Statistical Methods for Data Integration and Disease Classification

STATISTICAL METHODS FOR DATA INTEGRATION AND DISEASE CLASSIFICATION

BY

MOHAMMAD SHOFIQUL ISLAM,

B.Sc. (Honours Statistics), M.Sc. (Statistics)

A THESIS

SUBMITTED TO THE DEPARTMENT OF

HEALTH RESEARCH METHODS, EVIDENCE, AND IMPACT

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Doctor of Philosophy (2017) McMaster University
Health Research Methods, Evidence, and Impact Hamilton, Ontario, Canada

TITLE:	Statistical Methods for Data Integration and Disease Classification
AUTHOR:	Mohammad Shofiqul Islam B.Sc. (Honours), University of Rajshahi, Bangladesh M.Sc. (Statistics), University of Rajshahi, Bangladesh M.Sc. (Statistics), University of Calgary, Canada M.Sc. (Biostatistics), University of Alberta, Canada
SUPERVISOR:	Dr. Joseph Beyene
NUMBER OF PAGES:	xviii, 110

Dedicated to my wife Rutaba and daughter Tasfiah who encouraged and supported me at each and every single step of this journey.

Abstract

Classifying individuals into binary disease categories can be challenging due to complex relationships across different exposures of interest. In this thesis, we investigate three different approaches for disease classification using multiple biomarkers. First, we consider the problem of combining information from literature reviews and a real data set to determine the threshold of a biomarker for disease classification. We develop a Bayesian estimation procedure for this purpose that utilizes the conditional probability distribution of the biomarker. This method is flexible compared to the standard logistic regression approach and allows us to identify a more precise threshold of a biomarker. For example, higher levels of Apolipoprotein B and lower levels of Apolipoprotein A1 are well-known risk factors for myocardial infarction (MI), but the threshold at which maximum classification accuracy occurs is not clear. We illustrate the method finding thresholds of these biomarkers that utilizes the information from literature reviews and a large case-control study data set.

Second, we consider the problem of identifying a joint threshold for binary disease classification based on two dependent biomarkers. For example, the creatine kinase enzyme and cardiac troponin are often used to classify individuals who are at high risk of developing an acute MI. An independently identified threshold for this purpose usually leads to a conflicting classification for some individuals. Based on the probability distribution function of two dependent biomarkers, we develop and describe a new method for identifying the joint threshold. We also illustrate the method using a real data example. Thresholds determined using this approach may allow clinicians uniquely classify individuals at risk of developing the disease.

Third, we consider the problem of classifying an outcome based on multidimensional complex data sets. For example, gene and miRNA expression data are often used to classify individuals of having cancer. Linear principal component analysis (PCA) is a widely used approach to reduce the dimension of such data sets and utilize the reduced set in a subsequent procedure such as classification or to identify the association between disease and extracted components. Given that there may exist some degree of nonlinearity across variables in these data sets, many authors suggest a nonlinear approach such as kernel PCA for this purpose. However, the performance of kernel PCA over linear PCA in this context has not been well studied. Based on a real and simulated data sets, we compare these two approaches and assess the performance of components towards genetic data integration for an outcome classification. We also develop a simulation algorithm that takes into account the dependency and nonlinearity observed in these data sets. In general, the first few kernel principal components show similar performance compared to the linear principal components in this occasion. Reducing dimensions using linear PCA and a logistic regression model for classification seems acceptable to deal with high-dimensional gene or miRNA expression data sets. We also observe that integrating information from multiple data sets using either of these two approaches lead to a better performance of classification for the outcome of interest.

Acknowledgements

I would like to express my gratitude to Dr. Joseph Beyene for his outstanding supervision throughout the program and progress of this research work. His visionary guidance with friendly discussions allowed me to think through some of the critical steps and work independently to find some innovative ideas for this research. Without his continuous support and advice, this research might not have been possible.

I also would like to express my appreciation to the advisory committee members Dr. Sonia Anand, Dr. Jemila Hamid and Dr. Lehana Thabane for their help, support and advice during this journey.

Special thanks to our directors Dr. Salim Yusuf, Dr. Sonia Anand and Dr. Janice Pogue at the Population Health Research Institute, McMaster University, for their encouragement, and giving me the opportunity to pursue this research. This work might not have been possible without their advice and support.

A note of thanks to my former supervisor Professor MA Razzaque, Department of Statistics, University of Rajshahi, Bangladesh, who taught me how to start walking in the area of statistical research.

Last but not least, I owe my gratitude to my parents, wife, daughter, friends and colleagues for their care, kindness, encouragement and continuous support during the preparation of this thesis.

Declaration of Academic Achievements

Shofiqul Islam conceived the idea of determining the joint threshold for disease classification using conditional probability distribution of dependent biomarkers. Shofiqul programmed and ran all simulations contained within these papers and drafted the papers. My co-authors helped with setting and refining scenarios for the simulations, participated in drafting and critical revision of the manuscripts.

This dissertation is a sandwich thesis, composed of three papers. The first and second paper(Chapter 2 and Chapter 4) has been published, and the third paper(Chapter 3) has been submitted and received favourable reviews, and revised version is being re-reviewed by the journal.

Notations and Abbreviations

AUC	: Area Under the ROC Curve
CAMDA	: Critical Assessment of Massive Data Analysis
CNV	: Copy Number Variation
CDF	: Cumulative Distribution Function
CER	: Classification Error Rate
GE	: Gene Expression
ICD	: International Statistical Classification of Disease
ICGC	: International Cancer Genome Consortium
KPC	: Principal Component obtained using Kernel Approach
LPC	: Principal Component obtained using Linear Approach
MI	: Myocardial Infarction
miRNA	: Micro RNA
MSE	: Mean Squared Error
PCA	: Principal Component Analysis
PDF	: Probability Density Function
QQ	: Quantile-Quantile Plot
RNA	: Ribonucleic acid
ROC	: Receiver Operating Characteristic Curve
WHO	: World Health Organization

Table of Contents

Chapter 2: A Bayesian Approach of Developing Classification Rules(Continued)

Chapter 3: A Copula-based Method of Classification (Continued)

Chapter 4: Dimension Reduction and Disease Classification (Continued)

List of Tables

Tables: Chapter 2

Tables: Chapter 3

Tables: Chapter 3 (Continued)

Tables: Chapter 4

Tables: Chapter 4 (Supplementary Materials)

List of Figures

Figures: Chapter 2

Figures: Chapter 3

Figures: Chapter 3 (Continued)

Figures: Chapter 4

Figures: Chapter 4 (Supplementary Materials)

Chapter 1

Introduction and Problem Statement

In this chapter, we presented our findings from literature reviews and provided an introduction to the problem we have investigated for data integration and disease classification. We have also provided a brief outline of three research articles presented in this thesis.

Correct classification of a disease is essential to understand the causal pathway to survival or death. Although the concept of disease classification dates back to the early 15th century England, death registration first started in mid-15th century Italy (Moriyama et al., 2011). William Farr first introduced the importance of statistical classification of death in 1839 (Eyler, 1979). This procedure was formalized based on the classification rules developed in the early 19th century through the International Statistical Classification of Disease (ICD) (International Statistical Institute, 1899; US Bureau of the Census, 1901). Since its inception, numerous organizations have used these rules to classify different diseases for statistical or non-statistical purposes. As a result, these rules are expected to be updated based on new evidence.

The World Health Organization (WHO) and its collaborating centers continuously work to update disease classification rules based on revised criteria developed through ongoing research. Currently, researchers are using the 10th revised version of these rules, called ICD-10, which was endorsed by the Forty-Third World Assembly in 1990 (World Health Organization, 1992). The next revision is expected to be released in 2017. In this research, our goal is to shed some light on this area. In particular, we introduce some new ideas to determine the threshold for disease classification, as well as to compare and contrast some existing methods for this purpose. We hope that the findings from this research will be helpful to develop better classification rules for different diseases such as myocardial infarction (MI), cancer or death due to a specific disease.

The first step of disease classification is to identify markers associated with a particular disease. Next, it is necessary to determine the threshold at which these markers contribute to the causal pathway for the disease of interest. Markers obtained from biological specimens are commonly used for this purpose and referred to as biomarkers. Thresholds for different biomarkers are usually determined using a clinical or statistical approach (Vasan, 2006). Assuming that biomarkers are independent, these approaches identify the threshold for one biomarker at a time. However, classification of a disease often depends on the results obtained from multiple biomarkers called tests, and these tests may be interrelated. If the assumption of independence is incorrect, the result of two different dependent tests for a particular disease may lead to a wrong or conflicting classification. Thus, it is essential to develop a method that takes into account the dependency between tests to uniquely classify individuals at risk of developing a particular disease.

Disease-exposure relationships may also suffer from multidimensional complex data structures, and so it is necessary to consider more sophisticated statistical approaches to identify their association or for predicting future outcomes. The recent growth and development of computing and information management systems have allowed scientists to collect and store extensive amounts of data with complex multidimensional structures. As a result, summarizing or extracting information from these large data sets to use in subsequent processes is challenging. The number of tests also grows exponentially with the number of exposures or variables under consideration. For example, the progression of a disease can be related to biological, behavioral or genetic factors and these data sets often consist of many variables. The challenge here is to identify the best strategy for utilizing this information for an outcome classification.

Data integration is a process that allows one to combine information from such data sets. The concept of data integration varies within the context, such as business intelligence (Haque et al., 2014) or life sciences (Gomez-Cabrero et al., 2014) to obtain a meaningful summary of information. In either case, multiple sources of information called domains require data integration to perform a specific task. For example, integrated information can be used to classify different clinical outcomes such as cancer or death. In a recent article, Hamid et al. (2009) provided a conceptual framework for data integration and discussed some methodological challenges in the context of genomic data. In particular, genetic processes, such as the gene or miRNA expression data, appear in a very high dimension with a relatively large number of variables as compared to the number of subjects in the sample. These variables are often highly correlated within and across data sets. Statistical methodology to summarize this information is not well developed.

Due to the multivariate nature of the data, univariate statistical approaches are not optimal, and it is necessary to consider appropriate multivariate methodologies. For example, standard statistical procedures, like linear regression fail to utilize such information for classification, due to a large number of variables with complex relationships. As a result, reducing the dimension or summarizing a multidimensional correlated set of exposures is essential to develop the relationship between disease and risk factors of interest. Thus, integrating data sets in the context of an outcome classification can be accomplished in two steps: first, by reducing the dimension with meaningful features through a suitable statistical technique within a domain; second, by developing models to classify the outcome based on extracted features from the various domains. In the context of genomic data integration, many authors consider such an approach to dimension reduction and then utilize the reduced set to identify the association or disease classification (Chang and Keinan, 2014; Yi et al., 2012).

Depending on the relationship between variables within a domain, the two broad classes of dimension reduction techniques available for use are either the linear or the nonlinear approach. Some of the linear approaches include linear principal component analysis (PCA), latent class analysis (LCA) and canonical correlation analysis (CCA). While PCA and LCA can be used to reduce the dimension of a single set, CCA can be used to reduce dimensions of two correlated sets. The key characteristic of these approaches is to identify a smaller number of latent variables that can be expressed as a linear combination of observed variables with maximum variance or correlation. Similarly, some of the nonlinear approaches include Sammon's mapping, curvilinear component analysis, nonlinear PCA, and kernel PCA. These procedures can be considered as a nonlinear generalization of the standard PCA.

Many authors consider linear PCA to reduce the dimension of gene expression data and subsequently utilize the information to quantify the degree of association between disease and the extracted principal components (Chang and Keinan, 2014; Yi et al., 2012). This method was also used to identify a cluster of associated genes (Yeung, 2001), to correct for population stratification in genome-wide association studies (Price et al., 2006), or to predict an outcome based on the different types of clinical variables (Ahmadi et al., 2013; Korkeila et al., 2011). Estimation and test results related to this method depend on the linearity and multivariate normality assumption. However, gene and miRNA expression data often fail to satisfy these assumptions; as a result, nonlinear dimension reduction techniques may be optimal in this situation.

In the context of dimension reduction and pattern recognition, Schölkopf et al. (1998) suggested that pre-processing data using kernel PCA could improve the classification performance.

For example, this approach performs well for character or face recognition. The author also showed that a linear classifier is sufficient in this case, as long as features are extracted using the nonlinear approach. Recently, many authors also proposed kernel PCA to reduce the dimension of a genetic process (Gao et al., 2011; Liu et al., 2005), but this procedure requires identifying a suitable kernel for this computation. Based on the analysis of several data sets using different kernel PCA and logistic regression for classification, Liu et al. (2005) suggest that a polynomial kernel with a degree of two or three performs better to reduce the dimension of gene expression data. However, the performance of kernel PCA over linear PCA in the context of data integration and disease classification is yet to be explored and justified.

In this thesis, we investigate some of the issues discussed above, and a brief outline is given below. In Chapter 2, we develop and describe a Bayesian approach to determine the threshold of a biomarker for disease classification. We illustrate that this method utilizes information from literature reviews of selected biomarkers and a real data set. In particular, we consider the problem of classifying myocardial infarction (MI) based on Apolipoprotein B (ApoB), Apolipoprotein A1 (ApoA1) and the ratio of these two biomarkers. Higher levels of ApoB and lower levels of ApoA1 are well-known risk factors for MI. In a recent epidemiological study, called INTERHEART, elevated ApoB to ApoA1 ratio appeared to be the most influential predictor for MI (Yusuf et al., 2004). However, the thresholds at which these biomarkers become a risk for MI are not clear. Applying the method developed, we determine the threshold of these biomarkers for the classification of MI. During this process, we first construct prior distributions for location and scale parameters utilizing information from literature reviews, and then develop classification rules based on the posterior distribution for each of the biomarkers. We also use the classical and Bayesian approaches to identify the most informative predictor for MI among the three, as well as estimate the odds ratio for one standard deviation change in each of these risk factors.

In Chapter 3, we consider the problem of classifying disease based on two dependent biomarkers, where we develop a new method of identifying the joint threshold. This threshold allows one to classify uniquely patients into a binary disease group. However, this method requires constructing a joint probability distribution for these biomarkers. We use Frank's, Clayton's and Gumbel's copula for this purpose and construct the joint probability distribution with gamma marginals. Based on the joint probability distribution constructed through copula, we develop the method of classifying patients into binary disease categories, which takes into account the dependency between biomarkers and leads to a unique classification. We also develop a simulation algorithm for this purpose and conduct the simulation study in two steps. In the first step, we evaluate the performance of the joint probability distribution based on the relative bias and mean squared error for all parameters of interest. In the second step, we utilize the joint probability density function to determine the joint threshold of creatine kinase and cardiac troponin for acute MI classification, using a set of parameters identified through literature reviews. We also assess the classification accuracy of the method across different choices of copulas based on the empirical distribution of the area under the receiver operating characteristic curve. Finally, we illustrate the method with an example using a sub-set of the INTERHEART study data for MI. In this example, we demonstrate how to determine the joint threshold of Apolipoprotein B to Apolipoprotein A1 ratio and total cholesterol to high-density lipoprotein ratio for the classification of MI.

In Chapter 4, we consider the problem of data integration and an outcome classification based on high dimensional genomic data sets. In this chapter, we develop a copula-based simulation algorithm that takes into account the degree of dependence and nonlinearity observed in the expression data. Based on the algorithm developed, we conduct a simulation study to compare the performance of linear and kernel approaches in different scenarios. Subsequently, we demonstrate the data integration procedure using a real data set obtained from the international cancer genome consortium data portal (Zhang et al., 2011). During this process, we also compare the linear and kernel principal components for reducing the dimension of larger sets, and their performance towards data integration and death classification based on logistic regression models. We use percent of variance explained by the top three principal components, the classification error rate and the area under the receiver operating characteristic curve for this comparison.

Thus, three research projects we present in Chapter 2-4 describe methods for disease classification using real data examples and simulation. We consider the second project as a bivariate extension of the first project. In the third project, we compare methods dealing with multidimensional complex data sets for disease classification. Overall, we present these three chapters as a univariate, a bivariate and multivariate statistical methods for data integration and disease classification. Lastly, in Chapter 5 we present an overview of our main findings, recommendations for future research, and concluding remarks.

References

Ahmadi, H., Mitra, A. P., Abdelsayed, G. A., Cai, J., Djaladat, H., Bruins, H. M., and Daneshmand, S. (2013). Principal component analysis based pre-cystectomy model to predict pathological stage in patients with clinical organ-confined bladder cancer. *BJU International*, 111(4 Pt B):E173.

Chang, D. and Keinan, A. (2014). Principal component analysis characterizes shared pathogenetics from genome-wide association studies. *PLoS Computational Biology*, 10(9):e1003820.

Eyler, J. (1979). *Victorian social medicine: The ideas and methods of William Farr*. Baltimore, MD: Johns Hopkins University Press.

Gao, Q., He, Y., Yuan, Z., Zhao, J., Zhang, B., and Xue, F. (2011). Gene- or region-based association study via kernel principal component analysis. *BMC Genetics*, 12(1):75.

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

Hamid, J. S., Hu, P., Roslin, N. M., Ling, V., Greenwood, C. M. T., and Beyene, J. (2009). Data integration in genetics and genomics: methods and challenges. *Human Genomics and Proteomics*, 8690(1):1–13.

Haque, W., Urquhart, B., Berg, E., and Dhanoa, R. (2014). Using business intelligence to analyze and share health system infrastructure data in a rural health authority. *JMIR Medical Informatics*, 2(2):e16.

International Statistical Institute (1899). International Statistical Institute Meeting in Kristiana. In *Proceedings of the Meeting in Kristiana*, pages Vol II, Part I.

Korkeila, E. A., Sundstrom, J., Pyrhonen, S., and Syrjanen, K. (2011). Carbonic anhydrase IX, hypoxia-inducible factor-1alpha, ezrin and glucose transporter-1 as predictors of disease outcome in rectal cancer: multivariate Cox survival models following data reduction by principal component analysis of the clinicopathological. *Anticancer Research*, 31(12):4529–4535.

Liu, Z., Chen, D., and Bensmail, H. (2005). Gene expression data classification with kernel principal component analysis. *Journal of Biomedicine & Biotechnology*, 2005(2):155–159.

Moriyama, I. M., Loy, R. M., and Robb-Smith, A. H. (2011). *History of the Statistical Classification of Diseases and Causes of Death.* Rosenberg, H. M. and Hoyert, D. L. (Eds.). Hyattsville, MD: National Center for Health Statistics.

Price, A., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909.

Schölkopf, B., Smola, A., and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319.

US Bureau of the Census (1901). *Manual of international classification of causes of death. Adopted by the U.S. Census Office for the compilation of mortality statistics, for use beginning with the year 1900.* Washington, DC: Government Printing Office.

Vasan, R. S. (2006). Biomarkers of cardiovascular disease: Molecular basis and practical considerations. *Circulation*, 113(19):2335–2362.

World Health Organization (1992). International statistical classification of diseases and related health problems. *Occupational Health*, 41:1–201.

Yeung, K. Y. (2001). Principal component analysis for clustering gene expression data. *Bioinformatics*, 21(13):3009–3016.

Yi, H., Wo, H., Zhao, Y., Zhang, R., Bai, J., Wei, Y., and Chen, F. (2012). Gene-based principal component logistic regression model and its application on genome-wide association study. *Zhonghua liu xing bing xue za zhi*, 33(6):622–5.

Yusuf, S., Hawken, S., Ôunpuu, S., Dans, T., Avezum, A., Lanas, F., Mcqueen, M., Budaj, A., Pais, P., Ounpuu, S., Varigos, J., and Lisheng, L. (2004). Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): Case-control study. *The Lancet*, 364(9438):937–952.

Zhang, J., Baran, J., Cros, A., Guberman, J. M., Haider, S., Hsu, J., Liang, Y., Rivkin, E., Wang, J., Whitty, B., Wong-Erasmus, M., Yao, L., and Kasprzyk, A. (2011). International cancer genome consortium data portal: A one-stop shop for cancer genomics data. *Database*, 2011(1):1–10.

Chapter 2

A Bayesian Approach of Developing Classification Rules

The threshold of a biomarker for disease classification is usually determined using a classical approach. This approach does not allow one to use pre-existing evidence in the literature. This chapter includes an article published in the "Journal of Applied Statistics," where we developed a Bayesian approach to determine the threshold of a biomarker for disease classification. We illustrated this method utilizing information from literature reviews of selected biomarkers and a large case-control study data set. The approach developed in this article can be used to determine the threshold of any continuous biomarker for a binary disease classification.

Citation: S. Islam, S. Anand, M. McQueen, J. Hamid, L. Thabane, S. Yusuf & J. Beyene (2016): Classification rules for identifying individuals at high risk of developing myocardial infarction based on ApoB, ApoA1 and the ratio were determined using a Bayesian approach, Journal of Applied Statistics, DOI: 10.1080/02664763.2016.1270912.

Classification rules for identifying individuals at high risk of developing myocardial infarction based on ApoB, ApoA1 and the ratio were determined using a Bayesian approach

S. Islam[†§], S. Anand[†§¶], M. McQueen [†||], J. Hamid [§], L. Thabane [†§], S. Yusuf [†§¶] and J. Beyene[§*]

Abstract: **We have developed a new approach to determine the threshold of a biomarker that maximizes the classification accuracy of a disease, based on the conditional distribution of a biomarker. We consider a Bayesian estimation procedure for this purpose and illustrate the method using a real data set. In particular, we identify the threshold for Apolipoprotein B (ApoB), Apolipoprotein A1 (ApoA1) and the ratio of these two biomarkers for the classification of myocardial infarction (MI). We first conduct literature reviews and construct prior distributions for the location parameters of the distribution of these biomarkers. We then develop the classification rules based on the posterior distribution of the location and scale parameters for each of these biomarkers. During this process, we compute the posterior median and 95% credible interval for the location parameter of ApoB, ApoA1, and the ratio as 0.972 (0.966, 0.978), 1.119 (1.114, 1.124) and 0.903 (0.896, 0.906), respectively for cases. Finally, we identify the threshold for ApoB, ApoA1, and the ratio as 0.908 (gram/liter), 1.138 (gram/liter) and 0.808, respectively. We also observe that the threshold for disease classification varies slightly across different age and ethnic groups. In the next step, we identify the most informative predictor for MI among the three biomarkers. Based on this analysis, ApoA1 appeared to be a stronger predictor than ApoB for MI classification. Given that we have used this data set for illustration only, the results presented will require further investigation for use in clinical applications. However, the approach developed in this article can be used to determine the threshold for any continuous biomarker for a binary disease classification.**

Keywords: Bayesian Approach; Conditional Logistic Regression; ApoB; ApoA1; MI; AUC

[†] Population Health Research Institute, McMaster University and Hamilton Health Sciences, Hamilton, Ontario, Canada.

[§] Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada.

[¶] Department of Medicine, McMaster University, Hamilton, Ontario, Canada.

[||] Department of Pathology and Molecular Medicine, McMaster University, Hamilton, Ontario, Canada.

[*] Correspondence to: Joseph Beyene, Department of Health Research Methods, Evidence, and Impact, McMaster University
1280 Main Street West, Hamilton, ON L8S 4K1; Phone: 905-525-9140, ext. 21333; Email: beyene@mcmaster.ca

1. Introduction

The threshold of a biomarker for a disease classification is usually determined using classical approaches such as maximum likelihood estimator of a logistic regression model (Hosmer et al., 2013). These approaches do not allow one to use pre-existing evidence in the literature that may be useful to identify a more precise threshold. In this article, we combine information from literature reviews and a real data set to determine the threshold of a biomarker for binary disease classification. We develop a Bayesian estimation procedure for this purpose that utilizes the conditional distribution of the biomarker. We illustrate our method using different biomarkers related to myocardial infarction (MI) and a large case-control study data set, called INTERHEART (Yusuf et al., 2004).

Based on literature reviews, we identify that higher levels of Apolipoprotein B (ApoB) and lower levels of Apolipoprotein A1 (ApoA1) are well-known risk factors for MI (Lind et al., 2006; McQueen et al., 2008; Sabino et al., 2008; Holme et al., 2008). Based on the INTERHEART study, Yusuf et al. (2004) identified abnormal lipids, characterized by the ratio of ApoB to ApoA1 as the most influential risk factor for MI. The author also showed that the odds ratio increases with the increasing value of the ratio. However, the threshold at which maximum classification accuracy occurs for these biomarkers may vary across studies. Therefore, the cut-offs for ApoB, ApoA1 or the ratio of ApoB to ApoA1, which maximizes the classification accuracy of MI using INTERHEART data set would be clinically valuable. These thresholds may also vary across different age and ethnic groups.

The primary objective of this paper is to develop a Bayesian estimation procedure to determine the threshold of a biomarker that maximizes the classification accuracy of a disease. The secondary objective of this research is to identify the potential best predictors for the disease under consideration. We construct prior distribution for ApoB, ApoA1, and the ratio of ApoB to ApoA1 for MI classification and combine with the INTERHEART study data to obtain a posterior distribution for this purpose.

The remaining part of the paper is organized as follows. In the next Section, we describe the INTERHEART data set, where we present region-specific sample size, mean and standard deviation for ApoB, ApoA1, and the ratio. In Section 3, we develop and describe the threshold identification procedure using a Bayesian approach. We present the results of the analysis and discuss them in Section 4. Finally, we provide a summary of our findings in Section 5 with some concluding remarks in Section 6.

2. The data

INTERHEART is a case-control study, consisting of 15,152 age- and sex-matched MI cases and 14,280 controls from 52 countries around the world. Here, the term case-control refers to those subjects with the disease MI(cases) and those without the disease(controls). Due to various reasons, it was not possible to find exactly matched controls for a small number of cases. Among the 29,432 recruited individuals, blood samples were collected and analyzed for about 79% of the participants using standardized quality control criteria for sample transportation, storage and analysis at the central laboratory. As a result, ApoB (gram/liter) and ApoA1 (gram/liter) were missing for about 21% of the individuals which led to a breakdown in the matched case and control pair. Additional details on this data set can be found in Yusuf et al. (2004) and McQueen et al. (2008). Since the primary goal of this paper is to identify the cut-off for these biomarkers that maximizes the accuracy of classification, we use only a perfectly matched subset of the INTERHEART data within each region. This subset consists of age- (within five years), and sex-matched 8084 MI cases and 8084 controls from nine regions in the world. We identify this subset from the main INTERHEART database and a greedy matching algorithm implemented using a computer program developed in the SAS software. In this analysis, our interest lies in selected variables, such as case or control status, age, sex, region, ethnicity, ApoB, ApoA1, and the ratio of ApoB to ApoA1. Due to strict matching criteria, we have lost about one-fifth of the sample, but the matched subset allows us to rule out the confounding effect of age, sex, and region on the parameters of interest. In other words, this may help us identify more precise cut-offs as well as odds ratios of a unit change in each of these biomarkers.

We present region-specific sample size and the age distribution of selected participants overall, as well as by sex and their case or control status in Table 1. Note that there are the same number of cases and controls within each sex across nine regions. Age distribution within each stratum is approximately the same between cases and controls, representing an entirely balanced data set. We also compute and present the summary statistics, characterized by mean and standard deviation for each of these biomarkers in Table 2. In this table, we give the overall estimates, by sex and region, further stratified by their case or control status. Based on these estimates, we observe that there is a considerable amount of variation in the means across regions for ApoB, ApoA1, and the ApoB:ApoA1 ratio. We also observe that cases have a higher mean ApoB and the ratio but a lower mean ApoA1 than controls, an observation consistent with current literature. Density plots for each of the biomarkers among cases and controls were constructed and presented in Figure 1. Arrows in these figures indicate the cut-offs that maximize the classification accuracy for MI, and we discuss the related procedure in Section 3.

Table 1. Overall sample size, mean, and standard deviation of age by region, sex, and by case or control status

Region	Sex	Overall	Case	Control
All Region	All	16168 [57.4 (11.9)]	8084 [57.5 (12.0)]	8084 [57.4 (11.9)]
NAmerica/WEurope	All	894 [61.3 (11.6)]	447 [61.3 (11.7)]	447 [61.2 (11.6)]
	Women	260 [63.9 (12.0)]	130 [64.0 (12.1)]	130 [63.8 (11.9)]
	Men	634 [60.2 (11.3)]	317 [60.2 (11.4)]	317 [60.2 (11.3)]
Central/EastEurope	All	2056 [60.6 (12.1)]	1028 [60.6 (12.1)]	1028 [60.5 (12.0)]
	Women	656 [65.9 (10.8)]	328 [65.9 (10.8)]	328 [65.9 (10.7)]
	Men	1400 [58.1 (11.8)]	700 [58.2 (11.9)]	700 [58.0 (11.8)]
MiddleEast/Egypt	All	2184 [50.6 (9.95)]	1092 [50.7 (9.93)]	1092 [50.6 (9.97)]
	Women	248 [55.3 (9.54)]	124 [55.4 (9.59)]	124 [55.1 (9.53)]
	Men	1936 [50.0 (9.85)]	968 [50.1 (9.82)]	968 [50.0 (9.88)]
Africa	All	772 [54.0 (11.1)]	386 [54.1 (11.1)]	386 [53.9 (11.1)]
	Women	266 [57.0 (11.4)]	133 [57.0 (11.4)]	133 [56.9 (11.5)]
	Men	506 [52.4 (10.6)]	253 [52.5 (10.6)]	253 [52.4 (10.6)]
South Asia	All	2282 [53.1 (11.2)]	1141 [53.1 (11.3)]	1141 [53.0 (11.2)]
	Women	308 [57.7 (11.6)]	154 [57.7 (11.7)]	154 [57.7 (11.6)]
	Men	1974 [52.3 (11.0)]	987 [52.4 (11.0)]	987 [52.3 (10.9)]
China/HongKong	All	4588 [60.5 (11.2)]	2294 [60.6 (11.3)]	2294 [60.4 (11.1)]
	Women	1300 [65.6 (8.58)]	650 [65.7 (8.63)]	650 [65.5 (8.54)]
	Men	3288 [58.5 (11.4)]	1644 [58.6 (11.5)]	1644 [58.4 (11.3)]
SEAsia/Japan	All	1262 [56.9 (10.9)]	631 [56.9 (11.0)]	631 [56.8 (10.8)]
	Women	238 [61.9 (9.86)]	119 [61.9 (9.92)]	119 [61.8 (9.83)]
	Men	1024 [55.7 (10.8)]	512 [55.7 (10.9)]	512 [55.6 (10.7)]
Australia/NZ	All	344 [59.4 (12.2)]	172 [59.6 (12.3)]	172 [59.3 (12.2)]
	Women	66 [63.4 (12.5)]	33 [63.6 (12.5)]	33 [63.2 (12.7)]
	Men	278 [58.5 (11.9)]	139 [58.6 (12.0)]	139 [58.3 (11.9)]
SAmerica/Mexico	All	1786 [59.6 (12.2)]	893 [59.6 (12.3)]	893 [59.5 (12.2)]
	Women	436 [63.1 (11.8)]	218 [63.2 (11.8)]	218 [63.0 (11.8)]
	Men	1350 [58.4 (12.1)]	675 [58.4 (12.2)]	675 [58.4 (12.1)]

Table 2. Overall means and standard deviation of ApoB, ApoA1, and the ratio by region, sex, and by case or control status

Region		ApoB (gram/liter)			ApoA1 (gram/liter)			Ratio of ApoB to ApoA1		
		Overall	Case	Control	Overall	Case	Control	Overall	Case	Control
All Region	Overall	0.94(0.26)	0.97(0.27)	0.91(0.25)	1.17(0.27)	1.12(0.24)	1.21(0.28)	0.84(0.30)	0.90(0.31)	0.79(0.29)
NAmerica/WEurope	All	0.98(0.24)	0.99(0.24)	0.96(0.23)	1.26(0.27)	1.18(0.24)	1.35(0.27)	0.81(0.29)	0.88(0.32)	0.74(0.24)
	Women	0.98(0.24)	1.00(0.25)	0.96(0.23)	1.37(0.29)	1.26(0.25)	1.47(0.28)	0.75(0.28)	0.82(0.27)	0.68(0.27)
	Men	0.98(0.23)	0.98(0.24)	0.97(0.23)	1.22(0.25)	1.14(0.22)	1.30(0.25)	0.84(0.29)	0.90(0.34)	0.77(0.22)
Central/EastEurope	All	1.00(0.27)	1.04(0.28)	0.96(0.25)	1.23(0.26)	1.22(0.23)	1.24(0.28)	0.85(0.27)	0.87(0.26)	0.82(0.28)
	Women	1.02(0.27)	1.06(0.28)	0.97(0.24)	1.29(0.27)	1.29(0.24)	1.30(0.29)	0.82(0.26)	0.85(0.27)	0.79(0.25)
	Men	0.99(0.27)	1.03(0.28)	0.96(0.25)	1.20(0.25)	1.19(0.22)	1.21(0.27)	0.86(0.28)	0.88(0.26)	0.83(0.29)
MiddleEast/Egypt	All	1.02(0.27)	1.06(0.27)	0.97(0.26)	1.10(0.22)	1.06(0.22)	1.14(0.22)	0.96(0.34)	1.04(0.38)	0.87(0.28)
	Women	1.07(0.29)	1.11(0.29)	1.03(0.28)	1.23(0.25)	1.18(0.26)	1.28(0.23)	0.89(0.27)	0.96(0.26)	0.83(0.26)
	Men	1.01(0.27)	1.06(0.26)	0.96(0.26)	1.08(0.21)	1.04(0.21)	1.12(0.21)	0.96(0.35)	1.05(0.39)	0.88(0.28)
Africa	All	0.94(0.28)	0.99(0.28)	0.88(0.27)	1.15(0.30)	1.10(0.27)	1.21(0.32)	0.87(0.34)	0.94(0.32)	0.79(0.35)
	Women	0.97(0.27)	1.00(0.27)	0.93(0.26)	1.22(0.29)	1.15(0.27)	1.30(0.29)	0.83(0.29)	0.91(0.30)	0.75(0.26)
	Men	0.92(0.29)	0.99(0.28)	0.86(0.28)	1.12(0.30)	1.08(0.26)	1.16(0.33)	0.89(0.36)	0.96(0.32)	0.81(0.39)
South Asia	All	0.93(0.25)	0.96(0.26)	0.90(0.24)	1.04(0.24)	1.01(0.23)	1.07(0.24)	0.93(0.34)	0.99(0.34)	0.87(0.34)
	Women	0.95(0.28)	0.98(0.27)	0.93(0.29)	1.13(0.28)	1.09(0.26)	1.18(0.30)	0.89(0.36)	0.94(0.33)	0.85(0.38)
	Men	0.92(0.25)	0.96(0.25)	0.89(0.24)	1.03(0.23)	1.00(0.22)	1.06(0.23)	0.94(0.34)	0.99(0.34)	0.88(0.33)
China/HongKong	All	0.84(0.22)	0.86(0.23)	0.81(0.21)	1.21(0.27)	1.15(0.25)	1.26(0.29)	0.72(0.23)	0.77(0.23)	0.67(0.21)
	Women	0.88(0.24)	0.92(0.26)	0.84(0.22)	1.30(0.29)	1.24(0.27)	1.35(0.31)	0.71(0.24)	0.77(0.26)	0.65(0.20)
	Men	0.82(0.21)	0.83(0.22)	0.80(0.21)	1.17(0.26)	1.12(0.23)	1.23(0.27)	0.73(0.22)	0.77(0.23)	0.68(0.21)
SEAsia/Japan	All	1.02(0.26)	1.07(0.28)	0.97(0.23)	1.19(0.27)	1.12(0.24)	1.26(0.28)	0.90(0.30)	0.99(0.29)	0.81(0.27)
	Women	1.07(0.27)	1.11(0.32)	1.02(0.21)	1.29(0.29)	1.20(0.26)	1.38(0.31)	0.86(0.27)	0.95(0.30)	0.76(0.20)
	Men	1.01(0.26)	1.06(0.27)	0.96(0.23)	1.17(0.26)	1.10(0.23)	1.24(0.26)	0.91(0.30)	0.99(0.29)	0.82(0.29)
Australia/NZ	All	0.97(0.24)	0.97(0.26)	0.97(0.23)	1.27(0.25)	1.20(0.23)	1.35(0.25)	0.79(0.24)	0.84(0.26)	0.74(0.22)
	Women	0.97(0.27)	0.97(0.28)	0.97(0.25)	1.40(0.27)	1.31(0.27)	1.48(0.23)	0.72(0.24)	0.77(0.25)	0.68(0.23)
	Men	0.97(0.24)	0.97(0.25)	0.98(0.23)	1.24(0.24)	1.17(0.21)	1.32(0.24)	0.81(0.24)	0.85(0.26)	0.76(0.22)
SAmerica/Mexico	All	0.98(0.26)	1.00(0.27)	0.95(0.24)	1.14(0.26)	1.08(0.23)	1.19(0.27)	0.90(0.31)	0.96(0.30)	0.84(0.31)
	Women	1.03(0.27)	1.07(0.29)	0.98(0.24)	1.25(0.27)	1.16(0.25)	1.33(0.27)	0.87(0.29)	0.96(0.30)	0.78(0.26)
	Men	0.96(0.25)	0.98(0.26)	0.94(0.24)	1.10(0.24)	1.06(0.21)	1.15(0.26)	0.91(0.31)	0.96(0.30)	0.85(0.32)

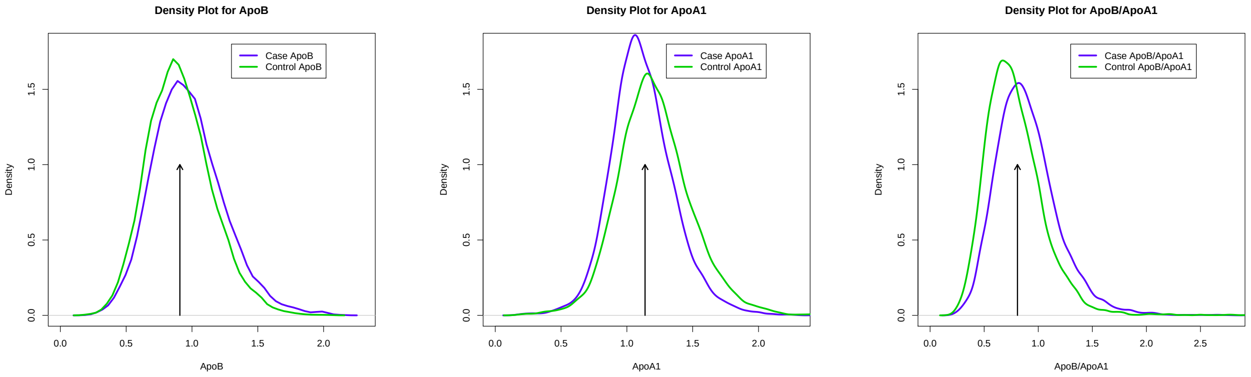

Figure 1. *Density plot of ApoB, ApoA1, and the ratio of ApoB to ApoA1 from INTERHEART cases and controls*

3. Methods

In this section, we develop and describe statistical methods related to two objectives stated in the introduction. For this purpose, we consider MI as the disease of interest as well as ApoB, ApoA1, and the ratio of ApoB to ApoA1 as biomarkers of interest. In Section 3.1, we describe the threshold identification procedure based on the conditional posterior distribution of each of these biomarkers. In Section 3.2, we provide a step-by-step computational algorithm. In Section 3.3, we present a Bayesian approach of estimating the odds ratio using a logistic regression model. Finally, in Section 3.4, we provide information on methods used for model diagnostics and related software.

3.1. Threshold identification method using a Bayesian approach

In this section, we develop and describe the classification rule for a disease based on a continuous exposure. Here, the term "classification rule" refers to identifying the threshold at which the maximum classification accuracy occurs for a given value of the exposure. To facilitate this process, let us suppose $D = 1$ indicates presence and $D = 0$ means the absence of a disease. Let us also assume e is the value of the continuous exposure E that can be used to classify individuals into binary disease categories. The cross tabulation of disease and exposure categories then gives us the frequencies a, b, c and d corresponding to four cells of a 2×2 table. Thus, the classification rules can be identified based on the following table.

Table 3. Cross tabulation of disease and exposure at a given threshold and corresponding cell frequencies

Classification Rule		Disease (D)		Total
		Yes (1)	No (0)	
Exposure (E)	$E \geq e$	a	b	a+b
	$E < e$	c	d	c+d
Total		a+c	b+d	N=a+b+c+d

Using the notation of Table 3, the sensitivity and specificity denoted by P_+ and P_- can respectively, be defined as:

$$P_+ = Pr\left[E \geq e|D = 1\right] = \frac{a}{(a+c)},$$
$$P_- = Pr\left[E < e|D = 0\right] = \frac{d}{b+d}.$$

Let us also define false negative Q_+ and false positive Q_- such that:

$$Q_+ = Pr\left[E < e|D = 1\right] = 1 - \frac{a}{a+c} = \frac{c}{a+c},$$
$$Q_- = Pr\left[E \geq e|D = 0\right] = 1 - \frac{d}{b+d} = \frac{b}{b+d}.$$

Note that our goal here is to find the threshold or cut-off e (the value of E) that maximizes classification accuracy. In the above formulation, we are assuming that a higher value of an exposure is related to a higher risk for the disease (e.g., ApoB). However, we can also use Table 3 in a situation where a lower value of a biomarker represents a risk (e.g. ApoA1), by just changing the inequality in the opposite direction. We refer to this as the classification rule, and the procedure is described below.

Let us assume the conditional probability distribution of each exposure is given by $[E|D = 1] \sim N(\mu_1, \sigma_1)$ and $[E|D = 0] \sim N(\mu_0, \sigma_0)$. Based on this assumption and the method described by Pepe (2003), we can express Q_- as:

$$Q_- = Pr\left[E \geq e|D = 0\right] = \Phi\left[\frac{\mu_0 - e}{\sigma_0}\right]. \tag{1}$$

For a given value of Q_-, say q_-, Equation 1 can be re-arranged as:

$$e = \mu_0 - \sigma_0\Phi^{-1}(q_-). \tag{2}$$

Using this information, P_+ can be re-expressed as:

$$P_+ = Pr[E \geq e|D = 1] = \Phi\left[\frac{\mu_1 - \mu_0}{\sigma_1} + \left(\frac{\sigma_0}{\sigma_1}\right)\Phi^{-1}(q_-)\right]. \tag{3}$$

Similarly, the area under the receiver operating characteristic (ROC) curve, denoted by AUC, can be expressed as:

$$AUC = Pr[(E|D = 1) > (E|D = 0)] = \Phi\left[\frac{(\mu_1 - \mu_0)/\sigma_1}{\sqrt{1 + (\sigma_0^2/\sigma_1^2)}}\right]. \tag{4}$$

In the above set of expressions, Φ is the cumulative distribution function of a standard normal variate. Note that we are assuming a conditional normal distribution for the exposure in this construction. We can use the maximum likelihood estimation procedure to solve for μ and σ used in the above set of expressions. However, it may be useful to utilize prior information from a literature review and we propose a Bayesian estimation procedure for this purpose. Assuming the prior distribution for the biomarker is known for both cases and controls, we can combine the prior and likelihood constructing posterior distributions for the location parameter μ and the scale parameter σ using the following equations:

$$p(\mu|E) = \frac{p(E|\mu)\, p(\mu)}{p(E)} \;\; and \;\; p(\sigma|E) = \frac{p(E|\sigma)\, p(\sigma)}{p(E)}. \tag{5}$$

We can use these general equations to construct posterior distribution for cases and controls separately as well as for each biomarker under consideration. In this construction, the location parameter μ can be considered to have conjugate family of distributions for a given scale parameter σ (Ntzoufras, 2009). Thus, we assume location parameter μ follows a normal distribution such that $\mu \sim N(\mu_I, \sigma_I)$. We determine hyper parameters μ_I and σ_I for each of these biomarkers based on literature reviews and provide specific values in Section 4.1.1. On the other hand, the scale σ is a nuisance parameter and it is reasonable to assume a flat prior such as uniform distribution. Based on literature reviews, we observe that the standard deviation of each of these biomarkers ranges between 0 and 1. Thus, a uniform distribution for σ within the interval 0 and 1 is a reasonable choice for these biomarkers.

3.2. Computation algorithm to identify the threshold

We now describe the simulation algorithm to find the threshold of a biomarker that maximizes the accuracy of classification of a disease. Let us suppose the prior information for location parameter follows a normal distribution such that $\mu \sim N(\mu_I, \sigma_I)$ and the scale parameter σ follows a uniform distribution on the interval 0 and 1. Utilizing this information and observed data, we can simulate random samples from the posterior distribution given in Equation 5 for cases and controls separately. For this purpose, we can use the Markov Chain Monte Carlo (MCMC) simulation procedure available in the WinBUGS software (Spiegelhalter et al., 2003) and the steps to do so are described below:

(1) Generate random samples from the prior distribution of the location parameter $\mu_1 \sim N(\mu_I, \sigma_I)$ and scale parameter $\sigma_1 \sim U(0, 1)$ for cases.

(2) Compute the likelihood based on case exposure data under normal distribution assumption such that: $[E|D = 1] \sim N(\mu_1, \sigma_1)$.

(3) Combine the prior information and the likelihood to construct posterior distribution for cases.

(4) Repeat steps 1-3 for control data such that $[E|D = 0] \sim N(\mu_0, \sigma_0)$.

(5) Specify a grid of points between 0 and 1 and consider those as a given set of points for Q_- and compute the specificity such that $p_- = 1 - q_-$.

(6) Compute e using the expression given in Equation (2).

(7) Compute the sensitivity p_+ for a given e and generated data from the posterior distribution of μ and σ using Equation (3).

(8) Identify the maximum sensitivity and specificity from the series obtained in steps 5 and 7.

(9) Identify the cut-off e that corresponds to the maximum sensitivity and specificity.

(10) For each set of generated μ and σ, compute AUC using Equation (4).

(11) Consider the computed set as a sample from the distribution of AUC and then compute median and 95% credible interval for the AUC.

3.3. Odds ratio estimation method using a Bayesian approach

Suppose that $D = 1$ indicates presence and $D = 0$ indicates the absence of a disease of interest. Let p_i be the conditional probability of the disease given the exposure. Then the logistic regression model with one continuous exposure is given by:

$$log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 * E_i, \quad i = 1, 2, ..., n, \tag{6}$$

where β_0 is the intercept and the slope β_1 is the log of the odds ratio for the exposure. In this equation, E could be any of the three exposures ApoB, ApoA1 or the ratio of ApoB to ApoA1. In this construction, the non-informative prior for the intercept and informative prior for the slope can be specified by the normal distribution (Ntzoufras, 2009) such that:

$$\beta_0 \sim N(\mu_{N0}, \sigma_{N0}) \;\; and \;\; \beta_1 \sim N(\mu_{I1}, \sigma_{I1}), \tag{7}$$

for a given hyperparameter mean μ_{N0} and a standard deviation σ_{N0} (here N denotes non-informative prior). A large standard deviation leads to a flat or non-informative prior for the intercept. We utilize information available in the literature to specify the hyperparameter μ_{I1} and σ_{I1} for the prior distribution of the slope (here I denotes an informative prior). The priors are then combined with the likelihood to form a posterior distribution for these parameters.

3.4. Model diagnostic methods and related software

We first compute the posterior distributions for each parameter using the MCMC simulation procedure. We then use autocorrelation plot, history plot and Brooks-Gelman-Rubin (BGR) plots (Brooks and Gelman, 1998; Gelman and Rubin, 1992) to identify if there is any problem with convergence of the posterior to a stationary distribution. We also computed BGR (Brooks and Gelman, 1998; Gelman and Rubin, 1992), Heideburger and Welch (Heidelberger and Welch, 1983) and Geweke statistic (Geweke, 1992) to test formally for non-convergence of the posterior distribution. During this process, we ran three chains with 3000 iterations each, in addition to 1000 burn-in for each parameter of interest. We also use the area under the ROC curve (AUC) to compare classification accuracies across different biomarkers. Based on the computed logistic regression, we also compare models for different biomarkers using Akaike Information Criteria (AIC) (Akaike, 1974) for the classical approach and Deviance Information Criteria (DIC) (Spiegelhalter et al., 2002) in the Bayesian approach. We use the WinBUGS (Spiegelhalter et al., 2003) and R (R Development Core Team, 2011) software for all analysis and prepare all figures. In particular, we use R2WinBUGS (Sturtz et al., 2005) and BOA (Smith, 2007) library to analyze the data simultaneously between these two software packages.

4. Results

In this section, we illustrate our methods described in the previous section using information obtained from literature reviews and the analysis of the INTERHEART data set. Note that our interest is classifying MI based on ApoB, ApoA1, and the ratio of ApoB to ApoA1. In Section 4.1, we present the result using a Bayesian approach and identify the threshold of these biomarkers that maximizes classification accuracy. In Section 4.2, we present the odds ratio estimates based on the conditional logistic regression model and corresponding information criteria using both Bayesian and classical approaches. Within the Bayesian approach, we use a random intercept model to take the effect of matching into account.

4.1. Classification rules for MI

Our goal in this section is to identify the threshold of ApoB, ApoA1, and the ratio of ApoB to ApoA1 that maximizes the classification accuracy for MI. In other words, we need to find the cut-off e, the value of the exposure E that maximizes the sensitivity and specificity using the algorithm developed in Section 3.2. Based on the histogram and Quantile-Quantile plot, we assume a normal distribution of each exposure given the case or control status of the participants. In the next sub-section, we construct the normal informative prior distribution for the location parameter of each biomarker. We then construct posterior distributions based on the normal informative prior for the location parameter μ and a uniform prior for the scale parameter σ for each of the biomarkers. Finally, we identify the threshold of each biomarker separately that maximizes the classification accuracy for MI. We also compute the AUC to identify the best predictor among these three biomarkers.

4.1.1. Prior distributions for the location and scale parameters: We review recent literature and construct the informative priors for the location and scale parameter of ApoB, ApoA1, and the ratio of ApoB to ApoA1. We identify a meta-analysis by Thompson and Danesh (2006), which presents the mean and standard deviation for each of these biomarkers from 23 prospective studies. In this article, the author considers coronary heart disease and MI as the disease of interest. Selected studies vary by location, sampling frame, baseline year, end point, age group, fasting status, assay, temperature, and different methods from manufacturers. Based on the summary of these studies, we construct the prior distribution for the location parameter μ related to ApoB as $\mu_{ApoB} \sim N(\mu_I = 1.07, \sigma_I = 0.1169)$. Similarly, for the location parameter for ApoA1 and the ratio of ApoB to ApoA1, the prior distributions are given by $\mu_{ApoA1} \sim N(\mu_I = 1.37, \sigma_I = 0.204)$ and $\mu_{ApoB/ApoA1} \sim N(\mu_I = 0.85, \sigma_I = 0.0786)$, respectively. Note that in this analysis scale, σ is a nuisance parameter. Thus, assuming a flat uniform prior U(0, 1) for each exposure seems to be adequate. Given that there is a small difference in the mean of these biomarkers between cases and controls, we use the same prior distribution for both case and control related model parameters.

4.1.2. Posterior distributions for the location and scale parameters: Using these prior distributions and INTERHEART data, we compute the posterior distributions based on Equation 5 and the MCMC simulation. Looking at the history, autocorrelation, and the BGR, Heideburger-Welch and Geweke tests, we do not see any problem with convergence of any of the posterior distributions. The computed values of the BGR test statistic for all parameters are very close to 1, reassuring that there is no problem with convergence in any of the posterior distributions. We present the density plot for all prior and posterior distributions associated with the location parameter of different exposures in Figure 2-4. The posterior median with 95% credible intervals for the location and scale parameters for each of the exposures are given in Table 4. Estimates obtained from ApoB, ApoA1, and the ratio among cases are 0.972 (0.966, 0.978), 1.119 (1.114, 1.124) and 0.903 (0.896, 0.906), respectively. We also observe that the median ApoB among cases is 0.064 of a unit higher, ApoA1 is 0.092 of a unit lower, and the ratio of ApoB to ApoA1 is 0.117 of a unit higher than controls. However, scale parameter estimates are slightly higher among cases than controls except for ApoA1, and we observe slightly lower standard deviations among cases than controls.

Table 4. Posterior median and 95% credible interval for the distribution of μ and σ

Parameters	ApoB (gram/liter)		ApoA1 (gram/liter)		ApoB/ApoA1	
	Case	Control	Case	Control	Case	Control
μ	0.972 (0.966, 0.978)	0.908 (0.903, 0.914)	1.119 (1.114, 1.124)	1.211 (1.205, 1.217)	0.903 (0.896, 0.909)	0.786 (0.780, 0.792)
σ	0.270 (0.266, 0.275)	0.245 (0.241, 0.249)	0.244 (0.240, 0.248)	0.279 (0.274, 0.283)	0.312 (0.308, 0.317)	0.285 (0.281, 0.290)

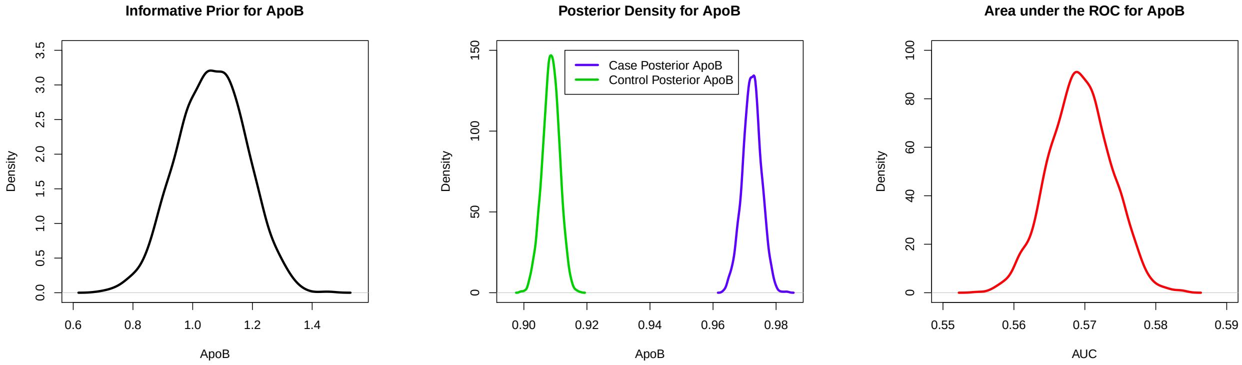

Figure 2. *Prior and posterior density for the location parameter of ApoB and corresponding AUC*

4.1.3. Threshold for ApoB, ApoA1, and the ratio of ApoB to ApoA1: Applying the algorithm described in Section 3.2, we identify the threshold of these biomarkers that maximizes the classification accuracy of MI. Threshold estimates and the corresponding AUC with their 95% credible intervals are given in Table 5. Since age, sex, and ethnicity are already known to play a significant role in modifying these biomarkers, we present the cut-off for those sub-groups, along with the overall estimates. In particular,

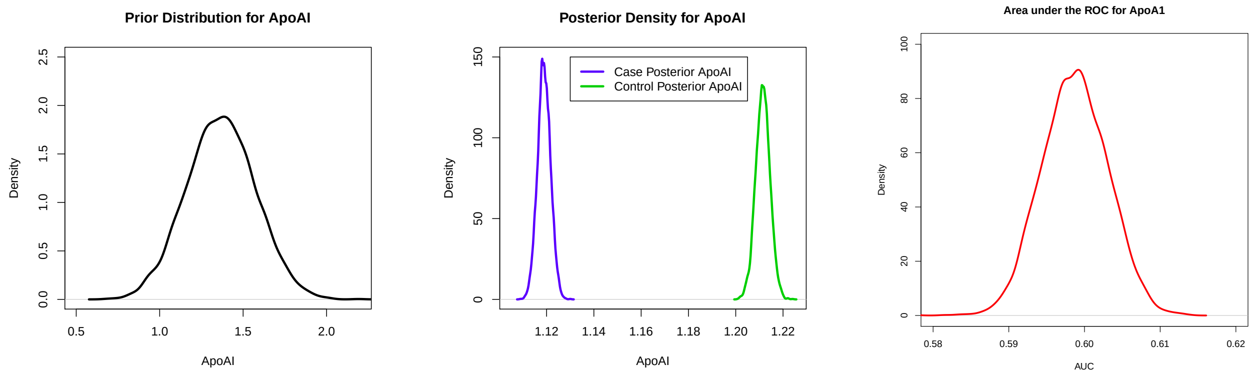

Figure 3. *Prior and posterior density for the location parameter of ApoA1 and corresponding AUC*

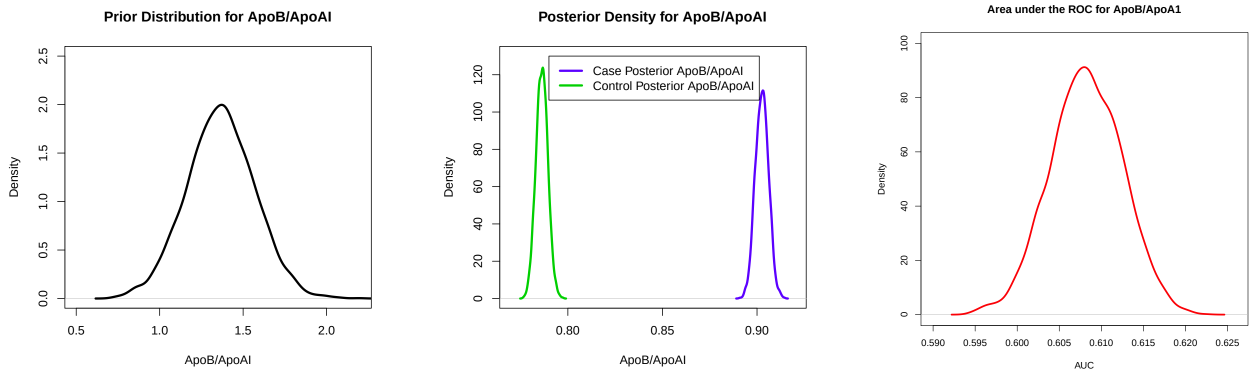

Figure 4. *Prior and posterior density for the location parameter of ApoB to ApoA1 ratio and corresponding AUC*

we provide the cut-off for younger (men $\leq$ 55 or women $\leq$ 65) and older (men $>$ 55 or women $>$ 65) individuals as well as for each ethnic group.

We identify the overall cut-off for ApoB 0.908 (gram/liter), corresponding to a sensitivity of 59% and a specificity of 50%. Similarly, for ApoA1 we identify the threshold as 1.138 (gram/liter) corresponding to a sensitivity of 60% and a specificity 53%. Finally, for ApoB to ApoA1 ratio the threshold is 0.808, corresponding to a sensitivity of 62% and a specificity 53%. Based on the estimates presented in Table 5, we also observe that the threshold varies slightly across the different ethnic groups as well as for younger and older individuals. Finally, the median AUC with 95% credible intervals associated with ApoB, ApoA1, and ApoB/ApoA1 are 0.570 (0.561, 0.578), 0.599 (0.590, 0.607) and 0.608 (0.600, 0.617), respectively.

Table 5. Posterior median and 95% credible interval for the cut-off of ApoB, ApoA1, and the ratio of ApoB to ApoA1, and corresponding AUC

Region	ApoB (gram/liter)		ApoA1 (gram/liter)		ApoB/ApoA1	
	Cut-off	AUC	Cut-off	AUC	Cut-off	AUC
Overall	0.908 (0.883, 0.935)	0.570 (0.561, 0.578)	1.138 (1.118, 1.159)	0.599 (0.590, 0.607)	0.808 (0.781, 0.838)	0.608 (0.600, 0.617)
Young	0.937 (0.915, 0.961)	0.605 (0.593, 0.617)	1.13 (1.1, 1.15)	0.601 (0.589, 0.613)	0.841 (0.816, 0.868)	0.636 (0.624, 0.648)
Old	0.877 (0.807, 0.912)	0.534 (0.522, 0.546)	1.15 (1.13, 1.17)	0.597 (0.585, 0.609)	0.752 (0.698, 0.808)	0.58 (0.568, 0.592)
European	0.965 (0.939, 0.99)	0.56 (0.542, 0.580)	1.22 (1.2, 1.24)	0.585 (0.566, 0.603)	0.794 (0.762, 0.828)	0.589 (0.570, 0.607)
Chinese	0.818 (0.797, 0.841)	0.557 (0.542, 0.573)	1.18 (1.159, 1.2)	0.614 (0.599, 0.630)	0.703 (0.687, 0.72)	0.618 (0.603, 0.634)
South Asian	0.92 (0.889, 0.957)	0.577 (0.556, 0.596)	1.02 (1.001, 1.05)	0.568 (0.548, 0.589)	0.859 (0.775, 0.927)	0.591 (0.571, 0.610)
Other Asian	0.988 (0.959, 1.013)	0.588 (0.553, 0.619)	1.17 (1.132, 1.19)	0.669 (0.637, 0.699)	0.843 (0.797, 0.883)	0.667 (0.636, 0.697)
Arab or Persian	0.94 (0.87, 0.992)	0.591 (0.564, 0.616)	1.04 (0.992, 1.09)	0.614 (0.588, 0.642)	0.909 (0.886, 0.929)	0.643 (0.618, 0.668)
Latin American	0.947 (0.917, 0.975)	0.567 (0.540, 0.593)	1.11 (1.089, 1.13)	0.609 (0.582, 0.634)	0.807 (0.725, 0.866)	0.605 (0.579, 0.631)
Black African	0.771 (0.63, 0.829)	0.598 (0.526, 0.664)	1.05 (0.764, 1.13)	0.578 (0.506, 0.645)	0.735 (0.552, 0.804)	0.598 (0.533, 0.663)
Coloured African	0.949 (0.872, 0.991)	0.613 (0.563, 0.663)	1.13 (1.095, 1.17)	0.614 (0.561, 0.665)	0.86 (0.817, 0.894)	0.659 (0.609, 0.707)

We also conduct a classical ROC analysis overall and for each strata of interest and for all three biomarkers using a R package called ‘OptimalCutpoints’, selecting the maximum sensitivity and specificity option. We observe that the threshold and AUC estimates are very similar to those we see in the Bayesian approach. For example, the overall threshold for ApoB, ApoA1 and the ratio using the classical approach are given by 0.919, 1.15 and 0.805, respectively. Similarly, the AUC estimates corresponding to ApoB, ApoA1 and the ratio are given by 0.567, 0.603 and 0.626, respectively.

4.2. Estimation of odds ratios and predictive ability of selected exposures

In this section, we first present the result from a logistic regression model using a Bayesian estimation procedure described in Section 3.3. We consider the following three biomarkers and standardize them by subtracting the overall mean and then dividing by the standard deviation: 1) ApoB (Mean=0.94, SD=0.26), 2) ApoA1 (Mean=1.17, SD=0.27) and 3) the ratio of ApoB to ApoA1 (Mean=0.84, SD=0.30).

The reason for choosing standardized variables here is to compare the odds ratio of one standard deviation change in different exposures of interest. We construct posterior distributions based on non-informative prior for the intercept and normal informative prior for the slope. We also use the classical approach to obtain parameter estimates based on a conditional logistic regression model and compare results with the Bayesian approach.

4.2.1. Prior distributions for the intercept and the slope: We consider the non-informative prior for the intercept β_0 and informative prior for the slope β_1 given in Equation 7, where the slope represents the log of the odds ratio for one standard deviation change in ApoB, ApoA1 or the ratio of ApoB to ApoA1. Reviewing current literature, we construct the informative prior for the slope. It was difficult to find a reasonable number of recent articles presenting the odds ratio of MI for one standard deviation change in those risk factors. Since the disease MI and stroke share risk factors and the degree of association with selected exposure are known to be approximately the same, we also consider stroke during this search. We identify some articles that report relative risk (Parish et al., 2009; Walldius et al., 2006, 2004), some with hazard ratio (Holme et al., 2008; Lind et al., 2006) and odds ratio (Sabino et al., 2008; Sniderman et al., 2006). The goal here is to identify a plausible center and spread for the log of the odds ratio of selected exposures to construct an informative prior. Based on the summary of selected studies, we construct prior for log odds ratio of standardized ApoB as $\beta_1 \sim N(\mu = 0.2231, \sigma = 0.1014)$. Similarly, for the log odds ratio of ApoA1 and the log odds ratio of ApoB to ApoA1 ratio, the prior distributions are given by $\beta_1 \sim N(\mu = -0.1625, \sigma = 0.0599)$ and $\beta_1 \sim N(\mu = 0.4055, \sigma = 0.1669)$, respectively. Since the intercept β_0 in each model is a nuisance parameter, we consider a non-informative prior such that $\beta_0 \sim N(0.0, 1.0E + 3)$.

4.2.2. Posterior distribution for the log of the odds ratio: Using the prior distribution specified above, as well as the data obtained from INTERHEART study, we compute posterior distributions for the log of the odds ratio applying the MCMC simulation. Based on the history, autocorrelation and the BGR plots, we do not find any problem with the convergence of any of these posterior distributions. The computed value of the BGR test statistic is very close to 1 for each distribution, indicating that there is no problem with convergence in any of these posterior distributions. We present the density plot for all posterior distributions associated with different exposure and priors in Figure 5.

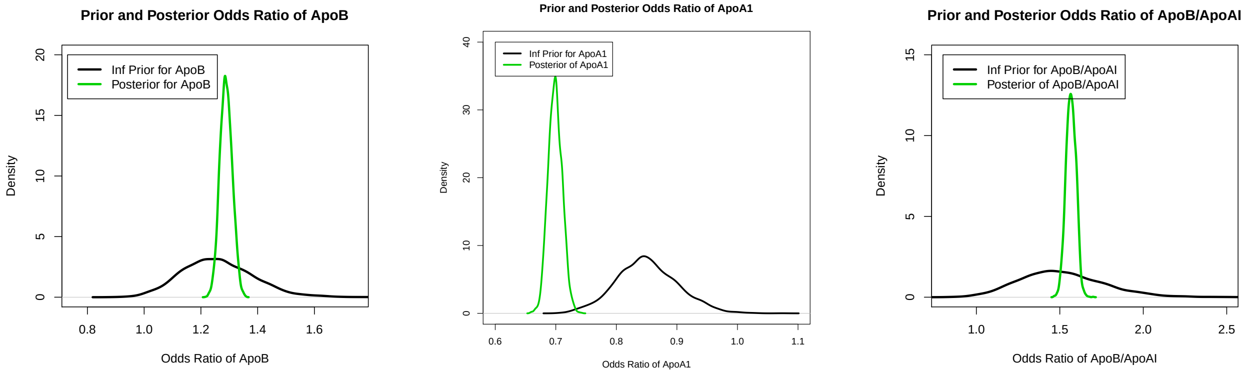

Figure 5. *Prior and posterior density for the odds ratio of ApoB, ApoA1, and the ratio of ApoB to ApoA1*

4.2.3. Odds ratio estimates and corresponding information criteria: We present the summary of the odds ratio estimates of standardized ApoB, ApoA1 and the ratio of ApoB to ApoA1, and corresponding information criteria in Table 6. We also present the odds ratio estimates obtained from the classical approach in the same table. The first column of this table shows the standard deviations, and the next two columns display the posterior medians with a 95% credible interval (Bayesian confidence interval) from the Bayesian approach for each exposure and selected priors as well as the corresponding DIC. In the last two columns, we present the odds ratio with 95% confidence interval (classical approach) using conditional logistic regression models for the same set of exposures and corresponding AIC. We observe that the estimates obtained using the Bayesian informative prior are slightly lower than those obtained from the classical approach.

We also observe that the model with ApoB and ApoA1 as independent covariates gave us the smallest DIC for the Bayesian approach and the lowest AIC for the classical approach. Based on the estimated odds ratio of one SD change in each of the exposures, we reconfirm that higher levels of ApoB lead to an increased risk for MI, and higher levels of ApoA1 has a protective effect. Note that ApoA1 appeared to be more predictive than ApoB for MI based on this analysis. Also, the ratio of ApoB to ApoA1 seemed to be less informative than both biomarkers in the model as independent predictors in the Bayesian as well as the classical approaches.

Table 6. Odds ratio for one standard deviation (SD) changes in ApoB, ApoA1, and the ratio of ApoB to ApoA1, and corresponding information criteria

Exposure	SD	Bayesian		Classical	
		Posterior Median and 95% Credible Interval		Point Estimates and 95% Confidence Interval	
		Odds ratio	DIC	Odds ratio	AIC
ApoB	0.260	1.29 (1.25, 1.33)	22,169	1.32 (1.28, 1.36)	21,737
ApoA1	0.267	0.70 (0.68, 0.72)	21,898	0.66 (0.64, 0.69)	21,448
ApoB/ApoA1	0.305	1.60 (1.55, 1.67)	21,735	1.65 (1.58, 1.71)	21,287
ApoB and	0.260	1.42 (1.37, 1.47)	21,462	1.45 (1.40, 1.50)	21,190
ApoA1	0.267	0.65 (0.63, 0.67)		0.62 (0.59, 0.64)	

5. Discussion

Thresholds for disease classification using biomarkers are typically identified using a classical approach. However, this approach does not allow one to use pre-existing evidence in the literature. We developed a Bayesian approach for this purpose using the conditional distribution of the biomarker. This approach combined information from literature reviews and real data set to determine the threshold. We illustrated this method using literature review of selected biomarkers related to MI and the INTERHEART study data set. Based on this study, elevated ApoB to ApoA1 ratio appeared to be the most influential risk factor for MI. It was also known that higher levels of ApoB or lower levels of ApoA1 are associated with increasing risk of MI. However, the threshold at which these biomarkers change from protection to risk was not clear.

To facilitate this process, we first developed a Bayesian estimation procedure to determine the threshold for ApoB, ApoA1, and the ratio of ApoB to ApoA1 with the maximum classification accuracy of MI. The overall threshold for ApoB, ApoA1, and ApoB/ApoA1 were estimated to be 0.908 (gram/liter), 1.138 (gram/liter) and 0.808, respectively. We also observed that the threshold varies slightly across different ethnic groups and age groups. A classical ROC analysis also produced a very similar result. However, it would be essential to conduct a systematic review to identify best possible information from the literature and reconstruct the prior for each of these biomarkers and revise the threshold accordingly. This procedure will allow us to update and proceed with confidence using these thresholds in a clinical practice.

We have also used a Bayesian and a classical approach to estimate the odds ratio corresponding to the standardized ApoB, ApoA1, and the ratio of ApoB to ApoA1. We compared estimates obtained from these two approaches and identified the most influential of the three predictors. We observed that estimates obtained from the Bayesian approach were slightly lower than those from the classical approach. Including ApoB and ApoA1 as independent covariates led to a better model fit for both the Bayesian and classical models, compared to the ratio of ApoB to ApoA1. Based on this analysis, higher levels of ApoB appear to be a risk factor and higher levels of ApoA1 seem to be a protective factor for MI, a finding consistent with current literature. Furthermore, our analysis showed that ApoA1 might be a better predictor than ApoB for MI. The ratio of ApoB to ApoA1 as a single exposure appeared to be less informative for predicting MI than the model where we used both of them as independent predictors.

6. Conclusion

In conclusion, the Bayesian approach can be useful to find the threshold of a biomarker combining information from literature reviews and a real data set. Success of this approach depends on identifying reliable prior information and corresponding probability distribution. Given that we have used the data set for illustration purpose only, the results presented will require further investigation for use in clinical applications. However, the approach developed in this article can be used to determine the threshold for any continuous biomarker for a binary disease classification.

7. Acknowledgements

The author gratefully acknowledges the help and support from the Population Health Research Institute and Sigma Laboratory of the Population Genomics Program, McMaster University during the preparation of this manuscript.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactionson Automatic Control*, 19(6):716–723.

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

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–511.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J. M. Bernado, J. O. Berger, A. P. David, and A. F. M. Smith (Eds.), Bayesian Statistics (Vol. 4). Oxford, UK: Oxford University Press.

Heidelberger, P. and Welch, P. D. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, 31(6):1109–1144.

Holme, I., Aastveit, A. H., Jungner, I., and Walldius, G. (2008). Relationships between lipoprotein components and risk of myocardial infarction: age, gender and short versus longer follow-up periods in the Apolipoprotein MOrtality RISk study (AMORIS). *Journal of Internal Medicine*, 264(1):30–38.

Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied Logistic Regression (3rd ed.)*. Wiley-Interscience, New York, USA.

Lind, L., Vessby, B., and Sundstro, J. (2006). The apolipoprotein B / AI ratio and the metabolic syndrome independently predict risk for myocardial infarction in middle-aged men. *Arteriosclerosis Thrombosis and Vascular Biology*, 26:406–410.

McQueen, M. J., Hawken, S., Wang, X., Ounpuu, S., Sniderman, A., Probstfield, J., Steyn, K., Sanderson, J. E., Hasani, M., Volkova, E., Kazmi, K., and Yusuf, S. (2008). Lipids, lipoproteins, and apolipoproteins as risk markers of myocardial infarction in 52 countries (the INTERHEART study): A case-control study. *The Lancet*, 372(9634):224–233.

Ntzoufras, I. (2009). *Bayesian Modeling Using WinBUGS (2nd ed.)*. John Wiley & Sons, Inc.: Hoboken, New Jersey.

Parish, S., Peto, R., Palmer, A., Clarke, R., Lewington, S., Offer, A., Whitlock, G., Clark, S., Youngman, L., Sleight, P., and Collins, R. (2009). The joint effects of apolipoprotein B, apolipoprotein A1, LDL cholesterol, and HDL cholesterol on risk: 3510 cases of acute myocardial infarction and 9805 controls. *European Heart Journal*, 30(17):2137–46.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Test for Classification and Prediction*. Oxford, UK: Oxford University Press.

R Development Core Team (2011). R: A language and environment for statistical computing, R foundation for statistical computing, Vienna, Austria, 2011, ISBN 3-900051-07-0.

Sabino, A. P., De Oliveira Sousa, M., Moreira Lima, L., Dias Ribeiro, D., Sant'Ana Dusse, L. M., Das Graças Carvalho, M., and Fernandes, A. P. (2008). ApoB/ApoA-I ratio in young patients with ischemic cerebral stroke or peripheral arterial disease. *Translational Research*, 152(3):113–118.

Smith, B. J. (2007). Boa : An R package for MCMC output convergence. *Journal of Statistical Software*, 21(11):1–37.

Sniderman, A. D., Jungner, I., Holme, I., Aastveit, A., and Walldius, G. (2006). Errors that result from using the TC/HDL C ratio rather than the apoB/apoA-I ratio to identify the lipoprotein-related risk of vascular disease. *Journal of Internal Medicine*, 259(5):455–461.

Spiegelhalter, D., Thomas, A., Best, N., and Way, R. (2003). WinBUGS User Manual Version 1.4. MRC Biostatistics Unit, Cambridge, UK.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B*, 64(4):583–639.

Sturtz, S., Ligges, U., and Gelman, A. (2005). R2WinBUGS: A package for running WinBUGS from R. *Journal of Statistical Software*, 12(3):1–16.

Thompson, A. and Danesh, J. (2006). Associations between apolipoprotein B, apolipoprotein AI, the apolipoprotein B/AI ratio and coronary heart disease: A literature-based meta-analysis of prospective studies. *Journal of Internal Medicine*, 259(5):481–492.

Walldius, G., Aastveit, A. H., and Jungner, I. (2006). Stroke mortality and the apoB/apoA-I ratio: Results of the AMORIS prospective study. *Journal of Internal Medicine*, 259(3):259–266.

Walldius, G., Jungner, I., Aastveit, A. H., Holme, I., Furberg, C. D., and Sniderman, A. D. (2004). The apoB/apoA-I ratio is better than the cholesterol ratios to estimate the balance between plasma proatherogenic and antiatherogenic lipoproteins and to predict coronary risk. *Clinical Chemistry and Laboratory Medicine*, 42(12):1355–1363.

Yusuf, S., Hawken, S., Ôunpuu, S., Dans, T., Avezum, A., Lanas, F., Mcqueen, M., Budaj, A., Pais, P., Ounpuu, S., Varigos, J., and Lisheng, L. (2004). Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): Case-control study. *The Lancet*, 364(9438):937–952.

Chapter 3

A Copula-based Method of Classification

Researchers often use more than one dependent biomarker to classify individuals who are at risk of developing a particular disease. The threshold identified independently for this purpose usually leads to a conflicting classification for some individuals. This chapter includes an article under review by the journal of "Statistical Methods and Applications", where we developed and described a new method for classifying individuals into binary disease categories using dependent biomarkers. This method allows us determine joint threshold values of two dependent biomarkers for an outcome classification. Using this information, clinicians may uniquely identify individuals who are at risk of developing the disease and plan for early intervention.

A copula-based method of classifying individuals into binary disease categories using dependent biomarkers

S. Islam[†§], S. Anand[†§¶], J. Hamid [§], L. Thabane [†§] and J. Beyene[§*]

Abstract: **Classification of a disease often depends on more than one test, and the tests can be interrelated. Under the incorrect assumption of independence, the test result using dependent biomarkers can lead to a conflicting disease classification. We develop a copula-based method for this purpose that takes dependency into account and leads to a unique decision. We first construct the joint probability distribution of the biomarkers considering Frank's, Clayton's and Gumbel's copulas. We then develop the classification method and perform a comprehensive simulation. Based on simulated data sets, we study the statistical properties of joint probability distributions and determine the joint threshold with maximum classification accuracy. Our simulation study results show that parameter estimates for the copula-based bivariate distributions are not biased. We observe that the thresholds for disease classification converge to a stationary distribution across different choices of copulas. We also observe that the classification accuracy decreases with the increasing value of the dependence parameter of the copulas. Finally, we illustrate our method with a real data example, where we identify the joint threshold of Apolipoprotein B to Apolipoprotein A1 ratio and total cholesterol to high-density lipoprotein ratio for the classification of myocardial infarction (MI). We conclude, copula-based method works well in identifying the joint threshold of two dependent biomarkers for an outcome classification. Our method is flexible and allows modeling broad classes of bivariate distributions that take dependency into account. The threshold may allow clinicians to classify uniquely individuals at risk of developing the disease and plan for early intervention.**

Keywords: Copula; Gamma Distribution; Biomarker; Sensitivity; Specificity; AUC

[†] Population Health Research Institute, McMaster University and Hamilton Health Sciences, Hamilton, Ontario, Canada.
[§] Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada.
[¶] Department of Medicine, McMaster University, Hamilton, Ontario, Canada.
[*] Correspondence to: Joseph Beyene, Department of Health Research Methods, Evidence, and Impact, McMaster University
1280 Main Street West, Hamilton, ON L8S 4K1; Phone: 905-525-9140, ext. 21333; Email: beyene@mcmaster.ca

1. Introduction

The classification of individuals at risk of developing a disease often depends on more than one test based on multiple biomarkers, and these biomarkers can be interrelated (Tzimas et al., 2008; Vasan, 2006). Each biomarker has a separate threshold for disease classification, which is determined using a clinical or statistical approach such as percentile (Vasan, 2006) or the minimum p-value (Mazumdar and Glassman, 2000). Assuming that the tests are independent, these methods identify the threshold for one biomarker at a time. If this assumption is incorrect, the test result using two dependent biomarkers can lead to a wrong or conflicting disease classification for some individuals.

Statistical methodologies that enable unique classification in this situation are not well developed. Disease classification rules ignoring dependency when it exists results in a conflicting classification for some individuals. Thus, the rules of classification need to be developed based on the joint probability distribution of selected biomarkers. In this paper, we consider a copula-based approach (Nelsen, 2006), which incorporate the dependency between biomarkers. This procedure requires identification of the marginal probability distribution of each biomarker separately, following which it combines the marginal distributions using a copula function to construct the joint probability density function.

The concept of copula was first introduced by the American mathematician Abe Sklar in 1959 (Sklar, 1959) and provided a theorem to build copula-based multivariate distributions. Subsequently, many authors developed different forms of copulas based on this theorem (Frank, 1979; Clayton, 1978; Gumbel, 1960). One of the most important features of the copula is that this approach allows one to construct a joint probability distribution function even if the marginals follow a different probability distribution. For example, this method is particularly useful when marginal distributions are skewed or non-normal. The copula has been widely used to develop models for financial risk management (Joe, 1997; Bouyé et al., 2001), but its application in medical research is infrequent, especially in the context of modeling biomarker dependencies.

Recently, Kuss et al. (2014) used this method to develop a joint probability distribution of sensitivity and specificity for the meta-analysis of diagnostic test accuracy studies. We are using a similar approach constructing the joint probability distribution for multiple biomarkers, but the goal here is to identify the threshold that maximizes the classification accuracy. Based on the joint probability distribution constructed through copula, we develop a new method for classifying individuals into binary diseased-categories that take the dependency between biomarkers into account and leads to a unique decision. Since there are different choices of copulas, it is also essential to understand the impact on the classification accuracy of these options.

Thus, the primary objective of this research is to develop a copula-based method of classifying individuals into binary disease categories based on dependent biomarkers. The secondary objective is to understand the effect of different values of the dependence parameter across different choices of copula-based bivariate distribution for this purpose.

We organize the remaining part of this article as follows: we provide two motivating examples in Section 2. We then describe the copula-based classification method of constructing a probability distribution followed by the process of identifying the joint threshold in Section 3. In Section 4, we present simulation results, where we first assess the statistical properties of the joint probability distribution and then determine the threshold of two biomarkers based on simulated data sets. We illustrate our method with a real data example in Section 5. Finally, we summarize our findings with a discussion in Section 6 and some concluding remarks in Section 7.

2. Motivating examples

In this section, we provide two examples, where clinicians usually encounter conflicting classification for some individuals with an independently developed threshold. However, the method we are proposing in this article can help researchers identify the joint threshold of two dependent biomarkers, which in turn may help clinicians uniquely classify individuals at risk of developing the disease.

2.1. Classification of STEMI using CK and cTn

Creatine kinase (CK) enzyme and cardiac troponin (cTn) are often independently used to classify individuals at risk of ST-elevation myocardial infarction (STEMI), but they are known to be dependent (Tzimas et al., 2008; Vasan, 2006). However, the classification rules for STEMI based on these biomarkers are usually developed separately assuming independence, and lead to a conflicting classification for some individuals. Thus, these rules need to be developed based on the joint probability distribution of these biomarkers. Applying the method we are proposing in this article and using simulated data sets, we demonstrate how to determine the threshold of these biomarkers for the classification of STEMI in Section 4.

2.2. Classification of MI using ApoB to ApoA1 ratio and TC to HDL ratio

Higher levels of Apolipoprotein B to Apolipoprotein A1 ratio (ApoB/ApoA1) and total cholesterol to high-density lipoprotein ratio (TC/HDL) are well-known risk factors for myocardial infarction (MI) (Walldius et al., 2004; Sniderman et al., 2006). These two biomarkers are often used to classify individuals at risk of developing the disease, and they are dependent on each other. However, thresholds are determined separately assuming they are independent, which may lead to a conflicting classification for some individuals. Thus, identifying a joint threshold for these biomarkers is essential and clinically valuable. In Section 5, we illustrate our method using a real data set to determine the threshold of these biomarkers, which may allow clinicians uniquely identify individuals at risk of developing MI with maximum classification accuracy.

3. Methods

In this section, we first derive the joint probability distribution of two dependent biomarkers based on copula. Biomarkers introduced in our motivating examples are positively dependent with long tail marginal distributions. Copulas with positive dependence parameter are adequate for this instance. Thus, we consider Frank's, Clayton's and Gumbel's copulas in this construction with gamma marginals. We then develop a new method of identifying the joint threshold for two biomarkers, which allows us to determine the disease status with maximum classification accuracy. This method takes into account the dependency between biomarkers and leads to a unique decision.

3.1. Constructing multivariate distribution based on copula

Let $X_1, X_2, ..., X_k$ be a set of k random variables with marginal cumulative distribution function (CDF) $F(x_1) = u_1, F(x_2) = u_2, ..., F(x_k) = u_k$, respectively, where u_i's are uniformly distributed random variables within the interval 0 and 1. Sklar's theorem (Sklar, 1959) allows one to construct the joint probability distribution of this set of random variables using the marginal CDF and a function called 'copula' that describes the dependence structure:

$$\begin{aligned} H(x_1, x_2, ..., x_k) &= C(F(x_1), F(x_2), ..., F(x_k)) \\ &= C(u_1, u_2, ..., u_k). \end{aligned}$$

In the above expression, H and F are the joint and marginal CDF of the set of random variables $X_1, X_2, ..., X_k$, respectively, and C represents the copula function. Thus, the copula is a function that allows one to construct the joint probability distribution of a set of uniformly distributed random variables (Nelsen, 2006; Jaworski et al., 2009).

Based on Sklar's theorem, many authors developed different forms of the copula function. Mainly, there are two families of copulas discussed in the literature: Gaussian and Archimedean, where a Gaussian copula can be constructed applying probability integral transformation of multivariate normal distribution. On the other hand, different choices of Archimedean copula have its expression

(Genest and Rivest, 1993). In this article, we consider the bivariate Archimedean copula, namely Frank's (Frank, 1979), Clayton's (Clayton, 1978) and Gumbel's (Gumbel, 1960). These copulas are further investigated by many authors including Joe (1997) and Nelsen (2006). We choose these copulas mainly due to their simplicity with one dependence parameter and because they allow developing models with long tail marginal distributions, as well as, the range and shape of real data examples that may fit the data well. These copulas also allow us to study the effect of different values of dependence parameter θ in terms of Kendalls τ, a well-understood measure of dependence.

In case of two random variables X_1 and X_2 such that $F(x_1) = u$ and $F(x_2) = v$, Frank's, Clayton's and Gumbel's family of copula are given by:

$$\begin{array}{llcll} Frank: & C_F(u,v) & = & log_{\theta_F}\{1+\frac{(\theta_F^u-1)(\theta_F^v-1)}{\theta_F-1}\}, & \theta_F \in (0,\infty), \theta_F \neq 1 \\ Clayton: & C_C(u,v) & = & (u^{-\theta_C}+v^{-\theta_C}-1)^{-\frac{1}{\theta_C}}, & \theta_C \in (0,\infty) \\ Gumbel: & C_G(u,v) & = & exp\{-((-log\ u)^{\theta_G}+(-log\ v)^{\theta_G}-1)^{\frac{1}{\theta_G}}\}, & \theta_G \in [1,\infty) \end{array}$$

In each of the above expressions, θ is the dependence parameter. Using these formulas, we construct the joint probability distributions. Suppose $X_i, i = 1, 2$, are two random variables that follow gamma distribution with shape parameters α_i and rate parameters β_i. Then the probability density function (PDF) of the i^{th} random variable can be written as:

$$f(x_i) = \frac{\beta_i^{\alpha_i}}{\Gamma(\alpha_i)} x_i^{\alpha_i - 1} e^{\beta_i x_i}, \qquad x_i \in (0,\infty), \;\; \alpha_i > 0, \;\; \beta_i > 0, i = 1, 2,$$

where Γ represent the gamma function. The cumulative distribution function of these random variables are given by:

$$F(x_i) = Pr(X_i < x_i) = \frac{1}{\Gamma(\alpha_i)} \gamma(\alpha_i, \beta_i x_i), i = 1, 2,$$

where $\gamma(.,.)$ is the incomplete gamma function. Note that we are constructing joint PDF of two random variables, say X_1 and X_2 such that $X_i \sim Gamma(\alpha_i, \beta_i), i = 1, 2$ with $F(x_1) = u$ and $F(x_2) = v$ and they are dependent. Thus, the bivariate distribution can be constructed using any of the three copula functions defined above with gamma marginals.

3.2. Relationship between Kendall's τ and the dependence parameter θ in copulas

Frank's, Clayton's and Gumbel's copula functions mainly differ regarding tail probability, and it is essential to understand the impact of different choices of copulas in this construction. Comparing the different values of the dependence parameter θ in different choices of copulas with other well-known rank-correlation based measures of dependence, such as Kendall's τ and Spearman's ρ, allows one to have a better understanding of the joint probability distribution. To facilitate this process, we numerically evaluate and plot the relationship between Kendall's τ and the dependence parameters θ from Frank's, Clayton's and Gumbel's copula. As discussed by Nelsen (2006), these relationships can be evaluated using the following expressions:

$$\begin{array}{llll} Frank: & \tau & = 1+\frac{4[D_1(\theta_F)-1]}{\theta_F}, & \theta_F \in (0,\infty), \theta_F \neq 1 \\ Clayton: & \tau & = \frac{\theta_C}{\theta_C+2}, & \theta_C \in (0,\infty) \\ Gumbel: & \tau & = \frac{\theta_G-1}{\theta_G}, & \theta_G \in [1,\infty), \end{array}$$

where $D_1(\theta_F)$ in Frank's copula is the Debey function (Luke, 1969) of order 1 as given below:

$$D_1(\theta_F) = \frac{1}{\theta_F}\int_0^{\theta_F} \frac{t}{e^t-1}dt.$$

Using the above relationships, we evaluate Kendall's τ for a series of the dependence parameter θ in different copulas and plot those in Figure 1.

We clearly observe that Kendall's τ varies substantially at a given value of the dependence parameter θ across different choices of copulas. For example, when $\theta = 5$, calculated value of Kendall's τ from Frank's, Clayton's and Gumbel's copula are given by 0.46, 0.71 and 0.80, respectively. Thus, different choices of copulas represent a different degree of dependence regarding Kendall's τ at a given value of the parameter θ. In general, Gumbel's copula represent a higher degree of dependence followed by Clayton's and Frank's for a given value of $\theta > 2$.

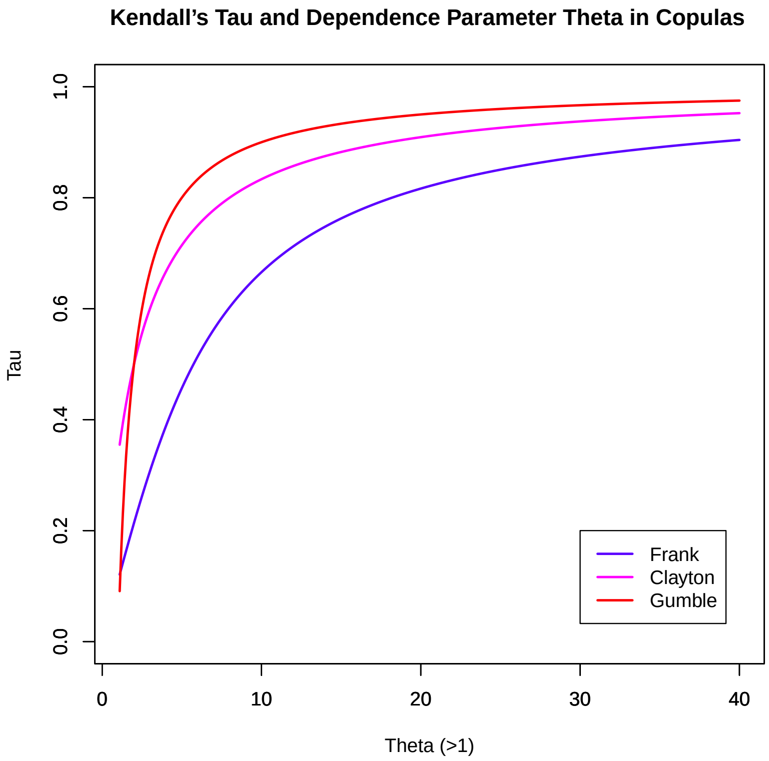

Figure 1. Relationship between Kendall's τ and the dependence parameter θ in different copulas

3.3. Classification rules

In this section, we build the classification rule for a disease D based on the joint probability distribution of two continuous random variables X_1 and X_2. Suppose $D = 1$ indicates the presence and $D = 0$ indicates the absence of the disease, and we consider two dependent biomarkers to classify disease status, denoted by X_1 and X_2, respectively. Our goal here is to find the threshold values of these biomarkers, say x_1 and x_2, that give us maximum classification accuracy for the disease of interest. We define cumulative probabilities from marginal distributions such that $Pr\left[X_1 < x_1\right] = p_1$, $Pr\left[X_2 < x_2\right] = p_2$ and cumulative probabilities from the joint probability distribution of X_1 and X_2 such that $Pr\left[X_1 < x_1 \cap X_2 < x_2\right] = p_{11}$, $Pr\left[X_1 < x_1 \cap X_2 \geq x_2\right] = p_{12}$, $Pr\left[X_1 \geq x_1 \cap X_2 < x_2\right] = p_{21}$, $Pr\left[X_1 \geq x_1 \cap X_2 \geq x_2\right] = p_{22}$. Note that these four joint probabilities defined here are mutually exclusive, and we present those in a 2x2 table below:

Table 1. Partitioned joint cumulative probability distribution of two biomarkers

	X_2		Marginal
X_1	$X_2 < x_2$	$X_2 \geq x_2$	
$X_1 < x_1$	p_{11}	p_{12}	p_1
$X_1 \geq x_1$	p_{21}	p_{22}	$1 - p_1$
Marginal	p_2	$1 - p_2$	1

Let us suppose both tests using these two biomarkers need to be weighted equally with a higher value of each biomarker represents a risk for the disease. Define E as the event that the test using these two biomarkers jointly indicate the risk of developing the disease with $Pr\left[E\right] = Pr\left[X_1 \geq x_1 \cap X_2 \geq x_2\right]$, and E^c be the event such that $Pr\left[E^c\right] = 1 - Pr\left[X_1 \geq x_1 \cap X_2 \geq x_2\right]$. Under this assumption, we need to find x_1 and x_2 such that $Pr\left[E\right] = Pr\left[X_1 \geq x_1 \cap X_2 \geq x_2\right]$ as the joint probability of increased risk of developing the disease using these biomarkers. However, this rule can be redefined based on the priority of tests and the tail probability for the biomarkers of interest. Using the definition given above, the cross tabulation of disease and biomarker categories produce frequencies a, b, c and d corresponding to the four cells, and we present those in the following table:

Table 2. Cross tabulation of disease and biomarkers at a given joint thresholds and corresponding cell frequencies

Classification Rule		Disease (D)		Total
		Yes (1)	No (0)	
Biomarkers X_1 and X_2	$\{x_1, x_2\} \in E$	a	b	a+b
	$\{x_1, x_2\} \in E^c$	c	d	c+d
Total		a+c	b+d	N=a+b+c+d

Using notations from the table above, we define and denote the sensitivity and specificity by P_+ and P_- as follows:

$$P_+ = Pr\left[\{x_1, x_2\} \in E | D = 1\right] = \frac{a}{(a+c)}, \tag{1}$$

$$P_- = Pr\left[\{x_1, x_2\} \in E^c | D = 0\right] = \frac{d}{b+d}, \tag{2}$$

respectively. Note that our goal here is to find the joint threshold values x_1 and x_2 that maximizes the classification accuracy, and we describe the procedure below:
Define Q_+ and Q_- such that

$$Q_+ = Pr\left[\{x_1, x_2\} \in E^c | D = 1\right] = 1 - \frac{a}{a+c} = \frac{c}{a+c}, \tag{3}$$

$$Q_- = Pr\left[\{x_1, x_2\} \in E | D = 0\right] = 1 - \frac{d}{b+d} = \frac{b}{b+d}. \tag{4}$$

Assume, $[X_1, X_2 | D = 1] \sim h(\Theta_1)$ and $[X_1, X_2 | D = 0] \sim h(\Theta_0)$, where $h(.)$ represent the joint probability density function of X_1 and X_2 with vector of parameters Θ_1 and Θ_0 corresponding to the diseased and non-diseased population, respectively. Under this assumption, we can express Q_- as:

$$Q_- = Pr\left[\{x_1, x_2\} \in E | D = 0\right] = 1 - F\left[x_1 | \Theta_0\right] - G\left[x_2 | \Theta_0\right] + H\left[x_1, x_2 | \Theta_0\right], \tag{5}$$

where $F(.)$ and $G(.)$ are marginal CDF of X_1 and X_2, respectively, with $H(.)$ as the joint CDF evaluating at Θ_0 for non-diseased population. For a given value of Q_-, say q_-, we need to solve Equation (5) for x_1 and x_2. We can easily evaluate this expression using a numerical integration technique or utilizing an R package called ‘copula’.

Once we know x_1 and x_2 form the previous step, P_+ can be re-expressed as:

$$P_+ = Pr[\{x_1, x_2\} \in E | D = 1] = 1 - F\left[x_1 | \Theta_1\right] - G\left[x_2 | \Theta_1\right] + H\left[x_1, x_2 | \Theta_1\right]. \tag{6}$$

We can also evaluate this equation using the ‘copula’ package available in R. Since maximum classification accuracy occurs when both true positive and true negative are maximum, we can find such joint threshold values x_1 and x_2, simultaneously solving Equation 5 and 6. A simulation algorithm to find such threshold using a copula-based joint probability distribution is also described in Appendix A.

We can also express the area under the receiver operating characteristics curve (AUC) as follows:

$$AUC = Pr[(\{x_1, x_2\} \in E|D = 1) > (\{x_1, x_2\} \in E|D = 0)]. \tag{7}$$

Given a complex form of the copula-based joint probability distribution, we are not able to find any exact solution of this equation. However, AUC represents the overall classification performance that can be obtained computing the area under the curve plotted based on a series of 1-specificity and sensitivity defined in Equation (5) and (6), respectively. Thus, we can apply numerical integration techniques such as the Trapezoidal or Simpson's rule to compute the area under the curve based on a series of sensitivity and 1-specificity evaluating over the range of possible thresholds.

4. Simulation

In this section, we describe the simulation procedure to find the joint threshold of two dependent biomarkers that maximizes the classification accuracy of a disease. As indicated in Section 2.1, we consider the problem of identifying the joint threshold of creatine kinase (CK) and cardiac troponin (cTn) to classify ST-elevation myocardial infarction (STEMI). To facilitate this process, we simulate multiple data sets using the joint probability distribution of these two biomarkers for cases and controls separately and place them together. Thus, each data set represents a case-control study of STEMI with 50% cases and 50% controls. We then apply the method to determine the joint threshold of these two biomarkers that maximizes the classification accuracy for STEMI.

4.1. Simulation design

Simulation procedure requires starting with a sensible parameter that reflects the disease and biomarkers of interest. We conduct a literature review and identify an article that describes a small case-control study for STEMI (Nusier and Ababneh, 2006). Using the information from this article, we specify the population means (standard deviation) for CK(U/l)/100 among STEMI cases and control as 8.76 (3.20) and 7.45 (2.74), respectively. Similarly, we specify the mean (standard deviation) for Cardiac Troponin - T (μg/l) among cases and controls as 2.76 (1.42) and 1.3 (0.5), respectively. We assume that marginal distribution of CK follows a gamma distribution with shape α_1, rate β_1, and cTn with shape α_2, rate β_2, along with a dependence parameter θ.

Applying the method of moments, we solve for the shape and rate parameter of the marginal probability distribution for CK and cTn. Thus, we identify the set of parameters α_1=7.5, β_1=0.86, α_2=3.8, β_2=1.4 for cases and α_1=7.4, β_1=0.99, α_2=6.8, β_2=4.2 for controls, to construct the bivariate distribution for this simulation. Since we do not have any information on the degree of dependence between these two biomarkers, we conduct the simulation assuming two values of the dependence parameter τ, and each one corresponds to different values of θ across copulas. Thus, we specify the Frank's, Clayton's and Gumbel's copula-based bivariate distribution using gamma marginals. We then determine the joint threshold for these two biomarkers and the algorithm for this simulation given in Appendix A.

Thus, we simulate case-control study data sets using three copula-based joint probability distribution. We repeat this simulation for 96 sets of parameter combinations using 8 parameter estimates (shape and rate parameters indicated in the previous paragraph for cases and controls), as well as two different values of the dependence parameter $\tau = 0.5$ and $\tau = 0.7$ (total number of parameter sets $= 8 \times 4 \times 3 = 96$). Based on the theoretical relationship presented in Section 3.2, when $\tau = 0.5$, the dependence parameter $\theta = 5.8$ in Frank's copula, $\theta = 2.0$ in Clayton's copula and $\theta = 2.0$ in Gumbel's copula. Similarly, for $\tau = 0.7$, the dependence parameter θ in Frank's, Clayton's and Gumbel's copula are given by 11.5, 4.7 and 3.4, respectively. We use the copula (Yan, 2007; Kojadinovic and Yan, 2010; Hofert and Mächler, 2011; Hofert et al., 2014) and CDVine (Brechmann and Schepsmeier, 2013) package in R for this simulation.

4.2. Evaluating performance

We numerically evaluate the performance of the method based on STEMI classification using CK and cTn in three steps. During this process, we first visually examine the shape of the marginal and copula-based joint probability distribution functions using a simulated data set. Next, we simulate 1000 random samples of size 1000 in each from three copula-based joint probability distribution functions. We then evaluate the performance of the joint probability density functions based the relative bias and mean squared error of parameter estimates. To determine the effect of different choices of parameters in this construction, we repeat this procedure for 96 sets of parameter combinations.

Finally, we utilize the probability distribution functions to determine the joint threshold of CK and cTn for STEMI classification. We evaluate the effect of different values of the dependence parameter across copulas based on the area under the receiver operating characteristics curve (AUC). We conduct this simulation for one set of parameters identified by literature review and two different values of the dependence parameter. We repeat this procedure for 5000 samples of size 1000 in each (500 cases and 500 controls) and three different choices of copula-based bivariate distributions. Using the Heidelberger and Welch diagnostic test (Heidelberger and Welch, 1981, 1983), we also assess the convergence of the threshold for disease classification across different choices of copulas.

4.3. Simulation results

In this section, we present the simulation results following the procedure described above. Here, the first sub-section includes some visual examination result of the marginal and joint probability distribution of CK and cTn. In the following sub-section, we present the performance evaluation result of the Frank's, Clayton's and Gumbel's copula-based bivariate distributions. In the last sub-section, we identify and present the joint threshold of these two biomarkers that maximizes the classification accuracy of STEMI across different choices of copulas.

4.3.1. Marginal and joint probability distributions constructed through copula: We prepare a series of plots using simulated data sets and examine the shape of the marginal and joint probability density functions built through different choices of copulas. We present some of these figures in Appendix B. For example, we prepare and submit a two-dimensional plot of marginal distributions for simulated CK and cTn in Appendix Figure 6. This figure shows how the distribution of each of these biomarkers overlaps between cases and controls. We then prepare a plot of Frank's copula-based joint probability distribution using both biomarkers separately for cases and controls and present in Figure 7. We also construct a contour plot superimposing bivariate joint probability density function of cases on top of controls and display in Figure 8. This figure clearly shows how joint probability distribution of these two biomarkers overlap between cases and controls in this experiment. We prepare and examine similar plot for the joint probability distribution function constructed through Clayton's and Gumbel's copula as well, and observe a slightly different shape regarding skewness.

4.3.2. Statistical properties of copula-based joint probability distribution function: During this step, we assess the unbiasedness of the maximum likelihood estimate (MLE) for each parameter from all three copula-based bivariate gamma distribution. We present the simulation results for the selected set of parameters in Table 3. We observe that mean of each parameter estimates is very close to the actual setting, but slightly higher value representing negligible positive bias for most of the scenarios. In general, absolute relative bias is less than 0.5%, but few are close to 0.8%. We also observe small MSE for each set of parameter estimates that represents a better fit of the data.

In addition to the MLE for the dependence parameter θ, we also compute the Kendall's τ and Spearman's rank-correlation coefficient ρ for each simulated sample. We present the summary of these estimates in Table 4 and observe that MLE of θ is very close to the actual value for all scenarios in different choices of copulas. Comparing the dependence parameter estimates θ in copulas with Kendall's τ, we observe that a given value of the parameter represents higher levels of dependence in Gumbel's, followed by Clayton's and Frank's. This result is consistent with the relationship observed in Figure 1. Spearman's rank-correlation ρ estimates obtained during this simulation also follow a similar pattern

Table 3. Relative bias in percent and MSE based on 1000 simulated samples of size 1000 in each using Frank's, Clayton's and Gumbel's copula and gamma marginals

	True Parameter					Relative Bias %					Mean Squared Error				
Copula	α_1	β_1	α_2	β_2	θ	$\hat{\alpha}_1$	$\hat{\beta}_1$	$\hat{\alpha}_2$	$\hat{\beta}_2$	$\hat{\theta}$	$\hat{\alpha}_1$	$\hat{\beta}_1$	$\hat{\alpha}_2$	$\hat{\beta}_2$	$\hat{\theta}$
Frank	7.5	0.86	3.8	1.4	5.8	0.20	0.24	0.15	0.19	-0.06	0.109	0.002	0.026	0.004	0.073
Frank	7.5	0.86	3.8	1.4	11.5	0.15	0.11	0.13	0.08	0.13	0.114	0.002	0.027	0.004	0.160
Frank	7.4	0.99	3.8	1.4	5.8	0.75	0.80	0.37	0.47	-0.12	0.110	0.002	0.027	0.004	0.065
Frank	7.4	0.99	3.8	1.4	11.5	-0.04	-0.01	0.11	0.15	0.05	0.095	0.002	0.027	0.004	0.158
Frank	7.5	0.86	6.8	4.2	5.8	0.31	0.32	0.40	0.44	-0.23	0.114	0.002	0.098	0.041	0.069
Frank	7.5	0.86	6.8	4.2	11.5	0.32	0.32	0.19	0.18	0.01	0.110	0.001	0.087	0.036	0.161
Frank	7.4	0.99	6.8	4.2	5.8	0.37	0.36	0.23	0.25	0.32	0.112	0.002	0.090	0.037	0.065
Frank	7.4	0.99	6.8	4.2	11.5	0.27	0.30	0.19	0.23	0.09	0.099	0.002	0.090	0.036	0.157
Clayton	7.5	0.86	3.8	1.4	2.0	0.20	0.24	0.06	0.09	0.07	0.109	0.002	0.027	0.004	0.013
Clayton	7.5	0.86	3.8	1.4	4.7	0.15	0.11	0.34	0.35	-0.16	0.114	0.002	0.027	0.004	0.056
Clayton	7.4	0.99	3.8	1.4	2.0	0.75	0.80	0.49	0.47	-0.26	0.110	0.002	0.026	0.004	0.014
Clayton	7.4	0.99	3.8	1.4	4.7	-0.04	-0.01	-0.06	-0.03	0.11	0.095	0.002	0.025	0.004	0.054
Clayton	7.5	0.86	6.8	4.2	2.0	0.31	0.32	0.22	0.20	-0.05	0.114	0.002	0.089	0.036	0.015
Clayton	7.5	0.86	6.8	4.2	4.7	0.32	0.32	0.40	0.42	-0.28	0.110	0.001	0.092	0.037	0.055
Clayton	7.4	0.99	6.8	4.2	2.0	0.37	0.36	0.54	0.53	-0.11	0.112	0.002	0.093	0.039	0.015
Clayton	7.4	0.99	6.8	4.2	4.7	0.27	0.30	0.15	0.14	-0.03	0.099	0.002	0.088	0.037	0.049
Gumbel	7.5	0.86	3.8	1.4	2.0	0.34	0.32	0.39	0.42	0.10	0.116	0.002	0.029	0.004	0.004
Gumbel	7.5	0.86	3.8	1.4	3.4	0.28	0.28	0.31	0.36	-0.04	0.102	0.001	0.027	0.004	0.014
Gumbel	7.4	0.99	3.8	1.4	2.0	0.43	0.51	0.52	0.62	-0.10	0.106	0.002	0.028	0.004	0.004
Gumbel	7.4	0.99	3.8	1.4	3.4	0.40	0.40	0.41	0.42	0.09	0.114	0.002	0.029	0.004	0.014
Gumbel	7.5	0.86	6.8	4.2	2.0	0.40	0.47	0.41	0.47	-0.13	0.105	0.001	0.087	0.036	0.004
Gumbel	7.5	0.86	6.8	4.2	3.4	0.41	0.46	0.58	0.64	0.08	0.109	0.002	0.088	0.036	0.015
Gumbel	7.4	0.99	6.8	4.2	2.0	0.31	0.35	0.22	0.31	0.06	0.105	0.002	0.086	0.037	0.004
Gumbel	7.4	0.99	6.8	4.2	3.4	0.42	0.39	0.45	0.39	0.11	0.110	0.002	0.093	0.038	0.015

as we see for Kendall's τ. These results ensure that simulation from the copula-based joint probability density functions is working fine with a negligible deviation from unbiasedness.

However, to check the complete unbiasedness of each parameter in different choices of copulas, we also run the simulation with varying sample size (10,15,20,...,2000), and observe that relative bias for each parameter converges to zero with the increase in sample size. Due to space limitation, we are not including these figures in the article.

4.3.3. Determining the joint threshold: During this step, we determine the joint threshold for CK and cTn together that maximizes the classification accuracy of STEMI, taking into account the dependence between these two biomarkers. Based on the relationship presented in Section 3.2 and the simulation results presented in the previous section, we observe that a fixed value of the dependence parameter

Table 4. Degree of dependence based on 1000 samples of size 1000 in each using Frank's, Clayton's and Gumbel's copula and gamma marginals

	True Parameter						Estimated Association Parameter		
Copula	α_1	β_1	α_2	β_2	τ	θ	$\hat{\theta}$	$\hat{\tau}$	$\hat{\rho}$
Frank	7.5	0.86	3.8	1.4	0.5	5.8	5.797	0.503	0.697
Frank	7.5	0.86	3.8	1.4	0.7	11.5	11.515	0.702	0.888
Frank	7.4	0.99	3.8	1.4	0.5	5.8	5.793	0.503	0.698
Frank	7.4	0.99	3.8	1.4	0.7	11.5	11.505	0.702	0.888
Frank	7.5	0.86	6.8	4.2	0.5	5.8	5.787	0.503	0.697
Frank	7.5	0.86	6.8	4.2	0.7	11.5	11.501	0.702	0.888
Frank	7.4	0.99	6.8	4.2	0.5	5.8	5.818	0.504	0.699
Frank	7.4	0.99	6.8	4.2	0.7	11.5	11.510	0.702	0.888
Clayton	7.5	0.86	3.8	1.4	0.5	2.0	2.000	0.500	0.680
Clayton	7.5	0.86	3.8	1.4	0.7	4.7	4.690	0.700	0.870
Clayton	7.4	0.99	3.8	1.4	0.5	2.0	1.990	0.500	0.680
Clayton	7.4	0.99	3.8	1.4	0.7	4.7	4.710	0.700	0.870
Clayton	7.5	0.86	6.8	4.2	0.5	2.0	2.000	0.500	0.680
Clayton	7.5	0.86	6.8	4.2	0.7	4.7	4.690	0.700	0.870
Clayton	7.4	0.99	6.8	4.2	0.5	2.0	2.000	0.500	0.680
Clayton	7.4	0.99	6.8	4.2	0.7	4.7	4.700	0.700	0.870
Gumbel	7.5	0.86	3.8	1.4	0.5	2.0	2.000	0.500	0.680
Gumbel	7.5	0.86	3.8	1.4	0.7	3.4	3.400	0.710	0.880
Gumbel	7.4	0.99	3.8	1.4	0.5	2.0	2.000	0.500	0.680
Gumbel	7.4	0.99	3.8	1.4	0.7	3.4	3.400	0.710	0.880
Gumbel	7.5	0.86	6.8	4.2	0.5	2.0	2.000	0.500	0.680
Gumbel	7.5	0.86	6.8	4.2	0.7	3.4	3.400	0.710	0.880
Gumbel	7.4	0.99	6.8	4.2	0.5	2.0	2.000	0.500	0.680
Gumbel	7.4	0.99	6.8	4.2	0.7	3.4	3.400	0.710	0.880

θ represents different degrees of dependency in different choices of copulas. Thus, we also conduct this part of the simulation assuming fixed value of Kendall's $\tau = 0.5$ and $\tau = 0.7$ along with other parameters discussed in Section 4.1. Using the relationship presented in Section 3.2, we then back calculate θ for three copulas under consideration. In particular, when $\tau = 0.5$, the dependence parameter $\theta = 5.8$ in Frank's copula, $\theta = 2.0$ in Clayton's copula and $\theta = 2.0$ in Gumbel's copula. Similarly for $\tau = 0.7$, the dependence parameter θ in Frank's, Clayton's and Gumbel's copula are given by 11.5, 4.7 and 3.4, respectively. These settings allow us to have a fare methodological comparison for threshold identification procedure using these three copula-based bivariate distributions.

Thus, we repeat the simulation procedure for 5000 samples and three different choices of copula-based bivariate distributions. Computing the median and 95% empirical confidence interval, we determine thresholds for CK (U/l)/100 and cTn (μg/l) based on Frank's, Clayton's and Gumbel's copula.

Applying numerical integration technique (Trapezoidal rule), we also compute the AUC based on each simulated sample and summarize across different choices of copulas. We present the summary of this simulation study in Table 5.

Based on the Heidelberger and Welch diagnostic test, we were not able to find any problem with the convergence of each series regardless of the choices of copulas ($p > 0.1$ for the test of each series). Assuming $\tau = 0.5$, threshold estimates of these biomarkers using Frank's, Clayton's and Gumbel's copula are given by (5.68, 1.49), (5.72,1.54) and (5.36, 1.50), respectively. Similarly, for $\tau = 0.7$, threshold estimates of these biomarkers are given by (6.85, 1.42), (6.71, 1.45) and (6.59, 1.42), respectively. We also compare the classification accuracy of these biomarkers across different choices of copulas. For this purpose, we prepare density plots of these AUC's from 5000 simulated samples using three different copulas and present them in Figure 2.

The left panel of this figure represents AUC considering $\tau = 0.5$ and the right panel assuming $\tau = 0.7$, along with other parameters. Results from this analysis indicate that all three copulas produce similar classification accuracy of the disease, with a moderate degree of dependence. However, at the higher levels of correlation, Clayton's copula shows slightly better performance compared to Frank's and Gumbel's copula. We also observe that classification accuracy decreases with increasing values of the dependence parameter, regardless of the choices of copulas.

However, in a real experiment the proportion of subjects with the disease may vary. To determine the effect of this parameter, we repeat the simulation using 20% cases and 80% controls as well, keeping all other parameters fixed and observe a very similar results compared to 50% cases and 50% controls. For example, assuming $\tau = 0.5$ in this case, threshold estimates of CK and cTn using Frank's, Clayton's and Gumbel's copula are given by (5.67, 1.49), (5.72,1.54) and (5.36, 1.50), respectively. Similarly, for $\tau = 0.7$, threshold estimates of these biomarkers are given by (6.85, 1.41), (6.71, 1.45) and (6.59, 1.42), respectively.

Table 5. Empirical median and 95 % confidence interval for the threshold and AUC based on 5000 simulated samples at a given value of Kendall's τ

	$\tau = 0.5$			$\tau = 0.7$		
	Threshold CK (U/l)/100	Threshold cTn (μg/l)	AUC	Threshold CK (U/l)/100	Threshold cTn (μg/l)	AUC
Frank	5.68 (5.33, 6.02)	1.49 (1.44, 1.55)	0.791 (0.764, 0.818)	6.85 (6.58, 7.11)	1.42 (1.36, 1.46)	0.733 (0.702, 0.764)
Clayton	5.72 (5.51, 5.94)	1.54 (1.46, 1.61)	0.812 (0.783, 0.835)	6.71 (6.52, 6.91)	1.45 (1.39, 1.53)	0.760 (0.728, 0.789)
Gumbel	5.36 (4.98, 5.73)	1.50 (1.45, 1.56)	0.783 (0.752, 0.811)	6.59 (6.29, 6.91)	1.42 (1.38, 1.47)	0.724 (0.692, 0.755)

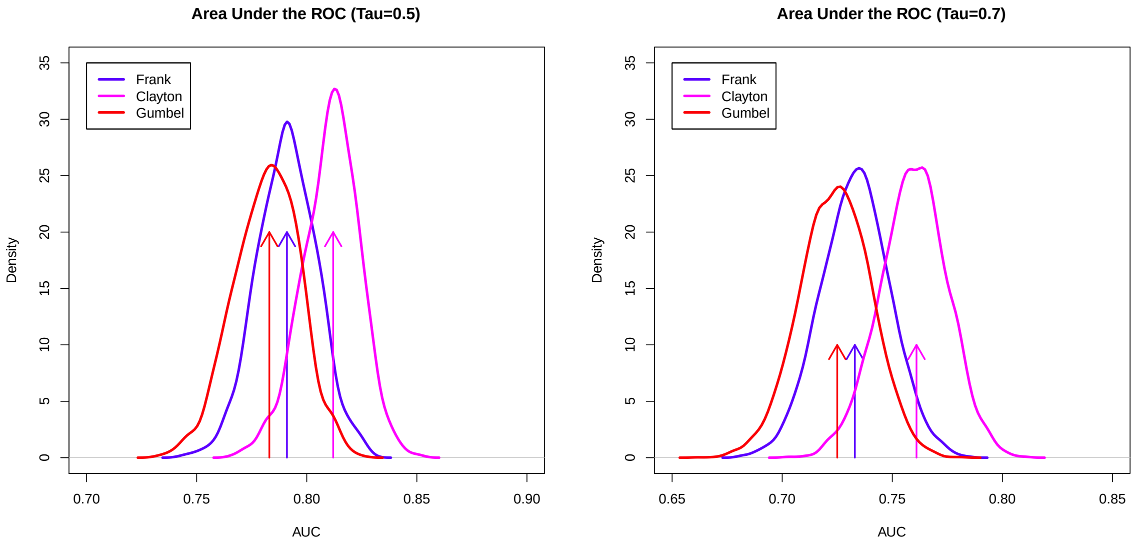

Figure 2. *Density plot for AUC based on 5000 simulated samples of size 1000 each assuming $\tau = 0.5$ (left) and $\tau = 0.7$ (right)*

5. Real data example

Higher levels of Apolipoprotein B to Apolipoprotein A1 ratio (ApoB/ApoA1) and total cholesterol to high-density lipoprotein ratio (TC/HDL) are well-known risk factors for myocardial infarction (MI). In this section, we illustrate the method using a real data set that includes the information on these biomarkers and the outcome. The data set consists of a subset (n=866) of the INTERHEART study sample, which is conducted in 52 countries around the world. This is an age- and sex-matched case-control study of MI, where biological samples are collected and analyzed using strict quality control criteria (McQueen et al., 2008). The subset we are using for this illustration consists of North American and European subjects only.

As a first step of the analysis, we apply a univariate approach called ‘minimum p-value’ and determine the threshold separately for ApoB/ApoA1 and TC/HDL ratio as 0.76 and 4.58, respectively. However, these two biomarkers are highly dependent, and the Spearman’s rank-correlation coefficient estimates among cases and controls are given by 0.86 and 0.91, respectively. Thus, applying these independently determined thresholds to the original sample, we observe that 12% (105/866) individuals lead to a conflicting classification using these two tests. To overcome this problem, we proceed with identifying the joint threshold applying the method developed in Section 3.3, which led to a unique classification for all individuals.

During this process, we first prepare density plots of these two biomarkers stratified by case-control status and present in Figure 3, which suggest that a gamma distribution for each of these biomarkers may be adequate. We also conduct Kolmogorov-Smirnov test for fitted gamma distribution of both biomarkers. P-values corresponding to Case:ApoB/ApoA1, Control:ApoB/ApoA1, Case:TC/HDL and Control:TC/HDL are given by 0.4, 0.7, 0.06 and 0.07, respectively. Thus we are not able to find sufficient evidence against the null hypothesis of the fitted gamma distribution. Next, we test the goodness of fit for three different choices of copulas using marginal gamma distribution for these two biomarkers. During this process, White’s goodness of fit test statistic values for Frank’s, Clayton’s and Gumbel’s copula are given by 0.925 (p=0.336), 3.78 (p=0.052) and 0.378 (p=0.528), respectively

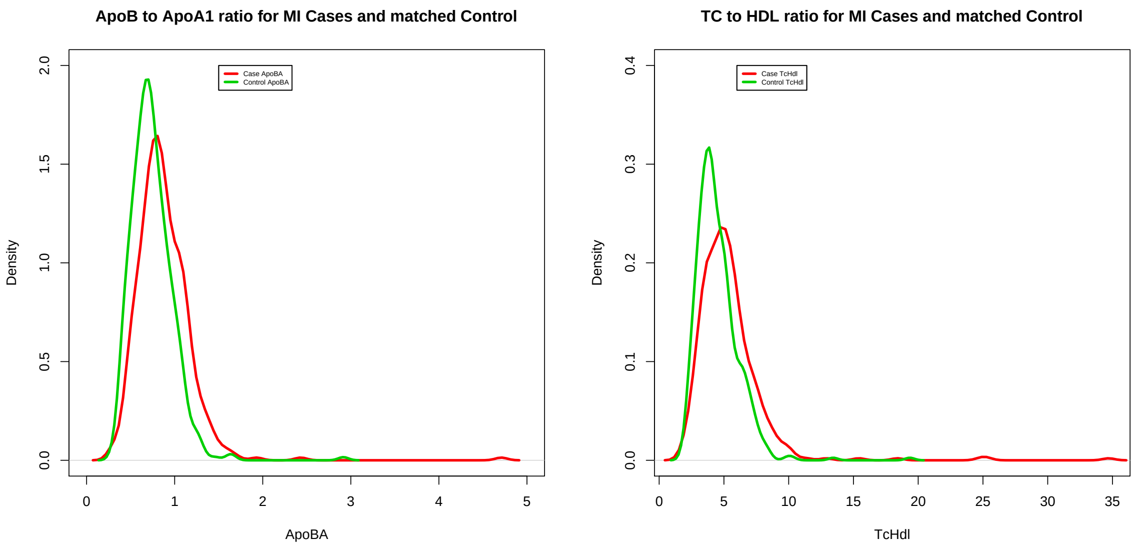

Figure 3. *Univariate density plot of ApoB to ApoA1 ratio and TC to HDL ratio separately by case and control*

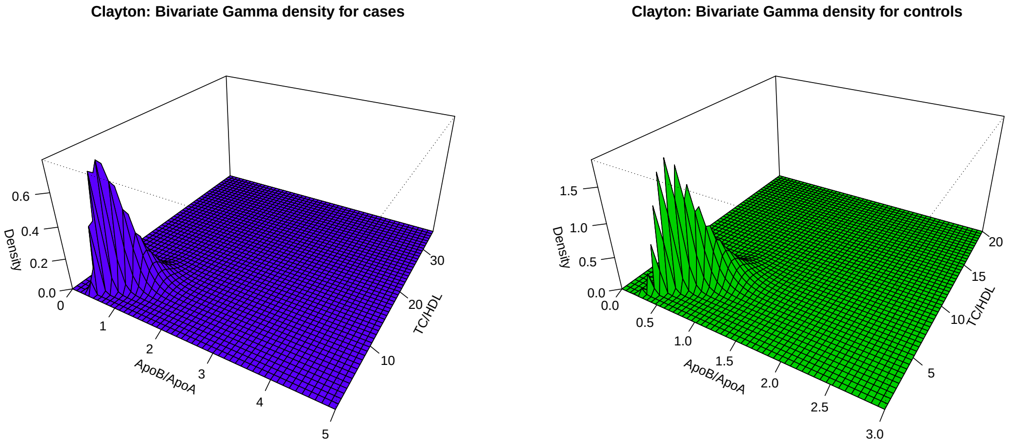

Figure 4. *Bivariate density plot of ApoB to ApoA1 ratio and TC to HDL ratio using Clayton's copula by case and control status respectively*

(Huang and Prokhorov, 2014). The result of this test suggests that any of these copula-based bivariate distributions may be appropriate for this purpose. However, examining the contour plot of each of these copula-based bivariate distributions, we feel that Clayton's copula fits the data better than other copulas. We provide the summary statistic for these biomarkers along with the parameter estimates in Table 6. We also present Clayton's copula-based joint density using gamma marginals in Figure 4, separately for cases and controls.

Table 6. Summary of ApoB to ApoA1 ratio and TC to HDL ratio with fitted parameters using gamma marginals and different copulas

	MI Cases (N=430)		Controls (N=436)	
	ApoB to ApoA1 ratio	TC to HDL ratio	ApoB to ApoA1 ratio	TC to HDL ratio
Mean (SD)	0.873 (0.323)	5.38 (2.73)	0.742 (0.238)	4.49 (1.63)
Median (IQR)	0.837 (0.339)	0.837 (0.339)	0.712 (0.288)	0.712 (0.288)
Gamma (Shape α)	9.7	6.5	11.3	9.5
Gamma (Rate β)	11.1	1.2	15.2	2.1
	Association parameters			
Sparman's ρ	0.86		0.91	
Kendall's τ	0.69		0.74	
Frank's Copula θ_F	11.8		13.4	
Clayton's Copula θ_C	3.71		4.22	
Gymbel's Copula θ_G	2.52		3.22	

Using the joint probability density function and the algorithm described in Appendix A, the thresholds for ApoB/ApoA1 ratio and TC/HDL ratio are given by 0.725 and 4.37. This joint threshold corresponds to the maximum classification accuracy of the disease, with a sensitivity and specificity of 60.3% and 58.4%, respectively. The area under the receiver operating characteristic curve to classify MI using these two biomarkers is given by 0.628. Under certain circumstances, a researcher may also want to determine the threshold at a given sensitivity or specificity, and this method allows us to identify such threshold as well. For example, using this approach we also find the threshold for ApoB/ApoA1 ratio and TC/HDL ratio at 80% specificity as 0.847 and 5.18, respectively, corresponding to 41% sensitivity. In Figure 5, we present a scatter plot (left panel) of the raw data and corresponding Clayton's copula-based bivariate contour plot (right panel) for these two biomarkers. Arrows in this plot represent the threshold with maximum classification accuracy as well as at a given specificity of 80% with 41% sensitivity. We also prepare the contour plot of Frank's and Gumbel's copula-based bivariate distribution for a comparison purpose and present in Appendix Figure 9. These figures ensure a better-fitted model for the data using Clayton's copula and the threshold of these two biomarkers for the classification of MI.

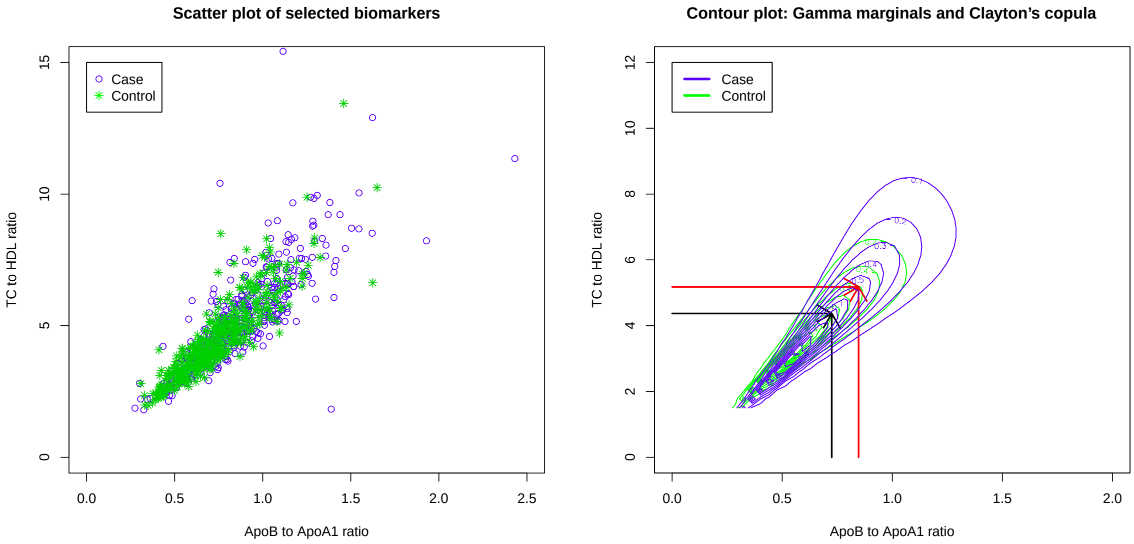

Figure 5. Scatter plot (left) and contour plot (right) for the joint PDF of ApoB to ApoA1 ratio and TC to HDL ratio based on Clayton's copula and by case and control status

6. Discussion

The classification rule for a disease using multiple biomarkers often developed ignoring dependency that leads to a conflicting classification for some individuals. We developed a new method of classifying individuals into binary disease groups using two dependent biomarkers. We first constructed copula-based bivariate distribution for selected biomarkers using Frank's, Clayton's and Gumbel's copula functions. We then developed the classification rule based on the joint probability distributions. Literature review and the density plot of the real data example used in this article motivated us to use marginal gamma distribution for these biomarkers.

We conducted a simulation study to evaluate the performance of the method and then illustrated with a real data example. To understand the effect of different choices of copulas for this purpose, we first evaluated and compared the dependence parameter θ in copulas with Kendall's τ. We observed that a given value of the parameter represented higher levels of dependence in Gumbel's, followed by Clayton's and Frank's copula. This result found to be consistent based on the theoretical relationship as well as computed values using simulated data sets. Relative bias and the mean squared error of all parameter estimates converged to zero with the increase of sample size and for each copula-based joint probability distributions.

We also determined the joint threshold for CK and cTn that maximizes the classification accuracy for STEMI using simulated data sets. The threshold for disease classification converged to a stationary distribution regardless of the choices of copulas. The parameter set-up in this simulation was determined based on a literature review, but we did not have access to the real data set for this experiment. Thus, the threshold identified in this section has a very limited clinical interpretation. Given that a fixed value of the dependence parameter θ represents different degree of dependency in these copulas, we conducted the simulation fixing Kendall's τ and back calculating θ across different choices of copulas. At a higher level of dependency, we observed slightly higher AUC estimate using Clayton's copula-based bivariate distribution compared to Frank's, followed by Gumbel's. However, these differences found to be statistically insignificant. We also observed a decreasing tendency in classification accuracy with the increasing value of the dependence parameter. This is most likely due to less information on the disease status with higher levels of dependency between biomarkers. We have also examined the effect of diseased proportion in this simulation and observed almost identical estimates using 20% cases and 80% controls compared to 50% cases and 50% controls.

Using a real data set, we identified the joint threshold of Apolipoprotein B to Apolipoprotein A1 ratio (ApoB/ApoA1) and total cholesterol to high-density lipoprotein ratio (TC/HDL) for the classification of MI. White's goodness of fit test suggested that any of the three copula-based bivariate distribution may be adequate for this data set. However, the bivariate contour plot with gamma marginals and Clayton's copula showed a better fit to the data, as compared to other copulas. Thus, we used Clayton's copula-based bivariate distribution to determine the threshold of these two biomarkers. The joint threshold for ApoB/ApoA1 ratio and TC/HDL ratio were given by 0.725 and 4.37, respectively, with a sensitivity and specificity of 60.3% and 58.4%, respectively. The area under the receiver operating characteristic curve was given by 0.628. Under certain circumstances, a researcher may also want to identify the threshold at a given sensitivity or specificity, and this method allows us to determine such threshold as well. For example, using this approach, we also found the threshold at 80% specificity as

0.847 and 5.18, respectively, corresponding to 41% sensitivity. Given that we have used a small data set for this purpose, result of this analysis may not lead to a generalizable conclusion.

In an experiment, the marginal distribution of different biomarkers may differ, and a standard set of bivariate distribution may not fit the data well. Thus, introducing copula to construct bivariate distribution with enormous flexibility is one of the main strength of this approach. This method allows one to construct broad classes of bivariate distributions with different marginals, which takes dependency into account and leads to an improved and unique classification. However, this method does not allow one to determine the joint threshold adjusting for additional confounding factors such as ethnicity or age group, which may be of interest for a clinician. We can overcome this limitation determining the threshold unique to each stratum of a confounding factor. The method developed in this article is also restricted to two dependent biomarkers, but we may encounter more than two biomarkers for a disease classification. We hope to extend this approach for three or more dependent biomarkers in future research.

7. Conclusion

In conclusion, the copula-based method works well in identifying the joint threshold of two dependent biomarkers for an outcome classification. This method is flexible and allows modeling broad classes of bivariate distributions that take dependency into account, which leads to an improved and unique classification. The threshold identified using this approach may allow clinicians uniquely classify individuals at risk of developing the disease and plan for early intervention.

8. Acknowledgements

The author gratefully acknowledges the help and support from the Population Health Research Institute and Sigma Laboratory of the Population Genomics Program, McMaster University during the preparation of this manuscript.

References

Bouyé, E., Durrleman, V., Nikeghbali, A., Riboulet, G., and Roncalli, T. (2001). Copulas: An open field for risk management. Retrieved from http://www.thierry-roncalli.com/download/copula-rm.pdf.

Brechmann, E. C. and Schepsmeier, U. (2013). Modeling dependence with C- and D-Vine copulas: The R packatge CDVine. *Journal of Statistical Software*, 52(3):1–27.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1):141–151.

Frank, M. J. (1979). On the simultaneous associativity of F(x, y) and x+y-F(x, y). *Aequationes Mathematicae*, 19(1):194–226.

Genest, C. and Rivest, L.-P. (1993). Statistical inference procedures for bivariate archimedian copulas. *Journal of the American Statistical Association*, 88(423):1034–1043.

Gumbel, E. (1960). Bivariate exponential distributions. *Journal of the American Statistical Association*, 55(292):698–707.

Heidelberger, P. and Welch, P. D. (1981). A spectral method for confidence interval generation and run length control in simulations. *Communications of the ACM*, 24(4):233–245.

Heidelberger, P. and Welch, P. D. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, 31(6):1109–1144.

Hofert, A. M., Kojadi, I., and Maech, M. (2014). Copula: Multivariate dependence with copulas. R package version 0.999-14. Retrieved from http://cran.r-project.org/package=copula.

Hofert, M. and Mächler, M. (2011). Nested archimedean copulas meet R: The nacopula package. *Journal of Statistical Software*, 39(9):1–20.

Huang, W. and Prokhorov, A. (2014). A goodness-of-fit test for copulas. *Econometric Reviews*, 33(7):751–771.

Jaworski, P., Durante, F., Härdle, W., and Rychlik, T. (2009). Copula Theory and Its Applications. In *Proceedings of the Workshop Held in Warsaw, 25-26 September*. Retrieved from http://www.springer.com/gp/book/9783642124648.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*. London, UK: Chapman & Hall.

Kojadinovic, I. and Yan, J. (2010). Modeling multivariate distributions with continuous margins using the copula R package. *Journal of Statistical Software*, 34(9):1–20.

Kuss, O., Hoyer, A., and Solms, A. (2014). Meta-analysis for diagnostic accuracy studies: a new statistical model using beta-binomial distributions and bivariate copulas. *Statistics in Medicine*, 33(1):17–30.

Luke, Y. (1969). *The Special Functions and their Approximations (Vol. II)*. New York: Academic Press, USA.

Mazumdar, M. and Glassman, G. (2000). Categorizing a prognostic value: Review of methods, code for easy implementation and applications to decision making about cancer treatments. *Statistics in Medicine*, 19(1):113–132.

McQueen, M. J., Hawken, S., Wang, X., Ounpuu, S., Sniderman, A., Probstfield, J., Steyn, K., Sanderson, J. E., Hasani, M., Volkova, E., Kazmi, K., and Yusuf, S. (2008). Lipids, lipoproteins, and apolipoproteins as risk markers of myocardial infarction in 52 countries (the INTERHEART study): A case-control study. *The Lancet*, 372(9634):224–233.

Nelsen, R. B. (2006). *An Introduction to Copulas, Springer Series in Statistics (2nd ed.).* New York, NY: Springer-Verlag.

Nusier, M. K. and Ababneh, B. M. (2006). Diagnostic efficiency of creatine kinase (CK), CKMB, troponin T and troponin I in patients with suspected acute myocardial infarction. *Journal of Health Science*, 52(2):180–185.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de lInstitut de Statistique de LUniversité de Paris*, 8:229–231.

Sniderman, A. D., Jungner, I., Holme, I., Aastveit, A., and Walldius, G. (2006). Errors that result from using the TC/HDL C ratio rather than the apoB/apoA-I ratio to identify the lipoprotein-related risk of vascular disease. *Journal of Internal Medicine*, 259(5):455–461.

Tzimas, P., Milionis, H., Arnaoutoglou, H., Kalantzi, K., Pappas, K., and Karfis, E. (2008). Cardiac troponin I versus creatine kinase-MB in the detection of post operative cardiac events after coronary artery bypass grafting surgery. *Journal of Cardiovascular Surgery*, 49(1):95–101.

Vasan, R. S. (2006). Biomarkers of cardiovascular disease: Molecular basis and practical considerations. *Circulation*, 113(19):2335–2362.

Walldius, G., Jungner, I., Aastveit, A. H., Holme, I., Furberg, C. D., and Sniderman, A. D. (2004). The apoB/apoA-I ratio is better than the cholesterol ratios to estimate the balance between plasma proatherogenic and antiatherogenic lipoproteins and to predict coronary risk. *Clinical Chemistry and Laboratory Medicine*, 42(12):1355–1363.

Yan, J. (2007). Enjoy the joy of copulas : With a package copula. *Journal of Statistical Software*, 21(4):1–21.

Appendix A: Simulation algorithm for the classification rule

In this section, we provide the simulation algorithm to find the threshold that maximizes the classification accuracy based on the joint probability distribution of two biomarkers, say X_1 and X_2. Let us assume the marginal distribution of two dependent biomarkers follow gamma distribution with shape α_1, rate β_1 for X_1 and shape α_2, rate β_2 for X_2, along with a dependence parameter θ for the copula under consideration. We describe the procedure for Frank's copula, and the other copulas will follow the same steps. Under this assumption, an algorithm to determine the joint threshold at which maximum classification accuracy occurs described below:

(1) Generate a random sample of desired size from the non-diseased(control) population using the Frank's copula-based joint probability distribution of X_1 and X_2 such that $[X_1\ X_2] \sim C_F(u_0, v_0|\theta_0)$. Here, $[X_1] \sim Gamma(\alpha_{10}, \beta_{10})$ with $u_0 = F(X_1 < x_1)$ and $[X_2] \sim Gamma(\alpha_{20}, \beta_{20})$ with $v_0 = F(X_2 < x_2)$.

(2) Compute the maximum likelihood estimate of all parameters for non-diseased sample.

(3) Follow Steps 1-2 for diseased(case) population such that $[X_1\ X_2] \sim C_F(u_1, v_1|\theta_1)$. Here, $[X_1] \sim Gamma(\alpha_{11}, \beta_{11})$ with $u_1 = F(X_1 < x_1)$ and $[X_2] \sim Gamma(\alpha_{21}, \beta_{21})$ with $v_1 = F(X_2 < x_2)$.

(4) Specify a sequence of q_-(1-specificity) within the interval 0 and 1 and then identify x_1 and x_2 that corresponds to each q_- using Equation (5) based on the set of parameters estimated from non-diseased sample.

(5) Compute the sensitivity P_+ using Equation (6) based on those x_1 and x_2 identified in Step 4 and the set of parameters estimated from the diseased sample.

(6) Based on the series of $\{x_1, x_2\}$ and corresponding specificity and sensitivity computed in Step 4 and 5, identify the joint threshold $\{x_1, x_2\}$ that corresponds to maximum classification accuracy(largest probability of both true positive and true negative).

(7) Based on the series of 1-specificity and sensitivity obtained in Step 4 and 5, compute AUC applying Trapezoidal or Simpson's rule of numerical integration.

(8) Repeat Steps 1-7, s times(number of simulated sample) and create three vectors of estimates such that $\{x_{11}, x_{12}, ..., x_{1s}\}$, $\{x_{21}, x_{22}, ..., x_{2s}\}$ and $\{auc_1, auc_2, ..., auc_s\}$.

(9) Perform Heidelberg and Welch diagnostic test to ensure each series converges to a stationary distribution with a sufficiently large value of s (Heidelberger and Welch, 1981, 1983).

(10) Compute the median and 95% confidence interval based on the empirical distribution of X_1, X_2 and AUC.

We prepared the R programming code for all analysis and figures presented in this paper utilizing some pre-existing packages. In particular, we used copula (Hofert et al., 2014; Yan, 2007; Kojadinovic and Yan, 2010; Hofert and Mächler, 2011) and CDVine (Brechmann and Schepsmeier, 2013) package for the simulation component of the article.

Appendix B: Additional figures

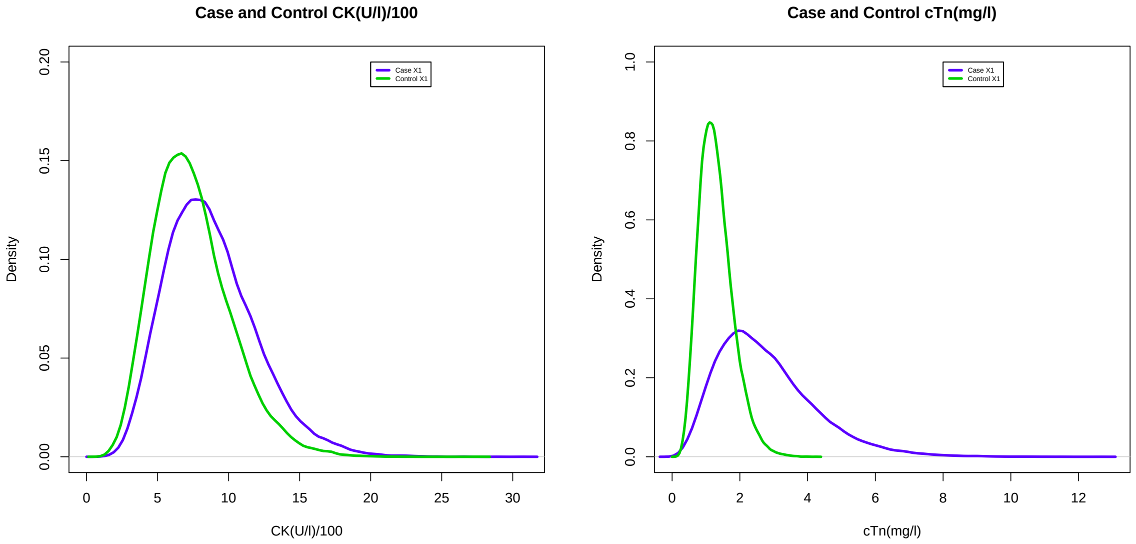

Figure 6. *Univariate density plot of a simulated sample by case and control data separately for CK and cTn*

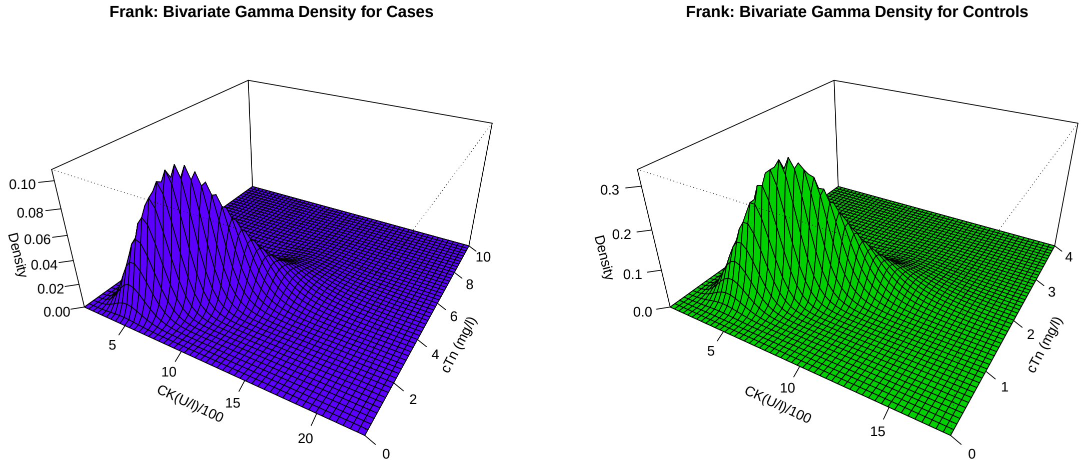

Figure 7. *Bivariate density plot of CK and cTn based on a simulated sample using Frank's copula by case and control status*

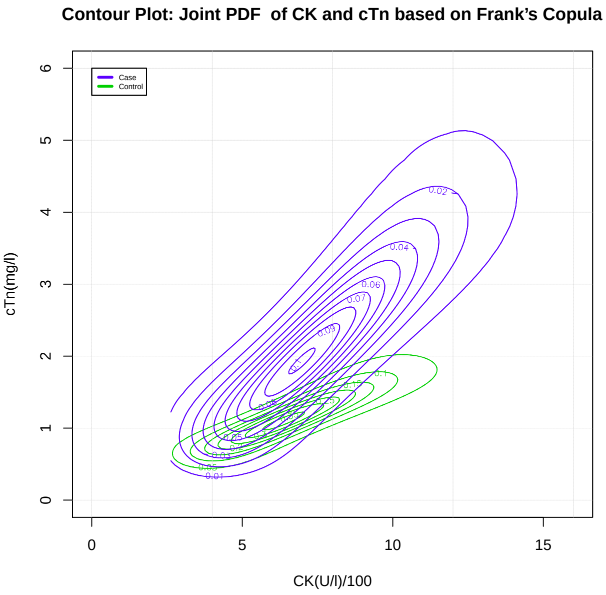

Figure 8. Contour plot for the joint probability distribution of CK and cTn based on a simulated sample using Frank's copula by case and control status

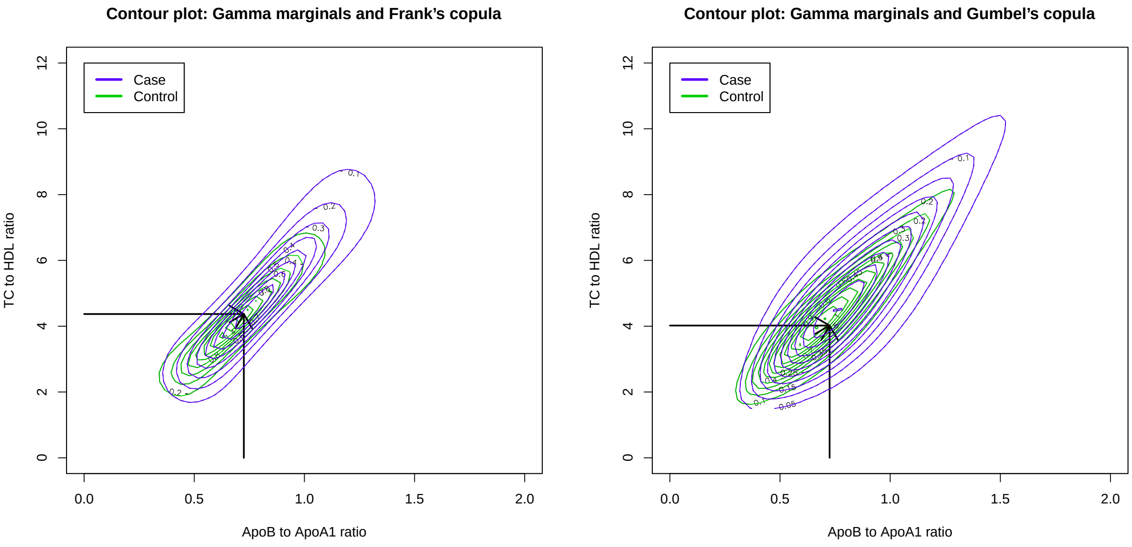

Figure 9. *Bivariate density plot of ApoB to ApoA1 ratio and TC to HDL ratio using Frank's and Gumbel's copula by case or control status*

Chapter 4

Methods for Dimension Reduction and Disease Classification

In general, researchers observe a certain degree of nonlinearity in the gene and miRNA expression data. In a genetic data integration aimed at disease classification, linear principal component analysis (PCA) is a widely used approach. However, a nonlinear PCA might be optimal. This chapter includes an article published in the journal of "Statistical Applications in Genetics and Molecular Biology", where we compared two dimension reduction methods and assessed the contribution of components towards genetic data integration and an outcome classification. We have also developed a simulation algorithm for this purpose and evaluated the performance of these methods with a varying degree of design-level parameters.

Citation: S. Islam, S. Anand, J. Hamid, L. Thabane, J. Beyene (2017): Comparing the performance of linear and nonlinear principal components in the context of high-dimensional genomic data integration. Stat Appl Genet Mol Biol. 16(3):199-216. doi:10.1515/sagmb-2016-0066.

Research Article

Comparing the performance of linear and nonlinear principal components in the context of high-dimensional genomic data integration

S. Islam[†§], S. Anand[†§¶], J. Hamid [§], L. Thabane [†§] and J. Beyene[§*]

Abstract: **Linear principal component analysis (PCA) is a widely used approach to reduce the dimension of gene or miRNA expression data sets. This method relies on the linearity assumption, which often fails to capture the patterns and relationships inherent in the data. Thus, a nonlinear approach such as kernel PCA might be optimal. We develop a copula-based simulation algorithm that takes into account the degree of dependence and nonlinearity observed in these data sets. Using this algorithm, we conduct an extensive simulation to compare the performance of linear and kernel principal component analysis methods towards data integration and death classification. We also compare these methods using a real data set with gene and miRNA expression of lung cancer patients. First few kernel principal components show poor performance compared to the linear principal components in this occasion. Reducing dimensions using linear PCA and a logistic regression model for classification seems to be adequate for this purpose. Integrating information from multiple data sets using either of these two approaches leads to an improved classification accuracy for the outcome.**

Keywords: Principal Component; Kernel PCA; AUC; Copula; Gamma Distribution

[†] Population Health Research Institute, McMaster University and Hamilton Health Sciences, Hamilton, Ontario, Canada.

[§] Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada.

[¶] Department of Medicine, McMaster University, Hamilton, Ontario, Canada.

[*] Correspondence to: Joseph Beyene, Department of Health Research Methods, Evidence, and Impact, McMaster University
1280 Main Street West, Hamilton, ON L8S 4K1; Phone: 905-525-9140, ext. 21333; Email: beyene@mcmaster.ca

1. Introduction

Disease exposure relationships often suffer from multidimensional complex data structures. For example, the progression of a disease can be related to biological, behavioral or genetic factors and some of these data sets usually consist of many variables. Data integration is a process that allows us to combine information from such data sets to perform a specific task. For example, integrated information can be used to classify different clinical outcomes, such as cancer or death. The concept of data integration varies with the context such as business intelligence (Eaton et al., 2008; Haque et al., 2014) or life sciences (Gomez-Cabrero et al., 2014; Reverter et al., 2014) to obtain a meaningful summary of information. In either case, multiple sources of information called domains require integration to perform a specific task. In an article, Hamid et al. (2009) provided a conceptual framework for data integration and discussed some methodological challenges in the context of genomic data. In particular, genetic processes such as the gene or miRNA expression data appear in a very high dimension with a relatively large number of variables, as compared to the number of subjects in the sample. These variables are often highly correlated within and across data sets. Recently, Khan et al. (2014) and Bunte et al. (2016) used a Bayesian group factor analysis approach for joint biclustering of multiple data sources, where exploring clusters of associated data is the main point of interest. However, our goal is to identify few linear combination of expressions that captures the majority of variation from multiple data sources, which can be used in a subsequent procedure such as classification or prediction.

Due to a large number of variables with complex relationships, most of the standard statistical procedures, such as linear regression, fail to utilize this information for classification. As a result, reducing the dimension or summarizing a multidimensional correlated set of exposures is essential. Thus, integrating data sets in the context of disease or death classification can be accomplished in two different steps. First, reduce the dimension with meaningful features through a suitable statistical technique within a domain. Second, develop models to classify an outcome based on extracted features from the various domains. In a recent article, Aguilera et al. (2006) introduced a similar approach known as the principal component logistic regression (PCLR). This approach allows one to remove the redundancy in the data matrix with a smaller number of latent variables, and use the uncorrelated reduced set for subsequent model building procedure. In the context of genomic data integration, many authors consider such an approach to reduce the dimension and subsequently utilize the reduced set to identify the association or disease classification (Chang and Keinan, 2014; Lee et al., 2012; Yi et al., 2012).

Depending on the relationship between variables within a domain, the two broad classes of dimension reduction techniques available are the linear or nonlinear approaches. Some of the linear approaches include linear Principal Component Analysis (PCA) (Pearson, 1901; Hotelling, 1933), Latent Class Analysis (LCA) (Gibson, 1959; Goodman, 1974; Hagenaars and McCutcheon, 2002) and Canonical Correlation Analysis (CCA) (Hotelling, 1936). While PCA and LCA can be used to reduce the dimension of a single set, CCA can be used to reduce the dimensions of two correlated sets. The key characteristic of these approaches is to identify a smaller number of latent variables that can be expressed as a linear combination of observed variables with maximum variance or correlation. Similarly, some of the nonlinear approaches include Sammon's mapping (Sammon, 1969), curvilinear component analysis (Demartines and Herault, 1997), nonlinear PCA (Scholz et al., 2005) and kernel PCA (Schölkopf et al., 1998). These procedures can be considered as nonlinear generalizations of standard PCA.

Many authors consider linear PCA to reduce the dimension of gene expression data and subsequently utilize the information to quantify the degree of association between disease and the extracted principal components (Lee et al., 2012; Yi et al., 2012; Chang and Keinan, 2014). This method has also been used to identify a cluster of associated genes (Yeung, 2001; Lu et al., 2011; Skov et al., 2012), to correct for population stratification in a genome-wide association studies (Price et al., 2006), or to predict an outcome based on different types of clinical variables (Korkeila et al., 2011; Ahmadi et al., 2013; Gloi and Buchanan, 2013). Estimation and test results related to this method depend on the linearity and multivariate normality assumption. However, gene and miRNA expression data often fail to satisfy these assumptions; as a result, nonlinear dimension reduction techniques may be optimal.

In the context of dimension reduction and pattern recognition, Schölkopf et al. (1998) suggested that pre-processing data using kernel PCA could improve the classification performance. For example, this approach performs very well for character or face recognition. The author also showed that a linear classifier is sufficient in this case, as long as features were extracted using the nonlinear approach. Recently, many authors also proposed kernel PCA to reduce the dimension of a genetic process (Liu et al., 2005; Reverter et al., 2010; Schaid, 2010a,b; Gao et al., 2011; Minnier et al., 2015), but this procedure requires to identify a suitable kernel for this computation. Based on the analysis of several data sets using different kernel PCA, Liu et al. (2005) suggested that a polynomial kernel with a degree of two or three performs well for gene expression data sets. However, the performance of kernel PCA over linear PCA in the context of data integration and an outcome classification in different scenarios need to be explored and justified.

The main objective of this research is to compare the performance of the linear and the kernel principal components towards genetic data integration and an outcome classification. We consider two steps for this comparison. First, we assess how well these two approaches extract information from a larger data set to a smaller number of latent variables. Second, we assess the performance of extracted components to classify an outcome. The secondary objective is to develop a simulation algorithm that takes into account the degree of dependence and nonlinearity observed in a genetic process. Using this algorithm, we first evaluate the performance of these methods based on simulated genetic data sets. This procedure allows us to identify the effect of varying the sample size, the proportion of deceased subjects in the sample and the degree of dependence within and across data sets. We then apply both the linear and kernel PCA with a polynomial kernel of degree three to reduce the dimension of gene and miRNA expression from an open source cancer genomic data set (Zhang et al., 2011). Subsequently, we integrate selected principal components from these two data sets along with age and sex, based on the logistic regression model for an outcome classification.

In the next section of this article, we describe the methodological details for dimension reduction, data integration, and simulation. Simulation results comparing linear and kernel approaches with varying design level parameters are presented in Section 3. Using the lung cancer genomic data set, we demonstrate and compare the integration procedure based on these two approaches in Section 4. Finally, we present a summary of our findings with discussion in Section 5 and some concluding remarks in Section 6.

2. Methods

In this section, we describe two dimension reduction methods, namely linear and kernel principal component analysis. We also briefly describe logistic regression model as a subsequent integration procedure for the purpose of classification. We then describe the development of a copula-based simulation algorithm, which allows us to generate data preserving the degree of dependence and nonlinearity observed within and across data sets.

2.1. Dimension reduction methods

Principal component analysis as it currently used was made popular by Hotelling in 1933. However, the idea of a principal axis was first introduced by Karl Pearson in 1901 (Pearson, 1901; Hotelling, 1933). Let X be the centered data matrix of continuous exposures with m rows and n columns, where x_{ij} represents the value for i^{th} gene or miRNA expression and j^{th} individual. Then principal components can be obtained applying the singular value decomposition (SVD) to the data matrix X, known as SVD PCA. With this approach, the matrix X can be uniquely expressed as $X_{m\times n} = U_{m\times m}\Lambda_{m\times n}V^T_{n\times n}$, where Λ is the matrix of singular values along with two orthogonal matrices U and V such that $U^TU = I$ and $V^TV = I$. Given that X is centered, the mean vector and covariance matrix can be written as:

$$\begin{aligned}\mu &= \tfrac{1}{m}\sum_{i=1}^{m} X_i = 0,\\ \Sigma &= \tfrac{1}{m}\sum_{i=1}^{m} X_i X_i^T.\end{aligned}$$

Under this assumption, principal components can also be obtained by diagonalizing the covariance matrix Σ such that:

$$\Sigma v = \lambda v,$$

where λ is the largest eigenvalue and v is corresponding eigenvector. Assuming the rank of Σ is k, one can extract as many as k eigenvalues such that $\lambda_1 > \lambda_2 > ... > \lambda_k$ and their corresponding eigenvectors.

In the above formulation, replacing the linear function with a nonlinear one leads to nonlinear PCA, known as an auto-associative neural network as introduced by Kramer (1991). This procedure requires one to specify mapping and demapping functions G and H, which are selected to minimize the Euclidean norm of the residual matrix E; this can be considered as a generalization of linear PCA. Cybenko (1989) provided details on a function that can be used for this purpose. However, this iterative procedure of constructing nonlinear mapping and demapping is computationally demanding. Following a similar procedure, Schölkopf et al. (1998) introduced another nonlinear approach, known as kernel PCA. This approach starts with mapping X into some high dimensional feature space using a mapping function $\Phi(X)$. We can then apply similar assumptions and procedures for linear PCA, to solve for the kernel principal components based on the following mean vector and covariance matrix:

$$\begin{aligned} \mu &= \tfrac{1}{m}\sum_{i=1}^{m}\Phi(X_i) = 0, \\ \Sigma &= \tfrac{1}{m}\sum_{i=1}^{m}\Phi(X_i)\Phi(X_i)^T. \end{aligned}$$

One of the main advantages of this approach is that the mapping function does not need to be known apriori. Those functions can be approximated using a dot product of features space such that $k(x, x^{'}) := \langle\Phi(x), \Phi(x^{'})\rangle$, known as a kernel trick. As a result, this procedure does not require one to consider any nonlinear optimization procedure. While, there are many different types of kernels available in the literature, the Gaussian or Polynomial kernels are commonly used for this purpose:

$$\begin{aligned} Gaussian \quad &: k(x, x^{'}) = && \exp(-\sigma||x - x^{'}||^2), \\ Polynomial \quad &: k(x, x^{'}) = && (scale.\ \langle x, x^{'}\rangle + c)^d. \end{aligned}$$

In the above expression, c is the offset and d is the degree or order of polynomial kernel. Gaussian is a general purpose kernel used when there is not much information available on the system, where σ is the scale parameter. Incase of polynomial kernel, higher values of d refers to increasing dimension of the polynomial that increases the computational complexity. Thus, required degree of the polynomial need to be carefully assessed and justified depending on the degree of nonlinearity present in a specific data set. Based on the analysis of several gene expression data sets using different kernel

PCA, Liu et al. (2005) suggested that polynomial kernel with a degree of two or three perform better compared to the Gaussian kernel. However, the performance of this approach compared to linear PCA in the context of the gene or miRNA expression data in different scenarios need to be explored.

Note that both of these approaches are known as an unsupervised learning algorithm, where we do not use class level information during feature extraction. As a result, first few principal components extracted from any of these approaches may not be the best for classification of a particular outcome. However, principal components extracted through these approaches are considered unbiased, and it provides a fair methodological comparison. We use the MASS package (Venables and Ripley, 2002) to reduce dimension through linear PCA and the kernlab package (Karatzoglou et al., 2004) to reduce dimension through kernel PCA.

2.2. Classification method

We use a logistic regression model to integrate information from different domains regarding linear or kernel principal components. A brief description of the model is given below. Let Y be the vector of binary responses with 0 represents alive and 1 represents deceased with proportion p. The matrix X of exposures with substantially reduced dimension consists of n rows and k columns, where n_1 observation from group one and n_2 observation from group two with $n_1 + n_2 = n$. Then the logistic regression model can be written as:

$$log(\tfrac{p}{1-p}) \quad = \quad \alpha + \boldsymbol{\beta}\, X.$$

In the above expression, α is the intercept and $\boldsymbol{\beta}$ is the slope coefficients associated with the design matrix X. We use the classification error rate (CER) and the area under the receiver operating characteristic curve (AUC) to assess the performance of linear versus kernel principal components towards data integration and death classification. We use the hmeasure package in R for this computation (Anagnostopoulos et al., 2012).

2.3. Simulation algorithm based on copula

Gene or miRNA expression data are expected to be skewed with a nonlinear relationship across variables. As a result, we need to consider constructing a multivariate distribution that takes into account long tail marginal distributions and the degree of dependence across variables. We propose a copula-based method to construct such distribution. Let $X_1, X_2, ..., X_m$ be a set of random variables such that $Pr(X_1 < x_1) = u_1, Pr(X_2 < x_2) = u_2, ..., Pr(X_m < x_m) = u_m$, respectively. Thereafter, the Frank's family (Frank, 1979) of multivariate distribution was further studied by many authors including Nelsen (2006). This can be constructed through the following expression:

$$C_F^m(u_1, u_2, ..., u_m) = -\frac{1}{\theta_{FM}} log\Big[1 + \frac{\prod_{i=1}^{m}(e^{-\theta_{FM}u_i} - 1)}{(e^{-\theta_{FM}} - 1)^{n-1}}\Big], \theta_{FM} > 0. \tag{1}$$

In the above expression, θ_{FM} represent the multivariate dependence parameter across variables. Suppose $X_i, i = 1, 2, ..., m$ follows a gamma distribution with shape α_i and rate β_i. Then the probability density function of i^{th} random variable can be written as:

$$f(x_i) = \frac{\beta_i^{\alpha_i}}{\Gamma(\alpha_i)} x^{\alpha_i - 1} e^{\beta_i x_i}, \qquad x_i \in (0, \infty), \;\; \alpha_i > 0, \;\; \beta_i > 0, i = 1, 2, ..., m,$$

where Γ represents the gamma function. The cumulative distribution function is given by:

$$F(x_i) = Pr(X_i < x_i) = \frac{1}{\Gamma(\alpha_i)} \gamma(\alpha_i, \beta_i x_i),$$

where $\gamma(., .)$ is the incomplete gamma function. Combining these marginals using the copula function defined above, we construct multivariate distribution with gamma marginals and a dependence parameter θ_{FM}. We prepare an R program utilizing the package called copula (Hofert et al., 2014) for this purpose.

However, the challenging part of this process is to generate data sets that reflect a real experiment, and we propose the following simulation algorithm for this purpose:

(1) Select an observed set of expression data as the basis of the simulation.

(2) Identify the marginal distribution for each expression and compute maximum likelihood estimates of all parameters.

(3) Compute cumulative probabilities based on the fitted marginal distributions.

(4) Compute the maximum likelihood estimate of the dependence parameter fitting Frank's copula, based on the cumulative probabilities obtained in the previous step.

(5) Set parameters obtained in step 2 and step 4 as true parameters for the simulation that reflect an observed experiment.

(6) Generate random data with the desired sample size, proportion deceased, and degree of dependence, based on the multivariate distribution constructed through Frank's copula.

Note that this procedure needs to be repeated separately for the deceased and non-deceased samples to ensure the discriminating ability observed in the real data set. Finally, combine the two sets and consider the set as a random sample from the target population. The Franks copula-based multivariate distribution with a gamma marginal approach we use in the simulation allows us to incorporate the degree of nonlinearity we observe in a real data set. In the supplementary materials of this article, we present a figure (Figure S1) that illustrates how different degree of nonlinear dependency can be incorporated and generated using the copula-based multivariate distribution.

3. Simulation

The main goal of the simulation study is to compare the performances of the linear and the kernel principal component analysis approaches reducing the dimension of a simulated genetic process with varying sample size, the degree of dependence and proportion of subjects deceased. To facilitate this process, we develop and apply a copula-based approach to generate data for this simulation. We also compute relative bias in percentage and corresponding mean squared error (MSE) of different values of dependence parameter with varying sample size and present in the supplementary materials (Table S2). Methodological details of this procedure along with a simulation algorithm are presented in Section 2.3.

3.1. Simulation design

Simulation study requires starting with a sensible set of parameters that reflect a real experiment. We consider a lung cancer data set as the basis of our simulation (ICGC, 2014). We present details and an exploratory analysis of this data set for selected domains in the next section (Section 4). Examining the histogram of each selected gene and miRNA expression data, we observe that most of the marginal distribution of expressions are highly skewed. A visual examination of the Spearman's rank correlation matrix within each domain suggests that running principal component analysis may allow us to remove redundancy in the observed data within each domain. To understand the relationship between different variables within each domain, we also examine the quantile-quantile plot for pairs of gene and miRNA expressions. We present some of these plots in Figure 1. Given that most of the pairs are nonlinearly related, we expect that nonlinear dimension reduction techniques may provide better information, as compared to the linear approach.

For the purpose of simulation setup, we select 20 gene and 20 miRNA expressions with a large interquartile range from the data set to identify suitable parameters for this simulation. We select this subset based on the largest inter-quartile range (IQR) from the entire gene and miRNA expression data set. Distance Correlation (DC) is a commonly used measure and test of significance of nonlinear dependency in gene expression data (Szekely et al., 2007; Guo et al., 2014). We compute each element and corresponding p-values of DC matrix for selected gene and miRNA expression data. We observe that 102 of 190 upper triangular elements of DC matrix based on 20 selected gene expression data satisfy significant ($p<0.05$) nonlinear dependency. Similarly, 69 of 190 elements of DC matrix based on 20 selected miRNA expression data satisfy significant ($p<0.05$) nonlinear dependency. We present a plot of these p-values as a supplementary material (Figure S2). Examining the univariate density plot of the selected set, we use a marginal gamma distribution for each expression. We then apply maximum likelihood method to estimate the shape and rate parameters for each marginal distribution. We also compute the cumulative probabilities for each expression and then estimate the dependence parameter based on Frank's copula. Finally, we use the estimated set of parameters as the basis of our simulation. We present the summary of selected gene, miRNA expression and corresponding parameter estimates in

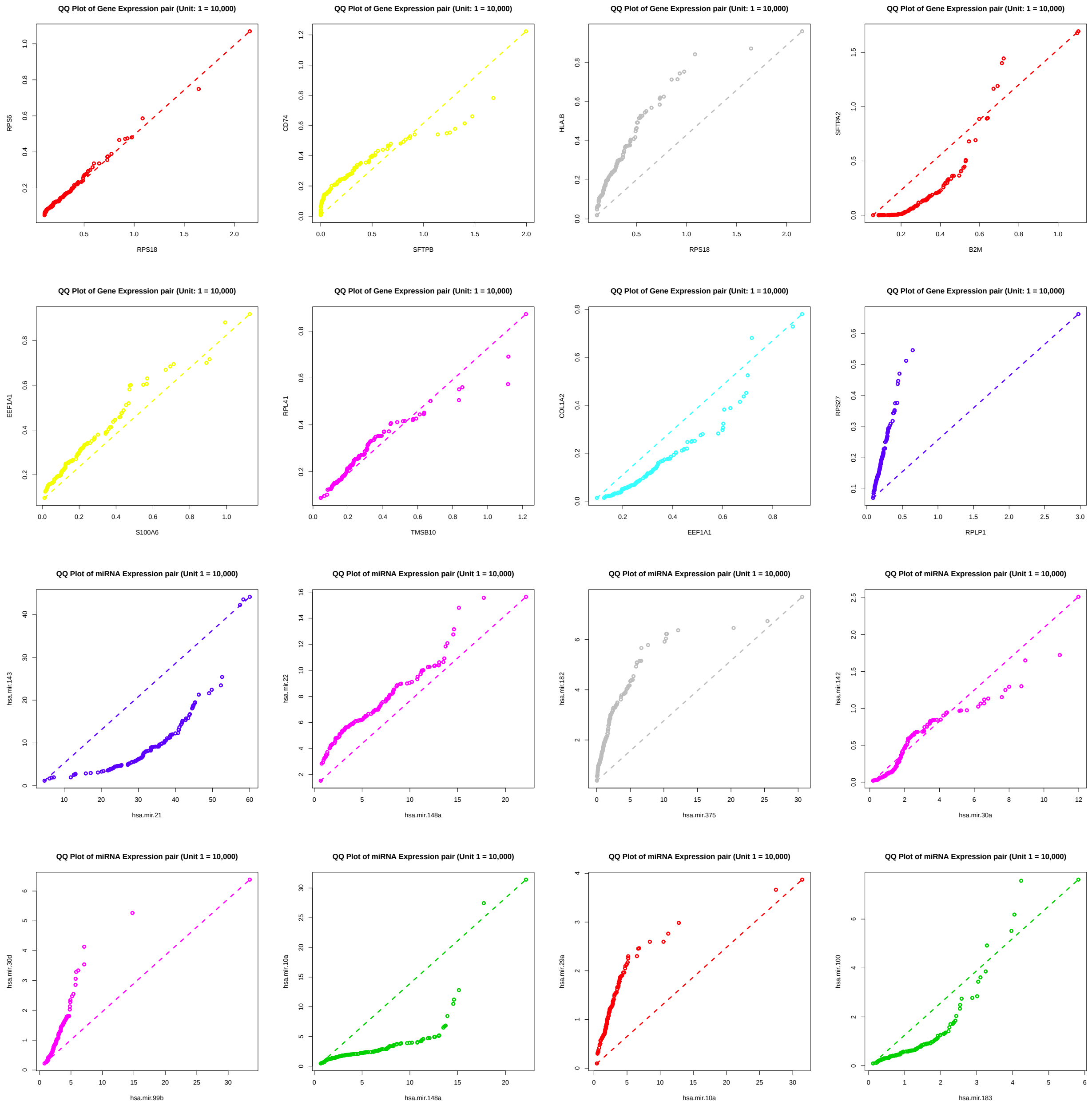

Figure 1. The relationship between selected pairs of gene and miRNA expressions from ICGC lung cancer data

the Table 1 and 2, respectively.

However, simulation result based on a single set of parameters may not be sufficient for this comparison. Thus, we conduct this simulation with varying design level parameters such as sample size, degree of dependence and proportion of deceased subjects in the sample. While varying the sample size ($n = 100, 120, ..., 400$), we keep the proportion of deceased subjects in the sample fixed ($p = 0.25$) and

Table 1. Summary of selected gene expression data (median and IQR: Interquartile range) from the deceased and alive sample and corresponding maximum likelihood estimate (MLE) of shape and rate parameter from the gamma distribution

	Deceased Sample (N=33)			Alive Sample (N=90)		
Gene Name	Median (IQR)	MLE: Shape	MLE: Rate	Median (IQR)	MLE: Shape	MLE: Rate
SFTPB	849 (2503)	0.37	1.92	2172 (4180)	0.64	1.90
RPS18	2827 (1789)	3.90	13.68	2848 (2469)	2.32	6.25
SFTPA1	550 (2001)	0.26	1.03	1278 (3417)	0.41	1.51
FTL	3113 (1843)	4.66	14.61	2984 (2382)	2.62	7.14
CD74	2275 (2004)	1.66	6.17	2156 (2197)	2.08	8.22
HLA.B	2501 (3003)	3.07	9.07	2447 (2033)	2.46	8.88
B2M	3093 (1320)	4.68	16.10	2559 (2304)	3.12	9.93
SFTPA2	304 (1759)	0.27	1.46	866 (2608)	0.42	2.04
TPT1	3230 (1979)	5.72	17.95	3289 (2498)	3.25	7.66
S100A6	1393 (1566)	1.60	7.61	1840 (1955)	1.44	6.11
TMSB10	2167 (1442)	2.78	9.56	2386 (1584)	2.87	9.80
RPL41	2159 (1060)	6.01	27.10	2530 (1564)	5.89	20.86
EEF1A1	2468 (1331)	6.63	25.50	2907 (1843)	4.58	13.73
LOC96610	653 (962)	0.93	8.46	740 (1026)	0.93	7.82
RPLP1	1715 (877)	5.28	26.66	2016 (1312)	2.97	11.59
RPS6	1229 (712)	4.42	30.78	1601 (1274)	2.76	13.88
ACTB	3105 (1468)	8.32	26.18	2386 (1027)	7.44	29.77
RPS27	1458 (1277)	6.07	34.46	1926 (1124)	5.42	24.82
COL1A2	1161 (1953)	1.23	6.89	822 (1013)	1.52	12.70
GAPDH	1625 (1696)	2.57	11.72	1378 (1019)	3.41	21.59

degree of dependence fixed (deceased sample $\theta_{FM} = 0.46$ and alive sample $\theta_{FM} = 0.49$). Similarly, while varying the degree of dependence in the sample ($\theta_{FM} = 0.5, 1.0, ..., 10.0$), we keep the sample size fixed ($n = 200$) and fixed proportion of deceased subjects in the sample ($p = 0.25$). Finally, while varying the proportion of deceased subjects in the sample ($p = 0.20, 0.22, ..., 0.50$), we keep the sample size fixed ($n = 200$) and fixed degree of dependence (deceased sample $\theta_{FM} = 0.46$ and alive sample $\theta_{FM} = 0.49$). The R code for this simulation can be downloaded from the following link: http://beyene-sigma-lab.com/code/.

3.2. Evaluating performance

At the first step of this simulation, we randomly split the generated data into training and test sets with 50% observations in each. Using the training data set, we ran linear and polynomial kernel of degree three principal component analysis to reduce the dimension and extract the first three principal components from each set, and then compute the percent of variance explained by extracted principal components.

Table 2. Summary of selected miRNA expression data (median and IQR: Interquartile range) from the deceased and alive sample and corresponding maximum likelihood estimate (MLE) of shape and rate parameter from the gamma distribution

	Deceased Sample (N=33)			Alive Sample (N=90)		
miRNA Name	Median (IQR)	MLE: Shape	MLE: Rate	Median (IQR)	MLE: Shape	MLE: Rate
hsa.mir.21	336217 (112690)	5.14	0.16	320967 (170376)	7.62	0.23
hsa.mir.143	92607 (104750)	2.02	0.17	81064 (72237)	2.48	0.27
hsa.mir.148a	36006 (55353)	1.63	0.29	63747 (51617)	2.43	0.37
hsa.mir.22	70687 (33276)	6.07	0.84	65080 (31945)	7.88	1.12
hsa.mir.375	8698 (14490)	0.62	0.29	19019 (32217)	1.06	0.32
hsa.mir.182	21039 (17700)	2.42	0.95	26661 (22633)	3.14	1.12
hsa.mir.30a	16180 (12392)	2.34	1.10	19423 (22965)	1.63	0.62
hsa.mir.99b	26873 (11872)	3.65	1.10	25451 (18786)	2.75	0.87
hsa.mir.10a	30856 (18756)	4.24	1.35	23650 (16648)	1.65	0.49
hsa.mir.183	11312 (12920)	1.72	1.23	11817 (11666)	2.58	1.89
hsa.let.7b	13774 (9824)	6.28	4.08	15827 (10158)	4.51	2.58
hsa.mir.30d	8186 (9415)	1.95	1.69	9222 (9179)	2.39	1.98
hsa.mir.200c	6490 (6966)	1.45	1.67	8535 (7977)	2.92	2.98
hsa.mir.10b	7323 (10592)	0.73	0.48	5669 (7178)	1.27	1.66
hsa.let.7a.2	15114 (8998)	4.62	2.73	13779 (8178)	3.75	2.46
hsa.mir.29a	11470 (9300)	3.58	3.02	12195 (7967)	3.99	2.99
hsa.mir.100	7839 (6757)	1.54	1.46	5967 (6390)	1.26	1.18
hsa.mir.30e	12690 (4426)	5.00	3.58	14692 (7815)	5.45	3.52
hsa.mir.101.1	11210 (5136)	6.33	5.76	11940 (7021)	4.42	3.28
hsa.mir.142	3197 (5592)	1.12	2.74	4091 (6053)	1.23	2.48

In the second step, we fit a logistic regression model using the principal components obtained from the training set and validate the performance of classification based on the principal components extracted from the test set. We use the percent of variance explained by the top three principal components, classification error rates, and the AUC to measure the performance of each method.

3.3. Simulation results

First, we present the density plot of the observed data sets for gene and miRNA expression (used to estimate parameters for simulation), separated by the deceased and non-deceased sample, along with a simulated set of expression data in Figure 2. We observe that the density plot of the simulated data set is very similar to those of real data for both the deceased and non-deceased sample, as well as for both the gene and miRNA expression, ensuring a good starting point for this simulation. The Kolmogorov-Smirnov test of gamma marginals using several simulated samples of different sizes also ensures a good fit. We present the test result for the fitted gamma distribution and a simulated data set in the supplementary materials (Table S3). Although, simulated data set expected to vary during each iteration but produces a stable result with repetitive sampling.

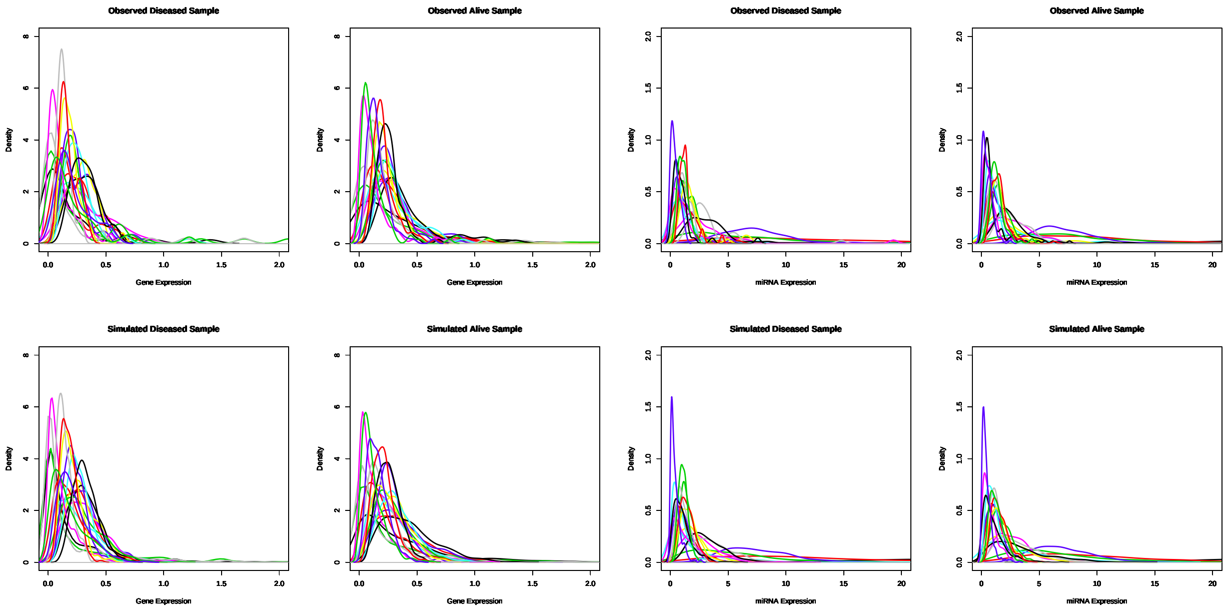

Figure 2. Density plot of a set of gene and miRNA expression from observed (top panel) and simulated (bottom panel) lung cancer data sets with gamma marginals

Based on the simulation results using linear and polynomial kernel approaches with varying sample size, we observe that percent of variance explained by the first three principal components using kernel approach is higher than the linear approach. We also observe that the classification error rates are very similar for both models using the linear and kernel principal components. The error rate also tends to decrease with an increase in the sample size. The median and 95% confidence interval of AUC estimate for death classification using linear or kernel principal components are given by 0.584 (0.582, 0.586) and 0.567 (0.565, 0.569), respectively. Thus, AUC obtained using linear principal components is slightly higher compared to kernel principal components (Figure 3).

Similarly, the results based on the simulation with varying degree of dependence shows that percent of variance explained by first three polynomial kernel principal components explain higher percent of variance compared to linear principal components. However, the classification error rate increases with increasing value of the dependence parameter. Following a similar trend, the median and 95% confidence interval of AUC estimate for death classification using linear or kernel principal

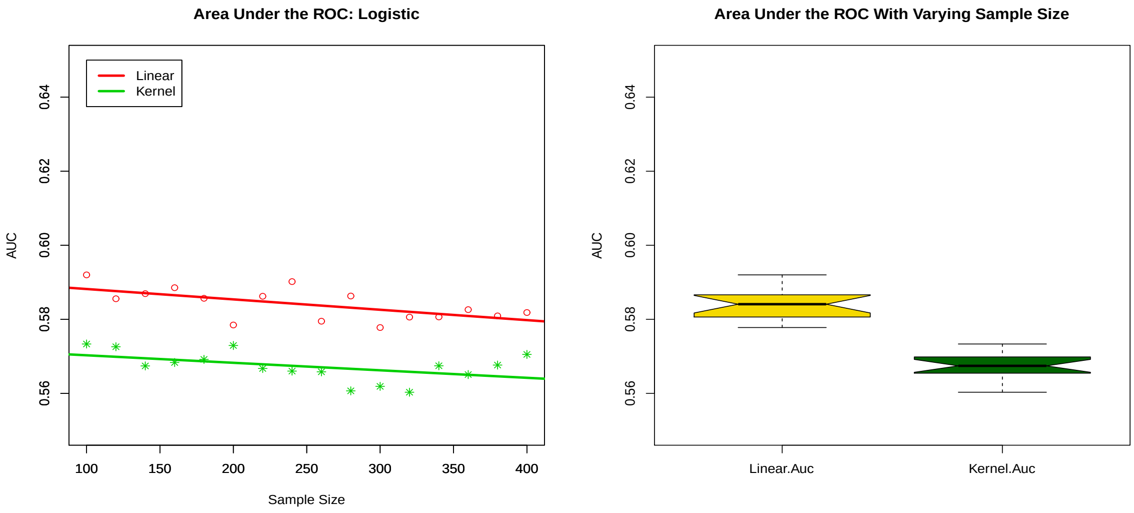

Figure 3. Trends in cross validated AUC and Notched Box Plot with varying sample size

components are given by 0.596 (0.586, 0.606) and 0.573 (0.571, 0.575), respectively. Thus, AUC from the linear principal components again produces slightly higher values with an increasing value of the dependence parameter, compared to the kernel principal components (Figure 4).

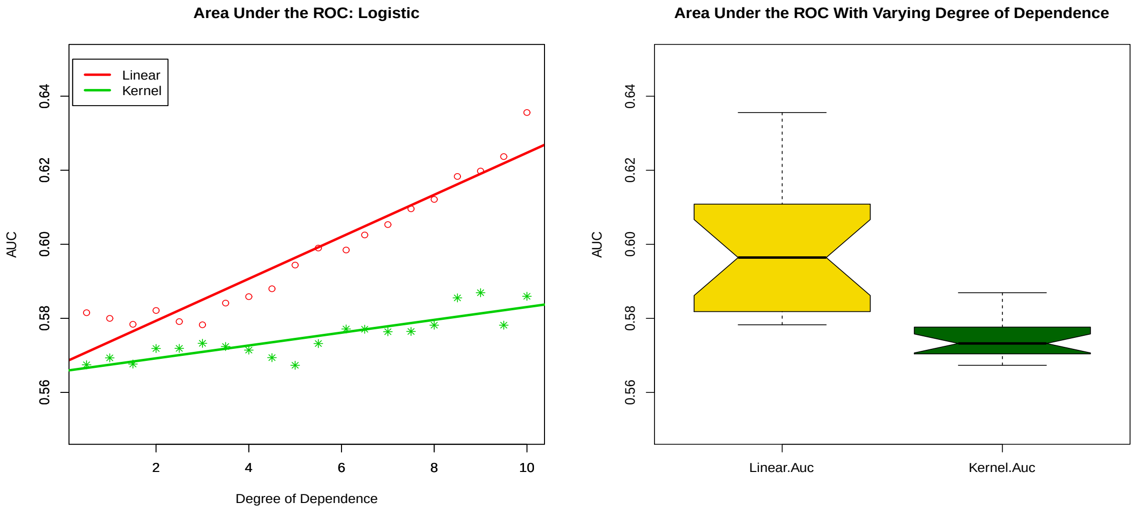

Figure 4. Trends in cross validated AUC and Notched Box Plot with varying degree of dependence

Finally, we conduct the simulation with varying proportion of deceased subjects in the sample. Percent of variance explained by polynomial kernel principal components remains higher compared to linear principal components. We also observe that the classification error rates using linear principal components are almost identical to the kernel principal components, but with an increasing trend with the increase of the deceased proportion in the sample. On the other hand, the trend in AUC is not affected by this variation, but we observe slightly higher values using the linear principal components, compared to kernel principal components (Figure 5). In this set up, the median and 95% confidence interval of AUC estimate for death classification using linear or kernel principal components are given by 0.583 (0.580, 0.586) and 0.563 (0.561, 0.565), respectively.

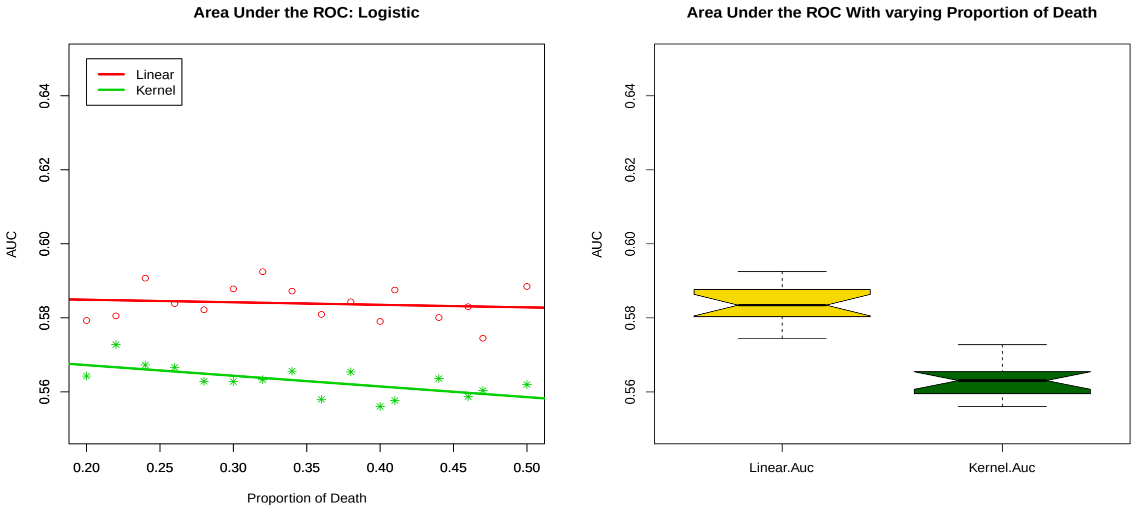

Figure 5. Trends in cross validated AUC and Notched Box Plot with varying proportion of deceased subjects

During this simulation, we also examine the effect of Gaussian kernel towards data integration and death classification. We observe that the Gaussian kernel performs slightly better compared to linear with a lower degree of dependency but performs poorly with a higher degree of dependencies. However, almost all cases, the percent of variance explained by first few Gaussian kernels are very low with higher classification error rates compared to the linear and polynomial kernel. We present the comparative study plot of AUC's using linear, polynomial and gaussian kernel in the supplementary materials (Figure S3).

Thus, in general, we observe a higher percent of variance explained by the first few polynomial kernel principal components, compared to the linear principal components in almost all instances. During the death classification using the linear or polynomial kernel principal components, we observe that the AUC obtained from either of these two approaches tend to increase with increasing value of the dependence parameter but remains constant with varying sample size or the proportion of deceased subjects in the sample. However, based on AUC, the performance of the linear principal components is slightly higher with similar error rates towards data integration and death classification. This result remains consistent with the varying sample size, the proportion of deceased subjects in the sample and the degree of dependence. Thus, principal components extracted using the linear approach shows slightly better performance of death classification in this simulation study with varying design level parameters.

4. Application to a real data set

To demonstrate the described method, we use some of the lung cancer data set introduced during the 13th annual international conference on critical assessment of massive data analysis (CAMDA 2014), as accessed through international cancer genome consortium data portal in January 2014 (ICGC, 2014). During this process, we compare the performances of the linear and kernel approaches towards data integration and death classification in two steps. First, we use the linear and polynomial kernel principal component analysis to reduce the dimension of larger sets. We then compute the percent of variance explained by first three principal components, a measure of performance to capture information from the larger set to a smaller number of latent variables. Second, we use a logistic regression model to integrate extracted features from different domains for classification. We compute the classification error rate (CER) and the area under the receiver operating characteristic curve (AUC) from each model. Finally, we compare the performance of the linear and kernel principal components towards data integration and death classification based on CER and AUC. We also validate results of this analysis using a split-half validation procedure.

Thus, we integrate information from different domains of lung cancer data set and compare the performance of these two approaches using the following steps: 1) Cleaning, transformation, and imputation of missing data. 2) Exploratory analysis such as descriptive statistics and visualization. 3) Dimension reduction of gene and miRNA expression data based on linear and kernel PCA. 4) Integration of the three domains based on the reduced set using logistic regression models for classification. 5)

Compare the performance of these two approaches based on the percent of variance explained by selected principal components, the classification error rate, and the AUC, with validation.

4.1. The lung cancer data set

This data set consist of information on gene expression (GE), micro RNA expression (miRNA), protein expression profiles, somatic copy number variation (CNV) and methylation from 395 lung cancer patients (215 women and 180 men). In addition to the genome sequence data, this data set includes clinical information such as age, sex and alive or death status as well as blood sample type. We restrict our analysis to three domains: gene expression, micro RNA, and age & sex from the clinical data set. In this analysis, death is an outcome of interest, and selected sets of exposures are expected to be highly associated with the outcome.

The main purpose of this analysis is to explore the performance of linear and kernel PCA in the context of high-dimensional genomic data integration and death classification. However, each of these domains consists of information on a different number of subjects. Since the ultimate goal of this analysis is to integrate information from different domains for classification, we identify and match subset across the different data sets. During this step, we identify 123 unique subjects with matched gene, miRNA and clinical data, and we restrict to this subset for the analysis. This sample consists of 33 dead and 90 alive individuals. Thus, the gene expression set leads to a 123x34047 dimensional data matrix and the miRNA expression set leads to a 123x869 dimensional data matrix. As a first step of the dimension reduction, we use an R package called genefilter (Gentleman et al., 2015) to filter out noisy genes and miRNA expressions, as well as the pcaMethods package in R (Stacklies et al., 2007) to impute missing observations in a data matrix.

4.2. Exploratory analysis of gene and miRNA expression data sets

As an exploratory analysis, first we partition the data sets by chromosome number and run both linear and kernel PCA to reduce the dimension of gene and miRNA expression data, separately for each chromosome. Subsequently, we run the logistic regression to identify the performance of the first principal component, as a predictor of death classification. We present the results of this analysis in Table 3, where we report the percent of variance explained by first principal component based on these two approaches, along with the AUC to classify death. We observe that the percent of variance explained by the first linear principal component is less than that of the nonlinear principal component for almost all chromosomes.

Table 3. Percent of variance and the AUC, classifying death based on first linear or polynomial kernel principal component from gene and miRNA expression data stratified by chromosome number

	Gene Expression					miRNA Expression				
		Variance%		AUC-Death			Variance%		AUC-Death	
Chr	Dimension	Linear	Kernel	Linear	Kernel	Dimension	Linear	Kernel	Linear	Kernel
1	123x3763	51	62	0.55	0.54	123x 73	57	77	0.61	0.60
2	123x2263	67	87	0.64	0.65	123x 46	85	93	0.69	0.69
3	123x2005	68	94	0.60	0.61	123x 44	96	99	0.57	0.57
4	123x1353	61	90	0.60	0.60	123x 33	77	73	0.64	0.64
5	123x1598	51	99	0.55	0.57	123x 34	100	100	0.57	0.57
6	123x2829	51	90	0.57	0.59	123x 24	100	100	0.57	0.57
7	123x1663	55	79	0.64	0.62	123x 37	81	95	0.59	0.59
8	123x1213	92	100	0.61	0.62	123x 37	100	100	0.56	0.56
9	123x1453	72	85	0.61	0.61	123x 36	57	67	0.57	0.57
10	123x1370	96	99	0.60	0.60	123x 34	55	80	0.54	0.57
11	123x2364	25	68	0.55	0.64	123x 42	54	52	0.52	0.51
12	123x1901	37	61	0.58	0.67	123x 32	93	100	0.57	0.57
13	123x606	87	72	0.59	0.59	123x 22	69	88	0.55	0.55
14	123x1151	27	90	0.61	0.51	123x 65	96	100	0.58	0.58
15	123x1113	63	100	0.61	0.60	123x 30	81	95	0.50	0.50
16	123x1481	34	91	0.58	0.53	123x 23	94	99	0.56	0.56
17	123x2128	39	82	0.59	0.62	123x 43	86	71	0.53	0.53
18	123x507	76	100	0.58	0.58	123x 15	89	100	0.57	0.57
19	123x2535	55	99	0.52	0.51	123x 84	99	100	0.56	0.56
20	123x1022	59	100	0.54	0.57	123x 24	82	68	0.51	0.51
21	123x435	52	91	0.59	0.54	123x 11	67	87	0.58	0.57
22	123x821	69	83	0.51	0.51	123x 21	80	91	0.56	0.57

We also observe that the AUC for death classification based on the first principal component from each chromosome ranges between 51 to 64%. Results from the gene expression data are very similar to the miRNA expression data and leads us to a conclusion that integrating information from different domains may lead to a better classification of death.

4.3. Lung cancer data integration results

In this sub-section we discuss the data integration results from different domains for death classification. Gene and miRNA expression values usually result in a data matrix with lots of noise. Pre-filtering is a standard approach to remove genes with very little or no information. Further matching and pre-filtering gene and miRNA expression data based on interquartile range (IQR) $>$ 1.5 lead us to a data matrix of dimension 123x1090 and 123x280, respectively.

At this stage, we apply both the linear and kernel approach to extract the top three principal components that explain most of the variation within each domain. We observe that the percent of variance explained by the first three linear principal components from gene and miRNA expressions are 54% and 78%, respectively. On the other hand, the first three polynomial kernel principal components explain about 80 and 83% variance from these two data sets. Next, we run the logistic regression using all three components together considering death as an outcome. We observe that the AUC estimate using linear principal components is higher compared to the estimate from polynomial kernel principal components. This result is counter-intuitive, given that first three polynomial kernel principal components explain more variance compared to the same for the linear principal components. However, the results of this analysis may not be surprising since principal components extracted from either approach did not use the class level information, known as an unsupervised learning algorithm.

We then proceed to the data integration and death classification using eight variables from the three domains: clinical (Age, Sex), the gene (GPC1, GPC2, and GPC3) and the miRNA (MPC1, MPC2, and MPC3). Using these variables, we compute the classification performance of gene and miRNA together and finally all three domains combined, based on logistic regression models. As mentioned in the methods section, principal components extracted from an unsupervised learning algorithm may not be the best for classification of a particular disease of interest or death. Therefore, we also identify the best three principal components, using forward selection procedure and repeat the analysis. During this step, we identify 34^{th}, 75^{th} and 94^{th} linear principal component as best three. Similarly, we identify 88^{th}, 64^{th} and 45^{th} polynomial kernel principal components as best three. We present the result of this analysis in Table 4.

Table 4. The area under the ROC (AUC) classifying death based on top three and best three linear or polynomial kernel principal components using logistic regression

Model with Multiple variables		AUC: Top 3 PC		AUC: Best 3 PC	
Domain	Variables	Linear	Kernel	Linear	Kernel
Gene: Model 1	GPC1+GPC2+GPC3	0.72	0.60	0.78	0.70
miRNA: Model 2	MPC1+MPC2+MPC3	0.57	0.55	0.74	0.76
Gene+miRNA: Model 3	GPC1+GPC2+GPC3+MPC1+MPC2+MPC3	0.73	0.60	0.83	0.85
Age,Sex+Gene+miRNA: Model 4	Age+Sex+GPC1+GPC2+GPC3+MPC1+MPC2+MPC3	0.81	0.76	0.86	0.87

Considering each domain separately, we observe that the top three principal components in multiple logistic regression models improve the performance of death classification with a slightly higher AUC from the linear, compared to the kernel approach. We also observe that integrating gene and miRNA leads to a better classification rate regarding AUC, and the conclusion remains the same as we observe from individual domains. Adding the clinical variables in the model completes the integration of all three domains, leading to a further improvement in the AUC. However, the classification performance of the polynomial kernel principal components remains lower than the linear principal components. We observe similar patterns utilizing the best three principal components in the model, but with a higher classification accuracy using either of these two approaches.

4.4. Cross validation results

All AUCs reported in Section 4.2 and 4.3 are based on the model classifying the same response used to develop the model, which may lead to a spurious result. Split half cross validation is standard procedure to rule out such doubt. To facilitate this process, we split the data into two sets (50% observations in each) with first part as a training set to develop the model and the second part as a test set for validation. To obtain a summary estimate of this validation procedure, we randomly split the data set 500 times. Each time we fit models with the six principal components (three from gene and three from miRNA) extracted from the training set and then classify using the same number of principal components obtained from the test set. During each step, we first compute the percent of variance explained by first three principal components. We observe that first three linear and polynomial kernel principal components explain about 57% (CI: 50%-63%) and 88% (CI: 70%-96%) variation in the gene expression data, respectively. Similarly, these two approaches explain about 80% (CI: 76%-87%) and 90% (CI: 76%-97%) variation in the miRNA expression data, respectively. We summarize and present the cross validated death classification performance result from Model 3 using top three and best three principal components in Figure 6. Left panel in this figure shows the result of top three principal components, where the median AUC using linear principal components is given by 0.574 (CI: 0.568, 0.580) and the AUC using polynomial kernel principal components is given by 0.558 (CI: 0.553, 0.563). Similarly, the right panel shows the result of best three principal components, where the median AUC using linear principal components is given by 0.651 (CI: 0.642, 0.660) and the AUC using polynomial kernel principal components is given by 0.626 (CI: 0.616, 0.636). Thus, the performance of death classification using best three polynomial kernel principal components remain lower than that of using the linear principal components, a result similar to what we observe using top three principal components.

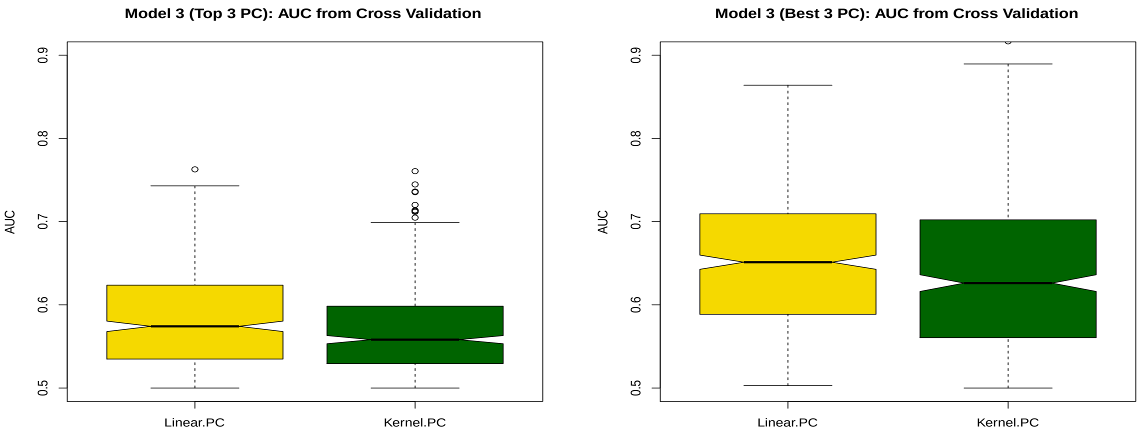

Figure 6. Cross validated AUC based on principal components extracted from gene and miRNA expression (Left panel: Top 3 PC, Right Panel: Best 3 PC)

During this step, we also examine the effect of Gaussian kernel principal components towards data integration and death classification. The Gaussian kernel shows poor classification performance with a very small percent of variance explained by first three components, compared to the same for linear and polynomial kernel. We present the comparative study plot of cross validated AUC's using linear, polynomial and gaussian kernel in the supplementary materials (Figure S4).

4.5. Summary of data integration

Data integration is the process of combining information from multiple sources called domains. During this process, we often encounter data sets with large dimensions such as the gene or miRNA expression of a genetic study. Linear principal component analysis is a widely used approach to reduce the dimension of such data sets. This method relies on the linearity assumption, which often fails to capture the pattern and relationship inherent in the data. As a result, nonlinear approaches might be optimal in this situation. However, the advantage of using nonlinear principal components in the context of genomic data integration and disease classification needs to be explored and justified.

We provided a step-by-step data integration procedure for three domains of a lung cancer data set obtained from the ICGC data portal. We explored and integrated gene expression, miRNA expression, age and sex to classify death due to the disease. Exploring the raw data through measures of location, association, and pairwise quantile-quantile plot suggested that there exists some degree of nonlinearity across many different pairs of gene and miRNA expression data. We applied the linear and polynomial kernel principal component analysis to reduce the dimension of these two data sets. We observed that the first few polynomial kernel principal component carry more information, using percent variance explained, as compared to the linear principal components. However, integrating information from different domains based on kernel principal components and using it for classification produced slightly lower AUC, as compared to the same using linear principal components. Cross validation based on the polynomial kernel principal components produced a similar result with lower AUC and similar error rate.

5. Discussion

The main goal of this research was to compare the performance of the linear and the kernel principal components towards genetic data integration and an outcome classification. We evaluated the performance of these two approaches based on simulated and real data sets. We developed a copula-based simulation algorithm for this purpose. Given that a simulation study is expected to start with sensible parameters, we used a subset of the lung cancer data set to identify parameters for this purpose. This procedure allowed us to simulate random samples, which preserved the degree of dependence and nonlinearity observed within and across gene and miRNA expression of the target population. This procedure also allowed us to generate data that reflects the gene and miRNA expression of deceased and non-disease sample related to lung cancer with additional scenarios. Based on the algorithm developed, we conducted an extensive simulation to compare the performances of these two approaches towards data integration and death classification. During this simulation, we varied the sample size, the proportion of deceased subjects in the sample and the degree of dependence, to determine the effect of each design level parameter towards the performance of these two approaches. We also demonstrated the data integration procedure and compared these two approaches using some of the lung cancer data set. We accessed the data set through the data portal of the international cancer genome consortium.

In general, results of this simulation indicated that the first three polynomial kernel principal components explain a higher percent of variance as compared to the same for the linear principal components. This result remained consistent with varying sample size, the proportion of deceased subjects in the sample and different values of the dependence parameter. Comparing the classification error rate using linear and kernel principal components, we observed that estimates are very similar across different scenarios. In general, the AUC obtained from either of these two approaches tended to increase with increasing value of the dependence parameter but remained similar with varying sample size or the proportion of deceased subject in the sample. However, the AUC obtained using the linear principal components was almost always higher regardless of sample size, the degree of dependence or the proportion of deceased subjects in the sample. We have also examined the effect of Gaussian kernel towards data integration and death classification. We observed slightly better performance of Gaussian kernel compared to linear with a lower degree of dependency but performed poorly with high dimension or a higher degree of dependencies.

During the analysis of the lung cancer data set, we observed a certain degree of nonlinearity across gene and miRNA expression data. We also observed that the first few polynomial kernel principal components carry more information on the expression level reducing the dimension of the genetic process. Integration part of this analysis showed that logistic regression models using the linear principal components provide better performance of classification, compared to the kernel principal components. We recognize that the first few principal components extracted from an unsupervised learning algorithm may not be the best for classification of a particular outcome. However, principal components extracted from this approach is considered unbiased, and it provides a fair methodological comparison for both the linear and kernel approaches. A researcher may also be interested in the best three principal component as opposed to the top three for classification purpose. Thus, we also identified the best three principal components by applying forward selection procedures and ran similar analysis as conducted for the top three principal components. As expected, the comparative study result remained unchanged with the revised set but provided better classification rates using either of these two approaches.

Thus, we observed consistent results based on the simulation study with different scenarios, as well as, based on the real data analysis. Despite having a larger percent of variance explained by first few principal components, we were not able to find any benefit of using polynomial kernel principal components during the disease or death classification, as compared to the linear principal components. Based on the analysis of lung cancer data, we also observed that integrating information from multiple data sets using either of these two approaches lead to an increased value of AUC for death classification.

We also recognize that there are several limitations of this study. For example, we were not able to find any advantages of using kernel principal components on this occasion, but this analysis is restricted to polynomial and Gaussian kernel only. There are may other kernels available in the literature that needs to be explored and compared to the linear approach. Kernel approach may also be useful for other purposes, such as identifying a cluster of genes or other types of data with higher degrees of nonlinearity. The sensitivity of the kernel approach for these purposes is yet to be compared with the linear approach. The data referred to this article includes a few other sets such as protein expression, somatic CNV, and methylation data. It would be interesting to see how well this information can be integrated to improve the performance of death classification. Depending on the marginal distributions of selected expressions, the copula-based simulation algorithm may need to be revised as well, along with different choices of copulas.

6. Conclusion

In general, the first few kernel principal components show poor performance compared to the linear principal components in this occasion. Reducing dimensions using linear PCA and a logistic regression model for classification seems to be adequate for this purpose. Integrating information from multiple data sets using either of these two approaches lead to an improved classification accuracy for the outcome of interest. However, the analysis presented in this article is restricted to the polynomial and Gaussian kernel only, but there are many other kernels available in the literature that needs to be explored and compared to the linear approach.

7. Acknowledgements

The author gratefully acknowledges the help and support from the Population Health Research Institute and Sigma Laboratory of the Population Genomics Program, McMaster University during the preparation of this manuscript.

References

Aguilera, A. M., Escabias, M., and Valderrama, M. J. (2006). Using principal components for estimating logistic regression with high-dimensional multicollinear data. *Computational Statistics & Data Analysis*, 50(8):1905–1924.

Ahmadi, H., Mitra, A. P., Abdelsayed, G. A., Cai, J., Djaladat, H., Bruins, H. M., and Daneshmand, S. (2013). Principal component analysis based pre-cystectomy model to predict pathological stage in patients with clinical organ-confined bladder cancer. *BJU International*, 111(4 Pt B):E173.

Anagnostopoulos, C., Hand, D. J., and Adams, N. M. (2012). Measuring classification performance : the hmeasure package. https://cran.r-project.org/web/packages/hmeasure/vignettes/hmeasure.pdf.

Bunte, K., Leppaaho, E., Saarinen, I., and Kaski, S. (2016). Sparse group factor analysis for biclustering of multiple data sources. *Bioinformatics*, 32(16):2457–2463.

Chang, D. and Keinan, A. (2014). Principal component analysis characterizes shared pathogenetics from genome-wide association studies. *PLoS Computational Biology*, 10(9):e1003820.

Cybenko, G. (1989). Degree of approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 9(3):303–314.

Demartines, P. and Herault, J. (1997). Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets. *IEEE Transactions on Neural networks*, 8(1):148–154.

Eaton, S., Ostrander, M., Santangelo, J., and Kamal, J. (2008). Managing data quality in an existing medical data warehouse using business intelligence technologies. *AMIA Annual Symposium Proceedings*, page 1076.

Frank, M. J. (1979). On the simultaneous associativity of F(x, y) and x+y-F(x, y). *Aequationes Mathematicae*, 19(1):194–226.

Gao, Q., He, Y., Yuan, Z., Zhao, J., Zhang, B., and Xue, F. (2011). Gene- or region-based association study via kernel principal component analysis. *BMC Genetics*, 12(1):75.

Gentleman, R., Carey, V., Huber, W., and Hahne, F. (2015). Genefilter: Methods for filtering genes from high-throughput experiments. R package version 1.53.0.

Gibson, W. A. (1959). Three multivariate models: Factor analysis, latent structure analysis, and latent profile analysis. *Psychometrika*, 24(3):229–252.

Gloi, A. M. and Buchanan, R. (2013). Dosimetric assessment of prostate cancer patients through principal component analysis (PCA). *Journal of Applied Clinical Medical Physics*, 14(1):3882.

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

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2):215.

Guo, X., Zhang, Y., Hu, W., Tan, H., and Wang, X. (2014). Inferring Nonlinear Gene Regulatory Networks from Gene Expression Data Based on Distance Correlation. *PLoS ONE*, 9(2):1–7.

Hagenaars, J. A. and McCutcheon, A. L. (2002). *Applied Latent Class Analysis*. Cambridge University Press.

Hamid, J. S., Hu, P., Roslin, N. M., Ling, V., Greenwood, C. M. T., and Beyene, J. (2009). Data integration in genetics and genomics: methods and challenges. *Human Genomics and Proteomics*, 8690(1):1–13.

Haque, W., Urquhart, B., Berg, E., and Dhanoa, R. (2014). Using business intelligence to analyze and share health system infrastructure data in a rural health authority. *JMIR Medical Informatics*, 2(2):e16.

Hofert, A. M., Kojadi, I., and Maech, M. (2014). Copula: Multivariate dependence with copulas. R package version 0.999-14. Retrieved from http://cran.r-project.org/package=copula.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *The Journal of Educational Psychology*, 24:417–441,498–520.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28(3):321–377.

ICGC (2014). International Cancer Genome Consortium Data Portal. Accessed in January 15, 2014.

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004). kernlab An S4 package for kernel methods in R. *Journal of Statistical Software*, 11(9):1–20.

Khan, S. A., Virtanen, S., Kallioniemi, O. P., Wennerberg, K., Poso, A., and Kaski, S. (2014). Identification of structural features in chemicals associated with cancer drug response: A systematic data-driven analysis. *Bioinformatics*, 30(17):497–504.

Korkeila, E. A., Sundstrom, J., Pyrhonen, S., and Syrjanen, K. (2011). Carbonic anhydrase IX, hypoxia-inducible factor-1alpha, ezrin and glucose transporter-1 as predictors of disease outcome in rectal cancer: multivariate Cox survival models following data reduction by principal component analysis of the clinicopathological. *Anticancer Research*, 31(12):4529–4535.

Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37(2):233–243.

Lee, S., Epstein, M. P., Duncan, R., and Lin, X. (2012). Sparse principal component analysis for identifying ancestry-informative markers in genome wide association studies. *Genetic Epidemiology*, 36(4):293–302.

Liu, Z., Chen, D., and Bensmail, H. (2005). Gene expression data classification with kernel principal component analysis. *Journal of Biomedicine & Biotechnology*, 2005(2):155–159.

Lu, J., Kerns, R. T., Peddada, S. D., and Bushel, P. R. (2011). Principal component analysis-based filtering improves detection for Affymetrix gene expression arrays. *Nucleic Acids Research*, 39(13):1–8.

Minnier, J., Yuan, M., Liu, J. S., and Cai, T. (2015). Risk Classification with an Adaptive Naive Bayes Kernel Machine Model. *Journal of American Stattistical Association*, 110(509):393–404.

Nelsen, R. B. (2006). *An Introduction to Copulas, Springer Series in Statistics (2nd ed.)*. New York, NY: Springer-Verlag.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559 – 572.

Price, A., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909.

Reverter, F., Vegas, E., and Oller, J. M. (2014). Kernel-PCA data integration with enhanced interpretability. *BMC Systems Biology*, 8 Suppl 2:S6.

Reverter, F., Vegas, E., and Sánchez, P. (2010). Mining gene expression profiles: An integrated implementation of kernel principal component analysis and singular value decomposition. *Genomics Proteomics & Bioinformatics*, 8(3):200–210.

Sammon, J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, C-18(5):401–409.

Schaid, D. J. (2010a). Genomic Similarity and Kernel Methods I : Advancements by Building on Mathematical and Statistical Foundations. *Human Heredity*, 70:109–131.

Schaid, D. J. (2010b). Genomic Similarity and Kernel Methods II : Methods for Genomic Information. *Human Heredity*, pages 132–140.

Schölkopf, B., Smola, A., and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319.

Scholz, M., Kaplan, F., Guy, C. L., Kopka, J., and Selbig, J. (2005). Non-linear PCA: A missing data approach. *Bioinformatics*, 21(20):3887–3895.

Skov, V., Thomassen, M., Riley, C. H., Jensen, M. K., Bjerrum, O. W., Kruse, T. A., Hasselbalch, H. C., and Larsen, T. S. (2012). Gene expression profiling with principal component analysis depicts the biological continuum from essential thrombocythemia over polycythemia vera to myelofibrosis. *Experimental Hematology*, 40(9):771–780.

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.

Szekely, G. J., Rizzo, M. L., and Baki, N. K. (2007). Measuring and Testing Dependence by Correlation of Distances. *The Annals of Statistics*, 35(6):2769–2794.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. New York: Springer, 4th edition.

Yeung, K. Y. (2001). Principal component analysis for clustering gene expression data. *Bioinformatics*, 21(13):3009–3016.

Yi, H., Wo, H., Zhao, Y., Zhang, R., Bai, J., Wei, Y., and Chen, F. (2012). Gene-based principal component logistic regression model and its application on genome-wide association study. *Zhonghua liu xing bing xue za zhi*, 33(6):622–5.

Zhang, J., Baran, J., Cros, A., Guberman, J. M., Haider, S., Hsu, J., Liang, Y., Rivkin, E., Wang, J., Whitty, B., Wong-Erasmus, M., Yao, L., and Kasprzyk, A. (2011). International cancer genome consortium data portal: A one-stop shop for cancer genomics data. *Database*, 2011(1):1–10.

Supplementary Materials

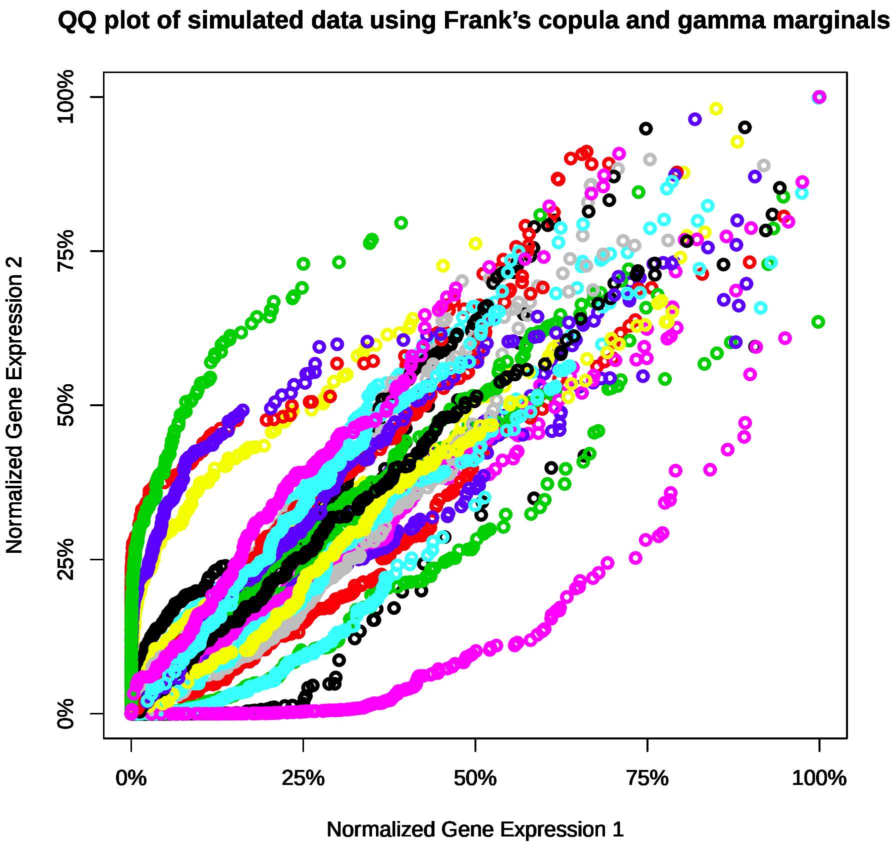

Figure S1. Pair-wise QQ plot of a set of simulated expression data with different shape, rate and dependence parameter using copula-based multivariate distribution

Note: The Franks copula-based multivariate distribution with a gamma marginal approach used in our simulation allows us to incorporate the degree of nonlinearity observed in the real data set. The above figure illustrates how different degree of nonlinear data can be incorporated and generated using copula-based multivariate distribution. In this illustration, we first randomly generate shape, rate and dependence parameter within the range of the observed set of parameters and then use those parameters to generate pairs of expression data. We then superimpose 30 pair-wise QQ plot in the same figure. Deviation of quantiles from the diagonal line in this plot shows the degree of nonlinearity in a specific pair. For example, the set of parameters (Theta=5.1, Shape1=1.7, Rate1=15.9, Shape2=2.2, Rate2=28.3) produce linearly dependent expression and the set of parameters (Theta=3.7, Shape1=0.2, Rate1=28.0, Shape2=3.9, Rate2=26.9) produce higher degree of nonlinear dependency. In the table below (Table S1), we also provide 30 set of parameters used in the simulation. Clearly, the degree of nonlinearity in this multidimensional complex data set depends on many parameters such as dependence, shape, and rate of each marginal.

Table S1. Randomly generated set of parameters within the observed range of parameters

Dependence Theta	Gamma Shape 1	Gamma Rate 1	Gamma Shape 2	Gamma Rate 2
7.0	0.9	25.8	2.1	3.4
8.9	4.1	26.5	3.8	32.2
3.5	2.6	27.2	0.8	26.9
4.2	2.7	33.7	3.9	30.5
3.2	2.6	15.8	4.8	26.8
9.0	1.2	24.4	1.3	11.5
9.1	0.5	3.4	5.6	10.3
9.8	2.3	21.9	2.5	31.1
7.0	0.5	2.4	2.2	4.2
7.7	3.8	15.6	1.4	18.1
5.1	1.7	15.9	2.2	28.3
5.7	1.8	33.2	1.9	18.1
8.1	2.5	27.4	0.6	10.3
3.6	3.1	5.1	2.6	33.0
6.0	1.1	22.6	2.8	22.8
0.2	3.0	8.3	1.8	7.4
2.5	4.9	6.5	4.8	8.7
3.4	0.1	26.5	4.6	30.2
8.0	2.5	30.6	3.0	30.6
5.6	0.4	8.1	4.9	16.3
8.5	4.0	2.1	2.5	25.4
7.0	5.1	5.1	0.3	16.5
0.9	4.7	22.5	3.4	17.5
6.2	2.6	8.6	4.3	17.4
7.5	3.2	28.7	4.2	14.6
2.7	5.1	12.3	4.4	21.1
3.7	0.2	28.0	3.9	26.9
2.0	3.9	13.5	5.9	27.3
9.2	2.6	24.6	5.1	21.1
4.4	2.6	16.8	5.0	31.5

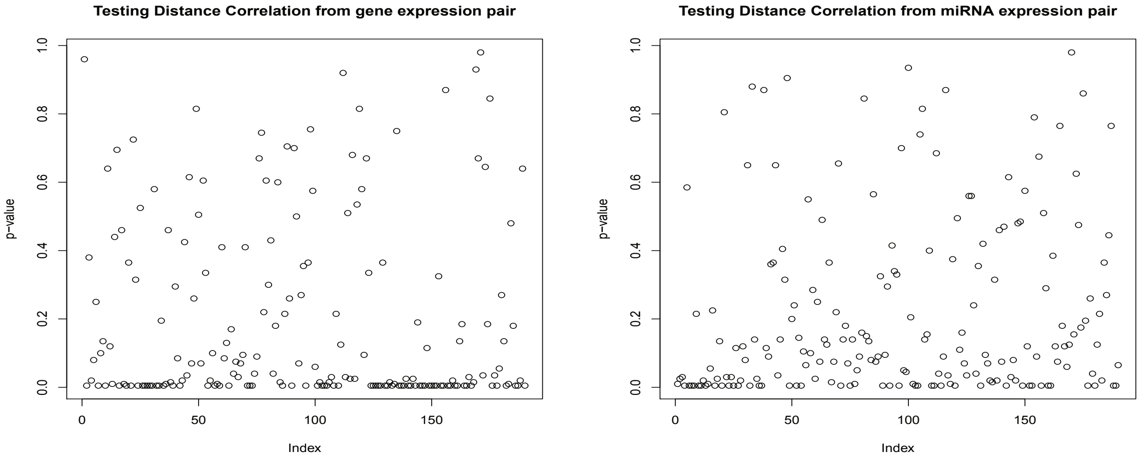

Figure S2. Plot of p-values for testing the significance of nonlinear dependency between pairs of selected gene and miRNA expression data

Table S2. Relative bias in percentage and corresponding mean squared error (MSE) for different values of dependence parameter in Franks copula-based multivariate distribution with varying sample size(N)

N=100			N=200			N=300			N=400		
Theta	Bias%	MSE	Theta	Bias%	MSE	Theta	Bias%	MSE	Theta	Bias%	MSE
1	-0.09	0.027	1	0.46	0.008	1	-0.89	0.007	1	-0.50	0.004
2	0.00	0.046	2	1.01	0.015	2	0.58	0.012	2	0.39	0.009
3	-1.13	0.057	3	-1.49	0.032	3	-0.23	0.021	3	-0.59	0.013
4	-0.42	0.070	4	-0.51	0.039	4	0.26	0.021	4	-0.24	0.018
5	-0.70	0.079	5	-0.58	0.050	5	-0.70	0.024	5	-0.91	0.021
6	-0.44	0.104	6	-0.92	0.056	6	-0.63	0.041	6	-0.58	0.027
7	-0.77	0.147	7	-0.94	0.082	7	-0.62	0.050	7	-0.58	0.039
8	-2.00	0.191	8	-1.36	0.107	8	-0.85	0.058	8	-0.66	0.046
9	-2.61	0.271	9	-1.48	0.129	9	-0.88	0.074	9	-0.82	0.053
10	-2.83	0.358	10	-1.66	0.149	10	-1.38	0.104	10	-1.01	0.068

Note: We have used 100 simulated samples from Franks copula-based multivariate distribution to compute average relative bias in percentage and mean squared error in each set-up. The relative bias ranges between 0% and 2.83% with smaller MSEs. Bias tends to increase with increasing vales of the dependence parameter, but decreases with increasing sample size.

Table S3. Kolmogorov-Smirnov test result for the fitted gamma distribution using selected set of gene and miRNA expression and a simulated data set

Gene Expression			miRNA Expression		
	Deceased(N=33)	Alive(N=90)		Deceased(N=33)	Alive(N=90)
Gene Name	p-value	p-value	miRNA Name	p-value	p-value
SFTPB	0.533	0.258	hsa.mir.21	0.877	0.431
RPS18	0.704	0.512	hsa.mir.143	0.255	0.422
SFTPA1	0.446	0.780	hsa.mir.148a	0.582	0.127
FTL	0.589	0.755	hsa.mir.22	0.328	0.935
CD74	0.489	0.135	hsa.mir.375	0.970	0.198
HLA.B	0.226	0.008	hsa.mir.182	0.374	0.374
B2M	0.049	0.032	hsa.mir.30a	0.034	0.966
SFTPA2	0.987	0.609	hsa.mir.99b	0.596	0.808
TPT1	0.908	0.748	hsa.mir.10a	0.433	0.750
S100A6	0.105	0.842	hsa.mir.183	0.746	0.651
TMSB10	0.043	0.094	hsa.let.7b	0.052	0.515
RPL41	0.427	0.916	hsa.mir.30d	0.128	0.810
EEF1A1	0.777	0.139	hsa.mir.200c	0.683	0.514
LOC96610	0.822	0.513	hsa.mir.10b	0.411	0.380
RPLP1	0.786	0.036	hsa.let.7a.2	0.031	0.338
RPS6	0.577	0.978	hsa.mir.29a	0.239	0.362
ACTB	0.817	0.923	hsa.mir.100	0.256	0.520
RPS27	0.350	0.054	hsa.mir.30e	0.485	0.111
COL1A2	0.807	0.484	hsa.mir.101.1	0.937	0.575
GAPDH	0.295	0.002	hsa.mir.142	0.071	0.703

Note: Resulting p-values for each gene and miRNA expression related to deceased and alive subjects are given in the above table. We observe that almost all p-values are large (>0.05) but few are small (<0.05). Thus we are not able to find sufficient evidence against the null hypothesis of the fitted gamma distribution at 5% level of significance.

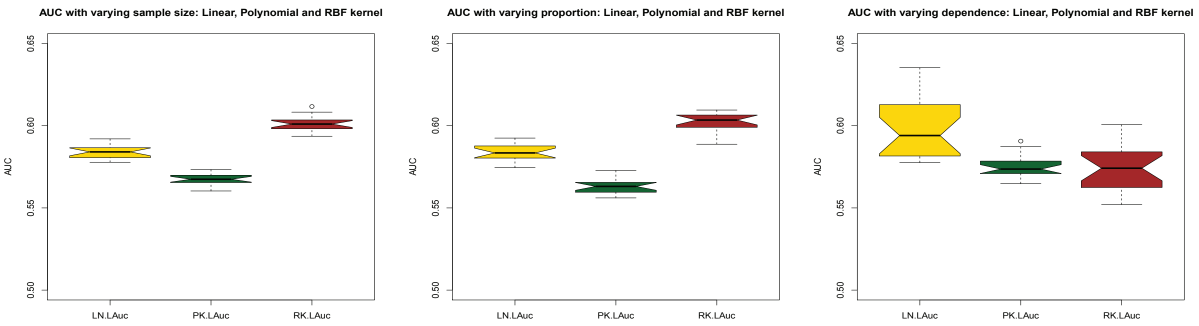

Figure S3. Notched Box plot of cross validated AUC for death classification using linear(LN.LAuc), polynomial(PK.LAuc) and Gaussian(RK.LAuc) kernel with varying sample size, proportion deceased and the degree of dependence using simulated data

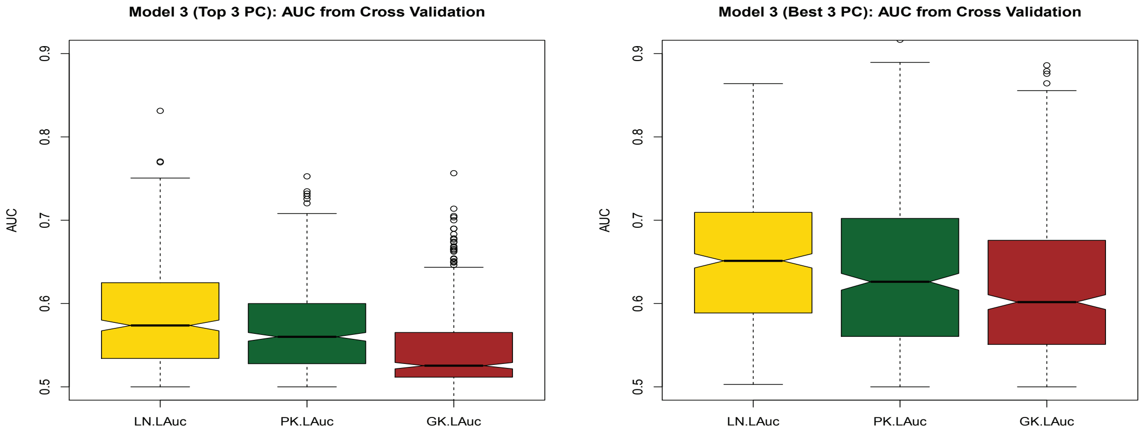

Figure S4. Cross validated AUC based on the logistic regression model using principal components extracted through linear(LN.LAuc), polynomial(PK.LAuc) and Gaussian(GK.LAuc) kernel from gene and miRNA expression real data

Chapter 5

Summary and Conclusions

In this chapter, we presented a brief summary of our findings with some concluding remarks and future directions.

Correct classification is an essential component of understanding the causal pathway to survival or death due to a specific disease. This procedure was formalized based on the classification rules developed in the early 19th century by the International Statistical Classification of Disease (ICD). However, these rules need to be updated based on new evidence from ongoing research in this area. Researchers often use a univariate approach for this purpose. Due to complex relationships between disease and exposures, it is essential to consider multivariate statistical approaches. This procedure may also require combining information from multiple sources, which we can accomplish through data integration.

In this research, we shed light on three different aspects of data integration and outcome classification. First, we have developed and described a method of integrating information from the published literature and a real data set; this has allowed us to determine the threshold of a biomarker for the classification of an outcome. Second, we have developed and described a new method of integrating information from two dependent biomarkers, allowing us to determine the joint threshold for the classification of an outcome. Third, we compared two dimension reduction methods and demonstrated how to integrate information from multiple data sets for an outcome classification.

To address the first issue, we have developed a Bayesian approach to determine the threshold of a biomarker for disease classification. We illustrated this method, utilizing information from literature reviews of selected biomarkers and a large case-control study data set. In particular, we considered the problem of classifying MI based on Apolipoprotein B (ApoB), Apolipoprotein A1 (ApoA1) and the ratio of these two biomarkers. Higher levels of ApoB and lower levels of ApoA1 are well-known risk factors for MI. However, the thresholds at which these markers are associated with the increased risk of MI were not clear. Applying the method developed, we have determined the threshold of ApoB, ApoA1, and the ratio of ApoB to ApoA1 that maximizes the classification accuracy for MI. During this process, we used the conditional distribution of these biomarkers given the case or control status of the participants. We first constructed prior distributions using information from literature reviews. We then constructed the posterior distributions for the location and scale parameters utilizing the prior distribution and the INTERHEART data set. Finally, thresholds for ApoB, ApoA1, and the ratio were identified as 0.908

(gram/liter), 1.138 (gram/liter) and 0.808, respectively, that maximized the classification accuracy of MI.

We have also used a classical and a Bayesian estimation procedure to quantify the odds ratio for one standard deviation change in ApoB, ApoA1, and the ratio, which enabled us to identify the most informative of the three predictors. Estimates obtained from the Bayesian approach with an informative prior were slightly lower than those obtained from the classical approach. Based on this analysis, higher levels of ApoB were shown to be a risk factor and higher levels of ApoA1 were a protective factor for MI, which is consistent with the current literature. Although ApoB is the most commonly used biomarker for MI, ApoA1 was more informative than ApoB based on this analysis. However, the clinical explanation for this finding needs to be explored. It is also noteworthy that the model with the ratio of ApoB to ApoA1 as a single exposure appeared to be less informative for MI than the model where we used both of them as independent predictors. Given that we have used this data set for illustration only, these results will require further investigation for use in clinical applications.

The threshold identification procedure developed in this article is flexible compared to the standard logistic regression approach and allows one to identify a more precise threshold of a biomarker. This method can be used to find the threshold of any continuous exposure for a binary disease classification. However, the ratio of two biomarkers may lead to a skewed distribution, and the normal distribution assumption needs to be carefully assessed and revised, particularly for a small sample.

To address the second issue, we considered the problem of classifying acute MI based on two dependent biomarkers, such as creatine kinase (CK) enzyme and cardiac troponin (cTn). Classification rules for a disease based on dependent biomarkers have often been developed ignoring the dependency that leads to a conflicting classification for some individuals. To overcome this, we have developed a new method of classifying individuals into binary disease groups that take into account the dependency between biomarkers and identifies a joint threshold, which in turn leads to a unique classification. For this purpose, we constructed bivariate distributions based on Frank's, Clayton's and Gumbel's copula functions, and then developed the rules for classification using the joint probability distributions.

To understand the effect of different choices of copulas, we first evaluated the relationship between Kendall's τ and the parameter θ in copulas, using the theoretical relationships as well as using simulated samples. We then conducted an extensive simulation to study the statistical properties of these joint probability distributions, constructed through different choices of copulas. Comparing the dependency parameter θ in copula with Kendall's τ, we observed that a given value of the dependency parameter represents higher levels of dependency in Gumbel's, followed by Clayton's and Frank's copula. The simulation study indicated that the relative bias and the mean squared error of all parameter estimate converged to zero as the sample size increased. Results of this simulation also indicated that the bivariate distribution constructed through Gumbel's copula represents a higher degree of dependency, followed by Clayton and Frank at a given value of the dependency parameter.

Applying the method developed, we have also determined the joint threshold for CK and cTn using simulated samples and compared the classification accuracy of MI using three copula-based bivariate distributions. Since a fixed value of the dependency parameter θ represents a different degree of dependency in these copulas, we conducted this comparison fixing Kendall's τ and back calculating θ across different choices of copulas. Repeating this procedure in thousands of simulated data sets, we observed that the threshold for disease classification converges to a stationary distribution, regardless of the choices of copulas.

Assuming $\tau = 0.5$, threshold estimates of these biomarkers using Frank's, Clayton's and Gumbel's copula were given by (5.68, 1.49), (5.72,1.54) and (5.36, 1.50), respectively. Similarly, for $\tau = 0.7$, threshold estimates of these biomarkers were given by (6.85, 1.42), (6.71, 1.45) and (6.59, 1.42), respectively. At a higher level of dependency, we observed slightly higher AUC estimate using Clayton's copula-based bivariate distribution compared to Frank's, followed by Gumbel's. However, these differences were not statistically significant. The classification accuracy also decreased with the increasing value of the dependency parameter, regardless of the choices of copulas. This was most likely due to less information on the disease status with higher levels of dependency between biomarkers. However, the most important feature of this approach is to allow modeling broad classes of bivariate distributions, which considers dependency and leads to an improved and unique classification.

Finally, we illustrated the method using a real data example, where we identified the joint threshold of Apolipoprotein B to Apolipoprotein A1 ratio (ApoB/ApoA1) and total cholesterol to high-density lipoprotein ratio (TC/HDL) to classify individuals at risk of developing myocardial infarction. Based on this analysis, the joint threshold for ApoB/ApoA1 ratio and TC/HDL ratio were given by 0.725 and 4.37, respectively, with a sensitivity and specificity of 60.3% and 58.4%, respectively. Under certain circumstances, a researcher may want to determine the threshold at a given sensitivity or specificity. Thus, we also identified the threshold at 80% specificity as 0.847 and 5.18, respectively, corresponding to 41% sensitivity.

To address the third issue, we considered the problem of classifying death due to cancer, based on gene and miRNA expression data sets. Data integration is a process of combining information from such data sets with large dimensions. However, it is usually necessary to reduce the dimension of gene and miRNA data before we can utilize the information in a standard statistical procedure like regression. Linear principal component analysis is a widely used approach for this purpose. This method relies on the linearity assumption, which often fails to capture the pattern and relationship inherent in the data. As a result, a nonlinear approach such as kernel principal component analysis might be optimal in this situation. However, the advantage of using kernel principal components in the context of genomic data integration and disease classification needs to be explored and justified.

In Chapter 4, we compared the performances of these two approaches towards data integration and an outcome classification, based on real and simulated data sets in two steps. First, we used the linear and kernel principal component analysis to reduce the dimension of larger data sets. Percent of variance explained by the first few principal components were used to assess how well these two approaches could extract information from a larger data set to a smaller number of latent variables. Second, we used a logistic regression model to integrate extracted features from different domains for classification. Finally, we compared the performance of the linear and kernel principal components based on the classification error rate and the area under the receiver operating characteristic curve. Results of this analysis were also validated using the split-half validation procedure.

In this chapter, we have developed a copula-based method of simulating random samples, which preserved the degree of dependence and nonlinearity observed within and across gene and miRNA expression of the target population. Based on the method developed, we conducted an extensive simulation to compare the performances of these two approaches towards data integration and death classification. During the simulation, we varied sample size, the proportion of deceased subjects in the sample and the degree of dependence, allowing us to identify the effect of each design level parameter towards the performance of these two approaches. Given that a simulation study is expected to start with sensible parameters, we used a subset of the lung cancer data set to identify parameters for this purpose. This procedure allowed us to generate data that reflected the gene and miRNA expression of deceased and non-deceased samples related to lung cancer with additional scenarios. After we generated each data set, we randomly split the full data set into a training and test set with 50 percent of the observations in each, for cross-validation.

Applying the linear and kernel principal component analysis, we extracted the first three principal components from the gene expression set, and the first three from the miRNA expression set and computed the percent of variance explained by these sets. During this process, we observed that a higher percent of variance explained by the first few polynomial kernel principal components compared to the linear principal components. Subsequently, we developed a logistic regression model, using those principal components obtained from the training set, and classified based on the principal components obtained from the test set. Comparing the classification error rate using the linear and polynomial kernel principal components, we observed that the estimates were very similar across different scenarios. However, the AUC obtained using the linear principal components was always higher regardless of sample size, the degree of dependence, or the proportion of deceased subjects in the sample. In general, the AUC obtained from either of these two approaches tended to increase with increasing value of the dependence parameter but remained constant with varying sample size or the proportion of deceased subjects.

Next, we provided a step-by-step data integration procedure for three domains of a lung cancer data set. In particular, we explored and integrated gene expression, miRNA expression, age, and sex to classify death due to the disease. Exploring the raw data suggested that some degree of nonlinearity exists across many different pairs of gene and miRNA expressions. We applied the linear and kernel principal component analysis to reduce the dimension of these two data sets. We observed that the first few polynomial kernel principal components carry more information, as compared to linear principal components, a result similar to what we have seen during simulation study. Integrating different domains for classification based on polynomial kernel principal components produced a slightly lower AUC, as compared to the linear principal components. The validation study based on the polynomial kernel principal components produced similar results with a slightly lower AUC and similar error rates.

Despite having a larger percent of variance explained by top three principal components, we were not able to find any advantages of using polynomial kernel principal components during the death classification, as compared to the linear principal components. We recognize that first three principal components extracted from an unsupervised learning algorithm may not be the most optimal for classification of a particular outcome. However, principal components extracted through these approaches is considered unbiased and provides a fair methodological comparison for both the linear and kernel approaches. A researcher may also be interested in the best three principal components as opposed to the top three for the purpose of classification. Thus, we also identified the best three principal components by applying the forward selection procedure and ran a similar analysis as conducted for the top three principal components. However, the comparative study results remained unchanged with the revised set but provided better classification rates from both approaches.

This analysis showed that first few polynomial kernel principal components carry more information on the expression level for reducing the dimension of a genetic process. However, logistic regression models using linear principal components provided better performance of classification in different scenarios. Thus, we were not able to find any benefits of using polynomial kernel principal components to classify death in this occasion, as compared to linear principal components. We also observed consistent results based on the simulation study with different scenarios, as well as during the analysis of a real data set. As a result, reducing dimensions using linear PCA and a logistic regression model for classification seems to be adequate for this purpose. We also observed that integrating information from multiple data sets using either of these approaches leads to a better classification rate for the outcome.

We hope that methods developed in this thesis will allow researchers identify more precise rules for the classification of an outcome. These rules, in turn, lead to a better strategy to improve our health through early stage intervention for a disease. We also recognize the several limitations of each method presented in each chapter that we hope to address in future research. In particular, the method developed in Chapter 2 depends on the normality assumption, but we might encounter a non-normal data set during this process, especially with small samples. A method that considers alternative distributions, such as gamma or Cauchy, needs to be developed for this purpose.

The method developed in Chapter 3 is limited to a bivariate case. However, the number of tests to identify a particular disease could be more than two, and it would be useful to extend the method to capture these scenarios. It is important to note that parameters used for the simulation component of this article were based on a literature review. As a result, the threshold identified for CK and cTn to classify acute MI based on simulated data sets require careful interpretation. We also illustrated the method with a real data set to determine the joint threshold of ApoB/ApoA1 ratio and TC/HDL ratio for the classification of MI, but the sample size was small. Thus, the result of this analysis may not be generalizable to the target population. However, we hope to use a large data set and apply this method in a subsequent article, which will allow us to derive a clinically applicable threshold for this purpose. Depending on selected biomarkers, we will need to identify the copula-based bivariate distribution that fits the data best.

In the article presented in Chapter 4, we were not able to find any advantages of using kernel principal components for the classification of death due to lung cancer. However, this procedure might be useful for other purposes, such as identifying a cluster of genes or other types of data with higher degrees of nonlinearity. The sensitivity of the kernel approach for these purposes is yet to be compared with the linear approach. Data used in this article includes a few other sets such as protein expression, somatic CNV, and methylation data. It might be worthwhile to see how well we could integrate this information to improve the performance of death classification. Depending on the marginal distributions of selected expressions, the copula-based simulation algorithm needs to be revised as well, along with an appropriate choice of copulas.

In the next step, we hope to develop a multidimensional probabilistic model based on copula functions to determine the probability of the disease using a set of latent variables from all domains. This procedure will allow us to determine the joint threshold of extracted features for the disease or death classification. We also hope to use some of these methods using subject level information, which might shed some light towards the goal of personalized medicine and rational drug treatment plans for an individual.