0.951	0.948
accuracy($\boldsymbol{\gamma}$)	0.926	0.909	0.894	0.926	0.907	0.898	0.931	0.923	0.893
sensitivity($\boldsymbol{\gamma}$)	1.000	0.970	0.865	0.990	0.970	0.850	1.000	0.945	0.840
specificity($\boldsymbol{\gamma}$)	0.924	0.907	0.895	0.924	0.906	0.900	0.929	0.923	0.894
MSE($\boldsymbol{\eta}$)	0.028	0.044	0.089	0.028	0.041	0.091	0.025	0.039	0.081
MSE($\boldsymbol{\mu}$)	0.001	0.000	0.000	0.011	0.007	0.003	0.236	0.219	0.141

The simulation studies indicate our proposed method is accurate for voxels classification and variable selection. For simulations using SSVS, even with the worse scenario when $SNR(\boldsymbol{\beta}) = 2$ and $SNR(\boldsymbol{\eta}) = 2$, the averaged accuracy, sensitivity and specificity for voxels classification are all above 0.94 and for variable selection are all 100%. As $SNR(\boldsymbol{\beta})$ increased to 5 and 8, classification performance improves as expected while $SNR(\boldsymbol{\eta})$ increased to 5 and 8, variable selection are all 100% accurate. For MSE of $\boldsymbol{\eta}$ and $\boldsymbol{\mu}$, it decreases in the general trend when $SNR(\boldsymbol{\beta})$ increases.

Compared to the results using SSVS, by applying non-sparse prior for variable selection, the voxels classification is robust but the variable selection generates worse performance. If we compare the scenario when $SNR(\boldsymbol{\beta}) = SNR(\boldsymbol{\eta}) = 2$, the accu-

racy, sensitivity and specificity decrease to 0.893, 0.840, 0.894 and MSE for $\boldsymbol{\eta}$ and $\boldsymbol{\mu}$ increases to 0.081 and 0.141. The proposed method does suffer a decrease performance as expected, but in general the results are acceptable. We recommend applying the fast algorithm when there is exceedingly large number of candidate SNPs in the study for fast computation purpose.

4.4 Real Data Application

We applied the proposed method in an imagine-genetic study to detect any strong associations between SNP genotypes and imaging phenotypes (both imaging intensities and activation shapes) in application to the Alzheimer's disease. To be specific, the primary goal is to determinate any specific gene markers that are correlated with regional activation levels in brains, which can serve as potential indication of disease with different levels of progression.

The data was collected by ADNI. There were three different cohorts: 69 normal cohort (NORM), 117 mild cognitive impairment subjects (MCI) and 49 Alzheimer's disease patients (AD), in total 235 subjects were included in the study. We selected 5 clinical factors for illustration purpose. They were: subjects' age (Subject.Age), gender (Subject.Sex), body weight in kilogram (Subject.WeightKg), neuropsychiatric inventory scores (NPISCORE) and functional activity questionnaire scores (FAQ.Total.Score). There were 2 missing values in NPISCORE and 4 missing values in FAQ.Total.Score, all were imputed by their individual mean values. Then we scaled each of them so that they had an average value of 0 and a variance of 1. For the regions we studied, we included 42 in total. There were 12 regions located in Frontal lobe including (Frontal_Sup_L, *etc*); 8 in Parietal lobe including Parietal_Sup_L, *etc*; 6 in Occipital lobe including Occipital_Sup_L, *etc* and 16 in Temporal lobe including Temporal_Sup_R, Hippocampus_L, *etc* (L: left hemisphere, R: right hemisphere). We

studied each of them at three different time points: baseline (bl), month 6 (m6) and month 12 (m12). As for SNPs, we selected top 614 SNPs for selection based on literature. Combining with cohort indicators, clinical factors as well as SNPs, the final input matrix $\mathbf{S}$ was of dimension 235×621.

We applied the proposed level set image segmentation for activation region fitting and utilizing the non-sparse prior for fast computation purpose. We applied our method to each of the brain anatomical regions and run in parallel. The objective was to learn the brain activation region changes over time and to select significant biomarkers that are related to activation intensities. There were some assumptions in the model in the way we implemented. First, we borrowed the anatomical structure information by assuming separate activation regions (two anatomical regions A and B, $\boldsymbol{\beta}$s are different: $\boldsymbol{\beta}(A) \neq \boldsymbol{\beta}(B)$), independent intensity levels within subject ($\mu_i(A) \neq \mu_i(B)$) and across subjects ($\mu_i(A) \neq \mu_j(B)$), individual set of influential SNPs ($\boldsymbol{\eta}(A) \neq \boldsymbol{\eta}(B)$). Second, we simplified our model by assuming the same level of activation within one anatomical brain region due to the fact that anatomical regions are usually small areas in the brain.

Across all regions in brain, the number of voxels ranges from 335 to 5104, with an average of 2134. We set $\epsilon = 1 \times 10^{-4}$, $\alpha = 0.75$ so the number of basis was 120. Then we run the proposed algorithm for 8000 iterations with 6000 burn-in. We presented all activation regions at brain-wide level presented at the axial, sagittal and coronal panel at the different time points. See Figure C.1, C.2, C.3. We observed that the activations follow the human brain structure symmetry. For each of the anatomical brain regions, we summarized the total number of spatial voxels inside as an indication of activation shape. Furthermore, by comparing the number of voxels across different time points, we observed brain activation shape changes over time. In general, compared month 6 to baseline, 20 regions were stable, 12 regions were enlarged but 10 shrank. Compared month 12 to month 6, 24 were stable, 8 were

enlarged and 10 shrank. See the changes of Hippocampus from the right hemisphere at month 6 and middle temporal gyrus from the left hemisphere at month 12 (see Figure 4.3).

For variable selection, there were 3, 2, 1 out of 42 regions at baseline, month 6, month 12 that the largest number of SNPs (4) were selected; 37, 36, 37 regions out of 42 at baseline, month 6, month 12 that at least one of the SNPs was selected . If we pooled the SNPs selected at the same time point together. From the Venn diagram (Figure 4.4), we observed that most of them stayed unchanged, which means most of them keep imposing a consistent impact on the brain activation shapes. However, we also observed that very few of them only functions at specific time points. For the SNPs only selected at baseline, they belong to the genes NEDD9, DAPK1, SORCS1, ADAM10 and for the SNPs only selected at month 6, they belong to one gene ADAM10. It is consistent with the results of Colciaghi et al. (2004); Gatta et al. (2002), where ADAM10 has shown some alterations in the early stages of Alzheimer's disease. Lastly but not least, there were no SNPs selected only at month 12 based on our study.

For better compare the changes of frequencies of SNPs selected from the perspective of genes, at different time point, we pooled all the selected SNPs together: SNPs were retained if they are correlated with from least one activation. Then we summarized the counts of these selected SNPs per gene. See Figure 4.5. We observed that the top four genes with the largest number of SNPs selected were: SORCS1, ADAM10, DAPK1, and NEDD9, which is consistent with Saykin et al. (1991) where they were significantly associated with the hippocampal volume or grey matter density changes after accounting for APOE. In general, the frequency pattern is similar at all time points with only a slight change. We observed that gene ACE, IL33, SORL1 only function at baseline and month 12, EXOC3L2 functions at baseline and month 6. From the literature, SORL1 and ACE were associated with the risk for late-onset

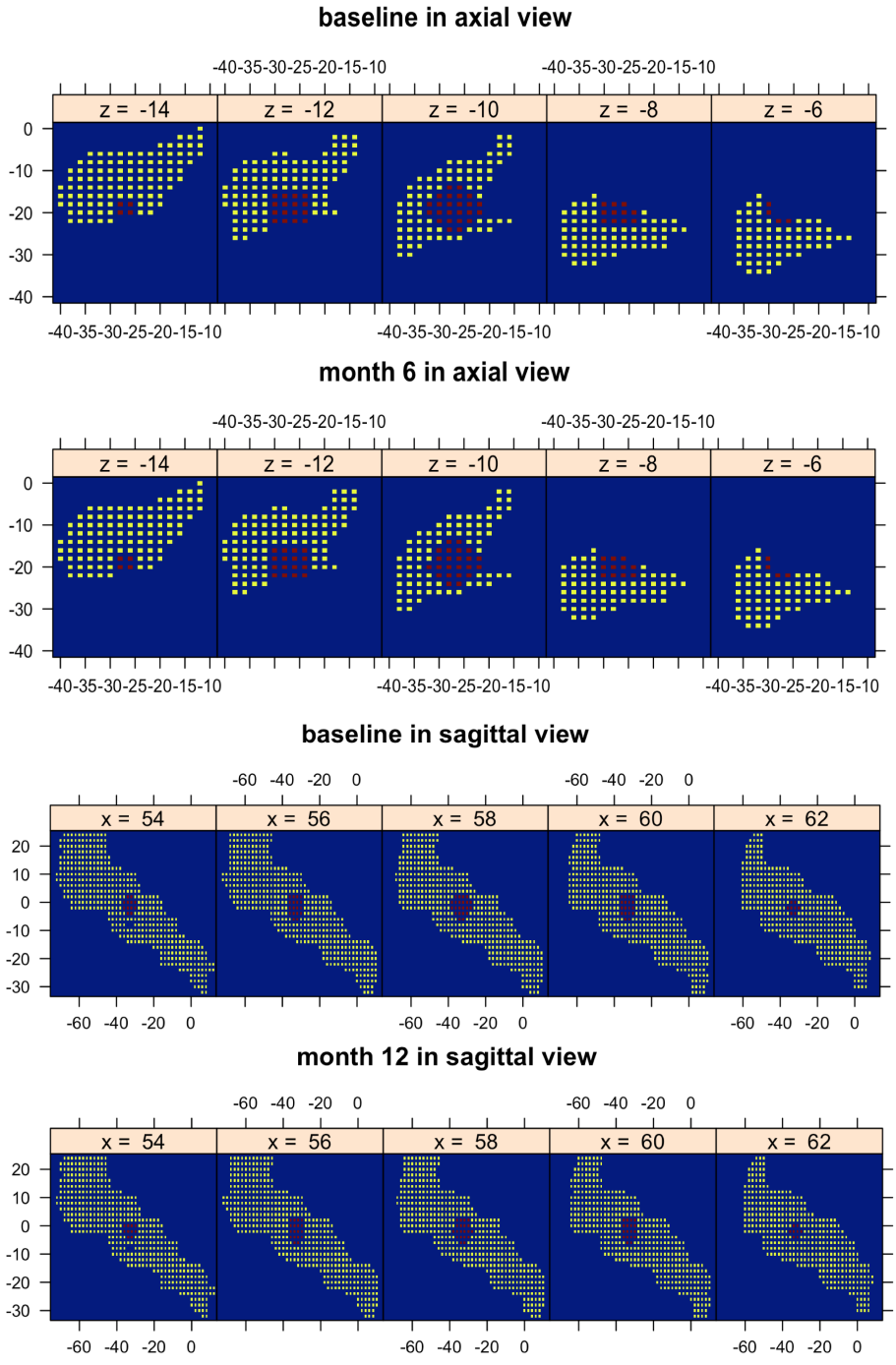


Figure 4.3: Changes of brian activation shapes. Top/bottom: the Hippocampus from the right hemisphere at month 6 from axial panel/ the middle temporal gyrus from the left hemisphere at month 12 at saggital paenl. Points: yellow, anatomical brain regions; red, activation regions

Alzheimer's disease (ThorntonWells et al., 2008; Lee et al., 2008; Edwards et al., 2009; Ning et al., 2010; Patel et al., 2011), IL33 was associated with pathological and cognitive decline (Chapuis et al., 2009; Fu et al., 2016), which might be associated with the late-onset of the Alzheimer's disease, but there still need additional affirmative investigation.

Table 4.3: Example SNPs with their gene, activation, lobes information.

SNP name	#regions	Gene	Lobes	Total regions over time
rs12209631	9	NEDD9	temporal lobe	Hippocampus_L, Hippocampus_R, Occipital_Inf_L
rs7095427	9	SORCS1	frontal lobe, temporal lobe	Frontal_Mid_L, Frontal_Mid_R, Temporal_Mid_L
rs2756271	7	PRNP	frontal lobe, occipital lobe, parietal lobe	Frontal_Sup_L, Precuneus_L, Temporal_Inf_R
rs1473180	6	DAPK1	parietal lobe, temporal lobe	Occipital_Mid_L, Precuneus_R
rs11193130	6	SORCS1	frontal lobe, temporal lobe	Frontal_Sup_Medial_R, Occipital_Mid_R
rs950809	4	SORCS1	frontal lobe, parietal lobe, temporal lobe	Cingulum_Ant_R, Occipital_Sup_L, Parietal_Inf_R
rs677066	3	CR1	frontal lobe	Frontal_Sup_R
rs2276575	3	BIN1	frontal lobe	Rectus_R
rs1427282	3	ADAM10	parietal lobe	Parietal_Inf_L
rs4353	2	ACE	parietal lobe	Parietal_Inf_R
rs10422797	2	EXOC3L2	temporal lobe	Temporal_Mid_R

We also ranked all SNPs based on their summed inclusion probability across all time points at different anatomical regions. Example SNPs are listed in Table 4.3 (complete information was presented in Appendix Table C.2). From Table 4.3, we observed that the temporal lobe is the most correlated lobes with Alzheimer's disease, where its functions mainly include sensory processing, visual memory, language and emotion comprehension (Smith and Kosslyn, 2013). Some of the SNPs selected based on our methods were consistent with previous studies while some of them were not. Specifically speaking, example SNPs that were also selected based on other methods: rs11193130 was selected from Kramer et al. (2012); Reitz et al. (2013), rs677066 was selected from Silver et al. (2012); McElroy (2013), rs7095427, rs10422797 were selected from Zhu et al. (2014a), rs12209631, rs1473180 were selected from Nathoo (2016), *etc.* However, some of the SNPs were not directly selected from others, but their functionalities can be inferred based on previous work. SNP rs2756271 was located in the promoter region of gene PRNP, where PRNP was the causative agent for transmissible spongiform encephalopathies (Moe Lee et al., 2012; Nathoo, 2016). The

encephalopathies contained a group of neurodegenerative disease where Alzheimer's disease was included. SNP rs950809 that was linked to gene SORCS1, was associated with the memory savings score (Reitz, Lee, Rogers and Mayeux, 2011). SNP rs2276575 (gene BIN1) was associated with the measure of cognitive aging (Hamilton et al., 2011). Besides SNPs, we also summarized the selection results based on anatomical regions in Table C.1 in the Appendix. For each anatomical region, we presented their number of voxels located inside the activation regions, the number of SNPs we selected, the genes related at each time point, and also the genes that are different (selected or not selected) compared between any two-time points for studying the changes of genetic functionality.

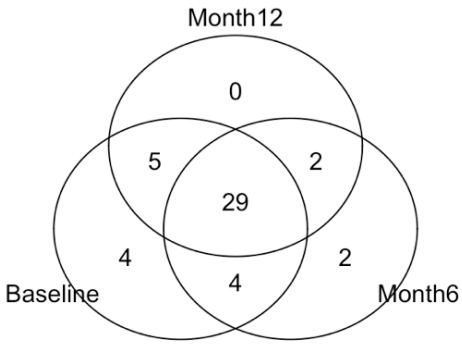


Figure 4.4: Venn diagram presenting how different total SNPs selected at each time point.

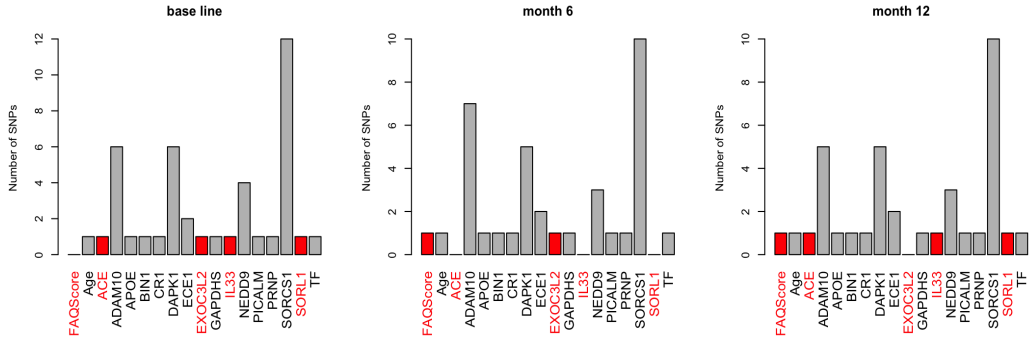


Figure 4.5: Pooled results of top 20 SNPs selected for each region at gene level. Red bars are genes that present a selected/ not selected pattern with some time points

4.5 Conclusion and Discussion

We have developed a novel Bayesian hierarchical model in imaging genetics studies for simultaneous activation shape estimation and variable selection. We applied to an ADNI dataset as real data application.

Our approach can jointly estimate the brain activation regions after accounting for external sources of clinical factors and genetic variation where currently there is no literature based on our knowledge share the same focus with us. Besides, our approach can detect important genetic and demographic factors associated with activation intensities inside activation regions. We also borrow the anatomical brain segmentation as prior information.

However, our method does suffer from some limitations. First, the assumptions that all averaged intensities inside are shared across all activation regions as long as they are anatomically the same are very strong. Mathematically speaking, the μ_i can be further extended to an activation-region-specific variable: $\mu_{i,r}$ where r can be pre-specified by some spatial clustering methods implemented as initial values. Second, due to computation, proposed method should be improved and optimized so that it can be scalable to thousands of SNPs which can be comparable to the popular GWAS studies.

Appendix A

Appendix for Chapter 2

Swendsen-Wang Suppose $\mathbf{W} = \{W_{ij}, i \sim j\}$ where the W_{ij} is defined only when gene pair i and j are connected. The distribution of W_{ij} is

$$P(W_{ij}|z_i, z_j) = \exp(-\rho_{z_i}\omega_j C_{ij} I[z_i = z_j]) \times I[0 \leq W_{ij} \leq \exp(\rho_{z_i}\omega_j C_{ij} I[z_i = z_j])]$$

Then the conditional distribution of $\mathbf{W}$ given $\mathbf{z}$ is:

$$P(\mathbf{W}|\mathbf{z}) \propto \exp\left(\sum_{i=1}\sum_{j\neq i} -\rho_{z_i}\omega_j c_{ij} I[z_i = z_j]\right) \prod_{i=1}\prod_{j\neq i} I[0 \leq W_{ij} \leq \exp(\rho_{z_i}\omega_j c_{ij} I[z_i = z_j])]$$

The full conditional distribution for $\mathbf{z}$ given $\mathbf{W}$ is:

$$P(\mathbf{z}|\mathbf{W}, \mathbf{r}, \widetilde{\boldsymbol{\theta}}) \propto P(\mathbf{W}|\mathbf{z}) P(\mathbf{r}|\mathbf{z}, \widetilde{\boldsymbol{\theta}}) P(\mathbf{z}) \propto P(\mathbf{r}|\mathbf{z}, \widetilde{\boldsymbol{\theta}}) \exp\left[\sum_{i=1}^{n} (\widetilde{\omega}_i \log(\pi_{z_i}))\right] \qquad \text{(A.1)}$$

DPM Density Updating Consider gene i with class label k and all the other genes with the same class label, if we integrate over $\mathbf{q}_k$, then the cluster index g_i has the following distribution:

$$\begin{aligned} P(g_i = g|g_1, g_2, \cdots, g_{i-1}) &= \frac{P(g_1, g_2, \cdots, g_{i-1}, g_i = g)}{P(g_1, g_2, \cdots, g_{i-1})} \\ &= \frac{\int_{(g_1, g_2, \cdots, g)} \Gamma(\tau_k)\Gamma(\tau_k/L_k)^{-L_k} g_1^{(\tau_k/L_k)-1} \cdots g_{L_k}^{(\tau_k/L_k)-1} dg_1 g_2 \cdots g_{L_k}}{\int_{(g_1, g_2, \cdots, g_{i-1})} \Gamma(\tau_k)\Gamma(\tau_k/L_k)^{-L_k} g_1^{(\tau_k/L_k)-1} \cdots g_{L_k}^{(\tau_k/L_k)-1} dg_1 g_2 \cdots g_{L_k}} \\ &= \frac{n_{i,g} + \tau_k/L_k}{i - 1 + \tau_k} \end{aligned}$$

where $n_{i,g} = \sum_{j=1}^{i-1} I[g_j = g]$ denotes the count of $g_j, j < i$ such that $g_j = g$.

Then let $L_k \to \infty$:

$$P(g_i = g, |g_1, g_2, \cdots, g_{i-1} \& g \in (g_1, \cdots, g_{i-1})) \to \frac{n_{i,g}}{i - 1 + \tau_k}$$
$$P(g_i = g, |g_1, g_2, \cdots, g_{i-1} \& g \notin (g_1, \cdots, g_{i-1})) \to \frac{\tau_k}{i - 1 + \tau_k} \tag{A.2}$$

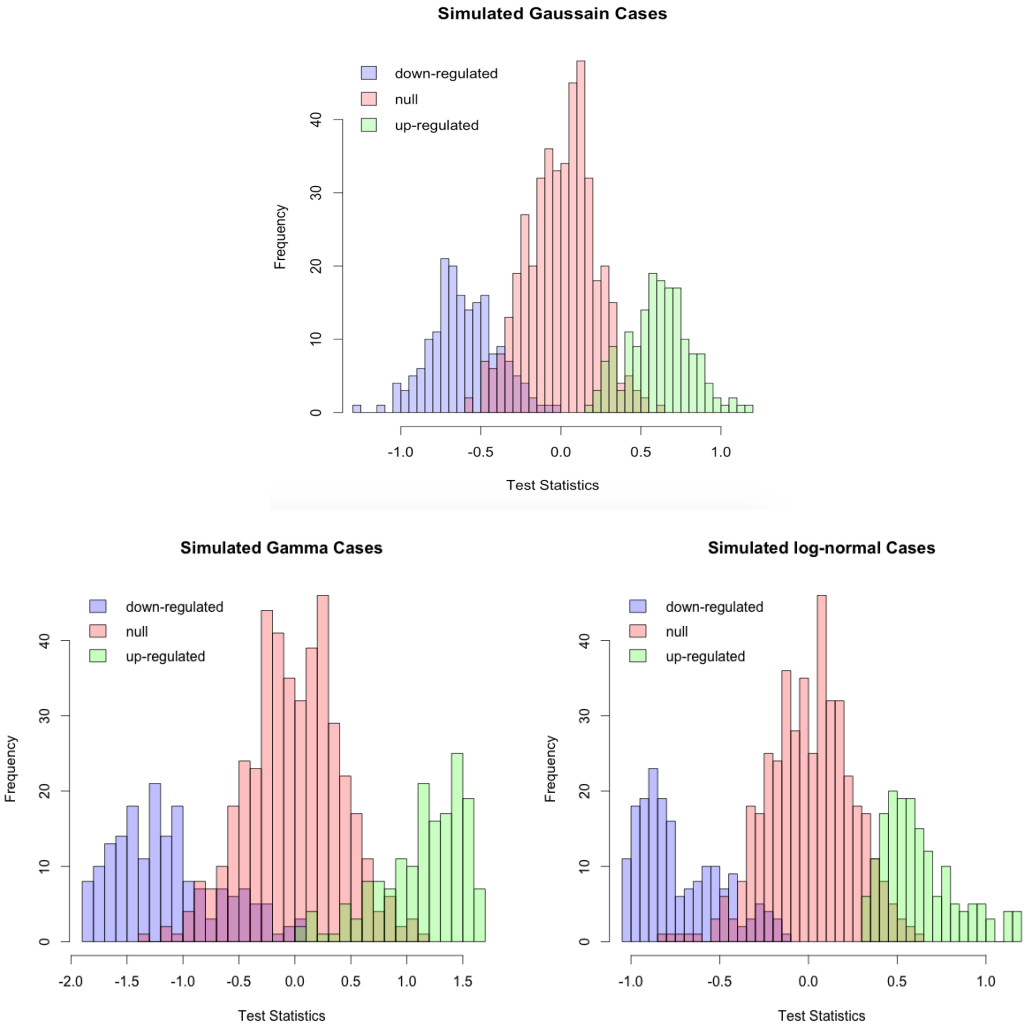


Figure A.1: An illustration of selected simulated datasets for the distributions of test statistics under each simulation setting.

Algorithm 1 Function: fully Bayesian posterior updating algorithm

Input observed test statistics $\mathbf{r} = (\mathbf{r}_{obs}, \mathbf{r}_{mis})$, adjacency matrix $\mathbf{C} = \{c_{ij}\}$, $\boldsymbol{\tau}$, $\boldsymbol{w}$, $\boldsymbol{\pi}$=NULL, $\boldsymbol{\rho}$=NULL, $\boldsymbol{\rho}_0$, $\mathbf{r}_0$, $\mathbf{z}$, PriorNullDensity=NULL, PriorForDPMDensityFitting, ParaForMCMC, rhoSD, rhoUpperBound, rhoLowerBound, piSD, piUpperBound, piLowerBound, MissingDataImputationMethod, TotalNumIterationsForDMH, nSaveForDMH, TotalNumIterations, nSave
Initialization:
if (is.null(PriorNullDensity)) **then**
PriorNullDensity ← BiGaussianDensityByCentralFitting($\mathbf{r}_{obs}$)
$(\mathbf{z}, \mathbf{g}, \widetilde{\boldsymbol{\theta}}, \mathbf{L})$ ← KL-HODC($\mathbf{r}_{obs}$, PriorForDPMDensityFitting, ParaForMCMC)
if (is.null($\boldsymbol{\pi}$) | is.null($\boldsymbol{\rho}$)) **then**
$(\boldsymbol{\pi}, \boldsymbol{\rho})$ ← DMH($\mathbf{C}$, $\mathbf{r}_{obs}$, $\boldsymbol{\rho}_0$, $\mathbf{r}_0$, $\mathbf{z}$, rhoSD, rhoUpperBound, rhoLowerBound, piSD, piUpperBound, piLowerBound, TotalNumIterationsForDMH, nSaveForDMH)
$\mathbf{r}_{mis}$ ← Mean($\mathbf{r}_{obs}$)
Loop:
zTrace ← $\mathbf{z}$
Iter ← 0
while (Iter < TotalNumIterations) **do**
$\mathbf{z}$ ← SW($\mathbf{C}, \mathbf{z}, \mathbf{r}_{obs}, \widetilde{\boldsymbol{\theta}}, \boldsymbol{\rho}, \boldsymbol{\pi}$)
$(\widetilde{\boldsymbol{\theta}}, \mathbf{g})$ ← DPMDensityFitting($\mathbf{C}, \mathbf{z}, \mathbf{r}$, PriorForDPMDensityFitting, ParaForMCMC)
$\mathbf{r}_{mis}$ ← MissingDataImputation(MissingDataImputationMethod, $\mathbf{C}, \mathbf{r}, \mathbf{g}, \widetilde{\boldsymbol{\theta}}$)
zTrace ← cbind(zTrace, $\mathbf{z}$)
Iter ← Iter+1
ClassIndicators ← ClassIndicatorsWithLocalFDRControl(zTrace, nSave)
return ClassIndicators

Algorithm 2 Function: prior null density fitted as bi-Gaussian density

function BiGaussianDensityByCentralFitting($\mathbf{r}$, QuantileForFitting=NULL)
if is.null(QuantileForFitting) **then**
QuantileForFitting ← $c(0.25, 0.75)$
CentralTestStat ← $\mathbf{r}$[which($\mathbf{r} \in$ QuantileForFitting)]
CutOff ← quantile($\mathbf{r}$,0.5)
NormalFitForUpRegulateClass ← NormalDensityFitting(CentralTestStat>CutoffWithItsReflected)
NormalFitForDownRegulateClass ← NormalDensityFitting(CentralTestStat<CutoffWithItsReflected) **return** CutOff, NormalDensityForUpRegulateClass, NormalDensityForDownRegulateClass

Algorithm 3 Function: initial values based on KL-HODC

function KL-HODC(**r**, PriorForDPMDensityFitting, ParaForMCMC, PriorNullDensity)
  $(\mathbf{g}, \widetilde{\boldsymbol{\theta}}) \leftarrow$ DPdensity(**r**, PriorForDPMDensityFitting, ParaForMCMC)
  $(\mathbf{g}, \widetilde{\boldsymbol{\theta}}) \leftarrow$ SortClusterByMeanLocation(**g**, $\widetilde{\boldsymbol{\theta}}$)
  procedure (initialize null class index)
    $D_{min} \leftarrow +\infty$
    NullClassIndex $\leftarrow \emptyset$
    DownRegulateClassIndex $\leftarrow \emptyset$
    UpRegulateClassIndex $\leftarrow \emptyset$
    for all $l_0 \in s$ **do**
      CandidateNullDensity $\leftarrow \{\widetilde{\theta}_{l_0}\}$
      D $\leftarrow$ KLDistance(CandidateNullDensity, PriorNullDensity)
      if $D < D_{min}$ **then**
        $D_{min} \leftarrow D$
        NullClassIndex $\leftarrow \{l_0\}$
        DownRegulateClassIndex $\leftarrow \{l'\}_{\forall l', 1 \le l' < l_0}$
        UpRegulateClassIndex $\leftarrow \{l'\}_{\forall l', l' > l_0}$;
  procedure (merge multiple clusters to null class index)
    $D_{\text{diff}} \leftarrow +\infty$
    while $D_{\text{diff}} > 0$ & DownRegulateClassIndex $\neq \emptyset$ & UpRegulateClassIndex $\neq \emptyset$ **do**
      CandidateNullClass $\leftarrow$ NullClassIndex $\cup \{l_0 + 1\}$
      CandidateNullDensity $\leftarrow$ CandidateNullDensity $\cup \{\widetilde{\theta}_{l_0+1}\}$
      $D_+ \leftarrow$ KLDistance(CandidateNullDensity, PriorNullDensity)
      CandidateNullClass $\leftarrow$ NullClassIndex $\cup \{l_0 - 1\}$
      CandidateNullDensity $\leftarrow$ CandidateNullDensity $\cup \{\widetilde{\theta}_{l_0+1}\}$
      $D_- \leftarrow$ KLDistance(CandidateNullDensity, PriorNullDensity)
      if $D_- \le D_+$ **then**
        NullClassIndex $\leftarrow$ NullClassIndex $\cup \{l_0 - 1\}$
        DownRegulateClassIndex $\leftarrow$ DownRegulateClassIndex$\{l'\}_{\forall l', 1 \le l' < (l_0 - 1)}$
        $D_{\text{diff}} \leftarrow D_{min} - D_-$
        $D_{min} = D_-$
      else
        NullClassIndex $\leftarrow$ NullClassIndex $\cup \{l_0 + 1\}$
        UpRegulateClassIndex $\leftarrow$ UpRegulateClassIndex$\{l'\}_{\forall l', l' > (l_0 +)}$;
        $D_{\text{diff}} \leftarrow D_{min} - D_+$
        $D_{min} = D_+$
  z $\leftarrow$ **z** $= (z_1, \cdots, z_n), \forall i \in$ NullClassIndex, $z_i = 0, \forall i \in$ DownRegulateClassIndex, $z_i = -1, \forall i \in$ UpRegulateClassIndex, $z_i = +1$
  g $\leftarrow$ **z**
  $\widetilde{\boldsymbol{\theta}} \leftarrow \widetilde{\boldsymbol{\theta}} = \{\widetilde{\theta}_{g_i}\}$
  L $\leftarrow$ **c**(|DownRegulateClassIndex|, |NullClassIndex|, |UpRegulateClassIndex|)
  return z, **g**, $\widetilde{\boldsymbol{\theta}}$, **L**

Algorithm 4 Function: hyperparameters by double Metropolis-Hasting

function DMH(Network, TestStat, $\boldsymbol{\rho}$, $\mathbf{r}$, $\mathbf{z}$, rhoSD, rhoUpperBound, rhoLowerBound, piSD, piUpperBound, piLowerBound, TotalNumIterations, nSave)
  rhoTrace $\leftarrow \boldsymbol{\rho}$
  piTrace $\leftarrow \mathbf{r}$
  Iter $\leftarrow 0$
  for (Iter< TotalNumIterations) **do**
    repeat
      $\boldsymbol{\rho}' = (\boldsymbol{\rho}'_1, \boldsymbol{\rho}'_2, \boldsymbol{\rho}'_3, \boldsymbol{\rho}'_4) \leftarrow$ rtruncnorm(1, $\boldsymbol{\rho}$, rhoSD, rhoLowerBound, rhoUpperBound)
      $\boldsymbol{\pi}' = (\boldsymbol{\pi}'_1, \boldsymbol{\pi}'_2, \boldsymbol{\pi}'_3) \leftarrow$ rtruncnorm(1, $\boldsymbol{\pi}$, rhoSD, rhoLowerBound, rhoUpperBound)
      $\boldsymbol{\pi}'_2 \leftarrow 1 - \boldsymbol{\pi}'_1 - \boldsymbol{\pi}'_3$
    until $\boldsymbol{\rho}'_1 > \boldsymbol{\rho}'_2$ & $\boldsymbol{\rho}'_3 > \boldsymbol{\rho}'_2$ & $\boldsymbol{\pi}'_2 > 0.5$
    $\mathbf{z}' \leftarrow$ DrawSampleFromPriorModel(Network, TestStat, $\boldsymbol{\rho}', \boldsymbol{\pi}'$)
    LogAcceptRate $\leftarrow$ LogDataLikelihood(Network, TestStat, $\mathbf{z}', \boldsymbol{\rho}, \boldsymbol{\pi}$) + LogDataLikelihood(Network, TestStat, $\mathbf{z}, \boldsymbol{\rho}', \boldsymbol{\pi}'$) -LogDataLikelihood(Network, TestStat, $\mathbf{z}, \boldsymbol{\rho}, \boldsymbol{\pi}$) - LogDataLikelihood(Network, TestStat, $\mathbf{z}', \boldsymbol{\rho}', \boldsymbol{\pi}'$)
    if (log(runif(1))< LogAcceptRate) **then**
      $\boldsymbol{\rho} \leftarrow \boldsymbol{\rho}'$
      $\boldsymbol{\pi} \leftarrow \boldsymbol{\pi}'$
      $\mathbf{z} \leftarrow \mathbf{z}'$
      rhoTrace $\leftarrow$ cbind(rhoTrace, $\boldsymbol{\rho}$)
      piTrace $\leftarrow$ cbind(piTrace, $\mathbf{r}$)
  $\boldsymbol{\rho} \leftarrow$ rowMeans(rhoTrace[, nSave])
  $\boldsymbol{\pi} \leftarrow$ rowMeans(piTrace[, nSave]) **return** $\boldsymbol{\pi}$

Algorithm 5 Function: updating $\mathbf{z}|\widetilde{\boldsymbol{\theta}}$ by Swendsen-Wang

function SW(Network, $\mathbf{z}, \mathbf{r}, \widetilde{\boldsymbol{\theta}}, \boldsymbol{\rho}, \boldsymbol{\pi}$)
  $G =< V, E > \leftarrow$ as.GraphObject(Network)
  procedure (graph clustering)
    $G \leftarrow G_{-1} \cup G_0 \cup G_1$; where $\forall$ node $i \in G_k =< V_k, E_k >$, $z_i = k$
    for $l \leftarrow \{-1, 0, 1\}$ **do**
      for all $e \in E_l$ **do**
        $W_e \leftarrow$ runif(1, 0, exp(ρ_{z_l}))
        if $(W_e < 1)$ **then** $e \leftarrow$ NULL
      $G_l \leftarrow \cup_{s=1}^{n_l} G_{ls}$, $G_{ls} =< V_{ls}, E_{ls} >$
    $G \leftarrow \cup_{l=-1}^{1} \cup_{s=1}^{n_l} G_{ls}$, $G_{ls} =< V_{ls}, E_{ls} >$
  procedure (graph relabling)
    for all $G_{cluster} =< V_{cluster}, E_{cluster} > \in \{G_{ls} =< V_{ls}, E_{ls} >, l = -1, 0, 1, s = 1, 2, \cdots, n_l\}$ **do**
      $\mathbf{z}'_{i \in G_{cluster}} \leftarrow$ SampleFromPosteriorDistributionOfZ($\mathbf{r}, \mathbf{z}, \widetilde{\boldsymbol{\theta}}$)
    return $\mathbf{z}$

Algorithm 6 Function: update $\widetilde{\boldsymbol{\theta}}|\mathbf{z}$ via DPM fitting

function DPMDensityFitting(Network, $\mathbf{z}, \mathbf{r}$, PriorForDPMDensityFitting, ParaForMCMC)
  for z in $\{-1, 0, 1\}$ **do**
    Nodes $\leftarrow \{i\}_{\forall i, z_i = z}$
    DPMFit $\leftarrow$ DPDensityFitting($\{r_i\}_{i \in Nodes}$, PriorForDPMDensityFitting, ParaForMCMC)
    DPMFitSort $\leftarrow$ DPMFitClusterSortByMeanLocation(DPMFit)
    $\widetilde{\boldsymbol{\theta}}_z \leftarrow$ DPMFitSort.Para
    $\{g_i\}_{\forall i, i \in \text{Nodes}} \leftarrow$ DPMFitSort.ClusterIndex
  return $\widetilde{\boldsymbol{\theta}}, \mathbf{g}$

Algorithm 7 Function: missing data imputation algorithm

function MissingDataImputation(MissingDataImputationMethod=c('BayesianPosteriorSampling', 'NearestNeighborImpute'), Network, $\mathbf{r}, \mathbf{g}, \widetilde{\boldsymbol{\theta}}$)

  if (MissingDataImputationMethod=='BayesianPosteriorSampling') **then**

    for *loc* in $\{i\}_{\forall i, r_i \in \mathbf{r}_{mis}}$ **do**

      $r_{loc} \leftarrow \text{rnorm}(\widetilde{\theta}_{g_{loc}})$

  if (MissingDataImputationMethod=='NearestNeighborImpute') **then**

    for *loc* in $\{i\}_{\forall i, r_i \in \mathbf{r}_{mis}}$ **do**

      Nbrs $\leftarrow$ ExtractNeighborsFromNetwork(Network)

      $r_{loc} \leftarrow \frac{1}{|\text{Nbrs}|} \sum_{k=1}^{|\text{Nbrs}|} r_k$

  return $\mathbf{r}_{mis}$

Appendix B

Appendix for Chapter 3

We subsampled at different sizes: 40, 80, 160, 200 from the total 415 samples in the CHD dataset. For each of the subdatasets, we generated data matrices with missing values at a missing rate ranging from 1% up to 40% by knocking out the missing locations following the real-data missing pattern. The algorithms for comparison included: KNN, BPCA, SLR, SVD and the proposed method Net_SVR. The SVI (Min/2, Mean and Median) methods were not considered as their overall performance compared to these five was relatively worse based on previous simulations. The results are shown in Supplementary Figure B.1. Compared to KNN, SLR and SVD had an NRMSE Ratio larger than 1 in most of the cases when the missing rate was larger, while BPCA and Net_SVR had NRMSE Ratios below 1 most of the time except that BPCA performed slightly worse than KNN when sample size was larger than 160 and the missing rate was 10% or 15%. Based on the simulation results, algorithm performances kept consistent at sample sizes as low as 40, and Net_SVR outperformed others in most cases.

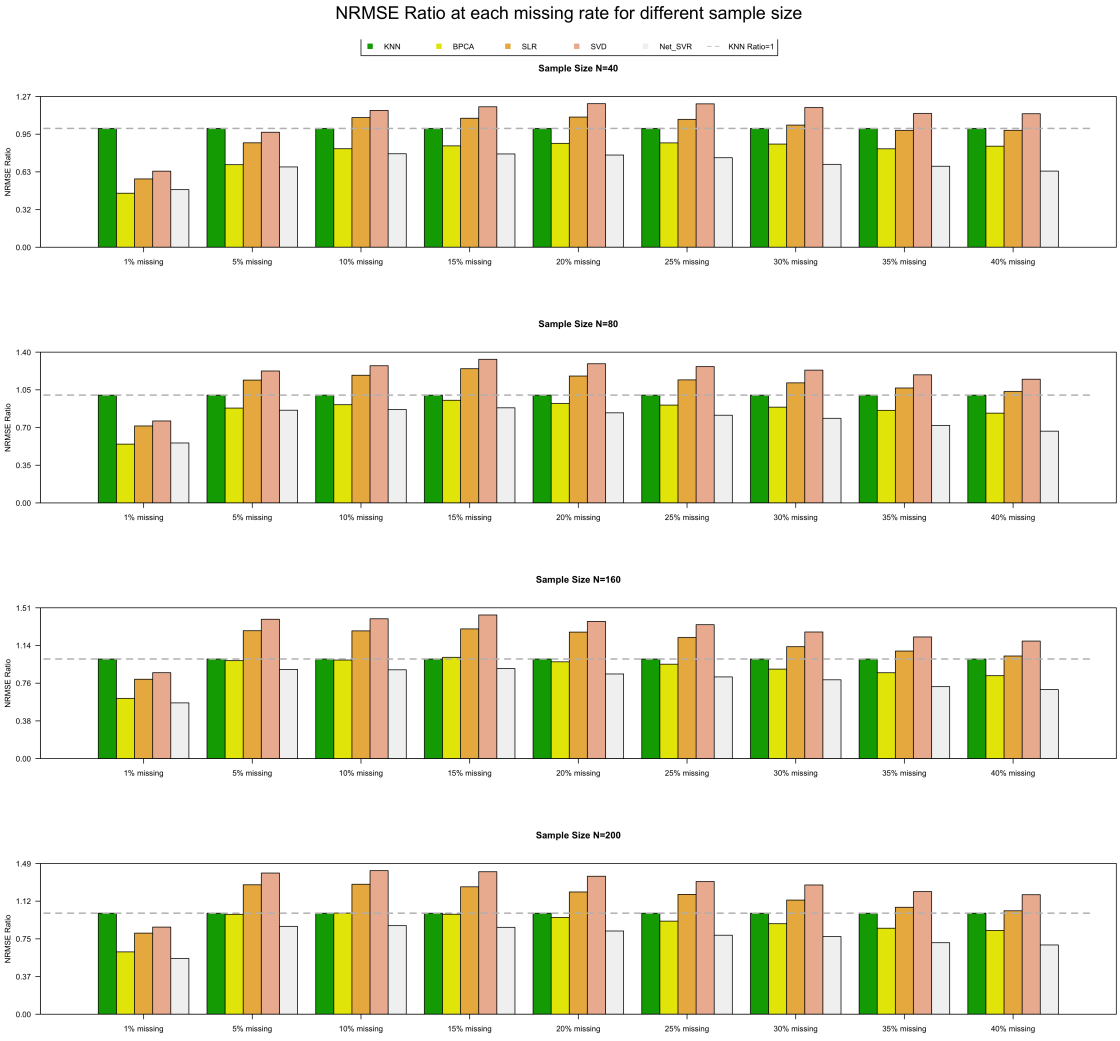


Figure B.1: Simulation results when varying sample size of the CHD dataset

Algorithm 8 Function: create the feature-level predictor network

Input data matrix $E_{m\times n}$; metabolite network G; adduct info-matrix $A = (A_{mz}, A_{rt})$; reference ions names I; tolerance level $tol.mz$, $tol.rt$; number of neighbors: n_1, n_2, n_3;
procedure CREATE THE FEATURE-LEVEL PREDICTOR NETWORK
    neighbors $N = list()$
    for feature i in $1 : m$ **do**
        nbrs.net=$\{j : i \sim j$ in $G\}$
        nbrs.ion=$\{j : \exists p, q \in I, s.t$

$$\frac{||A_{mz}[i] - A_{mz}[j]| - |A_{mz}[p] - A_{mz}[q]||}{|A_{mz}[p] - A_{mz}[q]|} \leq tol.mz$$

      and $||A_{rt}[i] - A_{rt}[j]| - |A_{rt}[p] - A_{rt}[q]|| \leq tol.rt\}$
        nbrs.corr=c()
          nbrs.corr1={ n_1 largest linear-correlated features with $i\}$
          nbrs.corr2={ n_2 largest DCOL-correlated features with $i\}$
          nbrs.corr3={ n_3 largest dCov-correlated features with $i\}$
        nbrs.corr=nbrs.corr1 $\cup$ nbrs.corr2 $\cup$ nbrs.corr3
        $N[[i]]$=nbrs.net $\cup$ nbrs.ion $\cup$ nbrs.corr
Return N

Algorithm 9 Function: rank features by averaged neighbood missigness

Input data matrix E; predictor network denoted as neighbors list N
procedure RANK FEATURES BY AVERAGED NEIGHBOOD MISSIGNESS
    impseq=c()
    avemiss=c()
    E.nmiss=apply(E,1,function(e){sum(is.na(e))})
    for feature i in $1 : m$ **do**
        nbrs.i=N[[i]]
        avemiss[i]=mean(E.nmiss[nbrs.i])
    impseq=rank $(1 : m)$ by avemiss
Return impseq

Algorithm 10 Function: MINMA imputation (Net_SVR)

Input data matrix E; metabolite network G; adduct info-matrix $A = (A_{mz}, A_{rt})$; reference ions names I; tolerance level $tol.mz$, $tol.rt$; number of neighbors: n_1, n_2, n_3;
procedure BUILD PREDICTOR NETWORK
    N=BUILD_NET$(E, G, A, I, tol.mz, tol.rt, n1, n2, n3)$
procedure CREAT AN IMPUTATION SEQUENCE
    $impseq$=IMP_SEQ(E, N)
procedure IMPUTATION
    Initialize $\hat{E} = E$
    for feature i in $impseq$ **do**
        create $e_i = E[i,]$, $e_{i,obs}, e_{i,mis}$
        extract neighbor locations from $N[[i]]$ as $nbr(i)$
        train a SVR model $e_{i,obs} \sim e_{nbr(i),obs}$
        predict $e_{i,mis}$ as $\hat{e}_{i,mis}$ using $e_{nbr(i),mis}$
        set $\hat{E}[i, mis] = \hat{e}_{i,mis}$
Return $\hat{E}$

Appendix C

Appendix for Chapter 4

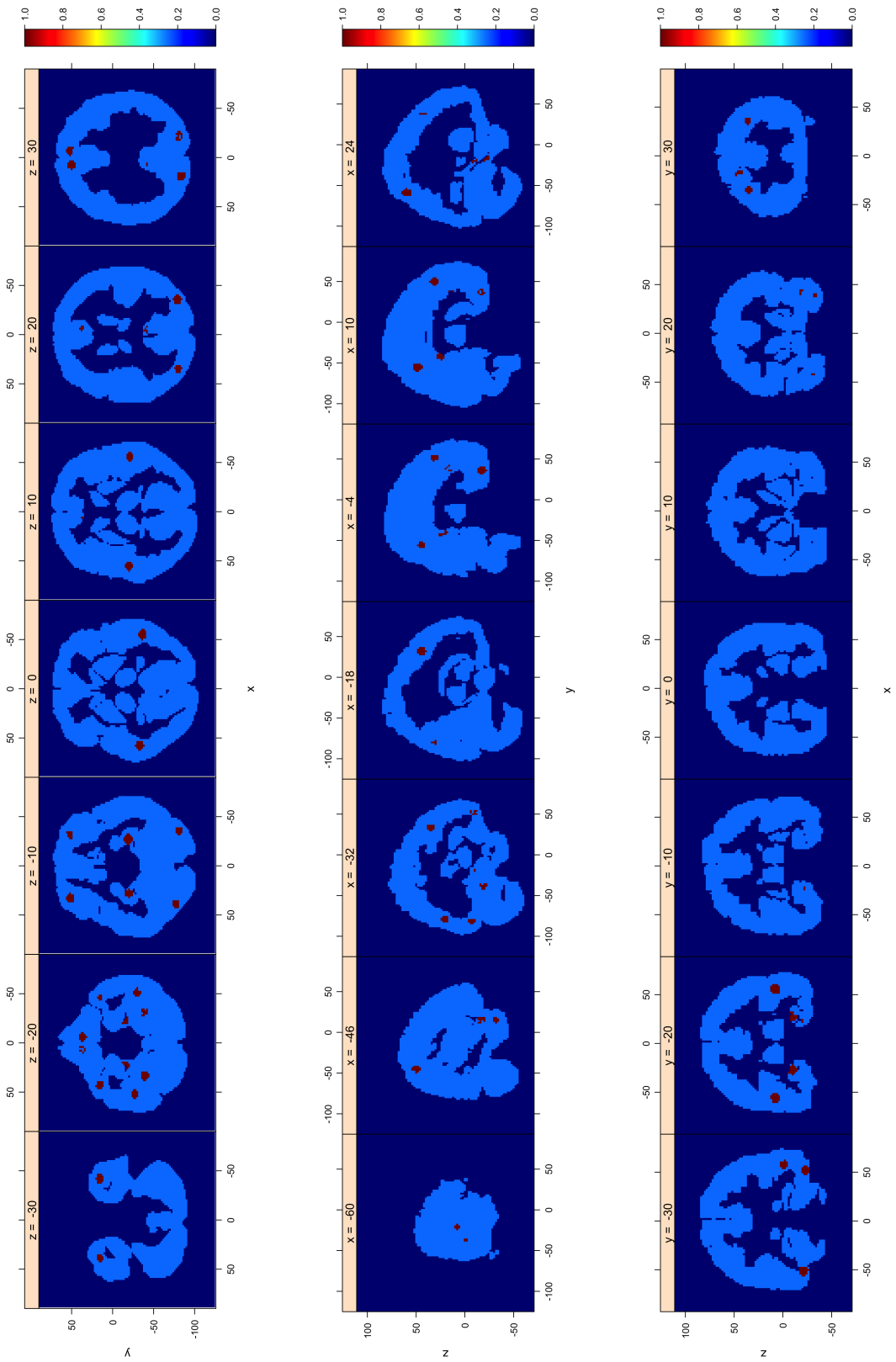


Figure C.1: Different views of brain-wide activation regions at baseline.

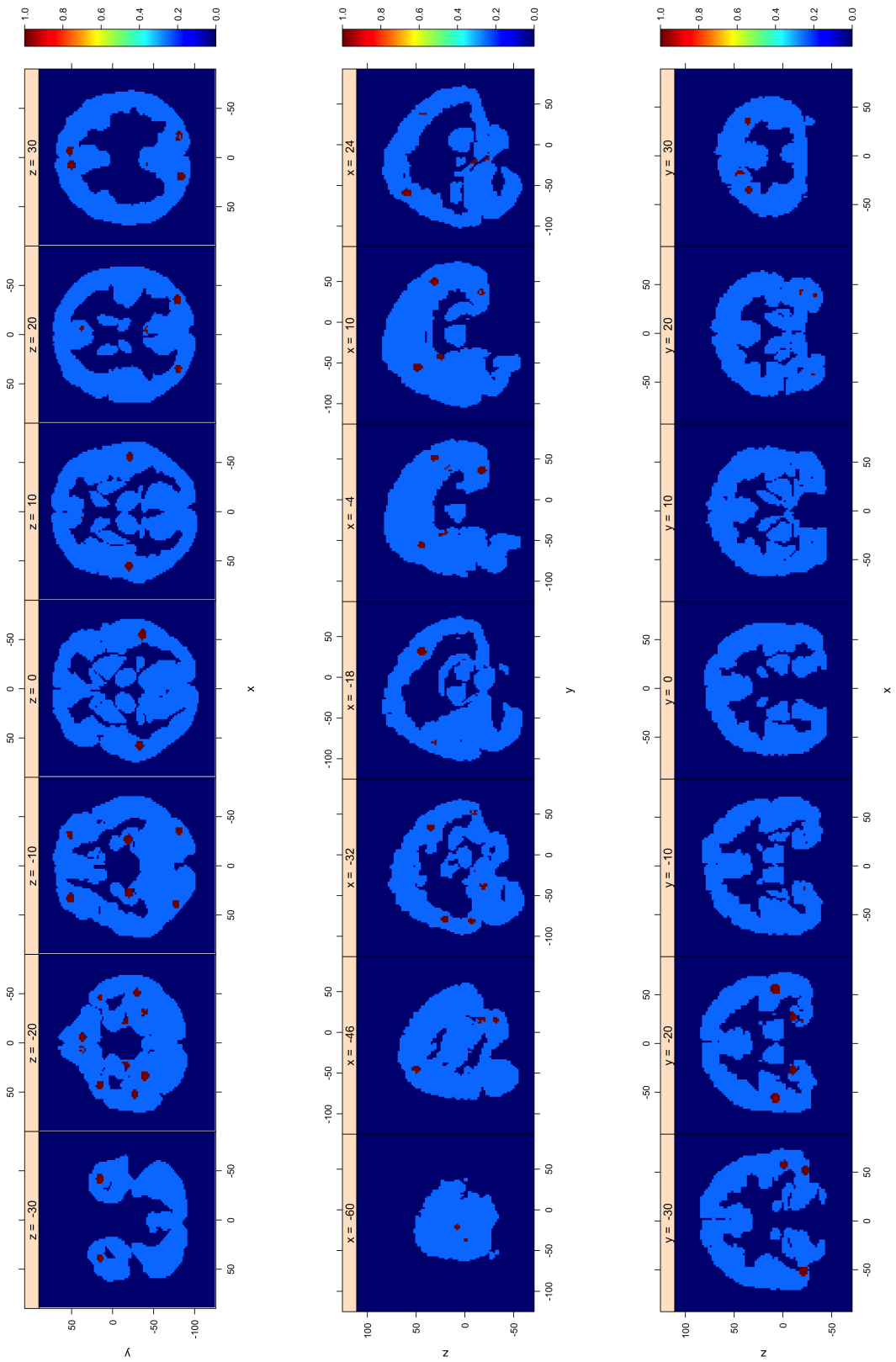


Figure C.2: Different views of brain-wide activation regions at month 6.

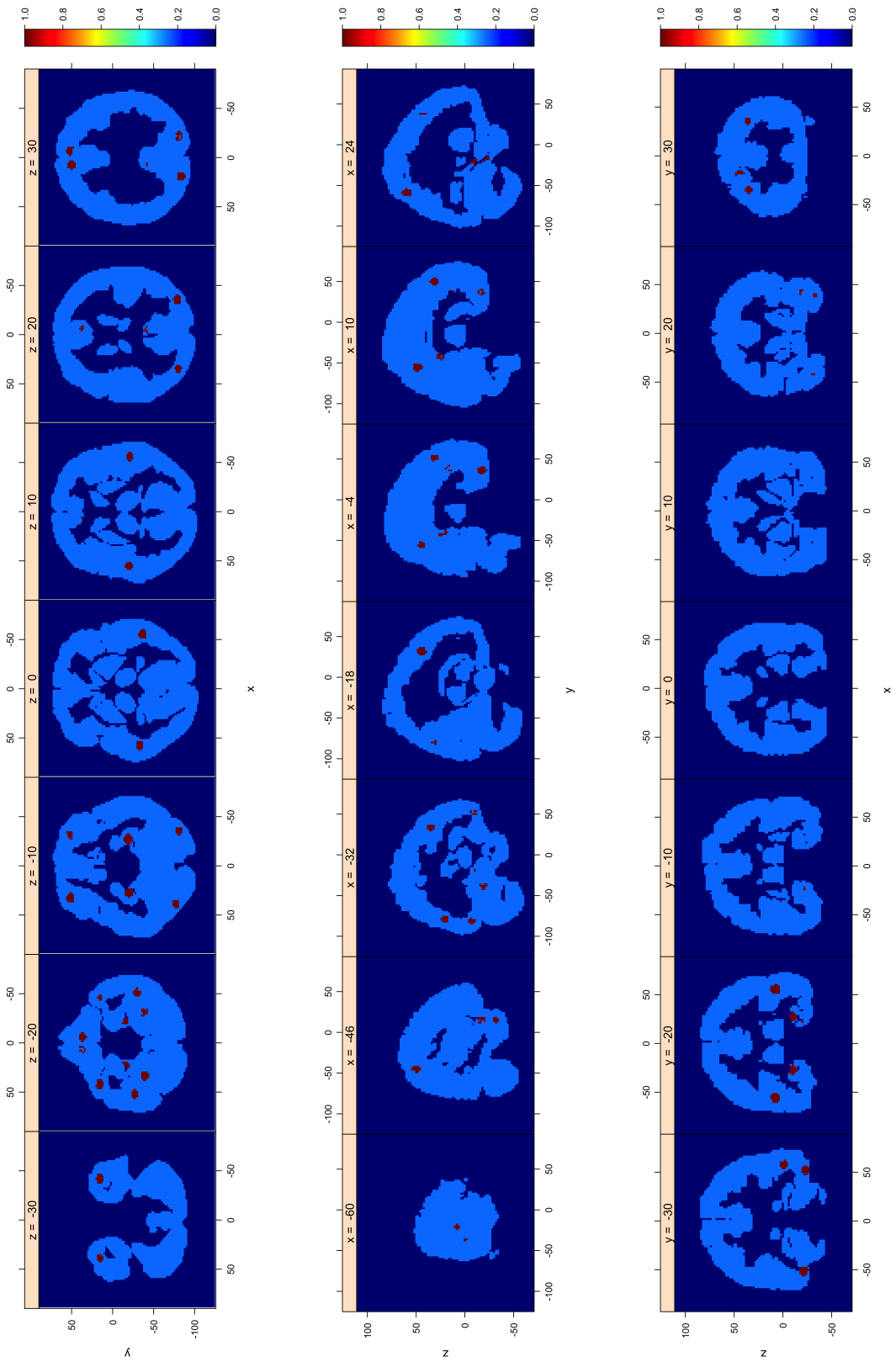


Figure C.3: Different views of brain-wide activation regions at month 12.

Table C.1: Anatomical region-wise results: number of voxels inside activation regions, related genes at all three different time points as well as genes with SNPs counts changed compared between any two time point.

	nvoxels.bl	nvoxels.m6	nvoxels.m12	nsnp.bl	nsnp.m6	nsnp.m12	Gene.bl	Gene.m6	Gene.m12	diffGene.blvsm6	diffGene.m6vsm12	diffGene.blvsm12
Frontal_Sup_L	60	62	61	3	3	3	ADAM10, APOE, PRNP	ADAM10, APOE, PRNP	ADAM10, APOE, PRNP			
Frontal_Sup_R	40	41	41	4	4	4	CR1, ECE1, ADAM10, APOE	CR1, ECE1, ADAM10, APOE	CR1, ECE1, ADAM10, APOE			
Frontal_Mid_L	73	73	74	1	1	2	SORCS1	SORCS1	Subject.Age, SORCS1		Subject.Age	Subject.Age
Frontal_Mid_R	74	71	72	2	1	1	SORCS1	SORCS1	SORCS1	SORCS1		SORCS1
Frontal_Sup_Medial_L	75	75	75	0	1	1		FAQ.Total.Score	FAQ.Total.Score	FAQ.Total.Score		FAQ.Total.Score
Frontal_Sup_Medial_R	66	66	66	2	2	2	SORCS1, GAPDHS	SORCS1, GAPDHS	SORCS1, GAPDHS			
Frontal_Mid_Orb_L	69	71	72	1	0	0	NEDD9			NEDD9		NEDD9
Frontal_Mid_Orb_R	31	32	32	1	1	1	NEDD9	NEDD9	NEDD9			
Rectus_L	35	35	35	0	1	1		SORCS1	SORCS1	SORCS1		SORCS1
Rectus_R	74	74	74	3	3	3	BIN1, SORCS1, ADAM10	BIN1, SORCS1, ADAM10	BIN1, SORCS1, ADAM10			
Cingulum_Ant_L	50	51	51	1	1	1	NEDD9	NEDD9	NEDD9			
Cingulum_Ant_R	30	30	30	3	2	3	DAPK1, SORCS1	DAPK1	DAPK1, SORCS1	SORCS1	SORCS1	
ParaHippocampal_L	36	36	36	1	1	0	SORCS1	SORCS1			SORCS1	SORCS1
ParaHippocampal_R	45	45	45	1	0	1	IL33		IL33	IL33	IL33	
Parietal_Sup_L	76	76	75	2	2	2	ADAM10	ADAM10	ADAM10			
Parietal_Sup_R	77	74	74	1	1	1	DAPK1	DAPK1	DAPK1			
Parietal_Inf_L	69	76	74	1	1	1	ADAM10	ADAM10	ADAM10			
Parietal_Inf_R	67	66	66	4	2	3	DAPK1, SORCS1, ACE	DAPK1	DAPK1, ACE	SORCS1, ACE	ACE	SORCS1
Precuneus_L	74	73	74	3	3	2	ADAM10, APOE, PRNP	APOE, PRNP, PICALM	ADAM10, APOE	ADAM10, PICALM	ADAM10, PRNP, PICALM	PRNP
Precuneus_R	76	76	76	3	2	1	DAPK1, SORCS1	DAPK1, SORCS1	DAPK1	DAPK1	SORCS1	DAPK1, SORCS1
Cingulum_Post_L	59	58	54	1	1	1	SORCS1	SORCS1	SORCS1			
Cingulum_Post_R	19	19	19	2	1	1	SORCS1, ADAM10	SORCS1	SORCS1	ADAM10		ADAM10
Temporal_Inf_L	73	73	74	1	1	1	SORCS1	SORCS1	SORCS1			
Temporal_Inf_R	60	62	63	3	4	3	PICALM, ADAM10, PRNP	PICALM, ADAM10, PRNP, APOE	PICALM, ADAM10, APOE	APOE	PRNP	PRNP, APOE
Fusiform_L	64	64	64	1	1	1	DAPK1	DAPK1	DAPK1			
Fusiform_R	29	29	29	0	0	0						
Occipital_Sup_L	71	67	65	3	2	2	DAPK1, SORCS1	DAPK1	DAPK1	SORCS1		SORCS1
Occipital_Mid_R	67	68	67	2	2	2	SORCS1, GAPDHS	SORCS1, GAPDHS	SORCS1, GAPDHS			
Occipital_Inf_L	70	70	70	1	1	1	NEDD9	NEDD9	NEDD9			
Occipital_Inf_R	67	68	68	0	0	0						
Temporal_Pole_Mid_L	54	55	55	1	0	1	SORL1		SORL1	SORL1	SORL1	
Temporal_Pole_Mid_R	68	68	68	2	2	2	TF, SORCS1	TF, SORCS1	TF, SORCS1			
Temporal_Pole_Sup_L	58	60	64	0	0	0						
Temporal_Pole_Sup_R	41	41	41	2	3	2	DAPK1	DAPK1, ADAM10	DAPK1	ADAM10	ADAM10	
Temporal_Mid_L	78	76	74	3	3	3	Subject.Age, SORCS1	Subject.Age, SORCS1	Subject.Age, SORCS1			
Temporal_Mid_R	74	74	72	3	3	2	ECE1, DAPK1, EXOC3L2	ECE1, DAPK1, EXOC3L2	ECE1, DAPK1		EXOC3L2	EXOC3L2
Hippocampus_L	42	52	52	1	1	1	NEDD9	NEDD9	NEDD9			
Hippocampus_R	59	58	59	1	1	1	NEDD9	NEDD9	NEDD9			
Temporal_Sup_L	73	72	71	1	1	1	DAPK1	DAPK1	DAPK1			
Temporal_Sup_R	72	72	72	2	1	1	SORCS1	SORCS1	SORCS1	SORCS1		SORCS1
Occipital_Sup_R	36	36	36	1	3	1	NEDD9	NEDD9, DAPK1, ADAM10	NEDD9	DAPK1, ADAM10	DAPK1, ADAM10	
Occipital_Mid_L	75	74	72	4	2	2	DAPK1, SORCS1	DAPK1, SORCS1	DAPK1, SORCS1	DAPK1, SORCS1		DAPK1, SORCS1

Table C.2: SNPs selected and ranked by their sum of inclusion probability across all regions and all time points

snps	nregions	gene	lobes	regions.bl	regions.m6	regions.m12
rs17477673	12	DAPK1	frontal lobe, parietal lobe, temporal lobe	Cingulum_Ant_R, Occipital_Sup_L, Parietal_Inf_R, Temporal_Pole_Sup_R	Cingulum_Ant_R, Occipital_Sup_L, Parietal_Inf_R, Temporal_Pole_Sup_R	Cingulum_Ant_R, Occipital_Sup_L, Parietal_Inf_R, Temporal_Pole_Sup_R
rs943855	12	DAPK1	frontal lobe, parietal lobe, temporal lobe	Cingulum_Ant_R, Occipital_Sup_L, Parietal_Inf_R, Temporal_Pole_Sup_R	Cingulum_Ant_R, Occipital_Sup_L, Parietal_Inf_R, Temporal_Pole_Sup_R	Cingulum_Ant_R, Occipital_Sup_L, Parietal_Inf_R, Temporal_Pole_Sup_R
rs12594742	11	ADAM10	frontal lobe, occipital lobe, parietal lobe	Frontal_Sup_L, Frontal_Sup_R, Precuneus_L, Temporal_Inf_R	Frontal_Sup_L, Frontal_Sup_R, Temporal_Inf_R	Frontal_Sup_L, Frontal_Sup_R, Precuneus_L, Temporal_Inf_R
rs769451	11	APOE	frontal lobe, occipital lobe, parietal lobe	Frontal_Sup_L, Frontal_Sup_R, Precuneus_L	Frontal_Sup_L, Frontal_Sup_R, Precuneus_L, Temporal_Inf_R	Frontal_Sup_L, Frontal_Sup_R, Precuneus_L, Temporal_Inf_R
rs10746816	10	DAPK1	occipital lobe, parietal lobe, temporal lobe	Fusiform_L, Parietal_Sup_R, Temporal_Sup_L	Fusiform_L, Occipital_Sup_R, Parietal_Sup_R, Temporal_Sup_L	Fusiform_L, Parietal_Sup_R, Temporal_Sup_L
rs12209631	9	NEDD9	temporal lobe	Hippocampus_L, Hippocampus_R, Occipital_Inf_L	Hippocampus_L, Hippocampus_R, Occipital_Inf_L	Hippocampus_L, Hippocampus_R, Occipital_Inf_L
7095427	9	SORCS1	frontal lobe, temporal lobe	Frontal_Mid_L, Frontal_Mid_R, Temporal_Mid_L	Frontal_Mid_L, Frontal_Mid_R, Temporal_Mid_L	Frontal_Mid_L, Frontal_Mid_R, Temporal_Mid_L
rs822326	8	SORCS1	frontal lobe, occipital lobe, temporal lobe	Rectus_R, Temporal_Sup_R	Rectus_R, Temporal_Inf_L, Temporal_Sup_R	Rectus_R, Temporal_Inf_L, Temporal_Sup_R
2756271	7	PRNP	frontal lobe, occipital lobe, parietal lobe	Frontal_Sup_L, Precuneus_L, Temporal_Inf_R	Frontal_Sup_L, Precuneus_L, Temporal_Inf_R	Frontal_Sup_L
rs1012503	6	NEDD9	frontal lobe, temporal lobe	Cingulum_Ant_L, Occipital_Sup_R	Cingulum_Ant_L, Occipital_Sup_R	Cingulum_Ant_L, Occipital_Sup_R
rs1473180	6	DAPK1	parietal lobe, temporal lobe	Occipital_Mid_L, Precuneus_R	Occipital_Mid_L, Precuneus_R	Occipital_Mid_L, Precuneus_R
rs11193130	6	SORCS1	frontal lobe, temporal lobe	Frontal_Sup_Medial_R, Occipital_Mid_R	Frontal_Sup_Medial_R, Occipital_Mid_R	Frontal_Sup_Medial_R, Occipital_Mid_R
rs11882238	6	GAPDHS	frontal lobe, temporal lobe	Frontal_Sup_Medial_R, Occipital_Mid_R	Frontal_Sup_Medial_R, Occipital_Mid_R	Frontal_Sup_Medial_R, Occipital_Mid_R
Subject.Age	4		frontal lobe, temporal lobe	Temporal_Mid_L	Temporal_Mid_L	Frontal_Mid_L, Temporal_Mid_L
rs12251340	4	SORCS1	parietal lobe, temporal lobe	Occipital_Mid_L, Precuneus_R	Occipital_Mid_L, Precuneus_R	
rs7896669	4	SORCS1	frontal lobe, temporal lobe	Frontal_Mid_R, Temporal_Mid_L	Temporal_Mid_L	Temporal_Mid_L
rs950809	4	SORCS1	frontal lobe, parietal lobe, temporal lobe	Cingulum_Ant_R, Occipital_Sup_L, Parietal_Inf_R		Cingulum_Ant_R
rs618679	4	PICALM	occipital lobe, parietal lobe	Temporal_Inf_R	Precuneus_L, Temporal_Inf_R	Temporal_Inf_R
rs677066	3	CR1	frontal lobe	Frontal_Sup_R	Frontal_Sup_R	Frontal_Sup_R
rs213052	3	ECE1	frontal lobe	Frontal_Sup_R	Frontal_Sup_R	Frontal_Sup_R
rs3026913	3	ECE1	temporal lobe	Temporal_Mid_R	Temporal_Mid_R	Temporal_Mid_R
rs2276575	3	BIN1	frontal lobe	Rectus_R	Rectus_R	Rectus_R
rs8177191	3	TF	temporal lobe	Temporal_Pole_Mid_R	Temporal_Pole_Mid_R	Temporal_Pole_Mid_R
rs1883238	3	NEDD9	frontal lobe	Frontal_Mid_Orb_R	Frontal_Mid_Orb_R	Frontal_Mid_Orb_R
rs1329600	3	DAPK1	temporal lobe	Temporal_Mid_R	Temporal_Mid_R	Temporal_Mid_R
rs1251753	3	SORCS1	occipital lobe	Cingulum_Post_R	Cingulum_Post_R	Cingulum_Post_R
rs2486154	3	SORCS1	temporal lobe	Temporal_Pole_Mid_R	Temporal_Pole_Mid_R	Temporal_Pole_Mid_R
rs822094	3	SORCS1	occipital lobe	Cingulum_Post_L	Cingulum_Post_L	Cingulum_Post_L
rs12908165	3	ADAM10	frontal lobe	Rectus_R	Rectus_R	Rectus_R
rs1427282	3	ADAM10	parietal lobe	Parietal_Inf_L	Parietal_Inf_L	Parietal_Inf_L
rs16940638	3	ADAM10	parietal lobe	Parietal_Sup_L	Parietal_Sup_L	Parietal_Sup_L
rs4238331	3	ADAM10	parietal lobe	Parietal_Sup_L	Parietal_Sup_L	Parietal_Sup_L
rs12685372	2	DAPK1	parietal lobe, temporal lobe	Occipital_Mid_L, Precuneus_R		
rs1929994	2	IL33	parietal lobe	ParaHippocampal_R		ParaHippocampal_R
rs10786998	2	SORCS1	parietal lobe	ParaHippocampal_L	ParaHippocampal_L	
rs12240947	2	SORCS1	temporal lobe	Occipital_Mid_L		Occipital_Mid_L
rs2418834	2	SORCS1	occipital lobe, temporal lobe	Temporal_Inf_L, Temporal_Sup_R		
rs7101373	2	SORL1	temporal lobe	Temporal_Pole_Mid_L		Temporal_Pole_Mid_L
rs4353	2	ACE	parietal lobe	Parietal_Inf_R		Parietal_Inf_R
rs10422797	2	EXOC3L2	temporal lobe	Temporal_Mid_R	Temporal_Mid_R	
FAQ.Score	2		frontal lobe		Frontal_Sup_Medial_L	Frontal_Sup_Medial_L
rs7897726	2	SORCS1	frontal lobe		Rectus_L	Rectus_L
rs4713379	1	NEDD9	frontal lobe	Frontal_Mid_Orb_L		
rs6494029	1	ADAM10	occipital lobe	Cingulum_Post_R		
rs12441313	1	ADAM10	temporal lobe		Occipital_Sup_R	
rs1427281	1	ADAM10	temporal lobe		Temporal_Pole_Sup_R	

Bibliography

Adolphs, R. (1999a), 'The human amygdala and emotion', Neuroscientist **5**, 125–137.

Adolphs, R. (1999b), 'The human amygdala and emotion', Neuroscientist **5**, 125–137.

Aggio, R. B. M., Ruggiero, K. and Villas-Bôas, S. G. (2010), 'Pathway activity profiling (papi): from the metabolite profile to the metabolic pathway activity', Bioinformatics **26**(23), 2969–76.

Agrawal, H. (2002), 'Extreme self-organization in networks constructed from gene expression data', Physical review letters **89**(26), 268702.

Akbani, R., Akdemir, K. C., Aksoy, B. A., Albert, M., Ally, A., Amin, S. B., Arachchi, H., Arora, A., Auman, J. T., Ayala, B., Baboud, J., Balasundaram, M., Balu, S., Barnabas, N., Bartlett, J., Bartlett, P., Bastian, B. C., Baylin, S. B., Behera, M., Belyaev, D., Benz, C., Bernard, B., Beroukhim, R., Bir, N., Black, A. D., Bodenheimer, T., Boice, L., Boland, G. M., Bono, R., Bootwalla, M. S., Bosenberg, M., Bowen, J., Bowlby, R., Bristow, C. A., Brockway-Lunardi, L., Brooks, D., Brzezinski, J., Bshara, W., Buda, E., Burns, W. R., Butterfield, Y. S., Button, M., Calderone, T., Cappellini, G. A., Carter, C., Carter, S. L., Cherney, L., Cherniack, A. D., Chevalier, A., Chin, L., Cho, J., Cho, R. J., Choi, Y.-L., Chu, A., Chudamani, S., Cibulskis, K., Ciriello, G., Clarke, A., Coons, S., Cope, L., Crain, D., Curley, E., Danilova, L., D'Atri, S., Davidsen, T., Davies, M. A., Delman, K. A., Demchok, J. A., Deng, Q. A., Deribe, Y. L., Dhalla, N., Dhir, R., DiCara, D.,

Dinikin, M., Dubina, M., Ebrom, J. S., Egea, S., Eley, G., Engel, J., Eschbacher, J. M., Fedosenko, K. V., Felau, I., Fennell, T., Ferguson, M. L., Fisher, S., Flaherty, K. T., Frazer, S., Frick, J., Fulidou, V., Gabriel, S. B., Gao, J., Gardner, J., Garraway, L. A., Gastier-Foster, J. M., Gaudioso, C., Gehlenborg, N., Genovese, G., Gerken, M., Gershenwald, J. E., Getz, G., Gomez-Fernandez, C., Gribbin, T., Grimsby, J., Gross, B., Guin, R., Gutschner, T., Hadjipanayis, A., Halaban, R., Hanf, B., Haussler, D., Haydu, L. E., Hayes, D. N., Hayward, N. K., Heiman, D. I., Herbert, L., Herman, J. G., Hersey, P., Hoadley, K. A., Hodis, E., Holt, R. A., Hoon, D. S., Hoppough, S., Hoyle, A. P., Huang, F. W., Huang, M., Huang, S., Hutter, C. M., Ibbs, M., Iype, L., Jacobsen, A., Jakrot, V., Janning, A., Jeck, W. R., Jefferys, S. R., Jensen, M. A., Jones, C. D., Jones, S. J., Ju, Z., Kakavand, H., Kang, H., Kefford, R. F., Khuri, F. R., Kim, J., Kirkwood, J. M., Klode, J., Korkut, A., Korski, K., Krauthammer, M., Kucherlapati, R., Kwong, L. N., Kycler, W., Ladanyi, M., Lai, P. H., Laird, P. W., Lander, E., Lawrence, M. S., Lazar, A. J., Łaźniak, R., Lee, D., Lee, J. E., Lee, J., Lee, K., Lee, S., Lee, W., Leporowska, E., Leraas, K. M., Li, H. I., Lichtenberg, T. M., Lichtenstein, L., Lin, P., Ling, S., Liu, J., Liu, O., Liu, W., Long, G. V., Lu, Y., Ma, S., Ma, Y., Mackiewicz, A., Mahadeshwar, H. S., Malke, J., Mallery, D., Manikhas, G. M., Mann, G. J., Marra, M. A., Matejka, B., Mayo, M., Mehrabi, S., Meng, S., Meyerson, M., Mieczkowski, P. A., Miller, J. P., Miller, M. L., Mills, G. B., Moiseenko, F., Moore, R. A., Morris, S., Morrison, C., Morton, D., Moschos, S., Mose, L. E., Muller, F. L., Mungall, A. J., Murawa, D., Murawa, P., Murray, B. A., Nezi, L., Ng, S., Nicholson, D., Noble, M. S., Osunkoya, A., Owonikoko, T. K., Ozenberger, B. A., Pagani, E., Paklina, O. V., Pantazi, A., Parfenov, M., Parfitt, J., Park, P. J., Park, W.-Y., Parker, J. S., Passarelli, F., Penny, R., Perou, C. M., Pihl, T. D., Potapova, O., Prieto, V. G., Protopopov, A., Quinn, M. J., Radenbaugh, A., Rai, K., Ramalingam, S. S., Raman, A. T., Ramirez, N. C., Ramirez, R., Rao, U., Rath-

mell, W. K., Ren, X., Reynolds, S. M., Roach, J., Robertson, A. G., Ross, M. I., Roszik, J., Russo, G., Saksena, G., Saller, C., Samuels, Y., Sander, C., Sander, C., Sandusky, G., Santoso, N., Saul, M., Saw, R. P., Schadendorf, D., Schein, J. E., Schultz, N., Schumacher, S. E., Schwallier, C., Scolyer, R. A., Seidman, J., Sekhar, P. C., Sekhon, H. S., Senbabaoglu, Y., Seth, S., Shannon, K. F., Sharpe, S., Sharpless, N. E., Shaw, K. R. M., Shelton, C., Shelton, T., Shen, R., Sheth, M., Shi, Y., Shiau, C. J., Shmulevich, I., Sica, G. L., Simons, J. V., Sinha, R., Sipahimalani, P., Sofia, H. J., Soloway, M. G., Song, X., Sougnez, C., Spillane, A. J., Spychała, A., Stretch, J. R., Stuart, J., Suchorska, W. M., Sucker, A., Sumer, S. O., Sun, Y., Synott, M., Tabak, B., Tabler, T. R., Tam, A., Tan, D., Tang, J., Tarnuzzer, R., Tarvin, K., Tatka, H., Taylor, B. S., Teresiak, M., Thiessen, N., Thompson, J. F., Thorne, L., Thorsson, V., Trent, J. M., Triche, T. J., Tsai, K. Y., Tsou, P., Van Den Berg, D. J., Van Allen, E. M., Veluvolu, U., Verhaak, R. G., Voet, D., Voronina, O., Walter, V., Walton, J. S., Wan, Y., Wang, Y., Wang, Z., Waring, S., Watson, I. R., Weinhold, N., Weinstein, J. N., Weisenberger, D. J., White, P., Wilkerson, M. D., Wilmott, J. S., Wise, L., Wiznerowicz, M., Woodman, S. E., Wu, C.-J., Wu, C.-C., Wu, J., Wu, Y., Xi, R., Xu, A. W., Yang, D., Yang, L., Yang, L., Zack, T. I., Zenklusen, J. C., Zhang, H., Zhang, J., Zhang, W., Zhao, X., Zhu, J., Zhu, K., Zimmer, L., Zmuda, E. and Zou, L. (2015), 'Genomic Classification of Cutaneous Melanoma', Cell **161**(7), 1681–1696.

URL: *http://www.sciencedirect.com/science/article/pii/S0092867415006340*

Alexander, G. E., Chen, K., Pietrini, P., Rapoport, S. I. and Reiman, E. M. (2002), 'Longitudinal pet evaluation of cerebral metabolic decline in dementia: a potential outcome measure in alzheimers disease treatment studies', American Journal of Psychiatry **159**(5), 738–745.

Alqallaf, F. and Gustafson, P. (2001), 'On cross-validation of bayesian models',

Canadian Journal of Statistics **29**, 333–340.

Ameis, S. H. and Szatmari, P. (2012), 'Imaging-genetics in autism spectrum disorder: advances, translational impact, and future directions', Frontiers in psychiatry **3**.

Andrawis, J. P., Hwang, K. S., Green, A. E., Kotlerman, J., Elashoff, D., Morra, J. H., Cummings, J. L., Toga, A. W., Thompson, P. M. and Apostolova, L. G. (2012), 'Effects of apoe4 and maternal history of dementia on hippocampal atrophy', Neurobiology of aging **33**(5), 856–866.

Antoniak, C. E. (1974a), 'Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems.', Institute of Mathematical Statistics. **2**(6), 1152–1174.

Antoniak, D. (1974b), 'Mixtures of dirichlet processes with applications to bayesian nonparametric problems', The Annals of Statistics **2**, 1152–1174.

Apolloni, J., Leguizamón, G. and Alba, E. (2016a), 'Two hybrid wrapper-filter feature selection algorithms applied to high-dimensional microarray experiments', Appl. Soft Comput. **38**, 922–932.

Apolloni, J., Leguizamón, G. and Alba, E. (2016b), 'Two hybrid wrapper-filter feature selection algorithms applied to high-dimensional microarray experiments', Applied Soft Computing **38**, 922–932.

Armitage, E. G., Godzien, J., Alonso-Herranz, V., López-Gonzálvez, Á. and Barbas, C. (2015), 'Missing value imputation strategies for metabolomics data', Electrophoresis **36**(24), 3050–3060.

Armstrong, M. (1997), Basic Linear Geostatistics, Springer Verlag, Berlin.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M.,

Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T. et al. (2000), 'Gene ontology: tool for the unification of biology', Nature genetics **25**(1), 25–29.

Ashford, J. W. and Mortimer, J. A. (2002), 'Non-familial alzheimer's disease is mainly due to genetic factors', Journal of Alzheimer's disease **4**(3), 169–177.

Atchade, Y. F. (2006), 'An adaptive version for the metropolis adjusted langevin algorithm with a truncated drift', Methodology and Computing in applied Probability **8**(2), 235–254.

Baddeley, A. (2010), Multivariate and marked point processes, in A. E. Gelfand, P. Diggle, P. Guttorp and M. Fuentes, eds, 'Handbook of Spatial Statistics', Chapman & Hall/CRC, chapter 21, pp. 371–402.

Baddeley, A., Turner, R., Mller, J. and Hazelton, M. (2005), 'Residual analysis for spatial point processes (with discussion)', Journal of the Royal Statistical Society, Series B **67**(5), 617–666.

Balafar, M. A., Ramli, A. R., Saripan, M. I. and Mashohor, S. (2010), 'Review of brain mri image segmentation methods', Artificial Intelligence Review **33**(3), 261–274.

Baldi, P. and Long, A. D. (2001), 'A bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes', Bioinformatics **17**(6), 509–519.

Barbu, A. and Zhu, S.-C. (2005), 'Generalizing swendsen-wang to sampling arbitrary posterior probabilities', IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(8), 1239–1253.

Barcella, W., De Iorio, M. and Baio, G. (2015a), 'Variable Selection for Covariate Dependent Dirichlet Process Mixtures of Regressions', p. 26.
URL: *http://arxiv.org/abs/1508.00129*

Barcella, W., De Iorio, M. and Baio, G. (2015b), 'Variable selection for covariate dependent dirichlet process mixtures of regressions', arXiv preprint arXiv:1508.00129 .

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M. et al. (2013), 'Ncbi geo: archive for functional genomics data setsupdate', Nucleic acids research **41**(D1), D991–D995.

Barupal, D. K., Haldiya, P. K., Wohlgemuth, G., Kind, T., Kothari, S. L., Pinkerton, K. E. and Fiehn, O. (2012), 'Metamapp: mapping and visualizing metabolomic data by integrating information from biochemical pathways and chemical and mass spectral similarity', BMC Bioinformatics **13**, 99.

Bdiri, T. and Bouguila, N. (2012), 'Positive vectors clustering using inverted Dirichlet finite mixture models', Expert Syst. Appl. **39**(2), 1869–1882.
URL: *http://www.sciencedirect.com/science/article/pii/S0957417411011754*

Bdiri, T., Bouguila, N. and Ziou, D. (2016), 'A statistical framework for online learning using adjustable model selection criteria', Engineering Applications of Artificial Intelligence **49**, 19–42.

Belacel, N., Wang, Q. and Cuperlovic-Culf, M. (2006), 'Clustering methods for microarray gene expression data.', OMICS **10**(4), 507–31.
URL: *http://online.liebertpub.com/doi/abs/10.1089/omi.2006.10.507*

Benjamini, Y. and Hochberg, Y. (1995a), 'Controlling the false discovery rate: a practical and powerful approach to multiple testing', Journal of the Royal Statistical Society, Series B, Methodological **57**, 289–300.

Benjamini, Y. and Hochberg, Y. (1995b), 'Controlling the false discovery rate: a

practical and powerful approach to multiple testing', Journal of the royal statistical society. Series B (Methodological) pp. 289–300.

Bergeest, J.-P. and Rohr, K. (2012), 'Efficient globally optimal segmentation of cells in fluorescence microscopy images using level sets and convex energy functionals', Medical image analysis **16**(7), 1436–1444.

Berger, B., Peng, J. and Singh, M. (2013a), 'Computational solutions for omics data', Nature reviews. Genetics **14**(5), 333.

Berger, B., Peng, J. and Singh, M. (2013b), 'Computational solutions for omics data', Nature Reviews Genetics **14**(5), 333–346.

Berger, B., Peng, J. and Singh, M. (2013c), 'Computational solutions for omics data', Nature reviews. Genetics **14**(5), 333.

Best, N. G., Ickstadt, K. and Wolpert, R. L. (2000), 'Spatial poisson regression for health and exposure data measured at disparate resolutions', Journal of the American Statistical Association **95**(452), 1076–1088.

Best, N. G., Ickstadt, K., Wolpert, R. L., Cockings, S., Elliott, P., Bennett, J., Bottle, A. and Reed, S. (2002), Modeling the impact of traffic-related air-pollution on childhood respiratory illness, in C. Gatsonis, R. Kass, B. Carlin, A. Carriquiry, A. Gelman., I. Verdinelli and M. West, eds, 'In Case Studies In Bayesian Statistics, Volume V', Springer-Verlag, pp. 183–259.

Bhat, M., Skill, N., Marcus, V., Deschenes, M., Tan, X. M., Bouteaud, J., Negi, S., Awan, Z., Aikin, R., Kwan, J., Amre, R., Tabaries, S., Hassanain, M., Seidah, N. G., Maluccio, M., Siegel, P. and Metrakos, P. (2015), 'Decreased pcsk9 expression in human hepatocellular carcinoma', BMC gastroenterology **15**, 176.

Bigos, K. L. and Weinberger, D. R. (2010), 'Imaging geneticsdays of future past', Neuroimage **53**(3), 804–809.

Bijlsma, S., Bobeldijk, I., Verheij, E. R., Ramaker, R., Kochhar, S., Macdonald, I. A., Van Ommen, B. and Smilde, A. K. (2006), 'Large-scale human metabolomics studies: a strategy for data (pre-) processing and validation', Analytical chemistry **78**(2), 567–574.

Blackwell, D. and MacQueen, J. B. (1973), 'Ferguson distributions via pólya urn schemes', The annals of statistics pp. 353–355.

Blei, D. M., Griffiths, T. L. and Jordan, M. I. (2010), 'The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies', Journal of the ACM **57**(2), 111–142.

Bondesson, L. (1982), 'On simulation from infinitely distributions', Advances in Applied Probability (14), 855–869.

Bookheimer, S. Y., Strojwas, M. H., Cohen, M. S., Saunders, A. M., Pericak-Vance, M. A., Mazziotta, J. C. and Small, G. W. (2000), 'Patterns of brain activation in people at risk for alzheimer's disease', New England journal of medicine **343**(7), 450–456.

Boros, G. and Moll, V. (2004), The expansion of the loggamma function., in 'Irresistible Integrals: Symbolics, Analysis and Experiments in the Evaluation of Integrals', Cambridge, England: Cambridge University Press, chapter 10.6, pp. 201–203.

Bowman, F. D. (2005), 'Spatio-temporal modeling of localized brian activity', Biostatistics **6**, 558–575.

Braak, H. and Braak, E. (1991), 'Neuropathological stageing of alzheimer-related changes', Acta neuropathologica **82**(4), 239–259.

Braak, H. and Braak, E. (1995), 'Staging of alzheimer's disease-related neurofibrillary changes', Neurobiology of aging **16**(3), 271–278.

Brigham, K. L. (2010), 'Predictive health: the imminent revolution in health care', J Am Geriatr Soc **58 Suppl 2**, S298–302.

Brix, A. and Møller, J. (2001), 'Space-time multitype log gaussian cox processes with a view to modeling weed data', Scandinavian Journal of Statistics **28**, 471–488.

Buescher, J. M. and Driggers, E. M. (2016), 'Integration of omics: more than the sum of its parts', Cancer & metabolism **4**(1), 4.

Cabeza, R. and Kingstone, A. (2006), Handbook of Functional Neuroimaging of Cognition, second edn, The MIT Press.

Cai, Q., Alvarez, J. A., Kang, J. and Yu, T. (2017a), 'Network marker selection for untargeted lc-ms metabolomics data', J Proteome Res **16**(3), 1261–1269.

Cai, Q., Alvarez, J. A., Kang, J. and Yu, T. (2017b), 'Network marker selection for untargeted lcms metabolomics data', Journal of Proteome Research **16**(3), 1261–1269.

Cerami, E. G., Gross, B. E., Demir, E., Rodchenkov, I., Babur, Ö., Anwar, N., Schultz, N., Bader, G. D. and Sander, C. (2011), 'Pathway commons, a web resource for biological pathway data', Nucleic acids research **39**(suppl 1), D685–D690.

Cerami, E., Gao, J. J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., Jacobsen, A., Byrne, C. J., Heuer, M. L., Larsson, E., Antipin, Y., Reva, B., Goldberg, A. P., Sander, C. and Schultz, N. (2012), 'The cbio cancer genomics

portal: an open platform for exploring multidimensional cancer genomics data', Cancer discovery **2**, 401–4.

Chapuis, J., Hot, D., Hansmannel, F., Kerdraon, O., Ferreira, S., Hubans, C., Maurage, C., Huot, L., Bensemain, F. and Laumet, G. (2009), 'Transcriptomic and genetic studies identify il-33 as a candidate gene for alzheimers disease', Molecular psychiatry **14**(11), 1004.

Chase, T. N., Foster, N. L., Fedio, P., Brooks, R., Mansi, L. and di Chiro, G. (1984), 'Regional cortical dysfunction in alzheimer's disease as determined by positron emission tomography', Annals of Neurology **15**(S1), 170–174.

Chi, E. C., Allen, G. I., Zhou, H., Kohannim, O., Lange, K. and Thompson, P. M. (2013), Imaging genetics via sparse canonical correlation analysis, in 'Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on', IEEE, pp. 740–743.

Chiang, H.-S. and Pao, S.-C. (2016), 'An eeg-based fuzzy probability model for early diagnosis of alzheimers disease', Journal of medical systems **40**(5), 125.

Chipman, H., George, E. I., McCulloch, R. E., Clyde, M., Foster, D. P. and Stine, R. A. (2001), 'The practical implementation of bayesian model selection', Lecture Notes-Monograph Series pp. 65–134.

Chu, C., Ni, Y., Tan, G., Saunders, C. J. and Ashburner, J. (2011), 'Kernel regression for fmri pattern prediction', NeuroImage **56**(2), 662–673.

Chuang, H.-Y., Lee, E., Liu, Y.-T., Lee, D. and Ideker, T. (2007), 'Network-based classification of breast cancer metastasis', Molecular systems biology **3**(1), 140.

Chumbley, J. and Friston, K. (2009), 'False discovery rate revisited: Fdr and topological inference using gaussian random fields', NeuroImage **44**(1), 62–70.

Chupin, M., Grardin, E., Cuingnet, R., Boutet, C., Lemieux, L., Lehricy, S., Benali, H., Garnero, L., Colliot, O. and Initiative, A. D. N. (2009), 'Fully automatic hippocampus segmentation and classification in alzheimers disease and mild cognitive impairment applied on data from adni', Hippocampus **19**(6), 579.

Chtelat, G., Fouquet, M., Kalpouzos, G., Denghien, I., De La Sayette, V., Viader, F., Mzenge, F., Landeau, B., Baron, J.-C. and Eustache, F. (2008), 'Three-dimensional surface mapping of hippocampal atrophy progression from mci to ad and over normal aging as assessed using voxel-based morphometry', Neuropsychologia **46**(6), 1721–1731.

Clauset, A., Newman, M. E. and Moore, C. (2004), 'Finding community structure in very large networks', Physical review E **70**, 066111.

Cohen, A. D. and Klunk, W. E. (2014), 'Early detection of alzheimer's disease using pib and fdg pet', Neurobiology of disease **72**, 117–122.

Colciaghi, F., Marcello, E., Borroni, B., Zimmermann, M., Caltagirone, C., Cattabeni, F., Padovani, A. and Di Luca, M. (2004), 'Platelet app, adam 10 and bace alterations in the early stages of alzheimer disease', Neurology **62**(3), 498–501.

Cong, S., Rizkalla, M., Du, E. Y., West, J., Risacher, S., Saykin, A. and Shen, L. (n.d.), Building a surface atlas of hippocampal subfields from mri scans using freesurfer, first and spharm, in 'Circuits and Systems (MWSCAS), 2014 IEEE 57th International Midwest Symposium on', IEEE, pp. 813–816.

Consortium, . G. P. et al. (2010), 'A map of human genome variation from population-scale sequencing', Nature **467**(7319), 1061–1073.

Consortium, . G. P. et al. (2012), 'An integrated map of genetic variation from 1,092 human genomes', Nature **491**(7422), 56–65.

Costafreda, S. G., Dinov, I. D., Tu, Z., Shi, Y., Liu, C.-Y., Kloszewska, I., Mecocci, P., Soininen, H., Tsolaki, M. and Vellas, B. (2011), 'Automated hippocampal shape analysis predicts the onset of dementia in mild cognitive impairment', Neuroimage **56**(1), 212–219.

Cox, D. R. (1955), 'Some statistical models related with series of events', Journal of the Royal Statistical Society, Series B **17**, 129–164.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehricy, S., Habert, M.-O., Chupin, M., Benali, H., Colliot, O. and Initiative, A. D. N. (2011a), 'Automatic classification of patients with alzheimer's disease from structural mri: a comparison of ten methods using the adni database', neuroimage **56**(2), 766–781.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehricy, S., Habert, M.-O., Chupin, M., Benali, H., Colliot, O. and Initiative, A. D. N. (2011b), 'Automatic classification of patients with alzheimer's disease from structural mri: a comparison of ten methods using the adni database', neuroimage **56**(2), 766–781.

Cun, Y. P. and Fröhlich, H. (2012), 'Biomarker gene signature discovery integrating network knowledge', Biology **1**, 5–17.

Cun, Y. P. and Fröhlich, H. (2013), 'Network and data integration for biomarker signature discovery via network smoothed t-statistics', PLoS One **8**, e73074.

Damien, P., Laud, P. W. and Smith, A. F. M. (1995), 'Approximate random variate generation from infinitely divisible distributions with applications to bayesian inference', Journal of the Royal Statistical Society, Series B **57**, 547–563.

Das, J. and Yu, H. Y. (2012), 'Hint: High-quality protein interactomes and their applications in understanding human disease', BMC systems biology **6**, 92.

Daye, Z. J., Xie, J. and Li, H. (2012), 'A sparse structured shrinkage estimator for nonparametric varying-coefficient model with an application in genomics', Journal of Computational and Graphical Statistics .

Dean, D. C., Jerskey, B. A., Chen, K., Protas, H., Thiyyagura, P., Roontiva, A., O'muircheartaigh, J., Dirks, H., Waskiewicz, N. and Lehman, K. (2014), 'Brain differences in infants at differential genetic risk for late-onset alzheimer disease: a cross-sectional imaging study', JAMA neurology **71**(1), 11–22.

Demidenko, R., Razanauskas, D., Daniunaite, K., Lazutka, J. R., Jankevicius, F. and Jarmalaite, S. (2015), 'Frequent down-regulation of abc transporter genes in prostate cancer', BMC cancer **15**, 683.

Derré, L., Corvaisier, M., Charreau, B., Moreau, A., Godefroy, E., Moreau-Aubry, A., Jotereau, F. and Gervois, N. (2006), 'Expression and release of hla-e by melanoma cells and melanocytes: potential impact on the response of cytotoxic effector cells', Journal of immunology (Baltimore, Md. : 1950) **177**, 3100–7.

Desikan, R. S., Sgonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P. and Hyman, B. T. (2006), 'An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest', Neuroimage **31**(3), 968–980.

Dice, L. R. (1945), 'Measures of the amount of ecologic association between species', Ecology **26**, 297–302.

Dickerson, B. C., Goncharova, I., Sullivan, M., Forchetti, C., Wilson, R., Bennett, D. and Beckett, L. A. (2001), 'Mri-derived entorhinal and hippocampal atrophy in incipient and very mild alzheimers disease', Neurobiology of aging **22**(5), 747–754.

Diggle, P. (1983), Statistical analysis of spatial point patterns, Academic Press.

Diggle, P. J. (1981), Some graphical methods in the analysis of spatial point patterns, in V. Barnett, ed., 'Interpreting multivariate data', New York: John Wiley & Sons, pp. 55–73.

Diggle, P. J. (1990), 'A point process modeling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point', Journal of the Royal Statistical Society, Series A **153**, 349–362.

Diggle, P. J. and Milne, R. K. (1983), 'Bivariate cox processes: some models for bivariate spatial point patterns', Journal of the Royal Statistical Society, Series B **45**, 11–21.

Diggle, P. J. and Rowlingson, B. (1994), 'A conditional approach to point process modeling of elevated risk', Journal of the Royal Statistical Society, Series A **157**(3), 433–440.

Diggle, P. J., Zheng, P. and Durr, P. (2005), 'Non-parametric estimation of spatial segregation in a multivariate point process', Applied Statistics **54**, 645–658.

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., Weingessel, A. and Leisch, M. F. (2009), 'Package 'e1071", R Software package, avaliable at http://cran. rproject. org/web/packages/e1071/index. html .

Ding, C. and Peng, H. (2005), 'Minimum redundancy feature selection from microarray gene expression data', Journal of bioinformatics and computational biology **3**(02), 185–205.

Do, K. A., Müller, P. and Tang, F. (2005), 'A bayesian mixture model for differential gene expression', Journal of the Royal Statistical Society: Series C (Applied Statistics) **54**, 627–644.

Dubins, L. and Freedman, D. (1964), 'Measurable sets of measures', Pacific Journal of Mathematics **14**(4), 1211–1222.

Dudoit, S., Fridlyand, J. and Speed, T. P. (2002), 'Comparison of discrimination methods for the classification of tumors using gene expression data', Journal of the American statistical association **97**(457), 77–87.

Dudoit, S., Shaffer, J. P. and Boldrick, J. C. (2003), 'Multiple hypothesis testing in microarray experiments', Statistical Science pp. 71–103.

Dudoit, S., van der Laan, M. J. and Pollard, K. S. (2004), 'Multiple testing. part i. single-step procedures for control of general type i error rates', Statistical Applications in Genetics and Molecular Biology **3**(1), 1–69.

Dunson, D. B. and Park, J.-H. (2008a), 'Kernel stick-breaking processes.', Biometrika **95**(2), 307–323.
URL: *http://biomet.oxfordjournals.org/content/95/2/307.short*

Dunson, D. B. and Park, J.-H. (2008b), 'Kernel stick-breaking processes', Biometrika **95**(2), 307–323.

Ecker, J. R., Bickmore, W. A., Barroso, I., Pritchard, J. K., Gilad, Y. and Segal, E. (2012), 'Genomics: Encode explained', Nature **489**(7414), 52–55.

Edgar, R., Domrachev, M. and Lash, A. E. (2002), 'Gene expression omnibus: Ncbi gene expression and hybridization array data repository', Nucleic acids research **30**(1), 207–210.

Edwards, T. L., PericakVance, M., Gilbert, J. R., Haines, J. L., Martin, E. R. and Ritchie, M. D. (2009), 'An association analysis of alzheimer disease candidate genes detects an ancestral risk haplotype clade in ace and putative multilocus association

between ace, a2m, and lrrtm3', American Journal of Medical Genetics Part B: Neuropsychiatric Genetics **150**(5), 721–735.

Efron, B. (2004a), 'Large-Scale Simultaneous Hypothesis Testing', J. Am. Stat. Assoc. **99**(465), 96–104.
URL: *http://amstat.tandfonline.com/doi/abs/10.1198/016214504000000089*

Efron, B. (2004b), 'Large-Scale Simultaneous Hypothesis Testing', J. Am. Stat. Assoc. **99**(465), 96–104.
URL: *http://amstat.tandfonline.com/doi/abs/10.1198/016214504000000089#.VvCQMcdQVFI*

Efron, B. (2005), Local false discovery rates.

Efron, B. (2007), 'Correlation and Large-Scale Simultaneous Significance Testing', J. Am. Stat. Assoc. **102**(477), 93–103.
URL: *http://amstat.tandfonline.com/doi/abs/10.1198/016214506000001211*

Efron, B. (2010), 'Correlated z-values and the accuracy of large-scale statistical estimates', Journal of the American Statistical Association **105**(491), 1042–1055.

Efron, B. (2012a), Large-scale inference: empirical Bayes methods for estimation, testing, and prediction, Vol. 1, Cambridge University Press.

Efron, B. (2012b), 'Large-scale simultaneous hypothesis testing', Journal of the American Statistical Association .

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. et al. (2004), 'Least angle regression', The Annals of statistics **32**(2), 407–499.

Efron, B., Storey, J. D. and Tibshirani, R. (2001), 'Microarrays empirical bayes methods, and false discovery rates', (Stanford University. Department of Statistics).

Efron, B. and Tibshirani, R. (2002), 'Empirical bayes methods and false discovery rates for microarrays', Genetic epidemiology **23**, 70–86.

Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. (2001), 'Empirical bayes analysis of a microarray experiment', Journal of the American statistical association **96**(456), 1151–1160.

Egan, M. F., Goldberg, T. E., Kolachana, B. S., Callicott, J. H., Mazzanti, C. M., Straub, R. E., Goldman, D. and Weinberger, D. R. (2001), 'Effect of comt val108/158 met genotype on frontal lobe function and risk for schizophrenia', Proceedings of the National Academy of Sciences **98**(12), 6917–6922.

Eickhoff, S. B., Laird, A., Grefkes, C., Wang, L. E., Zilles, K. and Fox, P. T. (2009a), 'Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: a random-effects approach based on empirical estimates of spatial uncertainty', Human Brain Mapping **30**(9), 2907–2926.

Eickhoff, S. B., Laird, A. R., Grefkes, C., Wang, L. E., Zilles, K. and Fox, P. T. (2009b), 'Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: a random-effects approach based on empirical estimates of spatial uncertainty', Human Brain Mapping **30**(9), 2907–2926.

Elsnerova, K., Mohelnikova-Duchonova, B., Cerovska, E., Ehrlichova, M., Gut, I., Rob, L., Skapa, P., Hruda, M., Bartakova, A., Bouda, J., Vodicka, P., Soucek, P. and Vaclavikova, R. (2016), 'Gene expression of membrane transporters: Importance for prognosis and progression of ovarian carcinoma', Oncology reports **35**, 2159–70.

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

Escobar, M. D. and West, M. (1995a), 'Bayesian density estimation and inference using mixtures', Journal of the American Statistical Association **90**, 577–588.

Escobar, M. D. and West, M. (1995b), 'Bayesian density estimation and inference using mixtures', Journal of the American Statistical Association **90**, 577–588.

Escobar, M. D. and West, M. (1995c), 'Bayesian Density Estimation and Inference Using Mixtures', J. Am. Stat. Assoc. .

URL: *http://www.tandfonline.com/doi/abs/10.1080/01621459.1995.10476550#.VvHrycdQXlI*

Estrach, S., Lee, S. A., Boulter, E., Pisano, S., Errante, A., Tissot, F. S., Cailleteau, L., Pons, C., Ginsberg, M. H. and Féral, C. C. (2014), 'Cd98hc (slc3a2) loss protects against ras-driven tumorigenesis by modulating integrin-mediated mechanotransduction', Cancer research **74**, 6878–89.

Everitt, B. and Bullmore, E. (1999), 'Mixture model mapping of brain activation in functional magnetic resonance images', Hum. Brain Map. **7**, 1–14.

Falcon, S. and Gentleman, R. (2007), 'Using gostats to test gene lists for go term association', Bioinformatics (Oxford, England) **23**, 257–8.

Fan, W. and Bouguila, N. (2013), 'Variational learning of a Dirichlet process of generalized Dirichlet distributions for simultaneous clustering and feature selection', Pattern Recognit. **46**(10), 2754–2769.

URL: *http://www.sciencedirect.com/science/article/pii/S0031320313001568*

Fan, W., Sallay, H., Bouguila, N. and Bourouis, S. (2015), 'A hierarchical Dirichlet process mixture of generalized Dirichlet distributions for feature selection', Comput. Electr. Eng. **43**, 48–65.

URL: *http://www.sciencedirect.com/science/article/pii/S0045790615001056*

Feng, C., Narayana, S., Lancaster, J. L., Jerabek, P. A., Arnow, T. L., Zhu, F., Tan, L. H., Fox, P. T. and Gao, J. (2004), 'Cbf changes during brain activation: fmri vs. pet', NeuroImage **22**(1), 443–446.

URL: *http://www.ncbi.nlm.nih.gov/pubmed/15110037*

Ferguson, T. S. (1973a), 'A bayesian analysis of some nonparametric problems', The Annals of Statistics **1**, 209–230.

Ferguson, T. S. (1973b), 'A bayesian analysis of some nonparametric problems', The annals of statistics pp. 209–230.

Ferguson, T. S. (1983), 'Bayesian density estimation by mixtures of normal distributions', Recent advances in statistics **24**(1983), 287–302.

Foster, N. L., Chase, T. N., Mansi, L., Brooks, R., Fedio, P., Patronas, N. J. and Di Chiro, G. (1984), 'Cortical abnormalities in alzheimer's disease', Annals of neurology **Psychiatry16**(6), 649–654.

Fox, E. B., Sudderth, E. B., Jordan, M. I. and Willsky, A. S. (2011), 'A sticky hdp-hmm with application to speaker diarization', The Annals of Applied Statistics p. To appear.

Fox, P. T., Lancaster, J. L., Parsons, L. M., Xiong, J. and Zamarripa, F. (1997a), 'Functional volumes modeling: theory and preliminary assessment', Human Brain Mapping **5**(4), 306–311.

Fox, P. T., Lancaster, J. L., Parsons, L. M., Xiong, J. and Zamarripa, F. (1997b), 'Functional volumes modeling: theory and preliminary assessment', Human Brain Mapping **5**(4), 306–311.

Fox, R. J. and Dimmic, M. W. (2006), 'A two-sample bayesian t-test for microarray data', BMC bioinformatics **7**(1), 1.

Fraley, C. and Raftery, A. E. (1999), 'Mclust: Software for model-based cluster analysis', Journal of Classification **16**, 297–306.

Friedman, J. H. (1991), 'Multivariate adaptive regression splines', The annals of statistics pp. 1–67.

Frisoni, G. B., Fox, N. C., Jack, C. R., Scheltens, P. and Thompson, P. M. (2010a), 'The clinical use of structural mri in alzheimer disease', Nature Reviews Neurology **6**(2), 67–77.

Frisoni, G. B., Fox, N. C., Jack, C. R., Scheltens, P. and Thompson, P. M. (2010b), 'The clinical use of structural mri in alzheimer disease', Nature Reviews Neurology **6**(2), 67–77.

Friston, K. J., Ashburner, J. T., Kiebel, S. J., Nichols, T. E. and Penny, W. D., eds (2007), Statistical Parametric Mapping: The Analysis of Functional Brain Images, Academic Press.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D. and Frackowiak, R. S. J. (1994), 'Statistical parametric maps in functional imaging: A general linear approach', Human Brain Mapping **2**(4), 189–210.
URL: *http://dx.doi.org/10.1002/hbm.460020402*

Friston, K., Worsley, K., Frackowiak, R., Mazziotta, J. and Evans, A. (1994), 'Assessing the significance of focal activations using their spatial extent', Human Brain Mapping **1**, 214–220.

Fu, A. K., Hung, K.-W., Yuen, M. Y., Zhou, X., Mak, D. S., Chan, I. C., Cheung, T. H., Zhang, B., Fu, W.-Y. and Liew, F. Y. (2016), 'Il-33 ameliorates alzheimers disease-like pathology and cognitive decline', Proceedings of the National Academy of Sciences **113**(19), E2705–E2713.

Gatta, L. B., Albertini, A., Ravid, R. and Finazzi, D. (2002), 'Levels of -secretase bace and -secretase adam10 mrnas in alzheimer hippocampus', Neuroreport **13**(16), 2031–2033.

Gelfand, A. E. (1996), Inference and monitoring convergence, in W. R. Gilks,

S. Richardson and D. J. Spiegelhalter, eds, 'Markov Chain Monte Carlo in Practice', Chapman& Hall, pp. 131–144.

Gelfand, A. E., Dey, D. K. and Chang, H. (1992), Model determination using predictive distributions with implementation via sampling-based methods (with discussion), in J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds, 'Bayesian Statistics', Vol. 4, Oxford University Press, pp. 147–167.

Gelman, A., Garlin, J. B., Stern, H. S. and Rubin, D. B. (2004), Bayesian Data Analysis, 2 edn, Chapman & Hall/CRC.

Genovese, C. R., Lazar, N. A. and Nichols, T. E. (2002), 'Thresholding of statistical maps in functional neuroimaging using the false discovery rate', Neuroimage **15**(4), 870–878.

George, E. I. and McCulloch, R. E. (1993), 'Variable selection via gibbs sampling', Journal of the American Statistical Association **88**(423), 881–889.

George, E. I. and McCulloch, R. E. (1997), 'Approaches for bayesian variable selection', Statistica sinica pp. 339–373.

Gerardin, E., Chtelat, G., Chupin, M., Cuingnet, R., Desgranges, B., Kim, H.-S., Niethammer, M., Dubois, B., Lehricy, S. and Garnero, L. (2009), 'Multidimensional classification of hippocampal shape features discriminates alzheimer's disease and mild cognitive impairment from normal aging', Neuroimage **47**(4), 1476–1486.

Gerlini, G., Tun-Kyi, A., Dudli, C., Burg, G., Pimpinelli, N. and Nestle, F. O. (2004), 'Metastatic melanoma secreted il-10 down-regulates cd1 molecules on dendritic cells in metastatic tumor lesions', The American journal of pathology **165**, 1853–63.

Ghosh, D. and Chinnaiyan, A. M. (2005), 'Classification and selection of biomarkers in genomic data using LASSO', J. Biomed. Biotechnol. **2005**(2), 147–154.

Ghosh, S., Matsuoka, Y., Asai, Y., Hsin, K.-Y. and Kitano, H. (2011), 'Software for systems biology: from tools to integrated platforms', Nature Reviews Genetics **12**(12), 821–832.

Gilks, W. R. and Wild, P. (1992), 'Adaptive rejection sampling for gibbs sampling', Applied Statistics **41**, 337–348.

Gillies, R. J., Kinahan, P. E. and Hricak, H. (2015), 'Radiomics: images are more than pictures, they are data', Radiology **278**(2), 563–577.

Girolami, M. and Calderhead, B. (2011), 'Riemann manifold langevin and hamiltonian monte carlo methods', Journal of the Royal Statistical Society: Series B (Statistical Methodology) **73**(2), 123–214.

Glahn, D. C., Paus, T. and Thompson, P. M. (2007), 'Imaging genomics: mapping the influence of genetics on brain structure and function', Human brain mapping **28**(6), 461–463.

Glahn, D. C., Thompson, P. M. and Blangero, J. (2007), 'Neuroimaging endophenotypes: strategies for finding genes influencing brain structure and function', Human brain mapping **28**(6), 488–501.

Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., Ballestar, E., Bongcam-Rudloff, E., Conesa, A. and Tegnr, J. (2014), 'Data integration in the era of omics: current and future challenges', BMC systems biology **8**(2), I1.

Gossl, C., Auer, D. and Fahrmeir, L. (2001), 'Bayesian spatiotemporal inference in functional magnetic resonance imaging', Biometrics **57**, 554–562.

Green, P., Hjort, N. and Richardson, S. (2003), Introducing Highly Structured Stochastic Systems, Oxford Statistical Science Series.

Greenlaw, K., Szefer, E., Graham, J., Lesperance, M., Nathoo, F. S. and Initiative, A. D. N. (2017), 'A bayesian group sparse multi-task regression model for imaging genetics', Bioinformatics **33**(16), 2513–2522.

Grenander, U. and Miller, M. (1994), 'Representations of knowledge in complex systems', Journal of the Royal Statistical Society. Series B **56**(4), 549–603.

Gromski, P. S., Xu, Y., Kotze, H. L., Correa, E., Ellis, D. I., Armitage, E. G., Turner, M. L. and Goodacre, R. (2014), 'Influence of missing values substitutes on multivariate analysis of metabolomics data', Metabolites **4**(2), 433–452.

Guha, S., Banerjee, S., Gu, C. and Baladandayuthapani, V. (2015), Nonparametric variable selection, clustering and prediction for large biological datasets, in 'Nonparametric Bayesian Inference in Biostatistics', Springer, pp. 175–192.

Guo, X., Li, H. W., Fei, F., Liu, B., Li, X. F., Yang, H. S., Chen, Z. N. and Xing, J. L. (2015), 'Genetic variations in slc3a2/cd98 gene as prognosis predictors in non-small cell lung cancer', Molecular carcinogenesis **54 Suppl 1**, E52–60.

Habes, M., Erus, G., Toledo, J. B., Zhang, T., Bryan, N., Launer, L. J., Rosseel, Y., Janowitz, D., Doshi, J. and Van der Auwera, S. (2016), 'White matter hyperintensities and imaging patterns of brain ageing in the general population', Brain **139**(4), 1164–1179.

Hamilton, G., Harris, S. E., Davies, G., Liewald, D. C., Tenesa, A., Starr, J. M., Porteous, D. and Deary, I. J. (2011), 'Alzheimer's disease genes are associated with measures of cognitive ageing in the lothian birth cohorts of 1921 and 1936', International journal of Alzheimers disease **2011**.

Hanley, J. A. and McNeil, B. J. (1982), 'The meaning and use of the area under a receiver operating characteristic (roc) curve.', Radiology **143**(1), 29–36.

Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., Klotzle, B., Bibikova, M., Fan, J.-B., Gao, Y. et al. (2013), 'Genome-wide methylation profiles reveal quantitative views of human aging rates', Molecular cell **49**(2), 359–367.

Hao, X., Yu, J. and Zhang, D. (n.d.), Identifying genetic associations with mri-derived measures via tree-guided sparse learning, in 'MICCAI (2)', pp. 757–764.

Harati Nejad Torbati, A. H. and Picone, J. (2016), 'A Doubly Hierarchical Dirichlet Process Hidden Markov Model with a Non-Ergodic Structure', IEEE/ACM Trans. Audio, Speech, Lang. Process. **24**(1), 174–184.
URL: *http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=7328698*

Hardoon, D. R., Ettinger, U., Mouro-Miranda, J., Antonova, E., Collier, D., Kumari, V., Williams, S. C. and Brammer, M. (2009), 'Correlation-based multivariate analysis of genetic influence on brain volume', Neuroscience letters **450**(3), 281–286.

Harold, D., Abraham, R., Hollingworth, P., Sims, R., Gerrish, A., Hamshere, M. L., Pahwa, J. S., Moskvina, V., Dowzell, K. and Williams, A. (2009), 'Genome-wide association study identifies variants at clu and picalm associated with alzheimer's disease', Nature genetics **41**(10), 1088–1093.

Hashemi, R. H., Bradley, W. G. J. and Lisanti, C. J. (2004), MRI: The Basics, 2 edn, Lippincott Williams & Wilkins.

Hasin, Y., Seldin, M. and Lusis, A. (2017), 'Multi-omics approaches to disease', Genome biology **18**(1), 83.

Hastie, T., Tibshirani, R. and Friedman, J. (2008), The elements of statistical learning: data mining, inference, and prediction, 2 edn, Springer.

Hastie, T., Tibshirani, R., Sherlock, G., Eisen, M., Brown, P. and Botstein, D. (1999), 'Imputing missing data for gene expression arrays'.

Hecker, M., Lambeck, S., Toepfer, S., Van Someren, E. and Guthke, R. (2009), 'Gene regulatory network inference: data integration in dynamic modelsa review', Biosystems **96**(1), 86–103.

Herholz, K., Salmon, E., Perani, D., Baron, J., Holthoff, V., Frlich, L., Schnknecht, P., Ito, K., Mielke, R. and Kalbe, E. (2002), 'Discrimination between alzheimer dementia and controls by automated analysis of multicenter fdg pet', Neuroimage **17**(1), 302–316.

Hibar, D. P., Stein, J. L., Kohannim, O., Jahanshad, N., Saykin, A. J., Shen, L., Kim, S., Pankratz, N., Foroud, T. and Huentelman, M. J. (2011), 'Voxelwise gene-wide association study (vgenewas): multivariate gene-based association testing in 731 elderly subjects', Neuroimage **56**(4), 1875–1891.

Hjort, N. L. (2003), "Topics in non-parametric Bayesian statistics" in Highly Structured Stochastic Systems edited by P. J. Green, N. L. Hjort and S. Richardson, Oxford University Press, chapter 15, pp. 455–478.

Ho, A. J., Stein, J. L., Hua, X., Lee, S., Hibar, D. P., Leow, A. D., Dinov, I. D., Toga, A. W., Saykin, A. J. and Shen, L. (2010), 'A commonly carried allele of the obesity-related fto gene is associated with reduced brain volume in the healthy elderly', Proceedings of the National Academy of Sciences **107**(18), 8404–8409.

Hoffman, M., Cook, P. and Blei, D. (2008), Data-driven recomposition using the hierarchical dirichlet process hidden markov model, in 'Proc. International Computer Music Conference'.

Holden, M., Deng, S., Wojnowski, L. and Kulle, B. (2008), 'Gsea-snp: applying gene set enrichment analysis to snp data from genome-wide association studies', Bioinformatics **24**(23), 2784–2785.

Holmes, A., Blair, R., Watson, J. and Ford, I. (1996), 'Nonparametric analysis of statistic images from functional mapping experiments', Journal of Cerebral Blood Flow and Metabolism **16**(1), 7–22.

Höogmander, H. and Särkkä, A. (1999), 'Multitype spatial point patterns with hierarchical interactions', Biometrics **55**, 1051–1058.

Hrydziuszko, O. and Viant, M. R. (2012), 'Missing values in mass spectrometry based metabolomics: an undervalued step in the data processing pipeline', Metabolomics **8**(1), 161–174.

Huang, J. F., Li, L., Lian, J. H., Schauer, S., Vesely, P. W., Kratky, D., Hoefler, G. and Lehner, R. (2016), 'Tumor-induced hyperlipidemia contributes to tumor growth', Cell reports **15**, 336–48.

Huang, M., Nichols, T., Huang, C., Yu, Y., Lu, Z., Knickmeyer, R. C., Feng, Q., Zhu, H., Initiative, A. D. N. et al. (2015), 'Fvgwas: Fast voxelwise genome wide association analysis of large-scale imaging genetic data', Neuroimage **118**, 613–627.

Huang, S., Li, J., Sun, L., Ye, J., Fleisher, A., Wu, T., Chen, K., Reiman, E., Initiative, A. D. N. et al. (2010), 'Learning brain connectivity of alzheimer's disease by sparse inverse covariance estimation', NeuroImage **50**(3), 935–949.

Huang, X. and Pan, W. (2003), 'Linear regression and two-class classification with gene expression data', Bioinformatics **19**(16), 2072–2078.

Hutchings, M. (1979), 'Standing crop and pattern in pure stands of mercurialis perennis and rubus fruticosus in mixed deciduous woodland', Oikos **31**, 351–357.

Ickstadt, K. and Wolpert, R. (1999), Spatial regression for marked point processes (with discussion), in J. Bernardo, J. Berger, A. Dawid and A. Smith, eds, 'Bayesian Statistics', Oxford: Oxford University Press, chapter 6, pp. 323–341.

Ikonomovic, M., Klunk, W., Abrahamson, E., Wuu, J., Mathis, C., Scheff, S., Mufson, E. and DeKosky, S. (2011), 'Precuneus amyloid burden is associated with reduced cholinergic activity in alzheimer disease', Neurology **77**(1), 39–47.

Illian, J., Penttinen, A., Stoyan, H. and Stoyan, D. (2008), Statistical Analysis and Modelling of Spatial Point Patterns, John Wiley & Sons.

Ishwaran, H. and James, L. F. (2001), 'Gibbs Sampling Methods for Stick-Breaking Priors', J. Am. Stat. Assoc. **96**(453), 161–173.

Ishwaran, H. and James, L. F. (2002a), 'Approximate Dirichlet Process Computing in Finite Normal Mixtures', J. Comput. Graph. Stat. **11**(3), 508–532.

Ishwaran, H. and James, L. F. (2002b), 'Approximate Dirichlet Process Computing in Finite Normal Mixtures', J. Comput. Graph. Stat. **11**(3), 508–532.
URL: *http://amstat.tandfonline.com/doi/abs/10.1198/106186002411#.VvHsmsdQXlI*

Ishwaran, H. and James, L. F. (2011), 'Gibbs sampling methods for stick-breaking priors', Journal of the American Statistical Association .

Ishwaran, H. and Rao, J. S. (2005a), 'Spike and slab gene selection for multigroup microarray data', Journal of the American Statistical Association **100**(471), 764–780.

Ishwaran, H. and Rao, J. S. (2005b), 'Spike and slab variable selection: frequentist and bayesian strategies', Annals of statistics pp. 730–773.

Ishwaran, H. and Rao, J. S. (2005c), 'Spike and slab variable selection: frequentist and bayesian strategies', Annals of statistics pp. 730–773.

Jack, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P. J., L Whitwell, J., Ward, C. et al. (2008),

'The alzheimer's disease neuroimaging initiative (adni): Mri methods', Journal of Magnetic Resonance Imaging **27**(4), 685–691.

Jacquelot, N., Enot, D. P., Flament, C., Vimond, N., Blattner, C., Pitt, J. M. and Zitvogel, L. (2016), 'Chemokine receptor patterns in lymphocytes mirror metastatic spreading in melanoma', The Journal of clinical investigation **126**, 921–37.

Jafari, P. and Azuaje, F. (2006), 'An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors.', BMC Med. Inform. Decis. Mak. **6**(1), 27.

URL: *http://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-6-27*

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F. and Gerstein, M. (2003), 'A bayesian networks approach for predicting protein-protein interactions from genomic data', Science **302**(5644), 449–453.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. and Barabási, A.-L. (2000), 'The large-scale organization of metabolic networks', Nature **407**(6804), 651–654.

Jezzard, P., Matthews, P. M. and Smith, S. M. (2001), Functional MRI: An Introduction to Methods, Oxford University Press.

Jian, X. (2010), A family-based association study of conduct disorder, Thesis.

Jirapech-Umpai, T. and Aitken, S. (2005), 'Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes', BMC bioinformatics **6**(1), 1.

Johnson, T. D. (2007), 'Analysis of pulsatile hormone concentration profiles with non-constant basal concentration: A bayesian approach', Biometrics **4**(63), 1207–1217.

Jollife, I. (1986), Principal component analysis, New York: Springer Verlag.

Jones, D. P., Park, Y. and Ziegler, T. R. (2012), 'Nutritional metabolomics: progress in addressing complexity in diet and health', Annual review of nutrition **32**, 183.

Joyce, A. R. and Palsson, B. Ø. (2006), 'The model organism as a system: integrating'omics' data sets', Nature Reviews Molecular Cell Biology **7**(3), 198–210.

Kanehisa, M. and Goto, S. (2000), 'Kegg: kyoto encyclopedia of genes and genomes', Nucleic acids research **28**(1), 27–30.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. and Hattori, M. (2004), 'The kegg resource for deciphering the genome', Nucleic acids research **32**(suppl 1), D277–D280.

Kang, J., Johnson, D. T. and Nichols, E. T. (2012), 'Generalized poisson/gamma random field models for multiple spatial point patterns', Working Paper .

Kang, J., Johnson, T. D., Nichols, T. E. and Wager, T. D. (2010), 'Meta analysis of functional neuroimaging data via bayesian spatial point processes', Manuscript .

Kang, J., Johnson, T. D., Nichols, T. E. and Wager, T. D. (2011), 'Meta analysis of functional neuroimaging data via bayesian spatial point processes', Journal of the American Statistical Association **106**(493), 124–134.

Karas, G., Scheltens, P., Rombouts, S., van Schijndel, R., Klein, M., Jones, B., van der Flier, W., Vrenken, H. and Barkhof, F. (2007), 'Precuneus atrophy in early-onset alzheimers disease: a morphometric structural mri study', Neuroradiology **49**(12), 967–976.

Karpievitch, Y., Stanley, J., Taverner, T., Huang, J., Adkins, J. N., Ansong, C., Heffron, F., Metz, T. O., Qian, W.-J., Yoon, H. et al. (2009), 'A statistical frame-

work for protein quantitation in bottom-up ms-based proteomics', Bioinformatics **25**(16), 2028–2034.

Karr, A. (1991), Point Processes and their Statistical Inference, New York: Marcel Dekker.

Katanoda, K., Matsuda, Y. and Sugishita, M. (2002), 'A spatio-temporal regression model for the analysis of functional mri data', NeuroImage **17**, 1415–1428.

Kendziorski, C., Newton, M., Lan, H. and Gould, M. (2003), 'On parametric empirical bayes methods for comparing multiple groups using replicated gene expression profiles', Statistics in medicine **22**(24), 3899–3914.

Kenny, L. C., Broadhurst, D. I., Dunn, W., Brown, M., North, R. A., McCowan, L., Roberts, C., Cooper, G. J., Kell, D. B., Baker, P. N. et al. (2010), 'Robust early pregnancy prediction of later preeclampsia using metabolomic biomarkers', Hypertension **56**(4), 741–749.

Kent, J., Dryden, I. and Anderson, C. (2000), 'Using circulant symmetry to model featureless objects', Biometrika **87**(3), 527–544.

Kessler, N., Neuweger, H., Bonte, A., Langenkämper, G., Niehaus, K., Nattkemper, T. W. and Goesmann, A. (2013), 'Meltdb 2.0-advances of the metabolomics software system', Bioinformatics **29**(19), 2452–9.

Kim, H., Golub, G. H. and Park, H. (2005), 'Missing value estimation for dna microarray gene expression data: local least squares imputation', Bioinformatics **21**(2), 187–198.

Kim, J., Basak, J. M. and Holtzman, D. M. (2009), 'The role of apolipoprotein e in alzheimer's disease', Neuron **63**(3), 287–303.

Kira, K. and Rendell, L. A. (1992), A practical approach to feature selection, in 'Proceedings of the ninth international workshop on Machine learning', pp. 249–256.

Kober, H., Barrett, L. F., Joseph, J., Bliss-Moreau, E., Lindquist, K. and Wager, T. D. (2008a), 'Functional grouping and corticalsubcortical interactions in emotion: A meta-analysis of neuroimaging studies', NeuroImage **42**, 998–1031.

Kober, H., Barrett, L. F., Joseph, J., Bliss-Moreau, E., Lindquist, K. and Wager, T. D. (2008b), 'Functional grouping and corticalsubcortical interactions in emotion: A meta-analysis of neuroimaging studies', NeuroImage **42**, 998–1031.

Kohavi, R. and John, G. H. (1997), 'Wrappers for feature subset selection', Artificial intelligence **97**(1), 273–324.

Kosorok, M. R. (2009), 'On brownian distance covariance and high dimensional data', Ann Appl Stat **3**(4), 1266–1269.

Kramer, P., Westaway, S. K., Zwick, M. and Shervais, S. (2012), Reconstructability analysis of genetic loci associated with alzheimer disease, in 'Soft Computing and Intelligent Systems (SCIS) and 13th International Symposium on Advanced Intelligent Systems (ISIS), 2012 Joint 6th International Conference on', IEEE, pp. 2104–2110.

Kuhl, C., Tautenhahn, R., Böttcher, C., Larson, T. R. and Neumann, S. (2012), 'Camera: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets', Anal Chem **84**(1), 283–9.

Kundu, S. and Kang, J. (2016), 'Semiparametric bayes conditional graphical models for imaging genetics applications', Stat **5**(1), 322–337.

Kuo, L. (1986), 'Computations of Mixtures of Dirichlet Processes', SIAM J. Sci. Stat. Comput. **7**(1), 60–71.
URL: *http://epubs.siam.org/doi/abs/10.1137/0907004*

Kwong, K. K., Belliveau, J. W., Chesler, D. A., Goldberg, I. E., Weisskoff, R. M., Poncelet, B. P., Kennedy, D. N., Hoppel, B. E., Cohen, M. S., Turner, R., Cheng, H., Brady, T. J. and Rosen, B. R. (1992), 'Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields', Proceedings of the National Academy of Sciences **89**, 5675–5679.

Lambert, J.-C., Ibrahim-Verbaas, C. A., Harold, D., Naj, A. C., Sims, R., Bellenguez, C., Jun, G., DeStefano, A. L., Bis, J. C. and Beecham, G. W. (2013), 'Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease', Nature genetics **45**(12), 1452–1458.

Lan, Z., Zhao, Y., Kang, J. and Yu, T. (2016), 'Banff: An r package for bayesian network feature finder', Bioinformatics p. In Press.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. et al. (2001), 'Initial sequencing and analysis of the human genome', Nature **409**(6822), 860–921.

Lao, Z., Shen, D., Xue, Z., Karacali, B., Resnick, S. M. and Davatzikos, C. (2004), 'Morphological classification of brains via high-dimensional shape transformations and machine learning methods', Neuroimage **21**(1), 46–57.

Lardone, R. D., Plaisier, S. B., Navarrete, M. S., Shamonki, J. M., Jalas, J. R., Sieling, P. A. and Lee, D. J. (2016), 'Cross-platform comparison of independent datasets identifies an immune signature associated with improved survival in metastatic melanoma', Oncotarget **7**, 14415–28.

Lawson, A. B., Biggeri, A. B., Boehning, D., Lesaffre, E., Viel, J.-F., Clark, A., Schlattmann, P. and Divino, F. (2000), 'Disease mapping models: an empirical evaluation', Statistics in Medicine **19**, 2217–2241.

Lazar, N. A. (2008), The Statistical Analysis of Functional MRI Data, Springer.

Le Floch, ., Guillemot, V., Frouin, V., Pinel, P., Lalanne, C., Trinchera, L., Tenenhaus, A., Moreno, A., Zilbovicius, M. and Bourgeron, T. (2012), 'Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse partial least squares', Neuroimage **63**(1), 11–24.

Lee, J. H., Cheng, R., Graff-Radford, N., Foroud, T. and Mayeux, R. (2008), 'Analyses of the national institute on aging late-onset alzheimer's disease family study: implication of additional loci', Archives of neurology **65**(11), 1518–1526.

Lee, J., Müller, P., Zhu, Y. and Ji, Y. (2013), 'A nonparametric bayesian model for local clustering with application to proteomics', Journal of the American Statistical Association **108**(503), 775–788.

Lerch, J. P., Pruessner, J., Zijdenbos, A. P., Collins, D. L., Teipel, S. J., Hampel, H. and Evans, A. C. (2008), 'Automated cortical thickness measurements from mri can accurately separate alzheimer's patients from normal elderly controls', Neurobiology of aging **29**(1), 23–30.

Li, C., Huang, R., Ding, Z., Gatenby, J. C., Metaxas, D. N. and Gore, J. C. (2011), 'A level set method for image segmentation in the presence of intensity inhomogeneities with application to mri', IEEE Transactions on Image Processing **20**(7), 2007–2016.

Li, C. and Li, H. (2010), 'Variable selection and regression analysis for graph-structured covariates with an application to genomics', The annals of applied statistics **4**(3), 1498.

Li, C. Y. and Li, H. Z. (2008), 'Network-constrained regularization and variable selection for analysis of genomic data', Bioinformatics **24**, 1175–1182.

Li, F. and Zhang, N. R. (2012), 'Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics', Journal of the American Statistical Association .

Li, S., Park, Y., Duraisingham, S., Strobel, F. H., Khan, N., Soltow, Q. A., Jones, D. P. and Pulendran, B. (2013), 'Predicting network activity from high throughput metabolomics', PLoS Comput Biol **9**(7), e1003123.

Li, S., Shi, F., Pu, F., Li, X., Jiang, T., Xie, S. and Wang, Y. (2007), 'Hippocampal shape analysis of alzheimer disease based on machine learning methods', American Journal of Neuroradiology **28**(7), 1339–1345.

Liang, F. (2010a), 'A double metropolis hastings sampler for spatial models with intractable normalizing constants', Journal of Statistical Computation and Simulation **80**(9), 1007–1022.

Liang, F. (2010b), 'A double metropolis hastings sampler for spatial models with intractable normalizing constants', Journal of Statistical Computation and Simulation **80**, 1007–1022.

Liao, J., Lin, Y., Selvanayagam, Z. E. and Shih, W. J. (2004), 'A mixture model for estimating the local false discovery rate in dna microarray analysis', Bioinformatics **20**(16), 2694–2701.

Lim, C. Y. and Dass, S. C. (2011), 'Assessing fingerprint individuality using epic: A case study in the analysis of spatially dependent marked processes', Technometrics **53**(2), 112–124.

Lin, J.-a., Zhu, H., Knickmeyer, R., Styner, M., Gilmore, J. and Ibrahim, J. G. (2012), 'Projection regression models for multivariate imaging phenotype', Genetic epidemiology **36**(6), 631–641.

Lindon, J. C., Holmes, E. and Nicholson, J. K. (2007), 'Metabonomics in pharmaceutical r & d', Febs Journal **274**(5), 1140–1151.

Lindquist, M. A. (2008), 'The statistical analysis of fmri data', Statistical Science **23**(4), 439–464.

Little, R. J. A. and Rubin, D. B. (2014), Statistical analysis with missing data, number John Wiley & Sons.

Liu, F., Chakraborty, S., Li, F., Liu, Y., Lozano, A. C. et al. (2014), 'Bayesian regularization via graph laplacian', Bayesian Analysis **9**(2), 449–474.

Liu, J. and Calhoun, V. D. (2000), 'A review of multivariate analyses in imaging genetics'.

Liu, J., Pearlson, G., Windemuth, A., Ruano, G., PerroneBizzozero, N. I. and Calhoun, V. (2009), 'Combining fmri and snp data to investigate connections between brain function and genetics using parallel ica', Human brain mapping **30**(1), 241–255.

Liu, W. T., Peng, Y. H. and Tobin, D. J. (2013), 'A new 12-gene diagnostic biomarker signature of melanoma revealed by integrated microarray analysis', PeerJ **1**, e49.

Lloyd, M. C., Szekeres, K., Brown, J. S. and Blanck, G. (2015), 'Class ii transactivator expression in melanoma cells facilitates t-cell engulfment', Anticancer research **35**, 25–9.

Lo, A. Y. (1984), 'On a Class of Bayesian Nonparametric Estimates: I. Density

Estimates', Ann. Stat. **12**(1), 351–357.
URL: *http://projecteuclid.org/euclid.aos/1176346412*

Lotwick, H. W. and Silverman, B. W. (1982), 'Methods for analysing spatial point processes of several types of points', Journal of the Royal Statistical Society, Series B **44**, 406–413.

Luo, J., Wu, M., Gopukumar, D. and Zhao, Y. (2016), 'Big data application in biomedical research and health care: A literature review', Biomedical informatics insights **8**, 1.

Luo, W. L. and Nichols, T. E. (2003), 'Diagnosis and exploration of massively univariate fmri models', NeuroImage **19**(3), 1014–1032.

Ma, S. and Huang, J. (2008), 'Penalized feature selection and classification in bioinformatics', Briefings in bioinformatics **9**(5), 392–403.

Maceachern, S. N. (1994), 'Estimating normal means with a conjugate style dirichlet process prior', Commun. Stat. - Simul. Comput. **23**(3), 727–741.
URL: *http://www.tandfonline.com/doi/abs/10.1080/03610919408813196*

MacEachern, S. N. (1999), Dependent nonparametric processes, in 'Proceedings of the Bayesian Statistical Science Section', American Statistical Association.

MacEachern, S. N. and Müller, P. (1998), 'Estimating mixture of dirichlet process models', Journal of Computational and Graphical Statistics **7**, 223–238.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979), Multivariate Analysis, Academic Press.

Marroquin, J., Arce, E. and Botello, S. (2003), 'Hidden markov measure field models for image segmentation', IEEE Trans. Pattern Anal. Machine Intell. (Special Issue

on Energy Minimization Methods in Computer Vi- sion and Pattern Recognition) **25**, 1380–1387.

Martinet, L., Le Guellec, S., Filleron, T., Lamant, L., Meyer, N., Rochaix, P., Garrido, I. and Girard, J. P. (2012), 'High endothelial venules (hevs) in human melanoma lesions: Major gateways for tumor-infiltrating lymphocytes', Oncoimmunology **1**, 829–839.

Mazziotta, J., Toga, A., Evans, A., Fox, P., Lancaster, J., Zilles, K., Woods, R., Paus, T., Simpson, G., Pike, B., Holmes, C., Collins, L., Thompson, P., MacDonald, D., Iacoboni, M., Schormann, T., Amunts, K., Palomero-Gallagher, N., Geyer, S., Parsons, L., Narr, K., Kabani, N., Le Goualher, G., Boomsma, D., Cannon, T., Kawashima, R. and Mazoyer, B. (2001), 'A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM).', Philosophical transactions of the Royal Society of London. Series B, Biological sciences **356**(1412), 1293–1322.

McElroy, J. J. (2013), Genetics of spontaneous idiopathic preterm birth: exploration of maternal and fetal genomes, Vanderbilt University.

Merelli, I., Pérez-Sánchez, H., Gesing, S. and DAgostino, D. (2014), 'Managing, analysing, and integrating big data in medical bioinformatics: open problems and future perspectives', BioMed research international **2014**.

Meyer-Lindenberg, A. (2010a), 'Imaging genetics of schizophrenia', Dialogues in clinical neuroscience **12**(4), 449.

Meyer-Lindenberg, A. (2010b), 'Imaging genetics of schizophrenia', Dialogues in clinical neuroscience **12**(4), 449.

Meyer-Lindenberg, A. and Weinberger, D. R. (2006), 'Intermediate phenotypes

and genetic mechanisms of psychiatric disorders', Nature Reviews Neuroscience **7**(10), 818–827.

Micheel, C. M., Nass, S. J. and Omenn, G. S. (2012), 'Committee on the review of omics-based tests for predicting patient outcomes in clinical trials'.

Mitchell, T. J. and Beauchamp, J. J. (1988), 'Bayesian variable selection in linear regression', Journal of the American Statistical Association **83**(404), 1023–1032.

Mitra, K., Carvunis, A.-R., Ramesh, S. K. and Ideker, T. (2013), 'Integrative approaches for finding modular structure in biological networks', Nature Reviews Genetics **14**(10), 719–732.

Moe Lee, S., Ran Ju, Y., Choi, B.-Y., Wook Hyeon, J., Sun Park, J., Kyeong Kim, C. and Yeon Kim, S. (2012), 'Genotype patterns and characteristics of prnp in the korean population', Prion **6**(4), 375–382.

Mohammadi, M., Sharifi Noghabi, H., Abed Hodtani, G. and Rajabi Mashhadi, H. (2016), 'Robust and stable gene selection via Maximum-Minimum Correntropy Criterion.', Genomics **107**(2-3), 83–7.
URL: *http://www.sciencedirect.com/science/article/pii/S0888754315300495*

Mller, C., Pijnenburg, Y. A., van der Flier, W. M., Versteeg, A., Tijms, B., de Munck, J. C., Hafkemeijer, A., Rombouts, S. A., van der Grond, J. and van Swieten, J. (2015), 'Alzheimer disease and behavioral variant frontotemporal dementia: automatic classification based on cortical atrophy for single-subject diagnosis', Radiology **279**(3), 838–848.

Nathoo, F. (2016), 'A bayesian group-sparse multi-task regression model for imaging genomics'.

Neal, R. M. (2000), 'Markov chain sampling methods for dirichlet process mixture models', Journal of computational and graphical statistics **9**, 249–265.

Neal, R. M. (2012), 'Markov chain sampling methods for dirichlet process mixture models', J. Comput. Graph. Stat. .

Network, C. G. A. (2015), 'Genomic classification of cutaneous melanoma', Cell **161**, 1681–96.

Neumann, J., von Cramon, D. Y. and Lohmann, G. (2008), 'Model-based clustering of meta-analytic functional imaging data', Human Brain Mapping **29**, 177–192.

Newton, M. A. and Kendziorski, C. (2003), Parametric empirical bayes methods for microarrays, in 'The Analysis of Gene Expression Data', Springer, pp. 254–271.

Nichols, T. E. and Hayasaka, S. (2003a), 'Controlling the familywise error rate in functional neuroimaging: A comparative review', Statistical Methods in Medical Research **12**, 419–446.

Nichols, T. E. and Hayasaka, S. (2003b), 'Controlling the familywise error rate in functional neuroimaging: a comparative review', Statistical Methods in Medical Research **12**(5), 419–446.

Nielsen, F. A. and Hansen, L. K. (2002), 'Modeling of activation data in the brainmap database: detection of outliers', Human Brain Mapping **15**(3), 146–156.

Niemi, A. and Fernández, C. (2010), 'Bayesian spatial point process modeling of line transect data', Journal of Agricultural, Biological, and Environmental Statistics **15**(3), 327–345.

Ning, M., Yang, Y., Zhang, Z., Chen, Z., Zhao, T., Zhang, D., Zhou, D., Xu, J., Liu, Z. and Wang, Y. (2010), 'Amyloid--related genes sorl1 and ace are genetically

associated with risk for late-onset alzheimer disease in the chinese population', Alzheimer Disease & Associated Disorders **24**(4), 390–396.

Oba, S., Sato, M.-a., Takemasa, I., Monden, M., Matsubara, K.-i. and Ishii, S. (2003), 'A bayesian missing value estimation method for gene expression profile data', Bioinformatics **19**(16), 2088–2096.

Ogawa, S. and Lee, T. M. (1992), 'Functional brain imagining with physiologically sensitive image signals', Journal of Magnetic Resonance Imaging **2(P)-WIP(Suppl)**, S22.

Ogawa, S., Lee, T. M., Nayak, A. S. and Glynn, P. (1990), 'Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields', Mangnetic Resonance Medicine **14**, 68–78.

Paisley, J., Wang, C., Blei, D. M. and Jordan, M. I. (2015), 'Nested Hierarchical Dirichlet Processes.', IEEE Trans. Pattern Anal. Mach. Intell. **37**(2), 256–70.
URL: *http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=6802355*

Pan, W. (2003), 'On the use of permutation in and the performance of a class of nonparametric methods to detect differential gene expression', Bioinformatics **19**(11), 1333–1340.

Pan, W., Xie, B. H. and Shen, X. T. (2010), 'Incorporating predictor network in penalized regression with application to microarray data', Biometrics **66**, 474–484.

Patel, A., Rees, S. D., Kelly, M. A., Bain, S. C., Barnett, A. H., Thalitaya, D. and Prasher, V. P. (2011), 'Association of variants within apoe, sorl1, runx1, bace1 and aldh18a1 with dementia in alzheimer's disease in subjects with down syndrome', Neuroscience letters **487**(2), 144–148.

Patel, R. S., Sun, Y. V., Hartiala, J., Veledar, E., Su, S., Sher, S., Liu, Y. X., Rahman, A., Patel, R., Rab, S. T., Vaccarino, V., Zafari, A. M., Samady, H., Tang, W. H. W., Allayee, H., Hazen, S. L. and Quyyumi, A. A. (2012), 'Association of a genetic risk score with prevalent and incident myocardial infarction in subjects undergoing coronary angiography', Circ Cardiovasc Genet **5**(4), 441–9.

Pati, D. and Dunson, D. B. (2014), 'Bayesian closed surface fitting through tensor products', Journal of Machine Learning Research, under revision .

Paul, P., Rouas-Freiss, N., Khalil-Daher, I., Moreau, P., Riteau, B., Le Gal, F. A., Avril, M. F., Dausset, J., Guillet, J. G. and Carosella, E. D. (1998), 'Hla-g expression in melanoma: a way for tumor cells to escape from immunosurveillance', Proceedings of the National Academy of Sciences of the United States of America **95**, 4510–5.

Pelckmans, K., De Brabanter, J., Suykens, J. A. and De Moor, B. (2005), 'Handling missing values in support vector machine classifiers', Neural Networks **18**(5), 684–692.

Penny, W., Trujillo-Barreto, N. and Friston, K. (2005), 'Bayesian fmri time series analysis with spatial priors', NeuroImage **24**, 350–362.

Pereira, F., Mitchell, T. and Botvinick, M. (2009), 'Machine learning classifiers and fmri: A tutorial overview', NeuroImage **45**, S199–S209.

Perman, M., Pitman, J. and Yor, M. (1992), 'Size-biased sampling of Poisson point processes and excursions', Probab. Theory Relat. Fields **92**(1), 21–39.
URL: *http://link.springer.com/10.1007/BF01205234*

Pharoah, P. D., Tsai, Y.-Y., Ramus, S. J., Phelan, C. M., Goode, E. L., Lawrenson, K., Buckley, M., Fridley, B. L., Tyrer, J. P. and Shen, H. (2013), 'Gwas meta-

analysis and replication identifies three new susceptibility loci for ovarian cancer', Nature genetics **45**(4), 362–370.

Phelps, E. A. and LeDoux, J. E. (2005a), 'Contributions of the amygdala to emotion processing: from animal models to human behavior', Neuron **48**(2), 175–187.

Phelps, E. A. and LeDoux, J. E. (2005b), 'Contributions of the amygdala to emotion processing: from animal models to human behavior', Neuron **48**(2), 175–187.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/16242399*

Pitman, J. and Yor, M. (1997), 'The two-parameter poisson-dirichlet distribution derived from a stable subordinator', The Annals of Probability pp. 855–900.

Ploner, A., Calza, S., Gusnanto, A. and Pawitan, Y. (2006), 'Multidimensional local false discovery rate for microarray studies', Bioinformatics **22**(5), 556–565.

Plummer, M., Best, N., Cowles, K. and Vines, K. (2006), 'CODA: Convergence diagnosis and output analysis for MCMC', R News **6**(1), 7–11.
URL: *http://CRAN.R-project.org/doc/Rnews/*

Poldrack, R. (2006), 'Can cognitive processes be inferred from neuroimaging data?', Trends in Cognitive Sciences **10**, 59–63.

Poldrack, R., Halchenko, Y. and Hanson, S. (2009), 'Decoding the large-scale structure of brain function by classifying mental states across individuals', Psychological Science **20**, 1364–1372.

Pollard, K. S., Dudoit, S. and van der Laan, M. J. (2005), Multiple testing procedures: the multtest package and applications to genomics, in 'Bioinformatics and computational biology solutions using R and bioconductor', Springer, pp. 249–271.

Potkin, S. G., Guffanti, G., Lakatos, A., Turner, J. A., Kruggel, F., Fallon, J. H., Saykin, A. J., Orro, A., Lupoli, S. and Salvi, E. (2009), 'Hippocampal atrophy as

a quantitative trait in a genome-wide association study identifying novel susceptibility genes for alzheimer's disease', PloS one **4**(8), e6501.

Potkin, S. G., Turner, J. A., Guffanti, G., Lakatos, A., Torri, F., Keator, D. B. and Macciardi, F. (2009), 'Genome-wide strategies for discovering genetic influences on cognition and cognitive disorders: methodological considerations', Cognitive neuropsychiatry **14**(4-5), 391–418.

Potkin, S. G., Turner, J., Fallon, J., Lakatos, A., Keator, D., Guffanti, G. and Macciardi, F. (2009), 'Gene discovery through imaging genetics: identification of two novel genes associated with schizophrenia', Molecular psychiatry **14**(4), 416–428.

Press, S. J. (1982), Applied Multivariate Analysis, 2 edn, Dover Publications.

Preston, C. J. (1975), 'Spatial birth-and-death processes', Bulletin of the International Statistical Institute **46**, 371–391.

Preston, C. J. (1977), 'Spatial birth-and-death processes', Bulletin of the International Statistical Institute **46**, 371–391.

Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., Sullivan, P. F., Sklar, P., Ruderfer, D. M., McQuillin, A., Morris, D. W. et al. (2009), 'Common polygenic variation contributes to risk of schizophrenia and bipolar disorder', Nature **460**(7256), 748–752.

Qiu, A., Fennema-Notestine, C., Dale, A. M., Miller, M. I. and Initiative, A. D. N. (2009), 'Regional shape abnormalities in mild cognitive impairment and alzheimer's disease', Neuroimage **45**(3), 656–661.

Qiu, A. and Miller, M. I. (2008), 'Multi-structure network shape analysis via normal surface momentum maps', NeuroImage **42**(4), 1430–1438.

Radua, J. and Mataix-Cols, D. (2009), 'Voxel-wise meta-analysis of grey matter changes in obsessive–compulsive disorder', The British Journal of Psychiatry **195**, 393–402.

Rahnenführer, J., Domingues, F. S., Maydt, J. and Lengauer, T. (2004), 'Calculating the statistical significance of changes in pathway activity from gene expression data', Statistical Applications in Genetics and Molecular Biology **3**(1), 1–29.

Raichle, M. (2003), 'Functional brain imaging and human brain function', The Journal of Neuroscience **23**, 3959–3962.

Raichle, M. E. (2006), Functional neuroimaging: A historical and physiological perspective, in R. Cabeza and A. Kingstone, eds, 'Handbook of Functional Neuroimaging of Cognition, 2nd Edition', Cambridge (MA): The MIT Press, chapter 1, pp. 3–20.

Rathore, S., Habes, M., Iftikhar, M. A., Shacklett, A. and Davatzikos, C. (2017), 'A review on neuroimaging-based classification studies and associated feature extraction methods for alzheimer's disease and its prodromal stages', NeuroImage .

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. and Barabási, A.-L. (2002), 'Hierarchical organization of modularity in metabolic networks', science **297**(5586), 1551–1555.

Reitz, C., Brayne, C. and Mayeux, R. (2011), 'Epidemiology of alzheimer disease', Nature Reviews Neurology **7**(3), 137–152.

Reitz, C., Lee, J. H., Rogers, R. S. and Mayeux, R. (2011), 'Impact of genetic variation in sorcs1 on memory retention', PloS one **6**(10), e24588.

Reitz, C., Tosto, G., Vardarajan, B., Rogaeva, E., Ghani, M., Rogers, R., Conrad, C., Haines, J., Pericak-Vance, M. and Fallin, M. (2013), 'Independent and epistatic effects of variants in vps10-d receptors on alzheimer disease risk and processing of the amyloid precursor protein (app)', Translational psychiatry **3**(5), e256.

Rentoft, M., Lindell, K., Tran, P., Chabes, A. L., Buckland, R. J., Watt, D. L., Marjavaara, L., Nilsson, A. K., Melin, B., Trygg, J., Johansson, E. and Chabes, A. (2016), 'Heterozygous colon cancer-associated mutations of samhd1 have functional significance', Proceedings of the National Academy of Sciences of the United States of America **113**, 4723–8.

Ripley, B. D. (1976), 'The second-order analysis of stationary point processes', Journal of Applied Probability **13**, 237–259.

Ripley, B. D. (1977a), 'Modelling spatial patterns (with discussion)', Journal of the Royal Statistical Society Series B **39**, 172–212.

Ripley, B. D. (1977b), 'Modelling spatial patterns (with discussion)', Journal of the Royal Statistical Society, Series B **39**, 172–212.

Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A. and Kim, D. (2015a), 'Methods of integrating data to uncover genotype-phenotype interactions', Nature Reviews Genetics **16**(2), 85–97.

Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A. and Kim, D. (2015b), 'Methods of integrating data to uncover genotype-phenotype interactions', Nature reviews. Genetics **16**(2), 85.

Ročková, V. and George, E. I. (2014), 'Emvs: The em approach to bayesian variable selection', Journal of the American Statistical Association **109**, 828–846.

Rodriguez, A., Dunson, D. B. and Gelfand, A. E. (2008), 'The nested dirichlet processes', Journal of the American Statistical Association **103**(483), 1131–1154.

Rosset, S. and Zhu, J. (2007), 'Piecewise linear regularized solution paths', The Annals of Statistics pp. 1012–1030.

Rubin, D. B. (1976), 'Inference and missing data', Biometrika **63**(3), 581–592.

Ruiz, R., Riquelme, J. C. and Aguilar-Ruiz, J. S. (2006), 'Incremental wrapper-based gene selection from microarray data for cancer classification', Pattern Recognition **39**(12), 2383–2392.

Saeys, Y., Inza, I. and Larrañaga, P. (2007a), 'A review of feature selection techniques in bioinformatics.', Bioinformatics **23**(19), 2507–17.
URL: *http://bioinformatics.oxfordjournals.org/content/23/19/2507.short*

Saeys, Y., Inza, I. and Larrañaga, P. (2007b), 'A review of feature selection techniques in bioinformatics', bioinformatics **23**(19), 2507–2517.

Sajda, P. (2006), 'Machine learning for detection and diagnosis of disease', Annu. Rev. Biomed. Eng. **8**, 537–565.

Salimi-Khorshidi, G., Smith, S. M., Keltner, J. R., Wager, T. D. and Nichols, T. E. (2009), 'Meta-analysis of neuroimaging data: a comparison of image-based and coordinate-based pooling of studies', NeuroImage **45**(3), 810–823.

Salli, E., Visa, A., Aronen, H., Korvenoja, A. and Katila, T. (1999), 'Statistical segmentation of fmri activations using contextual clustering', Med. Image Comp. Comput.-Assist. Intervention pp. 481–488.

Sartor, M. A., Tomlinson, C. R., Wesselkamper, S. C., Sivaganesan, S., Leikauf, G. D. and Medvedovic, M. (2006), 'Intensity-based hierarchical bayes method im-

proves testing for differentially expressed genes in microarray experiments', BMC bioinformatics **7**(1), 1.

Saykin, A. J., Gur, R. C., Gur, R. E., Mozley, P. D., Mozley, L. H., Resnick, S. M., Kester, D. B. and Stafiniak, P. (1991), 'Neuropsychological function in schizophrenia: selective impairment in memory and learning', Archives of general psychiatry **48**(7), 618–624.

Saykin, A. J., Shen, L., Foroud, T. M., Potkin, S. G., Swaminathan, S., Kim, S., Risacher, S. L., Nho, K., Huentelman, M. J., Craig, D. W. et al. (2010), 'Alzheimer's disease neuroimaging initiative biomarkers as quantitative phenotypes: Genetics core aims, progress, and plans', Alzheimer's & Dementia **6**(3), 265–273.

Scheltens, P., Leys, D., Barkhof, F., Huglo, D., Weinstein, H., Vermersch, P., Kuiper, M., Steinling, M., Wolters, E. C. and Valk, J. (1992), 'Atrophy of medial temporal lobes on mri in" probable" alzheimer's disease and normal ageing: diagnostic value and neuropsychological correlates', Journal of Neurology, Neurosurgery & Psychiatry **55**(10), 967–972.

Seab, J., Jagust, W., Wong, S., Roos, M., Reed, B. and Budinger, T. (1988), 'Quantitative nmr measurements of hippocampal atrophy in alzheimer's disease', Magnetic Resonance in Medicine **8**(2), 200–208.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D. and Friedman, N. (2003), 'Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data', Nature genetics **34**(2), 166–176.

Serrano-Pozo, A., Frosch, M. P., Masliah, E. and Hyman, B. T. (2011), 'Neuropathological alterations in alzheimer disease', Cold Spring Harbor perspectives in medicine **1**(1), a006189.

Sethuraman, J. (1994), 'A constructive definition of dirichlet priors', Statistica Sinica **4**(2), 639–650.
URL: *http://www.jstor.org/stable/24305538*

Shay, T. and Kang, J. (2013), 'Immunological genome project and systems immunology', Trends in immunology **34**(12), 602–609.

Shen, K.-k., Fripp, J., Mriaudeau, F., Chtelat, G., Salvado, O., Bourgeat, P. and Initiative, A. D. N. (2012), 'Detecting global and local hippocampal shape changes in alzheimer's disease using statistical shape models', Neuroimage **59**(3), 2155–2166.

Shen, L., Kim, S., Risacher, S. L., Nho, K., Swaminathan, S., West, J. D., Foroud, T., Pankratz, N., Moore, J. H. and Sloan, C. D. (2010), 'Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in mci and ad: A study of the adni cohort', Neuroimage **53**(3), 1051–1063.

Silva, R. R., Jourdan, F., Salvanha, D. M., Letisse, F., Jamin, E. L., Guidetti-Gonzalez, S., Labate, C. A. and Vêncio, R. Z. N. (2014), 'Probmetab: an r package for bayesian probabilistic annotation of lc-ms-based metabolomics', Bioinformatics **30**(9), 1336–7.

Silver, M., Janousova, E., Hua, X., Thompson, P. M., Montana, G. and Initiative, A. D. N. (2012), 'Identification of gene pathways implicated in alzheimer's disease using longitudinal imaging phenotypes with sparse regression', NeuroImage **63**(3), 1681–1694.

Sivachenko, A. Y., Yuryev, A., Daraselia, N. and Mazo, I. (2005), Identifying local gene expression patterns in biomolecular networks, in 'Computational Systems Bioinformatics Conference, 2005. Workshops and Poster Abstracts. IEEE', IEEE, pp. 180–181.

Smith, E. E. and Kosslyn, S. M. (2013), Cognitive Psychology: Pearson New International Edition: Mind and Brain, Pearson Higher Ed.

Smola, A. and Vapnik, V. (1997), 'Support vector regression machines', Advances in neural information processing systems **9**, 155–161.

Solorio-Fernández, S., Carrasco-Ochoa, J. A. and Martínez-Trinidad, J. F. (2016), 'A new hybrid filter–wrapper feature selection method for clustering based on ranking', Neurocomputing **214**, 866–880.

Soufan, O., Kleftogiannis, D., Kalnis, P. and Bajic, V. (2015), 'DWFS: A Wrapper Feature Selection Tool Based on a Parallel Genetic Algorithm', PLoS One .
URL: *http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0117988*

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. and Futcher, B. (1998), 'Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization', Molecular biology of the cell **9**(12), 3273–3297.

Springer, C., Patlak, C., Palyka, I. and Huang, W. (1999), Principles of susceptibility contrast-based functional mri: The sign of the functional mri response, in C. T. W. Moonen and P. A. Bandettini, eds, 'Functional MRI', Springer, chapter 9, pp. 90–102.

Stacklies, W., Redestig, H., Scholz, M., Walther, D. and Selbig, J. (2007), 'pcamethods—a bioconductor package providing pca methods for incomplete data', Bioinformatics **23**(9), 1164–1167.

Stein, J. L., Hua, X., Lee, S., Ho, A. J., Leow, A. D., Toga, A. W., Saykin, A. J., Shen, L., Foroud, T. and Pankratz, N. (2010), 'Voxelwise genome-wide association study (vgwas)', neuroimage **53**(3), 1160–1174.

Stein, J. L., Hua, X., Lee, S., Ho, A. J., Leow, A. D., Toga, A. W., Saykin, A. J., Shen, L., Foroud, T., Pankratz, N. et al. (2010), 'Voxelwise genome-wide association study (vgwas)', neuroimage **53**(3), 1160–1174.

Stein, J. L., Hua, X., Morra, J. H., Lee, S., Hibar, D. P., Ho, A. J., Leow, A. D., Toga, A. W., Sul, J. H. and Kang, H. M. (2010), 'Genome-wide analysis reveals novel genes influencing temporal lobe structure with relevance to neurodegeneration in alzheimer's disease', Neuroimage **51**(2), 542–554.

Stekhoven, D. J. and Bühlmann, P. (2012), 'Missforest—non-parametric missing value imputation for mixed-type data', Bioinformatics **28**(1), 112–118.

Stelling, J., Klamt, S., Bettenbrock, K., Schuster, S. and Gilles, E. D. (2002), 'Metabolic network structure determines key aspects of functionality and regulation', Nature **420**(6912), 190–193.

Stephens, M. (2000), 'Bayesian analysis of mixture models with an unknown number of components- an alternative to reversible jump methods', The Annals of Statistics **28**(1), 40–74.

Stingo, F. C., Chen, Y. A., Tadesse, M. G. and Vannucci, M. (2011), 'Incorporating biological information into linear models: A bayesian approach to the selection of pathways and genes', The annals of applied statistics **5**.

Stingo, F. C. and Vannucci, M. (2011), 'Variable selection for discriminant analysis with markov random field priors for the analysis of microarray data', Bioinformatics **27**, 495–501.

Storey, J. D. (2002a), 'A direct approach to false discovery rates', J. R. Stat. Soc. Ser. B (Statistical Methodol. **64**(3), 479–498.
URL: *http://doi.wiley.com/10.1111/1467-9868.00346*

Storey, J. D. (2002b), 'A direct approach to false discovery rates', Journal of the Royal Statistical Society: Series B (Statistical Methodology) **64**(3), 479–498.

Storey, J. D. and Tibshirani, R. (2003), 'Statistical significance for genomewide studies', Proceedings of the National Academy of Sciences **100**(16), 9440–9445.

Stoyan, D., Kendall, W. S. and Mecke, J. (1995), Stochastic geometry and its applications, second edn, John Wiley & Sons Ltd.

Stoyan, D. and Penttinen, A. (2000), 'Recent applications of point process methods in forestry statistics', Statistical Science **15**(1), 61–78.

Strother, S. C. (2006), 'Evaluating fmri preprocessing pipelines', Engineering in Medicine and Biology Magazine, IEEE **25**(2), 27–41.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. et al. (2005), 'Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles', Proceedings of the National Academy of Sciences **102**(43), 15545–15550.

Svensn, M., Kruggel, F. and von Cramon, D. (2000), 'Probabilistic modeling of single-trial fmri data', IEEE Trans. Med. Imag. **19**, 25–35.

T., T. M. (2015), 'A package for survival analysis in s', version 2.38 .

Tang, X., Qin, Y., Wu, J., Zhang, M., Zhu, W. and Miller, M. I. (2016), 'Shape and diffusion tensor imaging based integrative analysis of the hippocampus and the amygdala in alzheimer's disease', Magnetic resonance imaging **34**(8), 1087–1099.

Taylor, R. C., Patel, A., Panageas, K. S., Busam, K. J. and Brady, M. S. (2007), 'Tumor-infiltrating lymphocytes predict sentinel lymph node positivity in patients

with cutaneous melanoma', Journal of clinical oncology : official journal of the American Society of Clinical Oncology **25**, 869–75.

Taylor, S. L., Ruhaak, L. R., Kelly, K., Weiss, R. H. and Kim, K. (2016), 'Effects of imputation on correlation: implications for analysis of mass spectrometry data from multiple biological matrices', Briefings in bioinformatics p. bbw010.

Teh, Y. W. (2010), 'Dirichlet Process', Encycl. Mach. Learn. pp. 280–287.
URL: *http://dx.doi.org/10.1007/978-0-387-30164-8_219*

Teh, Y. W., Dilan, G. and Ghahramani, Z. (2007), 'Stick-breaking Construction for the Indian Buffet Process', Proc. Elev. Int. Conf. Artif. Intell. Stat. **2**, 556–563.

Teh, Y. W., Jordan, M. I., Beal, M. J. and Blei, D. M. (2006a), 'Hierarchical dirichlet processes', Journal of the American Statistical Association **101**, 1566–1581.

Teh, Y. W., Jordan, M. I., Beal, M. J. and Blei, D. M. (2006b), 'Hierarchical dirichlet processes', Journal of the American Statistical Association **101**(476), 1566–1581.

Teh, Y. W., Jordan, M. I., Beal, M. J. and Blei, D. M. (2006c), 'Hierarchical Dirichlet Processes', J. Am. Stat. Assoc. **101**(476), 1566–1581.
URL: *http://amstat.tandfonline.com/doi/abs/10.1198/016214506000000302#.VvHs28dQXlI*

Thirion, B., Pinel, P., Mriaux, S., Roche, A., Dehaene, S. and Poline, J.-B. (2007), 'Analysis of a large fmri cohort: Statistical and methodological issues for group analyses', Neuroimage **35**(1), 105–120.

Thomas, J. G., Olson, J. M., Tapscott, S. J. and Zhao, L. P. (2001), 'An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles', Genome Research **11**(7), 1227–1236.

Thompson, P. M., Cannon, T. D., Narr, K. L., Van Erp, T., Poutanen, V.-P., Huttunen, M., Lönnqvist, J., Standertskjöld-Nordenstam, C.-G., Kaprio, J., Khaledy,

M. et al. (2001), 'Genetic influences on brain structure', Nature neuroscience **4**(12), 1253–1258.

Thompson, P. M., Martin, N. G. and Wright, M. J. (2010), 'Imaging genomics', Current opinion in neurology **23**(4), 368.

ThorntonWells, T. A., Moore, J. H., Martin, E. R., PericakVance, M. A. and Haines, J. L. (2008), 'Confronting complexity in lateonset alzheimer disease: application of twostage analysis approach addressing heterogeneity and epistasis', Genetic epidemiology **32**(3), 187–203.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', Journal of the Royal Statistical Society. Series B (Methodological) pp. 267–288.

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002), 'Diagnosis of multiple cancer types by shrunken centroids of gene expression', Proceedings of the National Academy of Sciences **99**(10), 6567–6572.

Torgo, L. and Torgo, M. L. (2013), 'Package 'dmwr".

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R. B. (2001), 'Missing value estimation methods for dna microarrays', Bioinformatics **17**(6), 520–525.

Turkeltaub, P. E., Eden, G. F., Jones, K. M. and Zeffiro, T. A. (2002), 'Meta-analysis of the functional neuroanatomy of single-word reading: method and validation', NeuroImage **16**(31), 765–780.

Tusher, V. G., Tibshirani, R. and Chu, G. (2001), 'Significance analysis of microarrays applied to the ionizing radiation response', Proceedings of the National Academy of Sciences **98**(9), 5116–5121.

Tziortzi, A. C., Searle, G. E., Tzimopoulou, S., Salinas, C., Beaver, J. D., Jenkinson, M., Laruelle, M., Rabiner, E. A. and Gunn, R. N. (2010), 'Imaging dopamine receptors in humans with [(11)C]-(+)-PHNO: Dissection of D3 signal and anatomy.', NeuroImage .

Uppal, K., Soltow, Q. A., Strobel, F. H., Pittard, W. S., Gernert, K. M., Yu, T. and Jones, D. P. (2013), 'xmsanalyzer: automated pipeline for improved feature detection and downstream analysis of large-scale, non-targeted metabolomics data', BMC Bioinformatics **14**, 15.

Uppal, K., Walker, D. I. and Jones, D. P. (2017), 'xmsannotator: An r package for network-based annotation of high-resolution metabolomics data', Anal Chem **89**(2), 1063–1067.

van Lieshout, M. and Baddeley, A. J. (2001), Extrapolating and interpolating spatial patterns, in 'In Spatial cluster modelling, A.B. Lawson and D.G.T. Denison (Eds.) Boca Raton: Chapman and Hall/CRC', Press, pp. 61–86.

van Lieshout, M. N. M. and Baddeley, A. J. (1999), 'Standing crop and pattern in pure stands of mercurialis perennis and rubus fruticosus in mixed deciduous woodland', Scandinavian Journal of Statistics **26**, 511–532.

van Lieshout, M. N. M. and Baddeley, A. J. (2002), Extrapolating and interpolating spatial patterns, in A. B. Lawson and D. G. T. Denison, eds, 'Spatial Cluster Modelling', Chapman & Hall/CRC, chapter 4, pp. 61–86.

Vehtari, A. and Lampinen, J. (2002), 'Bayesian model assessment and comparison using cross-validation predictive densities', Neural Computation **14**, 2439–2468.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A. et al. (2001), 'The sequence of the human genome', science **291**(5507), 1304–1351.

Vounou, M., Nichols, T. E., Montana, G. and Initiative, A. D. N. (2010), 'Discovering genetic associations with high-dimensional neuroimaging phenotypes: a sparse reduced-rank regression approach', Neuroimage **53**(3), 1147–1159.

Vounou, M., Nichols, T. E., Montana, G., Initiative, A. D. N. et al. (2010), 'Discovering genetic associations with high-dimensional neuroimaging phenotypes: a sparse reduced-rank regression approach', Neuroimage **53**(3), 1147–1159.

Vul, E., Harris, C., Winkielman, P. and Pashler, H. (2009), 'Puzzlingly high correlations in fmri studies of emotion, personality, and social cognition', Perspectives on Psychological Science **4**(3), 274–290.

Wager, T. D., Hernandez, L., Jonides, J. and Lindquist, M. (2007), Elements of functional neuroimaging, in J. T. Cacioppo, L. G. Tassinary and G. G. Berntson, eds, 'Handbook of Psychophysiology', 4 edn, Cambridge University Press, pp. 19–55.

Wager, T. D., Jonides, J. and Reading, S. (2004), 'Neuroimaging studies of shifting attention: a meta-analysis', NeuroImage **22**(4), 1679–1693.

Wager, T. D., Lindquist, M. A., Nichols, T. E., Kober, H. and van Snellenberg, J. X. (2009a), 'Evaluating the consistency and specificity of neuroimaging data using meta-analysis', NeuroImage **45**(1), 210–221.

Wager, T. D., Lindquist, M. A., Nichols, T. E., Kober, H. and van Snellenberg, J. X. (2009b), 'Evaluating the consistency and specificity of neuroimaging data using meta-analysis', NeuroImage **45**(1), 210–221.

Wager, T. D., Lindquist, M. and Kaplan, L. (2007a), 'Meta-analysis of functional neuroimaging data: current and future directions', Social Cognitive and Affective Neuroscience **2**, 150–158.

Wager, T. D., Lindquist, M. and Kaplan, L. (2007b), 'Meta-analysis of functional neuroimaging data: current and future directions', Social Cognitive and Affective Neuroscience **2**(2), 150–158.

Wagner, A. and Fell, D. A. (2001), 'The small world inside large metabolic networks', Proceedings of the Royal Society of London B: Biological Sciences **268**(1478), 1803–1810.

Wan, J., Kim, S., Inlow, M., Nho, K., Swaminathan, S., Risacher, S. L., Fang, S., Weiner, M. W., Beg, M. F. and Wang, L. (n.d.), Hippocampal surface mapping of genetic risk factors in ad via sparse learning models, in 'International Conference on Medical Image Computing and Computer-Assisted Intervention', Springer, pp. 376–383.

Wang, H., Nie, F., Huang, H., Kim, S., Nho, K., Risacher, S. L., Saykin, A. J. and Shen, L. (2011), 'Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the adni cohort', Bioinformatics **28**(2), 229–237.

Wang, H., Nie, F., Huang, H., Kim, S., Nho, K., Risacher, S. L., Saykin, A. J., Shen, L. and Initiative, A. D. N. (2011), 'Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the adni cohort', Bioinformatics **28**(2), 229–237.

Wang, L., Beg, F., Ratnanather, T., Ceritoglu, C., Younes, L., Morris, J. C., Csernansky, J. G. and Miller, M. I. (2007), 'Large deformation diffeomorphism and momentum based hippocampal shape discrimination in dementia of the alzheimer type', IEEE transactions on medical imaging **26**(4), 462–470.

Wang, Z., Gerstein, M. and Snyder, M. (2009), 'Rna-seq: a revolutionary tool for transcriptomics', Nature Reviews Genetics **10**(1), 57–63.

Warnecke-Eberz, U., Metzger, R., Hölscher, A. H., Drebber, U. and Bollschweiler, E. (2016), 'Diagnostic marker signature for esophageal cancer from transcriptome analysis', Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine **37**, 6349–58.

Wei, P. and Pan, W. (2008), 'Incorporating gene networks into statistical tests for genomic data via a spatially correlated mixture model', Bioinformatics **24**(3), 404–411.

Wei, P. and Pan, W. (2010), 'Network-based genomic discovery: application and comparison of markov random-field models', Journal of the Royal Statistical Society: Series C (Applied Statistics) **59**(1), 105–125.

Wei, Z. and Li, H. (2008), 'A hidden spatial-temporal markov random field model for network-based analysis of time course gene expression data', The Annals of Applied Statistics pp. 408–429.

Wei, Z. and Li, H. Z. (2007), 'A markov random field model for network-based analysis of genomic data', Bioinformatics **23**, 1537–1544.

Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., Harvey, D., Jack, C. R., Jagust, W. and Liu, E. (2013), 'The alzheimer's disease neuroimaging initiative: a review of papers published since its inception', Alzheimer's & Dementia **9**(5), e111–e194.

Wennerberg, E., Kremer, V., Childs, R. and Lundqvist, A. (2015), 'Cxcl10-induced migration of adoptively transferred human natural killer cells toward solid tumors causes regression of tumor growth in vivo', Cancer immunology, immunotherapy : CII **64**, 225–35.

Wiesinger, M., Haiduk, M., Behr, M., de Abreu Madeira, H. L., Glöckler, G., Perco,

P. and Lukas, A. (2011), 'Data and knowledge management in cross-omics research projects', Bioinformatics for Omics Data: Methods and Protocols pp. 97–111.

Wills, A. P. and Hector, L. G. (1924), 'The magnetic susceptibility of oxygen, hydrogen and helium', Physical Review **23**(2), 209–220.

Winter, C., Kristiansen, G., Kersting, S., Roy, J., Aust, D., Knösel, T., Rümmele, P., Jahnke, B., Hentrich, V., Rückert, F., Niedergethmann, M., Weichert, W., Bahra, M., Schlitt, H. J., Settmacher, U., Friess, H., Büchler, M., Saeger, H.-D., Schroeder, M., Pilarsky, C. and Grützmann, R. (2012), 'Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes.', PLoS Comput. Biol. **8**(5), e1002511.
URL: *http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002511*

Wolpert, R. L. and Ickstadt, K. (1998a), 'Poisson/gamma random field models for spatial statistics', Biometrika **85**(2), 251–267.

Wolpert, R. L. and Ickstadt, K. (1998b), Simulation of lévy random fields, in D. Dey, P. Müller and D. Sinha, eds, 'Practical Nonparametric and Semiparametric Bayesian Statistics', New York: Springer Verlag, pp. 227–242.

Wood, A. T. A. and Chan, G. (1994a), 'Simulation of stationary gaussian processes in $[0, 1]^d$', Journal of Computational and Graphical Statistics **3**, 409–432.

Wood, A. T. A. and Chan, G. (1994b), 'Simulation of stationary gaussian processes in $[0, 1]^d$', Journal of Computational and Graphical Statistics **3**, 409–432.

Woodard, D. B., Wolpert, R. L. and OConnell, M. A. (2010), 'Spatial inference of nitrate concentrations in groundwater', Journal of Agricultural, Biological, and Environmental Statistics **15**(2), 209–227.

Woolrich, M. W., Behrens, T. E. J., Beckmann, C. F. and Smith, S. M. (2005a), 'Mixture models with adaptive spatial regularisa- tion for segmentation with an application to fmri data', Biometrika **24**, 1–11.

Woolrich, M. W., Behrens, T. E. J., Beckmann, C. F. and Smith, S. M. (2005b), 'Mixture models with adaptive spatial regularisation for segmentation with an application to fmri data', IEEE Transactions on Medical Imaging **24**, 1–11.

Xia, J., Psychogios, N., Young, N. and Wishart, D. S. (2009), 'Metaboanalyst: a web server for metabolomic data analysis and interpretation', Nucleic acids research **37**(suppl 2), W652–W660.

Xia, J. and Wishart, D. S. (2010), 'Metpa: a web-based metabolomics tool for pathway analysis and visualization', Bioinformatics **26**(18), 2342–4.

Xing, E. P. and Sohn, K. A. (2007), 'Hidden markov dirichlet process: Modeling genetic inference in open ancestral space', Bayesian Analysis **2**(3), 501–528.

Xiong, M., Fang, X. and Zhao, J. (2001), 'Biomarker identification by feature wrappers', Genome Research **11**(11), 1878–1887.

Yadrenko, M. I. (1983), Spectral theory of random fields, Optimization Software.

Yan, W. H., Lin, A. F., Chang, C. C. and Ferrone, S. (2005), 'Induction of hla-g expression in a melanoma cell line ocm-1a following the treatment with 5-aza-2'-deoxycytidine', Cell research **15**, 523–31.

Yarkoni, T. (2009), 'Big correlations in little studies: Inflated fmri correlations reflect low statistical powercommentary on vul et al. (2009)', Perspective on Psychological Science **4**, 294–298.

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C. and Wager,

T. D. (2011), 'Large-scale lexical decoding of human brain activity', Unpublished Manuscript .

Yarkoni, T., Poldrack, R. A., Van Essen, D. C. and Wager, T. D. (2010), 'Cognitive neuroscience 2.0: building a cumulative science of human brain function', Trends in Cognitive Sciences **14**(11), 489–496.

Yeung, K. Y., Bumgarner, R. E. and Raftery, A. E. (2005), 'Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data', Bioinformatics **21**(10), 2394–2402.

Yu, T., Park, Y., Johnson, J. M. and Jones, D. P. (2009), 'aplcms–adaptive processing of high-resolution lc/ms data', Bioinformatics **25**(15), 1930–6.

Yu, T., Park, Y., Li, S. and Jones, D. P. (2013), 'Hybrid feature detection and information accumulation using high-resolution lc-ms metabolomics data', J Proteome Res **12**(3), 1419–27.

Yu, T. and Peng, H. (2013), 'Hierarchical clustering of high-throughput expression data based on general dependences', IEEE/ACM Trans Comput Biol Bioinform **10**(4), 1080–5.

Yu, T., Peng, H. and Sun, W. (2011), 'Incorporating nonlinear relationships in microarray missing value imputation', IEEE/ACM Trans Comput Biol Bioinform **8**(3), 723–31.

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

Zhang, J. F., Chen, Y., Lin, G. S., Zhang, J. D., Tang, W. L., Huang, J. H., Chen, J. S., Wang, X. F. and Lin, Z. X. (2016), 'High ifit1 expression predicts improved

clinical outcome, and ifit1 along with mgmt more accurately predicts prognosis in newly diagnosed glioblastoma', Human pathology **52**, 136–44.

Zhang, W., Li, F. and Nie, L. (2010), 'Integrating multiple omics analysis for microbial biology: application and methodologies', Microbiology **156**(2), 287–301.

Zhao, Y., Kang, J. and Yu, T. W. (2014), 'A bayesian nonparametric mixture model for selecting genes and gene subnetworks', The annals of applied statistics **8**, 999.

Zhou, B., Xiao, J. F., Tuli, L. and Ressom, H. W. (2012), 'Lc-ms-based metabolomics', Molecular BioSystems **8**(2), 470–481.

Zhou, W., Feng, X. L., Li, H., Wang, L., Li, H., Zhu, B., Zhang, H. J., Yao, K. T. and Ren, C. P. (2007), 'Functional evidence for a nasopharyngeal carcinoma-related gene bcat1 located at 12p12', Oncology research **16**, 405–13.

Zhu, H., Khondker, Z., Lu, Z. and Ibrahim, J. G. (2014a), 'Bayesian generalized low rank regression models for neuroimaging phenotypes and genetic markers', Journal of the American Statistical Association **109**(507), 977–990.

Zhu, H., Khondker, Z., Lu, Z. and Ibrahim, J. G. (2014b), 'Bayesian generalized low rank regression models for neuroimaging phenotypes and genetic markers', Journal of the American Statistical Association **109**(507), 977–990.

Zhu, H., Williams, C. K., Rohwer, R. and Morciniec, M. (1998), 'Gaussian regression and optimal finite dimensional linear models'.

Zlotnik, A. and Yoshie, O. (2012), 'The chemokine superfamily revisited', Immunity **36**, 705–16.

Zou, H. (2006), 'The adaptive lasso and its oracle properties', Journal of the American statistical association **101**(476), 1418–1429.

Zou, Q., Zeng, J., Cao, L. and Ji, R. (2016), 'A novel features ranking metric with application to scalable visual and bioinformatics data classification', Neurocomputing **173**, 346–354.

URL: *http://www.sciencedirect.com/science/article/pii/S0925231215012801*