TOPICS IN MEASUREMENT ERROR ANALYSIS AND HIGH-DIMENSIONAL BINARY CLASSIFICATION

A Dissertation

by

TIANYING WANG

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Raymond J. Carroll
Co-Chair of Committee,	Irina Gaynanova
Committee Members,	Suojin Wang
	Hongwei Zhao
Head of Department,	Valen Johnson

August 2018

Major Subject: Statistics

ABSTRACT

We propose novel methods to tackle two problems: the misspecified model with measurement error and high-dimensional binary classification, both have a crucial impact on applications in public health.

The first problem exists in the epidemiology practice. Epidemiologists often categorize a continuous risk predictor since categorization is thought to be more robust and interpretable, even when the true risk model is not a categorical one. Thus, their goal is to fit the categorical model and interpret the categorical parameters. We address the question: with measurement error and categorization, how can we do what epidemiologists want, namely to estimate the parameters of the categorical model that would have been estimated if the true predictor was observed? We develop a general methodology for such an analysis, and illustrate it in linear and logistic regression. Simulation studies are presented, and the methodology is applied to a nutrition data set. Discussion of alternative approaches is also included.

For the second project, we consider the problem of high-dimensional classification between the two groups with unequal covariance matrices. Rather than estimating the full quadratic discriminant rule, we propose to perform simultaneous variable selection and linear dimension reduction on original data, with the subsequent application of quadratic discriminant analysis on the reduced space. In contrast to quadratic discriminant analysis, the proposed framework does not require estimation of precision matrices and scales linearly with the number of measurements, making it especially attractive for the use on high-dimensional datasets. We support the methodology with theoretical guarantees on variable selection consistency, and empirical comparison with competing approaches. We apply the method to gene expression data of breast cancer patients and confirm the crucial importance of the ESR1 gene in differentiating estrogen receptor status.

Further, we provide software support for the proposed methodology. We develop two

R packages, CCP and DAP, and present two vignettes as long-format illustrations for their usage.

DEDICATION

To my mother, my father, and my husband.

ACKNOWLEDGMENTS

Working as a Ph.D. student in Texas A&M university was a wonderful as well as challenging experience to me. During these four years, many people helped me in shaping up my academic career. Here is a tribute of all those people.

First, I would like to thank my committee chair, Dr. Carroll, not only for his tremendous academic support, but also for giving me so many great opportunities. It was only due to his valuable guidance, cheerful enthusiasm and continued patience that I was able to complete this work. Similar, profound gratitude goes to my co-chair, Dr. Gaynanova, who has been supportive and worked actively to provide me with the protected academic time throughout my course of research. Under her supervision, I was able to learn many valuable things and finish my research work. I am also grateful to my committee members, Dr. Wang and Dr. Zhao, for their generous guidance and support during my Ph.D. curriculum.

I am hugely appreciative to the department faculty for providing me with a fantastic professional training and nurturing my enthusiasm for statistics. I am also indebted to the department staff who have helped me for making my time at Texas A&M University a great experience. I have very fond memories of my time here.

Last but not least, I would like to express my deepest gratitude to all my close friends and family. Thanks for all your encouragement and support. I would like to thank my parents, whose love and guidance are with me in whatever I pursue. Most importantly, I wish to thank my loving husband, Zhihao, who provides continued patience and unending support.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a dissertation committee consisting of Professor Raymond J. Carroll, Irina Gaynanova and Suojin Wang of the Department of Statistics and Professor Hongwei Zhao of the Department of Epidemiology and Biostatistics.

All work for the dissertation was completed by the student, in collaboration with Dr. Raymond J. Carroll, Dr. Irina Gaynanova and Dr. Ya Su of the Department of Statistics, Dr. Betsabé G. Blas Achic of the Departamento de Estatística, Universidade Federal de Pernambuco, Dr. Victor Kipnis, Dr. Kevin Dodd of Division of Cancer Prevention, National Cancer Institute.

Funding Sources

This work was made possible in part by National Cancer Institute under Grant Number U01-CA057030.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

In this manuscript, we propose novel methods for solving problems in public health. To be more specific, we focus on nutrient-based analysis of disease risk and genetic-based discriminant analysis of complex human diseases. Although motivated by the realistic problems from public health, our approaches are general and can be adapted into different contexts and areas.

This manuscript contains my work for two major projects and their supportive software vignettes. In the first project, we propose a method to analyze the relationship between extrinsic factors, or called environmental factors, and diseases. In the second project, we propose a method to deal with the relationship between intrinsic factors, i.e., genomic information, and complex human diseases. The next two chapters provide a concrete illustration for the two R packages we built for the proposed methods. Combining the environmental factors and the genetic factors together to understand disease schemes is the future work. The idea of proposing a novel semiparametric method to improve current estimators, when the distributions of environmental and genetic factors are hard to model, is discussed in the conclusion part.

The motivation for the first project is that misspecified models are widely used in epidemiology, with measurement error existing in it. Epidemiologists tend to categorize a continuous risk predictor because the categorical model is thought to have better interpretation and robustness. For example, Reedy et al. (2008, 2010) categorize food scores, defined to measure diet habit, to analyze the dietary pattern with colorectal cancer risk; Arem et al. (2013) categorize the Healthy Eating Index 2005 into quintiles to analyze the relationship between dietary pattern and pancreatic cancer risk. Besides epidemiology, the categorical model is also used widely in many other research areas. In environmental health studies, Chaix et al. (2016) analyze the relationship between built environments and walking trips, in which they

categorized age, income, distance covered in the trip into categories; Evenson et al. (2016) analyze the association of physical activity and sedentary behavior with all-cause and cardiovascular mortality, in which age, household income, body mass index, minutes of physical activity per day and so on are categorized; Wang et al. (2016) investigate the association of long-term exposure to traffic pollution with markers of atherosclerosis in an all-African American cohort, where household income is categorized in the model.

However, such categorization makes the model misspecified: the specified parametric family of probability distribution may be incorrect, especially when there are other covariates than the categorized predictor, which are also related to the response in the continuous model. White (1982) shows that when the model is misspecified, the quasi-maximum likelihood estimator converges to a limit, which is what epidemiologists interested in. When measurement error exists within the observed predictor, however, things become complicated.

Measurement error is common in epidemiology, while ignoring it may lead to poor inference quality. For example, the data from Eating at America's Table Study (Subar et al., 2001) is collected by questionnaire, only observed in a short time period. Thus, measurement errors may come from inaccurate recalls and daily variations, and the true risk predictor - obtained by the daily average over a long term- is not feasible. Other nutrient-based data may also share the same problems, since the underlying true nutrition intake cannot be observed directly. Ignoring the errors and using observed data without adjustment may cause problems in the misspecified model. Thus, the goal of the first project is to study the effect of measurement error existing in a misspecified model, especially for categorizing a continuous predictor.

In Chapter 2, we show how to obtain consistent estimates of what epidemiologists would have obtained when the true risk predictor is observed, and develop consistent standard errors, thus correct inferences. Technical background, methodology, simulation studies and

application on EATS data are presented.

The second project is motivated by the high-dimensional data and the difficulties in its analysis, such as genetic data analysis for complex human diseases, e.g., cancers, diabetes and cardiovascular diseases. The major feature of this kind of data is small sample and high-dimensional, which means the number of features per observation is much more than the number of observations, or the total sample size. In this case, most of the classical statistical methods are challenged, either facing mathematical or computational issues, or cannot maintain the optimal results.

In this project, we focus on high-dimensional binary classification problem, a supervised learning. For example, given two groups of people, diseased and non-diseased, we are able to learn a classification rule through training the genomic information data with the group label and classify a new observation into one of the two groups based on the rule. Moreover, the proposed approach is general and can be used in any cases wherever binary classification is needed for high-dimensional data.

Classical methods achieve satisfactory results in the large sample, low-dimensional scenario, including quadratic discriminant analysis (QDA) and linear discriminant analysis (LDA). QDA and LAD are both generated from the Bayes rule, which assigns a new observation to the group that maximize the production of prior probability and population density. Under the normality assumption, assuming equal variability in two groups leads to the linear decision boundary (LDA), otherwise leading to a quadratic decision boundary (QDA). QDA is more flexible because it does not assume the equal covariance matrix in two groups. However, both QDA and LDA work poorly in high-dimensional cases.

Moreover, the classification rule of QDA is very likely to suffer from the singularity of sample covariance matrix when $p > n$, due to the inversion required in the classification rule. Recently, QDA and LDA have been extended by sparse and regularized techniques. However, those approaches either required specific assumptions on covariance matrices or

computationally slow, due to the need of estimating a $p \times p$ matrix.

In Chapter 3, we propose new sparse quadratic classification rules, which assume unequal covariance matrices for two groups, while maintaining the computation efficiency. Thus, we start from Fisher's LDA and extended the projection idea into two directions. Since the number of parameters need to be estimated is linearly in p but not p^2, we are able to derive efficient algorithm which estimate the projection direction as well as perform variable selection simultaneously. The proposed method only requires inverting a 2×2 matrix instead of $n \times n$, and thus it is very likely to be full rank. Technical background including optimization, algorithm, theoretical proof for variable selection consistency and empirical study results are presented in the chapter.

In Chapter 4 and 5, we present vignettes for two R packages built for the two projects: CCP (Categorizing a Continuous Predictor), and DAP (Discriminant Analysis via Projection). Though manuals are provided within R packages, these two vignettes illustrate the usages of the packages from different aspects. First, they provide very brief methodology summary for readers who would like to use the packages without look into details of the first two chapters, which is more convenient from a practical aspect. Second, through showing the real data examples reported in the first two chapters, the two vignettes offer more details to support the analysis and conclusions based on proposed methods. Further, the usage of functions are presented with more concrete explanations.

Finally, the overall summary and conclusions are presented, and the ongoing project is discussed. To analyze the effect of gene-environment interactions on complex human disease, we propose a novel semiparametric method to improve current estimators for the case-control study using retrospective likelihood framework, allowing the distributions of environmental and genetic factors to be nonparametric.

2. CATEGORIZING A CONTINUOUS PREDICTOR SUBJECT TO MEASUREMENT ERROR

2.1 Introduction

Fitting models by categorizing a continuous risk predictor is a common practice in epidemiology. Among many recent examples, see Reedy et al. (2008, 2010); Arem et al. (2013); Chaix et al. (2016); Evenson et al. (2016) and Wang et al. (2016). A look at current issues of epidemiology journals will uncover many more examples. An important issue is that, generally in these problems, there are many covariates other than the main risk predictor.

The appeal of categorization in interpreting results is clear. If we have a risk predictor X, and we categorize it into J levels $(C_1, ..., C_J)$, one can compare the highest level of the predictor, C_J, to the lowest level, C_1, and if they are statistically significantly different, one can then conclude that it is better to be in the class that has the lowest risk, and quantify how much better.

One technical issue about this approach concerns the case that there are other covariates than X, say Z. Consider a binary response, Y, let $H(\cdot)$ be the logistic distribution function, and suppose that the true risk model in the continuous scale is $\text{pr}(Y = 1|X, Z) = H\{m(X, Z, \boldsymbol{\beta})\}$ for some function $m(\cdot)$. Then, if any of the covariates Z are related to Y in this continuous model, categorizing X into J levels and plugging that into $m(X, Z, \boldsymbol{\beta})$ leads to a misspecified model. As White (1982) shows, this leads to the question of how the categorized model actually relates to disease, which is not the simple characterization given in the previous paragraph.

Our point is not to try to get epidemiologists to change their common practice. Instead, we study the effect of measurement error when a continuous predictor variable subject to measurement error is categorized. Our goal is to answer the question: with measurement

error in this context, how can we (a) obtain consistent estimates of what epidemiologists would have obtained if X were actually observed; and (b) develop consistent standard errors.

We answer the question above in a general way. Section 2.2 gives basic technical background. Section 2.3 provides a general methodology for answering questions (a) and (b) above. Section 2.4 presents simulation studies for linear and logistic regression that show the good behavior of our methodology, both in terms of bias and confidence interval coverage. Section 2.5 shows applications of our approach by using data from the Eating at America's Table Study (Subar et al., 2001). Section 2.6 presents a discussion about other potential approaches to categorization and how those approaches compare to ours. Sketches of technical arguments are in the supplementary material.

Remark 1. As discussed above, categorization leads to a misspecified model. It is also well-known that such categorization generally leads to differential measurement error (Flegal et al., 1991; Gustafson, 2004; Buonaccorsi, 2010), and thus additional complications over simply fitting a measurement error model. Chapters 6.1-6.2 of Gustafson (2004) has a detailed discussion when the continuous variable is dichotomized, calling the result *differential by dichotomization*. We are thus assuming that the true risk model in a continuous variable X is not categorical in X. If it were, consult Gustafson (2004) and Buonaccorsi (2010), who also discuss the issue of doing a measurement error analysis in this case, especially the difficult complex issues of computation and identifiability both theoretical and practical.

2.2 Data generating mechanism and basic Ideas

2.2.1 Illustration: a special case of linear regression

It is instructive to consider a special case, namely linear regression. Doing so will set the stage for our general method. The response is Y, the scalar predictor subject to error is X, the observed scalar predictor is W, there are predictors Z measured without error, and we define $\widetilde{Z} = (1, Z^{\mathrm{T}})^{\mathrm{T}}$ to allow for an intercept. The regression model in the continuous

predictor X is $Y = X\beta_1 + \widetilde{Z}^{\mathrm{T}}\beta_2 + \epsilon$, where ϵ is mean zero independent of (W, X, Z). There are $j = 1, ..., J$ categories $(C_1, ..., C_J)$, and $M(X, Z) = \{I(X \in C_1), ..., I(X \in C_J), Z^{\mathrm{T}}\}^{\mathrm{T}}$. If X could be observed, then we would also immediately obtain an estimate of $\boldsymbol{\beta} = (\beta_1, \beta_2^{\mathrm{T}})^{\mathrm{T}}$.

By White (1982), when X is observed, what epidemiologists estimate by using the categorized $M(X, Z)$ is $\boldsymbol{\Theta}$, where, based on the normal equations for the categorized predictor, $\boldsymbol{\Theta} = (\theta_1, ..., \theta_J, \Theta_{J+1}^{\mathrm{T}})^{\mathrm{T}}$ is the solution to

$$0 = E[M(X, Z)\{Y - M^{\mathrm{T}}(X, Z)\boldsymbol{\Theta}\}] = E[M(X, Z)\{X\beta_1 + \widetilde{Z}^{\mathrm{T}}\beta_2 - M^{\mathrm{T}}(X, Z)\boldsymbol{\Theta}\}]. \quad (2.1)$$

The estimate $\widehat{\boldsymbol{\Theta}}$ is the solution to $0 = n^{-1}\sum_{i=1}^{n} M(X_i, Z_i)\{Y_i - M^{\mathrm{T}}(X_i, Z_i)\boldsymbol{\Theta}\}$, and this is a consistent estimate of Θ. Comparisons between categories j and k for $j, k \leq J$, say, are $\widehat{\theta}_j - \widehat{\theta}_k$.

However, when X is not observable, estimating the solution to (2.1) has to be based solely on (Y, W, Z). In (2.1), it makes sense that if one believes the true regression model is linear in (X, Z), then, at some point, an estimate of $\boldsymbol{\beta}$ can be obtained via a measurement error analysis if there are sufficient data to do so.

Solving (2.1) based only on the observed W though is not so easy, and it is clear that some part of the relationship between W and X given Z is going to need to be specified, as it needs to be to do a general measurement error analysis. One way to do this is to define

$$\mathcal{G}(X, Z, \boldsymbol{\Theta}, \boldsymbol{\beta}) = M(X, Z)\{X\beta_1 + \widetilde{Z}^{\mathrm{T}}\beta_2 - M^{\mathrm{T}}(X, Z)\boldsymbol{\Theta}\}, \quad (2.2)$$

and then define $Q(W, Z, \boldsymbol{\Theta}, \boldsymbol{\beta}) = E\{\mathcal{G}(X, Z, \boldsymbol{\Theta}, \boldsymbol{\beta})|W, Z\}$. Since $0 = E\{Q(W, Z, \boldsymbol{\Theta}, \boldsymbol{\beta})\}$, $\boldsymbol{\Theta}$ can be estimated by solving

$$0 = n^{-1}\sum_{i=1}^{n}\left[E\{M(X, Z)(X\beta_1 + \widetilde{Z}^{\mathrm{T}}\beta_2)|W_i, Z_i\} - E[\{M(X, Z)M^{\mathrm{T}}(X, Z)\}|W_i, Z_i]\boldsymbol{\Theta}\right].$$

Hence, in this simple case, for $j = 1, ..., J$ we will need to be able to calculate expectations

of $XI(X \in C_j)$ given (W, Z) and the probability that $X \in C_j$ given (W, Z). As we will see, in general problems, we will need to estimate the expectations of other functions of X given (W, Z).

So, to summarize, to get a general solution, it appears that we will need to estimate (β_1, β_2) by a measurement error analysis and estimate expectations of specified functions of X given (W, Z).

Remark 2. Following on Remark 1, it is obvious that in the unlikely event that the true risk model is actually categorical in X, so that $E(Y|X, Z) = M^{\mathrm{T}}(X, Z)\boldsymbol{\beta}$, then model misspecification and differential measurement error both disappear, and one really needs just the probabilities that X is in the categories given (W, Z). As Gustafson (2004) and Buonaccorsi (2010) discuss in detail, estimating such models is difficult because of model identifiability concerns. Often, papers dealing with this issue assume the existence of a validation data set, where X is actually observed on a subset of the data. Gustafson (2004) is a particularly good source for the difficulties we have mentioned and remedies using replication data. Buonaccorsi (2010), page 314, who states that estimating the misclassification rates is "*most likely coming from internal validation data*" and also has a nice discussion.

2.2.2 Assumptions

Our algorithm is basically the same as in Section 2.2.1

Our work is very general, and requires three basic assumptions. We let X be the continuous predictor subject to measurement error, Z covariates measured exactly, W the mismeasured version of X, and Y the response.

Assumption 1. *When X is observed, the true response model in the continuous scale has parameters $\boldsymbol{\beta}$, such that there is an estimating function, $\Phi_{\mathrm{true}}(Y, X, Z, \boldsymbol{\beta})$ that identifies $\boldsymbol{\beta}$ and satisfies*

$$0 = E\{\Phi_{\mathrm{true}}(Y, X, Z, \boldsymbol{\beta})|X, Z\}. \tag{2.3}$$

Assumption 1 occurs in at least two circumstances.

Example 1.

(A) There are functions $m_1(X, Z, \boldsymbol{\beta})$ *and* $m_2(X, Z, \boldsymbol{\beta})$ *such that* $E(Y|X, Z) = m_1(X, Z, \boldsymbol{\beta})$ *and the unbiased estimating function that would be used if* X *were observable is*

$$\Phi_{\text{true}}(Y, X, Z, \boldsymbol{\beta}) = m_2(X, Z, \boldsymbol{\beta})\{Y - m_1(X, Z, \boldsymbol{\beta})\}. \tag{2.4}$$

(B) There is a parametric model for Y *given* (X, Z).

Example 1(A) is very general, in that it includes traditional quasilikelihood models, nonlinear regression, generalized linear models, probit regression, etc. Crucially, it does not require a fully parametric model for the distribution of Y given (X, Z).

In our approach, as in linear regression in Section 2.2.1, we may need to obtain information about moments of specified functions of X given (W, Z). To do this, we will consider the setting in which there may be an external data set of N observations giving information on one set of parameters of the joint distribution, $\boldsymbol{\Lambda}_{\text{ext}}$: if there is no external study, $N = 0$ and $\boldsymbol{\Lambda}_{\text{ext}}$ does not exist. In addition, there is another set of the parameters, $\boldsymbol{\Lambda}_{\text{int}}$, that is estimated from the n observations in the internal data set.

Assumption 2. *When* X *is not observed, either (a) the distribution of* X *given* (W, Z) *is known up to parameters* $\boldsymbol{\Lambda}_{\text{ext}}$ *and* $\boldsymbol{\Lambda}_{\text{int}}$ *as described above, or (b) there is a function,* $\mathcal{G}(X, Z, \boldsymbol{\Theta}, \boldsymbol{\beta})$ *defined at (2.11) below, whose conditional expectation given* (W, Z) *depends on parameters* $\boldsymbol{\Lambda}_{\text{ext}}$ *and* $\boldsymbol{\Lambda}_{\text{int}}$ *and can be estimated. The parameter* $\boldsymbol{\Lambda}_{\text{ext}}$ *cannot be estimated by internal data, while the parameter* $\boldsymbol{\Lambda}_{\text{int}}$ *can be estimated by internal data. For both, there are unbiased estimating functions* $V_{\text{ext},m}(\boldsymbol{\Lambda}_{\text{ext}})$ *for the external data and* $V_{\text{int},i}(\boldsymbol{\Lambda}_{\text{int}}, \boldsymbol{\Lambda}_{\text{ext}})$ *for the internal data such that* $E\{V_{\text{ext},m}(\boldsymbol{\Lambda}_{\text{ext}})\} = 0$ *and* $E\{V_{\text{int},i}(\boldsymbol{\Lambda}_{\text{int}}, \boldsymbol{\Lambda}_{\text{ext}})\} = 0$.

For linear regression, $\mathcal{G}(X, Z, \boldsymbol{\Theta}, \boldsymbol{\beta})$ is given in (2.2).

If there are external data and $N > 0$, we estimate $\mathbf{\Lambda}_{\text{ext}}$ by solving the estimating equation

$$0 = N^{-1} \sum_{m=1}^{N} V_{\text{ext},m}(\mathbf{\Lambda}_{\text{ext}}). \tag{2.5}$$

In the internal data set, we estimate $\mathbf{\Lambda}_{\text{int}}$ by solving an estimating equation

$$0 = n^{-1} \textstyle\sum_{i=1}^{n} V_{\text{int},i}(\mathbf{\Lambda}_{\text{int}}, \widehat{\mathbf{\Lambda}}_{\text{ext}}). \tag{2.6}$$

There is also a very subtle issue that needs to be made explicit.

Assumption 3. *If external data are necessary for model identification, the parameter* $\mathbf{\Lambda}_{\text{ext}}$ *is transportable in the sense that this parameter is the same in the external and internal data sets.*

The issue of when parameters are transportable from an external data set to the internal data set is discussed in Chapter 2.2.4-2.2.5 of Carroll et al. (2006). As they state, it is much better if there are sufficient internal data that external data need not be used, but this is not always the case.

2.2.3 General observations when X is observed

As argued in Section 2.1, the goal is to fit a model when X is categorized into J levels $(C_1, ..., C_J)$, and so we define the dummy variables and Z as $M(X, Z) = \{I(X \in C_1), ..., I(X \in C_J), Z^{\mathrm{T}}\}^{\mathrm{T}}$: our formulation allows more complex forms, including interactions. Suppose there are $i = 1, ..., n$ subjects in the primary/main/internal study, and suppose further that we observe (Y_i, X_i, Z_i). If X is observed, the analysis done on these categories will be based on replacing (X, Z) in (2.3)-(2.4) by $M(X, Z)$, and to make clear the categorization, we define a parameter $\mathbf{\Theta}$, set $\Phi_{\text{cat}}\{Y_i, M(X_i, Z_i), \mathbf{\Theta}\} = \Phi_{\text{true}}\{Y_i, M(X_i, Z_i), \mathbf{\Theta}\}$, and obtain

$\widehat{\boldsymbol{\Theta}}$ by solving

$$0 = n^{-1}\sum_{i=1}^{n}\Phi_{\text{cat}}\{Y_i, M(X_i, Z_i), \boldsymbol{\Theta}\}. \tag{2.7}$$

More complex forms of (2.7) are easily accommodated.

Unlike in Assumption 1 and (2.3)-(2.4), except in the rare case that the categorized model is actually true, it is easy to see that $0 \neq E[\Phi_{\text{cat}}\{Y, M(X, Z), \boldsymbol{\Theta}\}|X, Z]$, a *conditional* expectation. This is a key part of the work in White (1982).

Despite the fact that the categorized model does not fit the data conditional on (X, Z), by standard estimating equation theory (White, 1982), the estimate formed by solving (2.7) has a limit as $n \to \infty$, $\boldsymbol{\Theta}$, which is the solution to

$$0 = E[\Phi_{\text{cat}}\{Y, M(X, Z), \boldsymbol{\Theta}\}]. \tag{2.8}$$

It is important to observe that (2.8) is an *unconditional* expectation, not a conditional one.

If, instead of observing X, we observe its mismeasured version W, and if we replace X by W, we will of course generally inconsistently estimate both $\boldsymbol{\beta}$ and $\boldsymbol{\Theta}$.

2.2.4 Estimating the true parameter $\boldsymbol{\beta}$

In our approach, as in Section 2.2.1 for linear regression, we must estimate $\boldsymbol{\beta}$ in (2.3). There is of course a large literature on how to do this (Gustafson, 2004; Carroll et al., 2006; Buonaccorsi, 2010; Yi, 2017). Borrowing on that literature, from Assumptions 1-2, for an estimating function $\Phi(Y, W, Z, \boldsymbol{\beta}, \boldsymbol{\Lambda}_{\text{int}}, \boldsymbol{\Lambda}_{\text{ext}})$, the estimate, $\widehat{\boldsymbol{\beta}}$, is the solution to

$$0 \;\; = \;\; n^{-1}\sum_{i=1}^{n}\Phi(Y_i, W_i, Z_i, \boldsymbol{\beta}, \widehat{\boldsymbol{\Lambda}}_{\text{int}}, \widehat{\boldsymbol{\Lambda}}_{\text{ext}}), \tag{2.9}$$

where $(\widehat{\boldsymbol{\Lambda}}_{\text{int}}, \widehat{\boldsymbol{\Lambda}}_{\text{ext}})$ are obtained from equations (2.5) and (2.6), respectively. Of course, the details and the form of $\Phi(\cdot)$ differ from case-to-case.

2.3 Methodology and asymptotic theory

2.3.1 Methodology: general case

The methodology is simple to explain at the general level. The target $\boldsymbol{\Theta}$ is defined as the solution to (2.8). However, we can rewrite (2.8) as

$$0 = E\left(E[\Phi_{\text{cat}}\{Y, M(X,Z), \boldsymbol{\Theta}\}|W, Z]\right). \tag{2.10}$$

Define

$$\begin{aligned} \mathcal{G}(X, Z, \boldsymbol{\Theta}, \boldsymbol{\beta}) &= E\left[\Phi_{\text{cat}}\{Y, M(X,Z), \boldsymbol{\Theta}\}|X, Z\right]; & (2.11)\\ Q(W, Z, \boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{\Lambda}_{\text{int}}, \boldsymbol{\Lambda}_{\text{ext}}) &= E\{\mathcal{G}(X, Z, \boldsymbol{\Theta}, \boldsymbol{\beta})|W, Z\}. & (2.12) \end{aligned}$$

Making the usual nondifferential measurement error assumption, i.e., that Y and W are independent given (X, Z),

$$0 = E\left\{Q(W, Z, \boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{\Lambda}_{\text{int}}, \boldsymbol{\Lambda}_{\text{ext}})\right\}. \tag{2.13}$$

Critically, (2.13) depends only on the observed covariates. Thus, if we have consistent estimates $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}_{\text{int}}, \widehat{\boldsymbol{\Lambda}}_{\text{ext}})$ of $(\boldsymbol{\beta}, \boldsymbol{\Lambda}_{\text{int}}, \boldsymbol{\Lambda}_{\text{ext}})$, then a consistent estimate, $\widehat{\boldsymbol{\Theta}}$, of $\boldsymbol{\Theta}$ solves

$$0 = n^{-1}\textstyle\sum_{i=1}^{n} Q(Z_i, W_i, \boldsymbol{\Theta}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}_{\text{int}}, \widehat{\boldsymbol{\Lambda}}_{\text{ext}}). \tag{2.14}$$

In some cases, we do not have external data. Thus, we do not have V_{ext} and $\boldsymbol{\Lambda}_{\text{ext}}$, and V_{int} and $\boldsymbol{\Theta}$ only depend on $\boldsymbol{\Lambda}_{\text{int}}$.

Remark 3. The key question is how to compute $\mathcal{G}(X, Z, \boldsymbol{\Theta}, \boldsymbol{\beta})$ in (2.11). In the fully general case (2.3), we require a parametric model for the distribution of Y given (X, Z), as in Example 1(B). However, in standard regression models of the form in (2.4) in Example

1(A), great simplification occurs, because in that case,

$$\Phi_{\text{cat}}\{Y, M(X,Z), \boldsymbol{\Theta}\} = m_2\{M(X,Z), \boldsymbol{\Theta}\}\left[Y - m_1\{M(X,Z), \boldsymbol{\Theta}\}\right],$$

and thus

$$\mathcal{G}(X, Z, \boldsymbol{\Theta}, \boldsymbol{\beta}) = m_2\{(X,Z), \boldsymbol{\Theta}\}\left[m_1(X, Z, \boldsymbol{\beta}) - m_1\{M(X,Z), \boldsymbol{\Theta}\}\right].$$

Section 2.7.4 of the **Supplementary Material** gives detailed formulae for linear and logistic regression.

2.3.2 Asymptotic theory

Asymptotic theory for the parameter estimates is easily derived. Let $\boldsymbol{\Omega} = (\boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{\Lambda}_{\text{int}}, \boldsymbol{\Lambda}_{\text{ext}})$ and let the true values of the parameters be denoted by $\boldsymbol{\Omega}$.

It is neater notation in this section to let $i = 1, ..., n$ denote the internal data, and $i = n+1, ..., n+N$ denote the external data. For $i > n$, define $\Psi_i(\boldsymbol{\Omega}) = \{0, 0, 0, V^{\text{T}}_{\text{ext},i}(\boldsymbol{\Lambda}_{\text{ext}})\}^{\text{T}}$, while for $i \leq n$ define

$$\Psi_i(\boldsymbol{\Omega}) = \{Q^{\text{T}}(W_i, Z_i, \boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{\Lambda}_{\text{int}}, \boldsymbol{\Lambda}_{\text{ext}}), \Phi^{\text{T}}(Y_i, W_i, Z_i, \boldsymbol{\beta}, \boldsymbol{\Lambda}_{\text{int}}, \boldsymbol{\Lambda}_{\text{ext}}), V^{\text{T}}_{\text{int},i}(\boldsymbol{\Lambda}_{\text{int}}, \boldsymbol{\Lambda}_{\text{ext}}), 0\}^{\text{T}}.$$

If there are external data, the estimate $\widehat{\boldsymbol{\Omega}}$ solves $0 = \sum_{i=1}^{n+N} \Psi_i(\widehat{\boldsymbol{\Omega}})$. If there are no external data, then $N = 0$, $\boldsymbol{\Omega} = (\boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{\Lambda}_{\text{int}})$ and the zero element and $\boldsymbol{\Lambda}_{\text{ext}}$ in the definition of $\Psi_i(\boldsymbol{\Omega})$ are removed.

By standard estimating equation results, we have the following results, which are shown in Appendices 2.7.1.1 and 2.7.1.2.

Lemma 1. If there are external data, i.e., $N > 0$, make Assumptions 1-3. Suppose that

$N \to \infty$ and $n \to \infty$ such that $n/N \to c_{\lim}$, where $0 < c_{\lim} < \infty$. Then

$$(n+N)^{1/2}(\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}) \to \text{Normal}\{0, A^{-1}B(A^{-1})^{\mathrm{T}}\},$$

where $A = \{(1+c_{\lim})/c_{\lim}\}^{-1}E\{\partial\Psi_1(\boldsymbol{\Omega})/\partial\boldsymbol{\Omega}^{\mathrm{T}}\} + (1+c_{\lim})^{-1}E\{\partial\Psi_{n+N}(\boldsymbol{\Omega})/\partial\boldsymbol{\Omega}^{\mathrm{T}}\}$ and $B = \{(1+c_{\lim})/c_{\lim}\}^{-1}\text{cov}\{\Psi_1(\boldsymbol{\Omega})\} + (1+c_{\lim})^{-1}\text{cov}\{\Psi_{n+N}(\boldsymbol{\Omega})\}$. In the definitions of A and B, the expectation and covariance matrix for $\Psi_1(\boldsymbol{\Omega})$ are computed in the internal data, while the expectation and covariance matrix for $\Psi_{N+n}(\boldsymbol{\Omega})$ are computed in the external data. Let $\widehat{C}_{\text{ext}}$ be the sample covariance matrix of $\Psi_i(\widehat{\boldsymbol{\Omega}})$ for $i = n+1, ..., n+N$ and let $\widehat{C}_{\text{int}}$ be the sample covariance matrix of $\Psi_i(\widehat{\boldsymbol{\Omega}})$ for $i = 1, ..., n$. Consistent estimates of A and B are easily seen to be $\widehat{A} = (n+N)^{-1}\sum_{i=1}^{N+n}\partial\Psi_i(\widehat{\boldsymbol{\Omega}})/\partial\boldsymbol{\Omega}^{\mathrm{T}}$ and $\widehat{B} = \{n/(n+N)\}\widehat{C}_{\text{int}} + \{N/(n+N)\}\widehat{C}_{\text{ext}}$.

Lemma 2. If there are no external data, i.e., $N = 0$, make Assumptions 1-2. As $n \to \infty$,

$$n^{1/2}(\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}) \to \text{Normal}\{0, A^{-1}B(A^{-1})^{\mathrm{T}}\},$$

where $A = E\{\partial\Psi_1(\boldsymbol{\Omega})/\partial\boldsymbol{\Omega}^{\mathrm{T}}\}$ and $B = \text{cov}\{\Psi_1(\boldsymbol{\Omega})\}$. In the definitions of A and B, the expectation and covariance matrix for $\Psi_1(\boldsymbol{\Omega})$ are computed in the internal data. Let $\widehat{C}_{\text{int}}$ be the sample covariance matrix of $\Psi_i(\widehat{\boldsymbol{\Omega}})$ for $i = 1, ..., n$. Consistent estimates of A and B are easily seen to be $\widehat{A} = n^{-1}\sum_{i=1}^{n}\partial\Psi_i(\widehat{\boldsymbol{\Omega}})/\partial\boldsymbol{\Omega}^{\mathrm{T}}$ and $\widehat{B} = \widehat{C}_{\text{int}}$.

2.4 Simulations: logistic and linear regression

2.4.1 Logistic regression

2.4.1.1 Scenarios

For simplicity, we do our simulations in the case that there is no Z. For logistic regression, we assume that the true model is

$$\text{pr}(Y = 1|X) \;=\; H(\beta_0 + X\beta_1) = H\{(1, X)\boldsymbol{\beta}\}, \tag{2.15}$$

where $H(\cdot)$ is the logistic distribution function. Then we generate data as

$$W = X + U; \quad X = \text{Normal}(\mu_x, \sigma_x^2); \quad U = \text{Normal}(0, \sigma_u^2), \tag{2.16}$$

where X and U are independent. We set $\beta_0 = -0.42$ and set $\beta_1 = \log(1.5)$ in Table 2.1. We set $(\mu_x = 0, \sigma_x^2 = 1, \sigma_u^2 = 1)$, so that the measurement error variance is the same as the variance of X, and the classical attenuation coefficient is $\lambda = \sigma_x^2/(\sigma_x^2 + \sigma_u^2) = 0.50$. Solving (2.8) numerically, we find that $\boldsymbol{\Theta} = (-0.98, -0.64, -0.42, -0.21, 0.14)^{\mathrm{T}}$. In both cases, the main study sample size is $n = 500$: similar and even more impressive (in favoring our methodology) results were obtained for $n = 1,000, 2,000, 3,000$, but the main conclusions were very similar and so we do not display those results here.

We did simulations in two cases:

1. External-Internal Data: The internal data has no replicates and the external data set has size $N = 300$ and $K = 2$ replicates for each observation. The nuisance parameters are $\boldsymbol{\Lambda}_{\text{ext}} = \sigma_u^2$ and $\boldsymbol{\Lambda}_{\text{int}} = (\mu_x, \sigma_x^2)$. We estimated σ_u^2 from the external data with replicates, and estimated μ_x, σ_x^2 using the internal data without any replicates. Standard errors were computed as in Lemma 1.

2. Internal Data Only: The internal data has $R = 2$ replicates and there are no external data ($K = 0$). The nuisance parameters $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}_{\text{int}} = (\mu_x, \sigma_x^2, \sigma_u^2)$. We estimated $(\mu_x, \sigma_x^2, \sigma_u^2)$ from the internal data with replicates. Standard errors were computed as in Lemma 2.

Section 2.7.4 of the **Supplementary Material** provides details for implementation.

2.4.1.2 Results

The results given below are similar when the main study sample size n increases to $n = 1,000, 2,000$ and $3,000$, and thus these are not displayed here. The results are also

similar when β_1 is either smaller or larger. The same qualitative results are also found for $\mathbf{\Theta} = (\theta_1, ..., \theta_5)^{\mathrm{T}}$ individually (results not shown).

We fit the new approach and compare it with the naive method for the both cases described above. Our main interest is to estimate the log relative risk $\theta_5 - \theta_1$, which compares the effect of the category 5 with the effect of the category 1. In the two simulations, we computed (a) the log relative risk pretending that X is observed; (b) our method; and (c) the naive method that ignores measurement error. In the scenario of internal data with $R = 2$, the predictor used was the sample mean of the replicates.

Based on 1000 simulated data sets, in Table 2.1, we report the empirical average mean bias, asymptotic standard error, standard deviation, root mean squared error, and coverage rate of the nominal 95% confidence interval across the simulations.

		Log Relative Risk Analysis				
Data	Method	mean bias	Mean Estimated Std. Err.	Actual Standard Deviation	RMSE	Coverage
X observed		0.016	0.304	0.301	0.301	95.2%
Ext-Int Data						
	Our Method	-0.005	0.41	0.402	0.402	94.5%
	Ignore ME	-0.453	0.251	0.256	0.520	0%
Int Data						
	Our method	0.005	0.361	0.323	0.323	95.9%
	Ignore ME	-0.287	0.268	0.266	0.391	80.2%

Table 2.1: Simulation study for logistic regression in Section 2.4.1 with sample size $n = 500$ and, where applicable, the external study has sample size $N = 300$ and 2 replicates, while $\beta_0 = -0.42$, $\beta_1 = \log(1.5)$. The target parameter, $\Theta = (\theta_1, ..., \theta_5)^{\mathrm{T}}$, where θ_j is the parameter for the j^{th} category. Displayed are results for the estimation of the log relative risk, $\theta_5 - \theta_1$. *Ext-Int Data* is the case that external data are used to estimate the measurement error variance. *Int Data* is the case that the internal data have 2 replicates, and the *Ignore ME* estimator ignores the measurement error and is based on the mean of these replicates. *Coverage* is the coverage rate of nominal 95% confidence intervals. RMSE is the square root of the mean squared error.

From Table 2.1, we observe the following.

- The estimator using true X and our method both have little bias and provide near-nominal coverage.
- The naive estimator that ignores the measurement error is badly biased and attenuated towards zero. Consequently the coverage probabilities are near-zero and the root mean squared errors are quite inflated.
- With no internal replicates, i.e., $R = 1$, the root mean squared error of our method is naturally higher than if X had been observed, but not quite as high as would be expected in a continuous analysis. Indeed, in a continuous analysis with attenuation $\lambda = 0.50$, as in our simulation, one would expect a doubling of root mean squared error.

2.4.2 Linear regression

2.4.2.1 Scenarios

In this section, we do simulations based on simple linear regression with no Z, including homoscedastic and heteroscedastic cases.

We assume that the true model is

$$Y \;=\; \beta_0 + X\beta_1 + \epsilon = (1, X)\boldsymbol{\beta} + \epsilon, \tag{2.17}$$

Similarly, we generate data as

$$W = X + U; \quad X = \text{Normal}(\mu_x, \sigma_x^2); \quad U = \text{Normal}(0, \sigma_u^2).$$

We set $\beta_0 = 0$ and set $\beta_1 = 0.75$ and studied two cases: (a) homoscedastic with $\epsilon \sim N(0, 1)$; and (b) heteroscedastic with $\epsilon \sim N(0, 0.2 + 0.5x^2)$. The classical attenuation coefficient and

sample size are the same as in Section 2.4.1. Solving (2.8) numerically, we find that $\boldsymbol{\Theta} = (-1.04, -0.40, 0.00, 0.40, 1.05)^{\mathrm{T}}$. Section 2.7.3 of the **Supplementary Material** provides implementation details.

2.4.2.2 Results

Similarly as before, our main interest is to estimate $\theta_5 - \theta_1$, which compares the effect of the category 5 with the effect of the category 1. In the two simulations, we computed $\theta_5 - \theta_1$ (a) pretending that X is observed; (b) our methods; and (c) the naive method that ignores measurement error. For the naive method, in internal data with $R = 2$, the predictor used is the sample mean of the replicates.

Based on 1000 simulated data sets, in Table 2.2, we report the empirical average mean bias, asymptotic standard error, standard deviation, root mean squared error, and coverage rate of the nominal 95% confidence intervals across the simulations.

From Table 2.2, we see that similar conclusions can be drawn as in Section 2.4.1. However, an interesting thing is in the heteroscedastic case, when noise ϵ has its variance related to X. Assuming that X is observed, the coverage rate of nominal 95% confidence intervals is low, because the heteroscedasticity is ignored. Using our method, we can get close to nominal coverage without knowing any information about the noise ϵ. Thus, this example shows that our method is very general as we stated in Example 1(A).

2.5 Empirical example

2.5.1 Data description

We illustrate our methods using data from the Eating at America's Table (EATS) Study (Subar et al., 2001), in which 964 participants completed multiple 24-hour recalls of diet. We consider the variable Fat Density, which is the percentage of calories coming from Fat. The response Y is either (a) the indicator of obesity, which means that a subject's body mass index (BMI, weight in kilograms divided by the square of height in meters) is 30 or greater.

Data	Method	Results Analysis ($\theta_5 - \theta_1$) mean bias	Mean Estimated Std. Err.	Actual Standard Deviation	RMSE	Coverage
			Homoscedastic $\epsilon \sim N(0,1)$			
X observed		0.004	0.145	0.150	0.150	95.1%
Ext-Int Data						
	Our Method	0.013	0.249	0.233	0.233	95.8%
	Ignore ME	-0.814	0.139	0.142	0.826	0.1%
Int Data						
	Our method	-0.007	0.176	0.170	0.170	95.3%
	Ignore ME	-0.536	0.142	0.145	0.555	3.7%
			Heteroscedastic $\epsilon \sim N(0, 0.2 + 0.5x^2)$			
X observed		0.004	0.123	0.169	0.169	85.3%
Ext-Int Data						
	Our Method	0.011	0.261	0.245	0.245	95.9%
	Ignore ME	-0.814	0.122	0.135	0.825	0.1%
Int Data						
	Our Method	-0.010	0.197	0.189	0.189	95.9%
	Ignore ME	-0.537	0.123	0.141	0.555	1.8%

Table 2.2: Simulation study for linear regression in Section 2.4.2 with $n = 500$ and, where applicable, the external study has sample size $N = 300$ and 2 replicates, while $\beta_0 = 0$, $\beta_1 = 0.75$. The target parameter, $\Theta = (\theta_1, ..., \theta_5)^{\mathrm{T}}$, where θ_j is the parameter for the j^{th} category. Displayed are results for the estimation of $\theta_5 - \theta_1$. *Ext-Int Data* is the case that external data are used to estimate the measurement error variance. *Int Data* is the case that the internal data have 2 replicates, and the *Ignore ME* estimator ignores the measurement error and is based on the mean of these replicates. *Coverage* is the coverage rate of nominal 95% confidence intervals. RMSE is the square root of the mean squared error.

or (b) the actual body mass index. We assume that W, is unbiased for usual intake X, and that $W = X + U$. It is reasonable in these data to take (a) X to be normally distributed, (b) that U is normally distributed; and (c) that X and U are independent, as we now describe. We used the methods described in Chapter 1.7 of Carroll et al. (2006). Specifically, for (a), a qq-plot of the individual means for Fat Density looked acceptably normal, with skewness and kurtosis = -0.06 and 3.02, respectively, see the top panel of Figure 2.1. For (b), we took differences of the first and second Fat Density measurements, which had skewness and kurtosis = -0.14 and 3.40, respectively: the somewhat higher kurtosis here is seen to be

minor on the qq-plot, see the bottom panel of Figure 2.1. Finally, for (c), the correlation between the individual-level mean and standard deviation = 0.06, and there was no obvious strong pattern when we plotted the data the latter against the former, see Figure 2.2.

For numerical stability, our analysis in the continuous scale is uses centered and standardized W using $(15W - 5)/\sqrt{0.5}$. To illustrate an example of an internal and an external study, we randomly selected $N = 200$ subjects as the external study to have the first two 24-hour recalls, while using the remaining data as the main internal study. As in the simulation, we either set the number of recalls $R = 1,\ K = 2$, meaning the external study data were used to estimate the measurement error variance, for $R = 2,\ K = 0$, in which case the external data were not used.

2.5.2 Results

2.5.2.1 Logistic regression

As described in Section 2.4.1, we assume the true model defined by (2.15)-(2.16), and the respective two cases. In this application we again estimate the log relative risk $\theta_5 - \theta_1$. We fit both our new approach and the naive model that ignores measurement error when external data is and is not used.

In Table 2.3, we observe that when using the external data and only 1 observation in the internal data the estimate of the log relative risk $\theta_5 - \theta_1$ from our approach is 108% greater than the naive estimate, while when using internal data with two replicates our estimate of our approach is 32% greater than the naive estimate. This makes sense because the second case uses the mean of two replicates, hence has smaller measurement error variance, and thus the naive estimate will be closer to our method.

In both cases, the asymptotic standard error from our new method is greater than the naive method, which led to wider confidence intervals. This makes sense, because with a scalar covariate measured with error, correcting for measurement error bias usually increases

		Log Relative Risk Analysis			
			Asymptotic		
Data	Method	Estimate	Std. Err.	95% CI	p-value
Ext-Int Data					
	Our Method	0.98	0.47	$(0.06, 1.90)$	0.036
	Ignore ME	0.47	0.24	$(0.00, 0.95)$	0.049
Int Data					
	Our Method	1.10	0.34	$(0.43, 1.77)$	0.001
	Ignore ME	0.83	0.22	$(0.39, 1.26)$	0.000

Table 2.3: Data analysis for logistic regression in Section 2.5. The target parameter, $\Theta = (\theta_1, ..., \theta_5)^{\mathrm{T}}$, where θ_j is the parameter for the j^{th} category. Displayed are results for the estimation of the log relative risk, $\theta_5 - \theta_1$. *Ext-Int Data* is the case that external data are used only to estimate the measurement error variance, and the external data have 2 replicates. *Int Data* is the case that the internal data have 2 replicates, and the *Ignore ME* estimator ignores the measurement error and is based on the mean of these replicates. *Asymptotic Std. Err.* is the standard error estimate from the theory. *CI* is the nominal 95% confidence interval for the log relative risk. *p-value* is the p-value for the test that the log relative risk = 0.

estimated standard errors, while of course reducing bias.

2.5.2.2 Linear regression

Next we consider the linear model with body mass index as the response. All assumptions for W, X and U are the same as in Section 2.5.1. Moreover, we maintain the standardization and sampling scheme in Section 2.5.1: the results are presented in Table 2.4.

From Table 2.4, we observe similar conclusions as in logistic regression case. One point of particular interest is that in both scenarios (external-internal or internal data only), our estimator converges theoretically to the same value, and this is seen in the results. The naive method that ignores measurement error estimates different parameters because the measurement error variance is twice as large in the external-internal case as it is in the internal-only case.

		Results Analysis ($\theta_5 - \theta_1$)			
			Asymptotic		
Data	Method	Estimate	Std. Err.	95% CI	p-value
Ext-Int Data					
	Our Method	0.59	0.18	$(0.24, 0.95)$	0.001
	Ignore ME	0.28	0.10	$(0.09, 0.47)$	0.004
Int Data					
	Our Method	0.56	0.13	$(0.30, 0.81)$	0.000
	Ignore ME	0.35	0.09	$(0.18, 0.52)$	0.000

Table 2.4: Data analysis in for linear regression Section 2.5. The target parameter, $\Theta = (\theta_1, ..., \theta_5)^{\mathrm{T}}$, where θ_j is the parameter for the j^{th} category. Displayed are results for the estimation of $\theta_5 - \theta_1$. *Ext-Int Data* is the case that external data are used only to estimate the measurement error variance, and the external data have 2 replicates. *Int Data* is the case that the internal data have 2 replicates, and the *Ignore ME* estimator ignores the measurement error and is based on the mean of these replicates. *Asymptotic Std. Err.* is the standard error estimate from the theory. *CI* is the nominal 95% confidence interval for $\theta_5 - \theta_1$. *p-value* is the p-value for the test that $\theta_5 - \theta_1 = 0$.

2.6 Other approaches and the assumptions

2.6.1 Other approaches

We emphasize once more that it is common practice in epidemiology to categorize a continuous predictor, and we have given numerous citations of this practice. Generally, this practice results in a misspecified model.

Our goal is to correct the analysis so as to reproduce, asymptotically, the estimators that would have been obtained if there were no measurement error. The problem has not been considered previously in the context that a continuous predictor has been categorized. Such categorization generally leads to differential measurement error (Flegal et al., 1991; Gustafson, 2004; Buonaccorsi, 2010), and thus additional complications over simply fitting a measurement error model.

While our paper is the first to consider the issue of how to correct an analysis to account for a continuous predictor that is categorized, there are of course other possible approaches,

but none of them really avoids the basic issues we have discussed of what is needed to obtain consistent estimators with asymptotically correct inference in the case of measurement error.

- For example, one could assume that the true risk model is based upon the categorized truth, even if this is implausible in most contexts. One could further assume that the misclassification is nondifferential, which is incorrect if the true risk model is in the continuous scale (Flegal et al., 1991; Gustafson, 2004; Buonaccorsi, 2010). There is a small literature on this problem. Gustafson (2004), especially Chapter 6.1, has remarks on the bias induced when a binary predictor is misclassified. Buonaccorsi (2010), Chapter 6.7.7 and Chapter 6.14, has a detailed discussion of the issue, and provides a number of references to the problem. Both Gustafson (2004) and Buonaccorsi (2010) show that a measurement error correction will require a distribution for the categorical X given (W, Z), sometimes called the reclassification rate, and both indicate that there are substantive issues, including identifiability, involved with estimating these models. For replication studies wherein W is measured repeatedly on a subset of the data, there is some evidence that 3 replicates will result in identifiability. However, both books emphasize the use of internal validation substudies, wherein one actually observes X in a substudy.

  If X_{cat} is the categorized truth, then one might attempt an analysis based on assuming a joint distribution of (Y, W, X_{cat}) given Z, but as in any measurement error model Carroll et al. (2006), the joint distribution requires (a) a distribution for Y given (X_{cat}, W, Z), and (b) the distribution of (W, X_{cat}) given Z. However, (a) actually depends on W, and thus that the modeling presents additional complications. In addition, (b) is no easier than ours, can be implausible and does not make fewer assumptions than we have done.

- Simulation-extrapolation, or SIMEX, (Cook and Stefanski, 1994; Stefanski and Cook, 1995; Carroll et al., 2006) is a well-known approach to the creation of *approximately*,

but not fully, consistent estimators for additive measurement error models of the form $W = X + Z^{\mathrm{T}}\alpha + U$, where U is independent of Z and can be homoscedastic or heteroscedastic but has replicates (Devanarayan and Stefanski, 2002), and is generally taken to be normally distributed. This literature attempts to dispense with distributional assumptions for X for the continuous case, but is at best approximately correct. The fact that a categorized risk model is implausible, leading to differential measurement error, may also cause complications, but the use of SIMEX in this context is a worthwhile topic for further study. We also mention the MCSIMEX procedure (Lederer and Küchenhoff, 2006), which is appropriate for misclassified data where the misclassification probabilities can be estimated.

- It is also possible to change the paradigm entirely and avoid categorization, and all the issues related to categorization, by instead using Bsplines. Indeed, part of the reason sometimes given for categorizing a continuous predictor and not modeling a response linearly in the continuous X is that it could lead to unduly extreme comparisons for risk between the lowest and the highest values of X. The general thought is that this can be overcome by replacing the linear X by a Bspline in X. There are papers involving Bsplines and measurement error (Berry et al., 2002; Ganguli et al., 2005; Pham et al., 2013), and it appears that regression calibration can possible be used by calibrating each spline basis function. After the fitting, one could compare the Bspline fits at the 10^{th}, 30^{th}, 50^{th}, 70^{th} and 90^{th} percentiles of X to form versions of the tables found in epidemiology papers, but the interpretations are not fully comparable.

We showed how to solve this problem and given asymptotically consistent estimators with asymptotically correct standard errors. Assumption 2 is reasonable in other contexts than ours, for example, that X has a mixture-of-normals distribution and U is normally distributed (Cordy and Thomas, 1997).

2.6.2 Assumptions in the simulations and example

Readers of an initial version of this paper have noted that our simulations and data example use the assumption that the distribution of X given (W, Z) is normally distributed, but misinterpreted this fact into concluding that the approach is only applicable in that case. For the data example in Section 2.5, we justified the assumptions using known methods for model checking of measurement error models. Assumption 2 is widely used and reasonable in many other contexts than ours numerical work, for example, that X has a mixture-of-normals distribution and U is normally distributed (Cordy and Thomas, 1997). Modeling via mixture distributions is a reasonable way to extend what we have done in the classical error case. See also Sarkar et al. (2014) for the homoscedastic and heteroscedastic cases when the variance function and the distributions of X and U are modeled as mixture distributions.

Many papers in the literature also rely on the existence of validation data, where X is actually observed in a subset of the main data set. In that case, Assumption 2 is easily checked by model fitting and validation on the observed validation data subset.

2.7 Supplementary material

The **Supplementary Material** includes detailed formulae for the linear and logistic cases as mentioned in as mentioned in Sections 2.3.1, 2.4.1.1 and 2.4.2.1, and plots mentioned in Section 2.5.1.

2.7.1 Sketch of technical arguments

2.7.1.1 Argument for Lemma 1

We consider the case that there are external data used to estimate $\mathbf{\Lambda}_{\text{ext}}$ and that there are parameters $\mathbf{\Lambda}_{\text{int}}$. As in Section 2.3.2, the data for $i = 1, ..., n$ are for the internal data, while, for $i = n + 1, ..., n + N$, are for the external data if such external data exist and are used. The functions $\Psi_i(\mathbf{\Omega})$ are also defined in Section 2.3.2.

By a standard Taylor series argument,

$$\begin{aligned} 0 &= (n+N)^{-1/2}\textstyle\sum_{i=1}^{N+n}\Psi_i(\widehat{\mathbf{\Omega}}) \\ &= (n+N)^{-1/2}\textstyle\sum_{i=1}^{N+n}\Psi_i(\mathbf{\Omega}) \\ &\quad + \left\{(n+N)^{-1}\textstyle\sum_{i=1}^{N+n}\partial\Psi_i(\mathbf{\Omega})/\partial\mathbf{\Omega}\right\}(n+N)^{1/2}(\widehat{\mathbf{\Omega}}-\mathbf{\Omega}) + o_p(1), \end{aligned}$$

so that

$$\begin{aligned} (n+N)^{1/2}(\widehat{\mathbf{\Omega}}-\mathbf{\Omega}) &= -\left\{(n+N)^{-1}\textstyle\sum_{i=1}^{N+n}\partial\Psi_i(\mathbf{\Omega})/\partial\mathbf{\Omega}\right\}^{-1} \\ &\quad \times(n+N)^{-1/2}\textstyle\sum_{i=1}^{N+n}\Psi_i(\mathbf{\Omega}) + o_p(1). \end{aligned}$$

It is obvious that $(n+N)^{-1}\sum_{i=1}^{N+n}\partial\Psi_i(\mathbf{\Omega})/\partial\mathbf{\Omega} = A + o_p(1)$, and immediate that

$$(n+N)^{-1/2}\sum_{i=1}^{N+n}\Psi_i(\mathbf{\Omega}) \to \text{Normal}(0, B),$$

where A and B are defined in Lemma 1.

2.7.1.2 Argument for Lemma 2

We consider the case that there are only parameters $\mathbf{\Lambda}_{\text{int}}$. As in Section 2.3.2, the data for $i = 1, ..., n$ are for the internal data. The functions $\Psi_i(\mathbf{\Omega})$ are also defined in Section 2.3.2.

By a standard Taylor series argument,

$$\begin{aligned} 0 &= n^{-1/2}\textstyle\sum_{i=1}^{n}\Psi_i(\widehat{\mathbf{\Omega}}) \\ &= n^{-1/2}\textstyle\sum_{i=1}^{n}\Psi_i(\mathbf{\Omega}) \\ &\quad + \left\{n^{-1}\textstyle\sum_{i=1}^{n}\partial\Psi_i(\mathbf{\Omega})/\partial\mathbf{\Omega}\right\}n^{1/2}(\widehat{\mathbf{\Omega}}-\mathbf{\Omega}) + o_p(1), \end{aligned}$$

so that

$$n^{1/2}(\widehat{\mathbf{\Omega}} - \mathbf{\Omega}) = -\left\{n^{-1}\textstyle\sum_{i=1}^{n}\partial\Psi_i(\mathbf{\Omega})/\partial\mathbf{\Omega}\right\}^{-1} \times n^{-1/2}\textstyle\sum_{i=1}^{n}\Psi_i(\mathbf{\Omega}) + o_p(1).$$

It is obvious that $n^{-1}\sum_{i=1}^{n}\partial\Psi_i(\mathbf{\Omega})/\partial\mathbf{\Omega} = A + o_p(1)$, and immediate that

$$n^{-1/2}\sum_{i=1}^{n}\Psi_i(\mathbf{\Omega}) \to \text{Normal}(0, B),$$

where A and B are defined in Lemma 2.

2.7.2 Estimate nuisance parameter $\mathbf{\Lambda}$

Here we only consider two cases among numerous possibilities. One is that the internal data consists of (Y_i, W_i, Z_i) for $i = 1, ...n$ and σ_u^2 is estimated from the external data using replicates W_{ik} for $k = 1, ..., K$ and $i = n+1, ..., n+N$. The second case is that the replicates are in the internal data.

2.7.2.1 External-internal data

For specificity, we consider the first case that the external data have no responses Y, are independent of the internal data. Suppose that we use external data only to estimate σ_u^2, and we observe $W_{ik} = X_i + U_{ik}$ for $k = 1, ..., K$ and $i = n+1, ..., n+N$. We use internal data to estimate μ_x, σ_x^2 without replicates. In the external data, let $\overline{W}_{i\cdot} = K^{-1}\sum_{k=1}^{K} W_{ik}$. Define $\widehat{\sigma}_{u,i}^2 = (K-1)^{-1}\sum_{k=1}^{K}(W_{ik} - \overline{W}_{i\cdot})^2$ to be the sample variance of the W_{ik} for a given i. Because $E\{(W_i - \mu_x)^2) = \sigma_x^2 + \sigma_u^2$, unbiased estimating equations for $(\mathbf{\Lambda}_{\text{ext}}, \mathbf{\Lambda}_{\text{int}}) = (\mu_x, \sigma_x^2, \sigma_u^2)$ are

$$\begin{aligned}
\text{For } \mu_x: &\quad n^{-1}\textstyle\sum_{i=1}^{n}(W_i - \mu_x) = 0;\\
\text{For } \sigma_u^2: &\quad N^{-1}\textstyle\sum_{i=n+1}^{n+N}(\widehat{\sigma}_{u,i}^2 - \sigma_u^2) = 0;\\
\text{For } \sigma_x^2: &\quad n^{-1}\textstyle\sum_{i=1}^{n}\{(W_i - \mu_x)^2 - \sigma_x^2 - \sigma_u^2\} = 0.
\end{aligned}$$

2.7.2.2 Internal data only

Suppose there is no external data, and we have replicates W_{ir} for $r = 1, ..., R$ in the internal data. Now we use internal data to estimate $\Lambda = (\mu_x, \sigma_x^2, \sigma_{uR}^2)$, and we observe $W_{ir} = X_i + U_{ir}$ for $r = 1, ..., R$ and $i = 1, ..., n$.

Define $\overline{W}_{i\cdot} = R^{-1} \sum_{r=1}^{R} W_{ir}$. Define $\widehat{\sigma}_{u,i}^2$ to be the sample variance of the W_{ir} within subject i, and define $\sigma_u^2/R = \sigma_{uR}^2$. The estimating equations are

$$\begin{aligned}
\text{For } \mu_x: &\quad n^{-1}\textstyle\sum_{i=1}^{n}(\overline{W}_{i\cdot} - \mu_x) = 0; \\
\text{For } \sigma_{uR}^2: &\quad n^{-1}\textstyle\sum_{i=1}^{n}(\widehat{\sigma}_{u,i}^2/R - \sigma_{uR}^2) = 0; \\
\text{For } \sigma_x^2: &\quad n^{-1}\textstyle\sum_{i=1}^{n}\{(\overline{W}_{i\cdot} - \mu_x)^2 - \sigma_x^2 - \sigma_{uR}^2\} = 0.
\end{aligned}$$

Since the two cases we considered are the same as in linear regression and logistic regression, the way we estimate Λ_{int} and Λ_{ext} are exactly the same. Then we will only give details for the estimating equations about $\boldsymbol{\beta}$ and $\boldsymbol{\Theta}$ below.

2.7.3 Details for linear regression

2.7.3.1 Background

Here we give full details of our methodology for linear regression. As in Lemma 1, $\boldsymbol{\Omega} = (\boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{\Lambda}_{\text{int}}, \boldsymbol{\Lambda}_{\text{ext}})$.

Let $\widetilde{Z} = (1, Z^{\mathrm{T}})^{\mathrm{T}}$. Here we consider the simple case of linear regression with the classical measurement error model in both the external and internal data sets to be

$$\begin{aligned}
Y &= X\beta_1 + \widetilde{Z}^{\mathrm{T}}\boldsymbol{\beta}_2 = (X, \widetilde{Z}^{\mathrm{T}})\boldsymbol{\beta}; \\
W &= X + U; \quad X = \text{Normal}(\widetilde{Z}^{\mathrm{T}}\boldsymbol{\alpha}, \sigma_x^2); \quad U = \text{Normal}(0, \sigma_u^2).
\end{aligned}$$

2.7.3.2 The forms of $\Phi(\cdot)$

In this linear model, denote the estimating equations for $\boldsymbol{\beta}$ as $\Phi(\cdot)$, we consider

$$\Phi(Y, W, Z, \boldsymbol{\beta}, \boldsymbol{\Lambda}_{\text{int}}, \boldsymbol{\Lambda}_{\text{ext}}) \quad = (1, W)^{\mathrm{T}}(Y - W\beta_1 - \widetilde{Z}^{\mathrm{T}}\beta_2) + (0, \beta_1\sigma_u^2)^{\mathrm{T}}.$$

2.7.3.3 The forms of $\Phi_{\text{cat}}(\cdot)$ and $Q(\cdot)$

Since we assume the true model is $Y = (X, \widetilde{Z}^{\mathrm{T}})\boldsymbol{\beta}$, it is easy to see that categorical estimating function

$$\Phi_{\text{cat}}\{Y, M^{\mathrm{T}}(X, Z)\boldsymbol{\Theta}\} = M(X, Z)[Y - M^{\mathrm{T}}(X, Z)\boldsymbol{\Theta}].$$

Hence, by simple calculations and following Remark 3, with $\boldsymbol{\Omega} = (\boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{\Lambda}_{\text{int}}, \boldsymbol{\Lambda}_{\text{ext}})$,

$$Q(W, Z, \boldsymbol{\Omega}) \quad = \quad E\left[M(X, Z)\left\{(X, \widetilde{Z}^{\mathrm{T}})\boldsymbol{\beta} - M^{\mathrm{T}}(X, Z)\boldsymbol{\Theta}\right\}\middle| W, Z\right].$$

We used the integrate function in the R package stats to compute the integrals.

The estimating function for $\boldsymbol{\beta} = (\beta_0, \beta_1)$ is

$$\Phi(\boldsymbol{\beta}, \widehat{\boldsymbol{\Lambda}}) \quad = \quad n^{-1}\textstyle\sum_{i=1}^{n} E\left([Y_i - H\{m(X_i, \boldsymbol{\beta})\}]\partial m(X_i, \boldsymbol{\beta})/\partial\boldsymbol{\beta}^{\mathrm{T}}\middle| W_i\right).$$

The estimating function for $\boldsymbol{\Theta}$ is

$$Q(W_i, \boldsymbol{\Theta}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}) \quad = \quad E\left[\begin{array}{c} m(X_i, \widehat{\boldsymbol{\beta}})I(X_i \in C_1) - \boldsymbol{\Theta}_1 I(X_i \in C_1) \\ \vdots \\ m(X_i, \widehat{\boldsymbol{\beta}})I(X_i \in C_J) - \boldsymbol{\Theta}_J I(X_i \in C_J) \end{array}\middle| W_i\right].$$

Asymptotic standard errors were estimated as in Lemma 1 and Lemma 2.

2.7.4 Details for logistic regression

2.7.4.1 Background

Here we give full details of our methodology for logistic regression. As in Lemma 1, $\boldsymbol{\Omega} = (\boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{\Lambda}_{\text{int}}, \boldsymbol{\Lambda}_{\text{ext}})$.

As before, let $H(\cdot)$ denote the logistic distribution function and let $\widetilde{Z} = (1, Z^{\text{T}})^{\text{T}}$. Here we consider the special case of linear logistic regression with the classical measurement error model in both the external and internal data sets to be

$$\begin{aligned} \text{pr}(Y = 1|X, Z) &= H(X\beta_1 + \widetilde{Z}^{\text{T}}\boldsymbol{\beta}_2) = H\{(X, \widetilde{Z}^{\text{T}})\boldsymbol{\beta}\}; \\ W &= X + U; \quad X = \text{Normal}(\widetilde{Z}^{\text{T}}\boldsymbol{\alpha}, \sigma_x^2); \quad U = \text{Normal}(0, \sigma_u^2). \end{aligned}$$

Different from the linear case in Section 2.7.3, we consider the case where X depends on another covariate Z. There are numerous data structures possible, but we here present the external-internal and internal data only cases.

2.7.4.2 Settings

There are two settings of interest.

- There is no information about σ_u^2 in the internal data, so that the external parameter is the measurement error variance, $\boldsymbol{\Lambda}_{\text{ext}} = \sigma_u^2$, while the internal parameters are $\boldsymbol{\Lambda}_{\text{int}} = (\boldsymbol{\alpha}^{\text{T}}, \sigma_x^2)^{\text{T}}$.

- There are no external data, so that $\boldsymbol{\Lambda}_{\text{ext}}$ is null, and the internal data with replicates allow estimation of $\boldsymbol{\Lambda}_{\text{int}} = (\boldsymbol{\alpha}^{\text{T}}, \sigma_u^2, \sigma_x^2)^{\text{T}}$.

In both case, σ_u^2 (or σ_{uR}^2 in the internal data only case) are estimated the same as in 2.7.2.1 and 2.7.2.2, while the estimating function for $(\boldsymbol{\alpha}, \sigma_x^2)$ is

$$V_{\text{int},i}(\boldsymbol{\Lambda}_{\text{int}}, \boldsymbol{\Lambda}_{\text{ext}}) = \left\{ \widetilde{Z}_i^{\mathrm{T}}(W_i - \widetilde{Z}_i^{\mathrm{T}}\boldsymbol{\alpha}), (W_i - \widetilde{Z}_i^{\mathrm{T}}\boldsymbol{\alpha})^2 - \sigma_x^2 - \sigma_u^2 \right\},$$

where $i = 1, ..., n$.

2.7.4.3 Estimating $\boldsymbol{\beta}$

In this section, we implement our method and give all estimating equations in the case where we have both external and internal data. In another case, where we only use internal data with replicates, all results below are still valid by removing $\boldsymbol{\Lambda}_{\text{ext}}$.

Define $\lambda = \sigma_x^2/(\sigma_x^2 + \sigma_u^2)$. Then, given (W, Z), X follows a normal distribution with mean $\mu(W, Z, \boldsymbol{\Lambda}_{\text{ext}}, \boldsymbol{\Lambda}_{\text{int}}) = \widetilde{Z}^{\mathrm{T}}\boldsymbol{\alpha} + \lambda(W - \widetilde{Z}^{\mathrm{T}}\boldsymbol{\alpha})$ and variance $\lambda\sigma_u^2$. We write this conditional density as $f_{x|w,z}(x, w, z, \boldsymbol{\beta}, \boldsymbol{\Lambda}_{\text{int}}, \boldsymbol{\Lambda}_{\text{ext}})$.

There are multiple ways to estimate $\boldsymbol{\beta}$ from the observed data. Here we describe two of them.

- The first is regression calibration, in which X is replaced by its mean given (W, Z) and the linear logistic model is fit. Thus the regression calibration method has

$$\begin{aligned}\Phi(Y, W, Z, \boldsymbol{\beta}, \boldsymbol{\Lambda}_{\text{int}}, \boldsymbol{\Lambda}_{\text{ext}}) &= \{\mu(W, Z, \boldsymbol{\Lambda}_{\text{ext}}, \boldsymbol{\Lambda}_{\text{int}}), \widetilde{Z}\}^{\mathrm{T}} \\ &\quad \times [Y - H\{\mu(W, Z, \boldsymbol{\Lambda}_{\text{ext}}, \boldsymbol{\Lambda}_{\text{int}})\beta_1 + \widetilde{Z}^{\mathrm{T}}\boldsymbol{\beta}_2\}].\end{aligned}$$

- A second possibility, one that we used, is the following. By simple calculations, $\text{pr}(Y = 1|W, Z) = p(W, Z, \boldsymbol{\beta}, \boldsymbol{\Lambda}_{\text{int}}, \boldsymbol{\Lambda}_{\text{ext}})$, where

$$p(W, Z, \boldsymbol{\beta}, \boldsymbol{\Lambda}_{\text{int}}, \boldsymbol{\Lambda}_{\text{ext}}) = \int H\{(x, \widetilde{Z}^{\mathrm{T}})\boldsymbol{\beta}\} f_{x|w,z}(x, W, Z, \boldsymbol{\Lambda}_{\text{int}}, \boldsymbol{\Lambda}_{\text{ext}}) dx, \quad (2.18)$$

a quantity that is easily computed in R using the integrate function in the R package stats. Denote $p_i = \text{pr}(Y_i = 1|W_i, Z_i)$. Thus, the loglikelihood $\propto n^{-1}\sum_{i=1}^{n} Y_i \text{log}(p_i) + (1 - Y_i)\text{log}(1 - p_i)$. We then use optim function in the R package stats to minimize negative loglikelihood to estimate $\boldsymbol{\beta}$.

2.7.4.4 The forms of $\Phi_{\text{cat}}(\cdot)$ and $Q(\cdot)$

Since we assume the true model is $\text{pr}(Y = 1|X, Z) = H\{(X, \widetilde{Z}^{\text{T}})\boldsymbol{\beta}\}$, it is easy to see that categorical estimating function

$$\Phi_{\text{cat}}\{Y, M^{\text{T}}(X, Z)\boldsymbol{\Theta}\} = M(X, Z)[Y - H\{M^{\text{T}}(X, Z)\boldsymbol{\Theta}\}].$$

Hence, by simple calculations and following Remark 3, with $\boldsymbol{\Omega} = (\boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{\Lambda}_{\text{int}}, \boldsymbol{\Lambda}_{\text{ext}})$,

$$Q(W, Z, \boldsymbol{\Omega}) \;=\; E\left(M(X, Z)\left[H\{(X, \widetilde{Z}^{\text{T}})\boldsymbol{\beta}\} - H\{M^{\text{T}}(X, Z)\boldsymbol{\Theta}\}\right] \middle| W, Z \right).$$

We used the integrate function in the R package stats to compute the integrals.

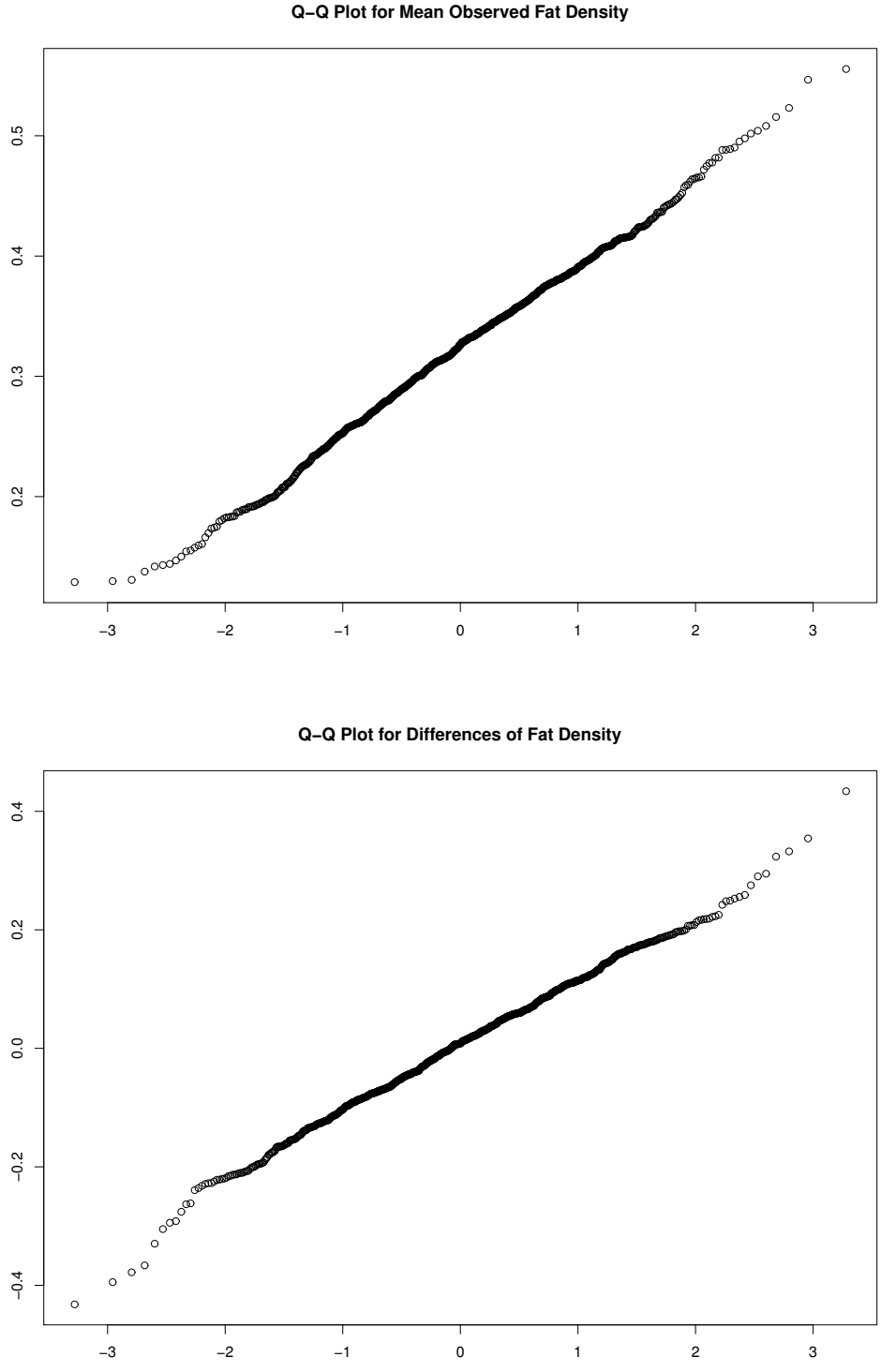

Figure 2.1: EATS data of Section 2.5. Top panel: Normal qq-plot of the mean Fat Density over 4 recalls. This indicates that the mean Fat Density is approximately normally distributed and qualifies for the assumptions in our numerical example. Bottom panel: Normal qq-plot of differences of observed Fat density, as a diagnosis that U is approximately normally distributed.

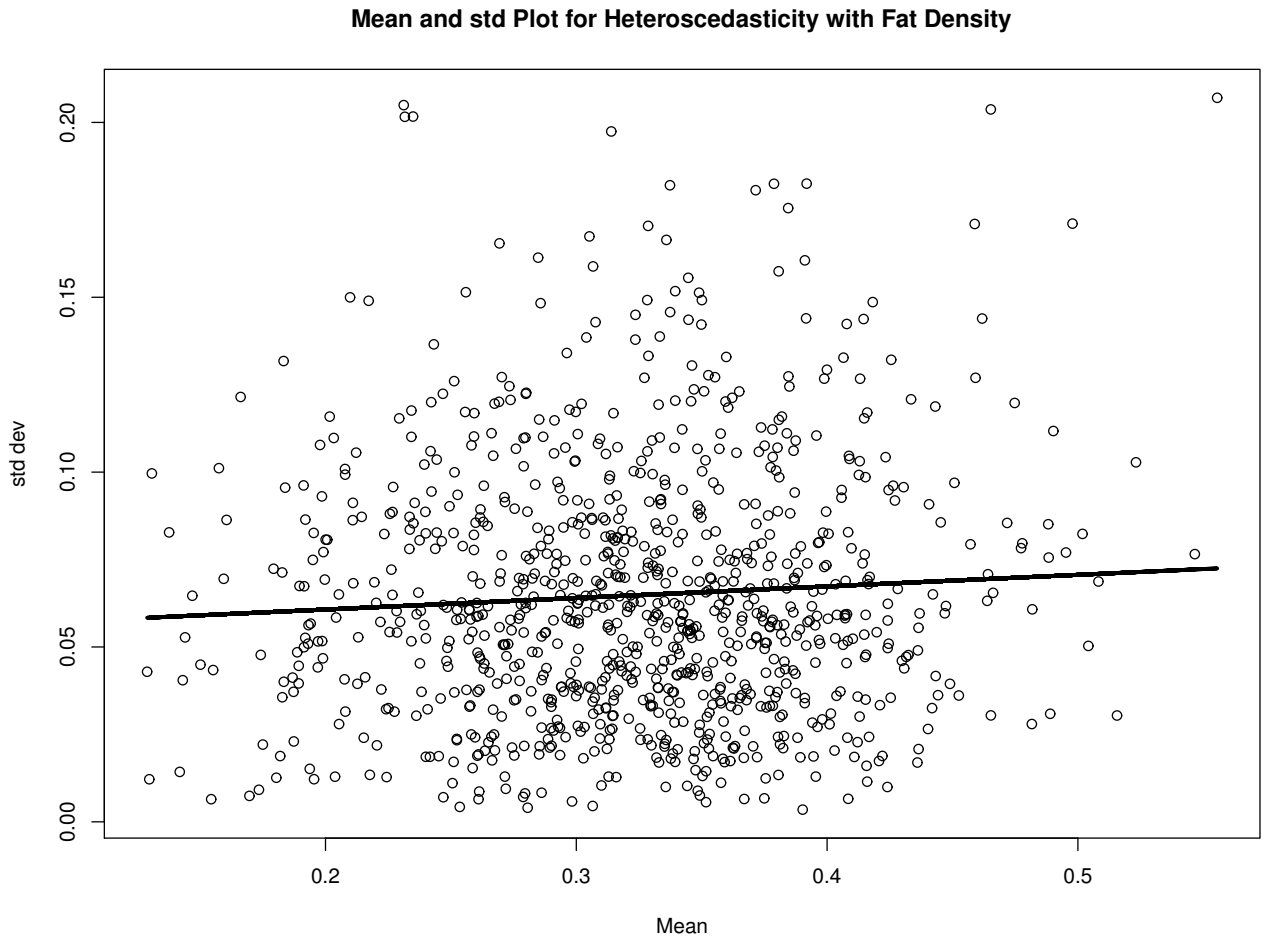

Figure 2.2: EATS data of Section 2.5. Mean and standard deviation plot to diagnose heteroscedasticity, showing that there is little heteroscedasticity in the measurement errors.

3. SPARSE QUADRATIC CLASSIFICATION RULES VIA LINEAR DIMENSION REDUCTION

3.1 Introduction

We consider a binary classification problem: given n independent pairs (X_i, Y_i) from a joint distribution (X, Y) on $\mathbb{R}^p \times \{1, 2\}$, our goal is to both learn a rule that will assign one of two labels to a new data point $X \in \mathbb{R}^p$, and determine the subset of p variables that influences the rule. One of the popular classification tools is linear discriminant analysis, or LDA (Mardia et al., 1979, Chapter 11). While it gives unsatisfactory results when applied to high-dimensional datasets (Dudoit et al., 2002), recent work suggests that additional regularization, variable selection in particular, leads to dramatic performance improvements. Earlier approaches perform variable selection and regularize the sample covariance matrix by treating it as diagonal (Tibshirani et al., 2003; Witten and Tibshirani, 2011). More recent methods directly estimate the discriminant directions by using convex optimization framework with sparsity-inducing penalties (Cai and Liu, 2011; Mai et al., 2012; Gaynanova et al., 2016).

Despite these significant advances, a key underlying assumption of linear discriminant analysis is the equality of covariance matrices between the groups, $\Sigma_1 = \Sigma_2$. This assumption is unlikely to be satisfied in practice, leading to suboptimal performance of linear rule. When the measurements are normally distributed, $X_i|Y_i = g \sim \mathcal{N}(\mu_g, \Sigma_g)$, $g \in \{1, 2\}$, with $\Sigma_1 \neq \Sigma_2$, the Bayes rule is quadratic, leading to quadratic discriminant analysis, or QDA. As with linear case, the quadratic discriminant analysis performs poorly when p is large. This unsatisfactory performance is largely due to the estimation of precision matrices Σ_1^{-1} and Σ_2^{-1}, a task that is extremely challenging when $p \gg n$. In fact, even when $p = n/2$ and the assumption of equal covariance matrices is violated, the misclassification error rate of sample QDA is worse than the rates of regularized linear discriminant methods (Gaynanova

et al., 2016, supplement).

Several extensions of sample QDA have been proposed. A common strategy is to jointly estimate Σ_1^{-1} and Σ_2^{-1}. Friedman (1989); Ramey et al. (2016) regularize sample covariance matrices by shrinkage. Wu et al. (2018) impose equicorrelation structure on each covariance matrix by pooling both the diagonal and off-diagonal elements. Danaher et al. (2014); Guo et al. (2011); Price et al. (2014); Simon and Tibshirani (2011) use a penalized likelihood technique, where the penalty enforces similarity either between the covariance matrices Σ_g or the precision matrices Σ_g^{-1}. While these methods perform better than quadratic rules based on sample covariance matrices, they again rely on estimating two precision matrices. As such, additional assumptions on Σ_g^{-1} such as sparsity are usually enforced, and the estimation procedure scales quadratically with the number of measurements p. Moreover, the resulting classification rules still rely on all p variables, and therefore can not be used for both classification and variable selection.

Li and Shao (2015) address the variable selection problem by enforcing sparsity in both the covariance matrices and the vector of mean differences via thresholding. The method comes with strong theoretical guarantees on classification consistency and promising empirical performance. Nevertheless, it again requires additional assumptions on Σ_g, and is computationally prohibitive for large p due to required matrices inversion together with a 3-dimensional search over tuning parameter values.

In summary, a significant progress in linear discriminant methods made it possible to apply them to large datasets and perform variable selection. In practice, however, the covariance matrices are often unequal, but the existing quadratic methods typically can not perform variable selection, and are computationally prohibitive for large p. In this work we bridge the gap between the linear and the quadratic methods by developing a new classification rule that takes into account unequal covariance matrices without sacrificing either variable selection or the computational speed.

Our key methodological contribution is a different approach for constructing quadratic rule in high-dimensional settings compared to the ones taken in the literature. The existing methods rely on improved estimation of the full Bayes quadratic discriminant rule by exploring additional structural assumptions on Σ_g or Σ_g^{-1} (Simon and Tibshirani, 2011; Price et al., 2014; Le and Hastie, 2014; Li and Shao, 2015; Wu et al., 2018). In contrast, we modify the Fisher's formulation of linear discriminant analysis for the case of unequal covariance matrices. The resulting method performs simultaneous variable selection and projection of original data on a lower-dimensional space, with the subsequent application of quadratic discriminant analysis. We call this approach discriminant analysis via projections, or DAP.

Unlike the existent quadratic methods, our rule is linear in p, which allows us to devise a very efficient optimization procedure to simultaneously estimate the projection directions and perform variable selection. For $p = 500$, it takes around 1.5 seconds to implement our method, whereas the closest competing sparse quadratic method takes 30 minutes. This makes it possible to apply our approach in situations where other quadratic methods are computationally infeasible. Moreover, we connect the variables in our rule with the nonzero variables in the linear part of Bayes quadratic rule, and prove the variable selection consistency of our method in high-dimensional settings. Empirical studies confirm that for large values of p the proposed rule leads to competitive, and often smaller, misclassification error rates than the existing approaches. At the same time, our method consistently selects the sparsest models thus achieving the best balance between model complexity and misclassification error rate. Finally, the application to gene expression data of breast cancer patients (Chin et al., 2006) confirms the crucial importance of ESR1 gene in differentiating estrogen receptor status; an insight that is not possible with other approaches due to much higher complexity of corresponding classification rules.

The rest of this paper is organized as follows. In Section 3.2, we describe a new quadratic classification rule, discriminant analysis via projections. We connect the proposed approach to both linear and quadratic discriminant analysis, and derive an efficient optimization al-

gorithm for sparse estimation. In Section 3.3, we provide theoretical guarantees on the variable selection consistency of our method in high-dimensional settings. In Section 3.4, we conduct empirical studies on both simulated and real data. In Section 3.5, we discuss possible extensions in future work.

Notation: For a vector $v \in \mathbb{R}^p$, we let $\|v\|_1 = \sum_{i=1}^p |v_i|$, $\|v\|_2 = (\sum_{i=1}^p v_i^2)^{1/2}$, $\|v\|_\infty = \max_i |v_i|$. We use e_j to denote a unit norm vector with jth element being equal to one, and 1_p to denote the vector of ones of length p. For a matrix $M \in \mathbb{R}^{n\times p}$, we let $\|M\|_{\infty,2} = \max_{1\le i\le n}(\sum_{j=1}^p m_{ij}^2)^{1/2}$, $\|M\|_2 = \sup_{x:\|x\|_2=1} \|Mx\|_2$ and $|M|$ be the determinant of M. Given an index set A, we use M_A to denote the submatrix of M with columns indexed by A. For a square matrix M, we use M_{AA} to denote the submatrix of M with both rows and columns indexed by A. We use I to denote the identity matrix. We use $a_n \lesssim b_n$ to denote that there exists a constant $C > 0$ such that $a_n \le Cb_n$ for n sufficiently large. We also let $a \vee b = \max(a, b)$.

3.2 Discriminant analysis via projections

3.2.1 Review of Fisher's discriminant analysis

Consider n independent pairs (X_i, Y_i) from a joint distribution (X, Y) on $\mathbb{R}^p \times \{1, 2\}$. Let $\Sigma_g = \mathrm{cov}(X|Y = g)$, $g = 1, 2$ and assume the covariance matrices are equal, $\Sigma_1 = \Sigma_2$. Fisher's discriminant analysis seeks a linear combination of p measurements that maximize between group variability with respect to within group variability (Mardia et al., 1979, Chapter 11):

$$\underset{v\in\mathbb{R}^p}{\text{maximize}} \left\{ \frac{v^{\mathrm{T}}(\bar{x}_1 - \bar{x}_2)(\bar{x}_1 - \bar{x}_2)^{\mathrm{T}} v}{v^{\mathrm{T}} W v} \right\}, \tag{3.1}$$

where $W = (n-2)^{-1} \sum_{g=1}^2 (n_g - 1) S_g$ is the pooled sample covariance matrix, S_g is the sample covariance matrix for group g, n_g is the number of samples in group g, and $\bar{x}_g$ is the sample mean for group g. Letting $\widehat{v}$ be a vector at which the maximum above is achieved,

the resulting classification rule for a new observation with observed value $x \in \mathbb{R}^p$ is

$$h_{\widehat{v}}(x) = \underset{g\in\{1,2\}}{\operatorname{argmin}} \left\{ (x^{\mathrm{T}}\widehat{v} - \bar{x}_g^{\mathrm{T}}\widehat{v})^{\mathrm{T}} (\widehat{v}^{\mathrm{T}} W \widehat{v})^{-1} (x^{\mathrm{T}}\widehat{v} - \bar{x}_g^{\mathrm{T}}\widehat{v}) - 2\log(n_g/n) \right\}. \tag{3.2}$$

Hence, both the new observation $x \in \mathbb{R}^p$ and the data $X \in \mathbb{R}^{n\times p}$ are projected onto the line determined by $\widehat{v}$, and the classification is performed according to Mahalanobis distance to the class means in the projected space. Since both the objective function in (3.1) and the classification rule (3.2) are invariant to the scaling of discriminant vector $\widehat{v}$, it can be expressed as $\widehat{v} = cW^{-1}(\bar{x}_1 - \bar{x}_2)$ for any constant $c \neq 0$. Moreover, the Fisher's rule (3.2) coincides with sample plug-in Bayes rule under the normality assumption, that is $X_i|Y_i = g \sim N(\mu_g, \Sigma)$.

3.2.2 Modification of Fisher's rule

Our proposal is based on the modification of criterion (3.1) to the case of unequal covariance matrices. Specifically, we consider two discriminant directions instead of one

$$\widehat{v}_g = \underset{v_g\in\mathbb{R}^p}{\operatorname{argmax}} \left\{ \frac{v_g^{\mathrm{T}}(\bar{x}_1 - \bar{x}_2)(\bar{x}_1 - \bar{x}_2)^{\mathrm{T}} v_g}{v_g^{\mathrm{T}} S_g v_g} \right\} \quad (g = 1, 2). \tag{3.3}$$

Similar to Fisher's criterion, the solutions to (3.3) can be expressed as $\widehat{v}_g = c_g S_g^{-1}(\bar{x}_1 - \bar{x}_2)$ for any $c_g \neq 0$, $g = 1, 2$. Subsequently, given matrix $\widehat{V} = [\widehat{v}_1\ \widehat{v}_2]$, we modify rule (3.2) to take into account unequal covariance matrices as

$$h_{\widehat{V}}(x) = \underset{g\in\{1,2\}}{\operatorname{argmin}} \left\{ (x - \bar{x}_g)^{\mathrm{T}} \widehat{V} (\widehat{V}^{\mathrm{T}} S_g \widehat{V})^{-1} \widehat{V}^{\mathrm{T}} (x - \bar{x}_g) + \log|\widehat{V}^{\top} S_g \widehat{V}| - 2\log(n_g/n) \right\}. \tag{3.4}$$

Remark 1. *If $\widehat{v}_1$ and $\widehat{v}_2$ are linearly dependent, then $\widehat{V}$ has rank one, and $\widehat{V}^{\mathrm{T}} S_1 \widehat{V}$ and $\widehat{V}^{\mathrm{T}} S_2 \widehat{V}$ are both singular. In this case the subspace spanned by the columns of $\widehat{V}$ is the same as the subspace spanned by only one column, and we use $\widehat{V} = \widehat{v}_1$ in* (3.4).

Rule (3.4) is equivalent to applying quadratic discriminant rule to $\widehat{V}^{\mathrm{T}} x$ instead of applying

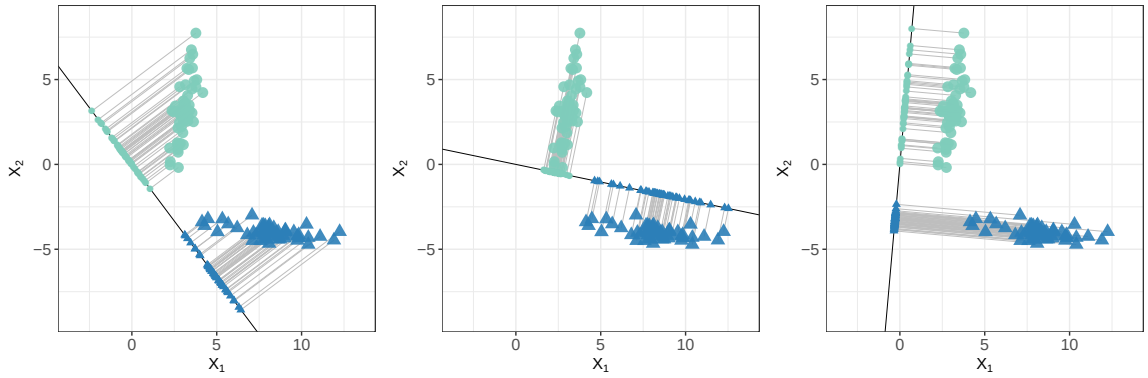

Figure 3.1: Two-group classification problem with $p = 2$ and unequal covariance matrices. *Left:* Projection using Fisher's discriminant vector. *Middle:* Projection using the covariance structure from the 1st group (circles). *Right:* Projection using the covariance structure from the 2nd group (triangles).

it directly to x. Unlike the equivalence between the Fisher's rule and the linear discriminant rule, in Section 3.2.6 we show that rule (3.4) is generally not equivalent to quadratic discriminant analysis. Nevertheless, formulation (3.4) allows to overcome possible rank degeneracy of S_g as well as perform variable selection. First, rule (3.4) requires inversion of 2×2 matrices $\widehat{V}^{\mathrm{T}} S_g \widehat{V}$, which are likely to be positive definite, in contrast to S_g. Secondly, since (3.4) effectively applies quadratic rule to $\widehat{V}^{\mathrm{T}} x$ instead of x, it only relies on those variables for which the corresponding rows of $\widehat{V}$ are nonzero. Hence, performing variable selection is equivalent to using row-sparse matrix $\widehat{V}$. Figure 3.1 shows that each $\widehat{v}_g$ from (3.3) can be viewed as a basis vector for the reduced space, and coincides with discriminant vector $\widehat{v}$ in Fisher's rule (3.1) if the pooled sample covariance matrix $W = S_1 = S_2$. Therefore, we call rule (3.4) the discriminant analysis via projections.

3.2.3 Sparse estimation

While rule (3.4) allows to overcome the potential singularity of sample covariance matrices, it still requires estimation of $\mathcal{O}(p)$ parameters and therefore may lead to poor performance in the high-dimensional settings when $p \gg n$. At the same time, in the context of linear discriminant analysis the classification performance can be significantly improved

by directly estimating the discriminant vector with sparsity regularization (Cai and Liu, 2011; Mai et al., 2012). Guided by this intuition, our goal is to obtain sparse estimates of $\psi_1 = c_1\Sigma_1^{-1}\delta$ and $\psi_2 = c_2\Sigma_2^{-1}\delta$ with $\delta = \mu_1 - \mu_2$, which are the population counterparts of $\widehat{v}_1$ and $\widehat{v}_2$ in (3.3). This approach leads to regularized row-sparse $\widehat{V}$ that can be used directly in rule (3.4).

To produce sparse estimates of ψ_1 and ψ_2, we consider penalized empirical risk minimization framework:

$$\widehat{V} = [\widehat{v}_1\ \widehat{v}_2] = \underset{v_1, v_2 \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \widehat{L}_{\psi_1}(v_1) + \widehat{L}_{\psi_2}(v_2) + \lambda \operatorname{Pen}(V) \right\},$$

where $\widehat{L}_{\psi_1}(v_1)$, $\widehat{L}_{\psi_2}(v_2)$ are empirical loss functions associated with ψ_1, ψ_2; $\lambda > 0$ is the tuning parameter, and $\operatorname{Pen}(V)$ is the sparsity-inducing penalty.

Remark 2. *Another possibility is to add sparse penalization directly within criterion* (3.3)*. In linear discriminant analysis, this approach leads to significant improvement over sample plug-in rule (Witten and Tibshirani, 2011). However, it also leads to nonconvex optimization problem and potential difficulties in obtaining very sparse solutions (Gaynanova et al., 2017). Therefore, we do not pursue the direct penalization here.*

First, we discuss our choice of penalty. As we are interested in simultaneous variable selection, that is row-sparsity of $\widehat{V}$, we propose to use group penalty. Specifically, we choose group-lasso, $\operatorname{Pen}(V) = \sum_{j=1}^{p}(v_{1j}^2 + v_{2j}^2)^{1/2}$ (Yuan and Lin, 2006), due to its convexity. Other possibilities include nonconvex group penalties, we refer the reader to Huang et al. (2012) for the review.

Next, we discuss our choice of empirical loss functions $\widehat{L}_{\psi_1}(v_1)$ and $\widehat{L}_{\psi_2}(v_2)$. Both the criterion (3.3) and the rule (3.4) are invariant to the scale of $\widehat{V}$, that is to the choice of constants c_1 and c_2. While the naive approach is to fix $c_1 = c_2 = 1$, we use $c_1 = \pi_2/(1 + \pi_2^2\delta^{\mathrm{T}}\Sigma_1^{-1}\delta)$, $c_2 = \pi_1/(1 + \pi_1^2\delta^{\mathrm{T}}\Sigma_2^{-1}\delta)$, which lead to lower-bounded empirical loss function as

well as significant computational savings. To be specific, we take advantage of the following equivalence due to the Sherman–Morrison formula:

Proposition 1. *For any $\rho \neq 0$, any non-singular matrix $M \in \mathbb{R}^{p\times p}$ and any vector $a \in \mathbb{R}^p$*

$$(M + \rho^2 aa^{\mathrm{T}})^{-1}\rho a = \rho M^{-1}a(1 + \rho^2 a^{\mathrm{T}}M^{-1}a)^{-1} \propto M^{-1}a.$$

Our choice of c_1 and c_2 leads to $\psi_1 = (\Sigma_1 + \pi_2^2\delta\delta^{\mathrm{T}})^{-1}\pi_2\delta$ and $\psi_2 = (\Sigma_2 + \pi_1^2\delta\delta^{\mathrm{T}})^{-1}\pi_1\delta$. Consider the following quadratic loss function associated with ψ_1

$$L_{\psi_1}(v_1) = (v_1 - \psi_1)^{\mathrm{T}}(\Sigma_1 + \pi_2^2\delta\delta^{\mathrm{T}})(v_1 - \psi_1)/2 = v_1^{\mathrm{T}}\Sigma_1 v_1/2 + (\pi_2\delta^{\mathrm{T}}v_1 - 1)^2/2 + C,$$

where C is a constant independent of v_1. Consider the empirical version of this loss function

$$\widehat{L}_{\psi_1}(v_1) = v_1^{\mathrm{T}}S_1 v_1/2 + \left(n^{-1}n_2 d^{\mathrm{T}}v_1 - 1\right)^2/2 + C, \tag{3.5}$$

where $d = \bar{x}_1 - \bar{x}_2$. First, $\widehat{L}_{\psi_1}(v_1)$ is invariant under linear transformation of the data (Rukhin, 1992). Secondly, $\widehat{L}_{\psi_1}(v_1)$ is always bounded from below by C, even when S_1 is singular. This ensures guaranteed convergence of the block-coordinate descent algorithm without the need to regularize S_1, and in particular, is not the case for $c_1 = 1$.

Furthermore, let $X_1 \in \mathbb{R}^{n_1\times p}$ be the submatrix of X corresponding to the 1st group, and $X_2 \in \mathbb{R}^{n_2\times p}$ be the one corresponding to the 2nd group. Let X be column-centered so that $\bar{x} = n^{-1}(n_1\bar{x}_1 + n_2\bar{x}_2) = 0$, and hence $d = n_2^{-1}n\bar{x}_1$. Then the loss (3.5) can be rewritten as

$$\begin{aligned}\widehat{L}_{\psi_1}(v_1) &= v_1^{\mathrm{T}}S_1 v_1/2 + \left(\bar{x}_1^{\mathrm{T}}v_1 - 1\right)^2/2 + C = n_1^{-1}v_1^{\mathrm{T}}X_1^{\mathrm{T}}X_1 v_1/2 - v_1^{\mathrm{T}}\bar{x}_1 + C \\ &= n_1^{-1}\|X_1 v_1 - 1_{n_1}\|_2^2/2 + C.\end{aligned}$$

That is, the loss function can be expressed as the linear regression loss function. Similarly,

$$\widehat{L}_{\psi_2}(v_2) = n_2^{-1}\|X_2 v_2 + 1_{n_2}\|_2^2/2 + C.$$

Therefore, our choice of c_1 and c_2 allows to re-express the problem of estimating ψ_1 and ψ_2 as a regression problem. This leads to efficient optimization algorithm described in Section 3.2.4.

In summary, given the column-centered data matrix $X \in \mathbb{R}^{n\times p}$ with submatrices $X_1 \in \mathbb{R}^{n_1\times p}$, $X_2 \in \mathbb{R}^{n_2\times p}$ corresponding to two groups, we find $\widehat{V} = [\widehat{v}_1\ \widehat{v}_2] \in \mathbb{R}^{p\times 2}$ as the solution to

$$\underset{V=[v_1,v_2]\in\mathbb{R}^{p\times 2}}{\text{minimize}} \left\{ n_1^{-1}\|X_1 v_1 - 1_{n_1}\|_2^2/2 + n_2^{-1}\|X_2 v_2 + 1_{n_2}\|_2^2/2 + \lambda \sum_{j=1}^{p} (v_{1j}^2 + v_{2j}^2)^{1/2} \right\}. \tag{3.6}$$

If $\lambda = 0$, $\widehat{V}$ coincides with the solution to (3.3) up to the choice of scaling. If $\lambda > 0$, then $\widehat{V}$ is row-sparse leading to variable selection. Given $\widehat{V}$, we apply rule (3.4) for classification.

3.2.4 Optimization algorithm

In this section we derive a block-coordinate descent algorithm to solve (3.6). Consider the optimality conditions with respect to each block $v_j = (v_{1j}, v_{2j})^{\mathrm{T}}$ (Boyd and Vandenberghe, 2004, Chapter 5):

$$\begin{aligned} n_1^{-1} X_{1j}^{\mathrm{T}} X_{1j} v_{1j} &= n_1^{-1} X_{1j}^{\mathrm{T}} (1_{n_1} - \sum_{k\neq j} v_{1k} X_{1k}) - \lambda u_{1j}, \\ n_2^{-1} X_{2j}^{\mathrm{T}} X_{2j} v_{2j} &= n_2^{-1} X_{2j}^{\mathrm{T}} (-1_{n_2} - \sum_{k\neq j} v_{2k} X_{2k}) - \lambda u_{2j}; \end{aligned}$$

where $u_j = (u_{1j}, u_{2j})^{\mathrm{T}}$ is the subgradient of $(v_{1j}^2 + v_{2j}^2)^{1/2}$

$$u_j = \begin{cases} v_j/\|v_j\|_2, & \text{if } \|v_j\|_2 \neq 0; \\ \in \{u : \|u\|_2 \leq 1\}, & \text{if } \|v_j\|_2 = 0. \end{cases} \tag{3.7}$$

In general, $n_1^{-1}X_{1j}^{\mathrm{T}}X_{1j} \neq n_2^{-1}X_{2j}^{\mathrm{T}}X_{2j}$, hence the block-update is not available in closed form and requires a line search (Barber and Drton, 2010). However, guided by the computational considerations as well as the ideas of standardized group lasso (Simon and Tibshirani, 2012), we pre-standardize X_1 and X_2 so that $n_1^{-1}\mathrm{diag}(X_1^{\mathrm{T}}X_1) = n_2^{-1}\mathrm{diag}(X_2^{\mathrm{T}}X_2) = 1_p$, and then perform the back-scaling of $\widehat{v}_1$, $\widehat{v}_2$. This ensures that the penalization of different variables is independent of their relative scales. Finally, we are ready to present the algorithm.

Define the residual vectors r_1, r_2 as

$$r_{1j} = n_1^{-1}X_{1j}^{\mathrm{T}}(1_{n_1} - \sum_{l=1}^{p} v_{1l}X_{1l}), \quad r_{2j} = n_2^{-1}X_{2j}^{\mathrm{T}}(-1_{n_2} - \sum_{l=1}^{p} v_{2l}X_{2l});$$

with $r_j = (r_{1j}, r_{2j})^{\mathrm{T}}$. From the optimality conditions, the equations for the jth block $v_j = (v_{1j}, v_{2j})^{\mathrm{T}}$ can be rewritten as

$$v_j = (1 - \lambda/\|v_j + r_j\|_2)_+ (v_j + r_j),$$

where $a_+ = \max(0, a)$. Starting with some initial value $V^{(0)}$, the block-coordinate descent algorithm proceeds by iterating the updates of v_1, v_2 with updates of residuals r_1, r_2 until convergence. Due to convexity of (3.6), the boundedness of the objective function from below, and the separability of the penalty with respect to block updates, the global optimum is finite and the algorithm is guaranteed to converge to the global optimum from any starting point (Tseng, 2001).

3.2.5 Connection with sparse linear discriminant analysis

We show that the sparse linear discriminant analysis can be viewed as a very special case of the proposed approach.

Proposition 2. *Consider the sparse discriminant analysis in Gaynanova et al. (2016) that finds the discriminant vector $\widetilde{v}(\lambda)$ for a given value of tuning parameter $\lambda > 0$. Define $c = (n_1/n)^{1/2} + (n_2/n)^{1/2}$. Under the constraint $(n/n_1)^{1/2}v_1 = (n/n_2)^{1/2}v_2$, the solution*

to (3.6) *satisfies*

$$(n/n_1)^{1/2}\widehat{v}_1(\lambda) = (n/n_2)^{1/2}\widehat{v}_2(\lambda) = c\,\widetilde{v}\,(\lambda/c)\,.$$

When v_1 and v_2 are restricted to be in the same direction, (3.6) gives the same solution as the sparse linear discriminant analysis up to scaling.

3.2.6 Connection with quadratic discriminant analysis

Let Y be a group indicator such that $P(Y = 1) = \pi_1$ and $P(Y = 2) = 1 - \pi_1 = \pi_2$, and consider $X|Y = g \sim N(\mu_g, \Sigma_g)$ $(g = 1, 2)$. The Bayes rule assigns a new observation with observed value $x \in \mathbb{R}^p$ to group one if and only if

$$\begin{aligned} &x^{\mathrm{T}}(\Sigma_2^{-1} - \Sigma_1^{-1})x - 2x^{\mathrm{T}}(\Sigma_2^{-1}\mu_2 - \Sigma_1^{-1}\mu_1) \\ &\qquad + \log\Big(|\Sigma_2|/|\Sigma_1|\Big) - \mu_1^{\mathrm{T}}\Sigma_1^{-1}\mu_1 + \mu_2^{\mathrm{T}}\Sigma_2^{-1}\mu_2 + 2\log(\pi_1/\pi_2) > 0. \end{aligned} \tag{3.8}$$

Consider centering x by the overall mean $\mathbb{E}(X) = \mu = \pi_1\mu_1 + \pi_2\mu_2$.

Proposition 3. *Let* $\delta = \mu_1 - \mu_2$. *The Bayes rule* (3.8) *can be written as*

$$\begin{aligned} &(x-\mu)^{\mathrm{T}}(\Sigma_2^{-1} - \Sigma_1^{-1})(x - \mu) + log\Big(|\Sigma_2|/|\Sigma_1|\Big) \\ &\quad + 2(x - \mu)^{\mathrm{T}}(\pi_1\Sigma_2^{-1}\delta + \pi_2\Sigma_1^{-1}\delta) + \pi_1^2\delta^{\mathrm{T}}\Sigma_2^{-1}\delta - \pi_2^2\delta^{\mathrm{T}}\Sigma_1^{-1}\delta + 2log(\pi_1/\pi_2) > 0. \end{aligned} \tag{3.9}$$

Consider the population version of the proposed discriminant analysis via projections, that is applying Bayes rule to $\Psi^{\mathrm{T}}X$ with $\Psi^{\mathrm{T}}X|Y = g \sim N(\Psi^{\mathrm{T}}\mu_g, \Psi^{\mathrm{T}}\Sigma_g\Psi)$ and $\Psi = [\psi_1,\ \psi_2] = [c_1\Sigma_1^{-1}\delta,\ c_2\Sigma_2^{-1}\delta]$, $c_1, c_2 \neq 0$.

Proposition 4. *Consider the population version of rule* (3.4)*, that is substituting* Ψ *for* $\widehat{V}$*,* Σ_g *for* S_g*,* μ_g *for* $\bar{x}_g$ *and* π_g *for* n_g/n*. A new observation with value* x *is assigned to group*

one if and only if

$$(x-\mu)^{\mathrm{T}}\Psi\Big\{(\Psi^{\mathrm{T}}\Sigma_2\Psi)^{-1}-(\Psi^{\mathrm{T}}\Sigma_1\Psi)^{-1}\Big\}\Psi^{\mathrm{T}}(x-\mu)+log\Big(|\Psi^{\mathrm{T}}\Sigma_2\Psi|/|\Psi^{\mathrm{T}}\Sigma_1\Psi|\Big) \\ +2(x-\mu)^{\mathrm{T}}(\pi_1\Sigma_2^{-1}\delta+\pi_2\Sigma_1^{-1}\delta)+\pi_1^2\delta^{\mathrm{T}}\Sigma_2^{-1}\delta-\pi_2^2\delta^{\mathrm{T}}\Sigma_1^{-1}\delta+2log(\pi_1/\pi_2)>0. \tag{3.10}$$

The only difference between the rules in Proposition 3 and 4 is on the first line, which involves the quadratic and the log terms. The linear terms and the remaining constant terms are identical. Therefore, rule (3.10) can be viewed as an approximation to rule (3.9).

While rule (3.10) is not the same as the Bayes rule, and therefore will lead to inferior performance on the population level, in Section 3.4 we see this relationship to be reversed when the corresponding regularized sample versions are considered and p is large relative to the sample size n. The main advantage of rule (3.10) comes from the significant reduction in the number of parameters to be estimated. Specifically, matrix Ψ has $p\times 2$ elements leading to $\mathcal{O}(p)$ parameters in rule (3.10). In contrast, the Bayes rule requires estimation of the $\Sigma_2^{-1}-\Sigma_1^{-1}$ leading to $\mathcal{O}(p^2)$ parameters in total.

3.3 Variable selection consistency in high-dimensional settings

We establish the variable selection consistency of estimator in (3.6) under the following assumptions.

Assumption 1 (Normality)**.** $X_i|Y_i=g\sim\mathcal{N}(\mu_g,\Sigma_g)$, $pr(Y_i=g)=\pi_g$ *for* $g=1,2$ *with* $0<\pi_{\min}\le\pi_1/\pi_2\le\pi_{\max}<1$.

Assumption 2 (Sparsity)**.** *Let* $\delta=\mu_1-\mu_2$, $A=\{i:(e_i^{\mathrm{T}}\Sigma_1^{-1}\delta)^2+(e_i^{\mathrm{T}}\Sigma_2^{-1}\delta)^2\neq 0\}$, $A^c=\{1,\ldots,p\}/A$ *and* $\mathrm{card}(A)=s$. *That is,* A *is the index set of nonzero variables in* $\Sigma_1^{-1}\delta$ *or* $\Sigma_2^{-1}\delta$.

Assumption 3 (Irrepresentability)**.** *There exist* $\alpha\in(0,1]$ *such that*

$$\max_{\substack{u_1,u_2\in\mathbb{R}^s\\ u_{1i}^2+u_{2i}^2\le 1\ \forall i}}\|\Sigma_{1A^cA}\Sigma_{1AA}^{-1}u_1,\Sigma_{2A^cA}\Sigma_{2AA}^{-1}u_2\|_{\infty,2}\le 1-\alpha.$$

Assumption 4. $0 < c \leq \lambda_{\min}(\Sigma_{gAA}) \leq \lambda_{\max}(\Sigma_{gAA}) \leq C$ *and* $e_j^{\mathrm{T}} \Sigma_g e_j \leq M$ *for* $j = 1, \ldots, p$ *and* $g = 1, 2$.

Assumption 1 is a standard assumption in the context of discriminant analysis (Mai et al., 2012; Kolar and Liu, 2015; Gaynanova and Kolar, 2015), and Assumptions 2–3 are typical in establishing variable selection consistency of penalized estimators in high-dimensional settings (Bach, 2008; Wainwright, 2009; Obozinski et al., 2011). We use Assumption 4 for convenience of treating the parameters depending on Σ_g as constants and presenting the rates in Theorems 1 and 2 through only n, p and s. We refer the reader to the Supplementary material for the more general statements of Theorems 1 and 2 without the use of Assumption 4. To prove variable selection consistency of estimator in (3.6), we use the primal-dual witness technique (Wainwright, 2009). First, we prove that under the appropriate scaling of the sample sizes, and sufficiently large value of the tuning parameter λ, the variables in A^c are set to zero with high probability. Let $\widehat{A} = \{i : \widehat{v}_{1i}^2 + \widehat{v}_{2i}^2 \neq 0\}$ denote the support of the solution to (3.6).

Theorem 1. *Let Assumptions 1–4 hold, the sample sizes satisfy* $\min_g n_g \gtrsim s log\{(p-s)\eta^{-1}\}$ *for some* $\eta \in (0, 1)$*, and the tuning parameter satisfy* $\lambda \gtrsim [log\{(p-s)\eta^{-1}\}/n]^{1/2}$*. Then* $pr(\widehat{A} \subseteq A) \geq 1 - \eta$.

Next, we show that under the additional assumption on the minimal signal strength defined as

$$\psi_{\min} = \min_{j \in A} \left\{ \pi_2^2 (e_j^{\mathrm{T}} \Sigma_1^{-1} \delta)^2 + \pi_1^2 (e_j^{\mathrm{T}} \Sigma_2^{-1} \delta)^2 \right\}^{1/2},$$

the true variables are nonzero with high probability leading to perfect recovery. In sparse linear models this assumption is often called β-min condition (Wainwright, 2009). According to Proposition 3, $\psi_{\min}$ can be interpreted as the smallest magnitude of the nonzero variables in the linear part of the Bayes quadratic discriminant rule.

Theorem 2. *Let the conditions of Theorem 1 hold and* $\psi_{\min} \gtrsim \lambda s^{1/2}(\max_g \delta_A^{\mathrm{T}} \Sigma_{gAA}^{-1} \delta_A \vee 1)$. *Then* $pr(\widehat{A} = A) \geq 1 - \eta$.

Theorem 2 reveals the advantage of using the group penalty in joint sparse estimation of ψ_1 and ψ_2. If variable j is nonzero in both ψ_1 and ψ_2, then it is sufficient to have large signal in only one of ψ_g for minimal signal strength condition to hold. In contrast, separate estimation via the lasso penalty will lead to the requirement of sufficiently large signal in both ψ_1 and ψ_2 simultaneously.

3.4 Empirical studies

3.4.1 Simulated data

We compare the misclassification error rates and variable selection performance of the following methods: (i) Sample QDA, rule (3.8) with plug-in estimates $\bar{x}_1$, $\bar{x}_2$, S_1, S_2; (ii) Sparse QDA of Le and Hastie (2014); (iii) Sparse QDA of Li and Shao (2015); (iv) Sparse QDA via ridge fusion (Price et al., 2014); (v) Logistic regression with pairwise interactions and lasso penalty on the vector of coefficients; (vi) Regularized discriminant analysis (Friedman, 1989); (vii) Sparse LDA (Mai et al., 2012; Gaynanova et al., 2016); (viii) Discriminant analysis via projections proposed in this paper, that is rule (3.4) with estimator from (3.6). The details of all methods' implementation together with tuning parameter selection criteria are described in Supplementary materials.

We fix the sample sizes $n_1 = n_2 = 100$, the dimension $p \in \{100, 500\}$, and the group means $\mu_1 = 0_p$ and $\mu_2 = (1_5, -1_5, 0_{p-10})$. We consider the following types of covariance structures:

1. Block-equicorrelation with block size $b \in \{10, 100\}$ and $\rho \in [0, 1]$:

$$\Sigma_g = \begin{pmatrix} \rho I_b + (1-\rho) 1_b 1_b^{\mathrm{T}} & 0 \\ 0 & I_{p-b} \end{pmatrix}.$$

Table 3.1: List of considered models for Σ_1 and Σ_2

Model	Σ_1	Σ_2
1	equicorrelation, $b = 100$, $\rho = 0.5$	equicorrelation, $b = 100$ $\rho = 0.5$
2	autocorrelation, $b = 100$ $\rho = 0.8$	equicorrelation, $b = 100$, $\rho = 0.5$
3	autocorrelation, $b = 10$, $\rho = 0.5$	equicorrelation, $b = 10$, $\rho = 0.8$
4	spiked, $b = 10$	spiked, $b = 10$ (q_1 and q_2 reversed)
5	spiked, $b = 100$	spiked, $b = 10$ (q_1 and q_2 reversed)
6	spiked, $b = 10$	equicorrelation, $b = 10$, $\rho = 0.8$
7	spiked, $b = 10$	equicorrelation, $b = 100$, $\rho = 0.3$
8	spiked, $b = 100$	equicorrelation, $b = 100$, $\rho = 0.3$

2. Block-autocorrelation with block size $b \in \{10, 100\}$ and $\rho \in [0, 1]$:

$$\Sigma_g = \{\Sigma_g\}_{i,j}, \quad \{\Sigma_g\}_{i,j} = \begin{cases} \rho^{|i-j|}, & (1 \leq i, j \leq b); \\ \mathbb{1}\{i = j\}, & (\text{otherwise}). \end{cases}$$

3. Spiked with parameters $q_1, q_2 \in \mathbb{R}^p$: $\Sigma_g = 30 q_1 q_1^{\mathrm{T}} + 2 q_2 q_2^{\mathrm{T}} + I$.

   (a) Block size $b = 10$: $q_1 = (1_5/\sqrt{5}, 0_{p-5})$, $q_2 = (0_{p-5}, 1_5/\sqrt{5}, 0_{p-10})$.

   (b) Block size $b = 100$: $q_1 = (1, \ldots, 100, 0_{p-100})^{\mathrm{T}}$ normalized so that $q_1^{\mathrm{T}} q_1 = 1$; $q_2 = (I - q_1 q_1^{\mathrm{T}})(100, \ldots, 1, 0_{p-100})^{\mathrm{T}}$ normalized so that $q_2^{\mathrm{T}} q_2 = 1$.

These structures are commonly used to assess the performance of discriminant analysis methods (Mai et al., 2012; Le and Hastie, 2014; Ramey et al., 2016). We use 8 combinations as described in Table 3.1, and fix the block sizes to make the Bayes error rate independent of p.

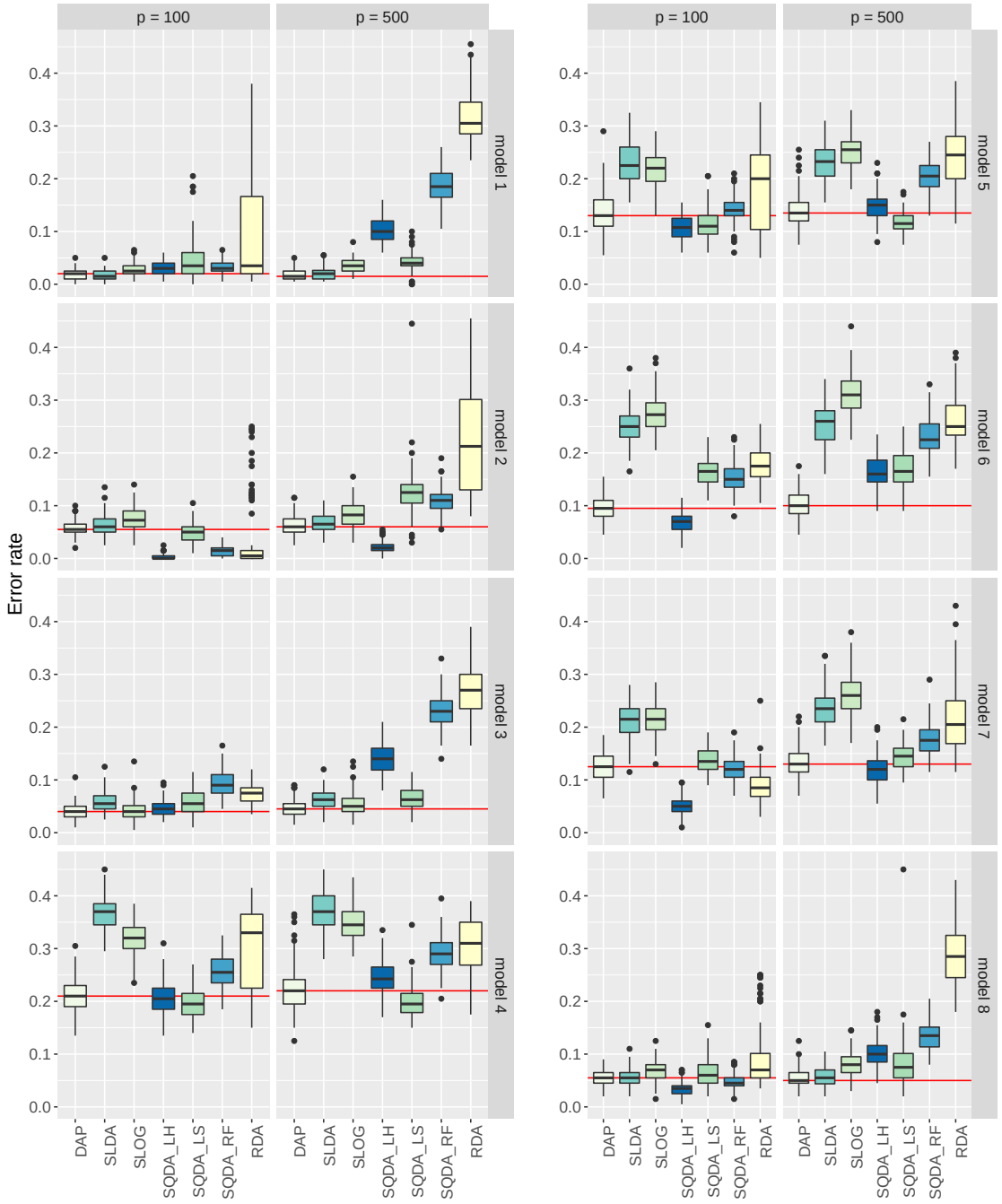

Figure 3.2: Misclassification error rates over 100 replications, the horizontal lines show the median errors of the proposed DAP, discriminant analysis via projections. SLDA: Sparse linear discriminant analysis; SLOG: Sparse logistic regression with interactions; SQDA_LH: Sparse QDA of Le and Hastie (2014); SQDA_LS: Sparse QDA of Li and Shao (2015); SQDA_RF: Sparse QDA via ridge fusion; RDA: Regularized discriminant analysis.

As expected, the sample QDA performs the worst, with misclassification error rates being larger than 40% consistently across all replications and models. Therefore, in Figure 3.2 we only present the rates for the other methods. First, we compare the proposed approach with sparse LDA. While in models 1, 2 and 8 they perform similarly, accounting for unequal covariance matrices results in drastic improvements on models 4–7. When comparing our approach to sparse QDA methods, the relative ranking often depends on p. For example, when $p = 100$, ridge fusion of Price et al. (2014) is better than our proposal on models 2 and 8, but is significantly worse on the same models when $p = 500$. Similarly, sparse QDA of Le and Hastie (2014) is significantly better than our proposal on models 6 and 8 when $p = 100$, but significantly worse on the same models when $p = 500$. This confirms that the proposed rule is well-suited to high-dimensional settings. Among the sparse QDA approaches, we find that the method of Li and Shao (2015) is most consistent across dimensions. In particular, it leads to better error rates on models 4 and 5 (2% difference in median error rates). Nevertheless, it still leads to significantly worse error rates on models 1, 2, 6 and 8. Finally, the proposed approach performs better than regularized discriminant analysis in all cases but model 2, $p = 100$, and performs as well or better than the sparse logistic regression in all scenarios.

Overall, we found that no method is universally the best in terms of error rates since the relative ranking depends on the particular model and the underlying dimension. This is consistent with previous research. In the words of Wu et al. (2018), "it is difficult to imagine that there could be a universally optimal discriminant analysis method for high-dimensional data. Almost every method can enjoy some advantages under certain circumstances." Nevertheless, three methods stand out as the best across all models and dimensions: our proposal and sparse QDA methods of Le and Hastie (2014) and Li and Shao (2015). Moreover, our proposal achieves comparable, and in certain scenarios significantly better, error rates than the best other methods in all the cases with $p = 500$ except model 2.

In summary, Figure 3.3 shows that the proposed discriminant analysis via projections

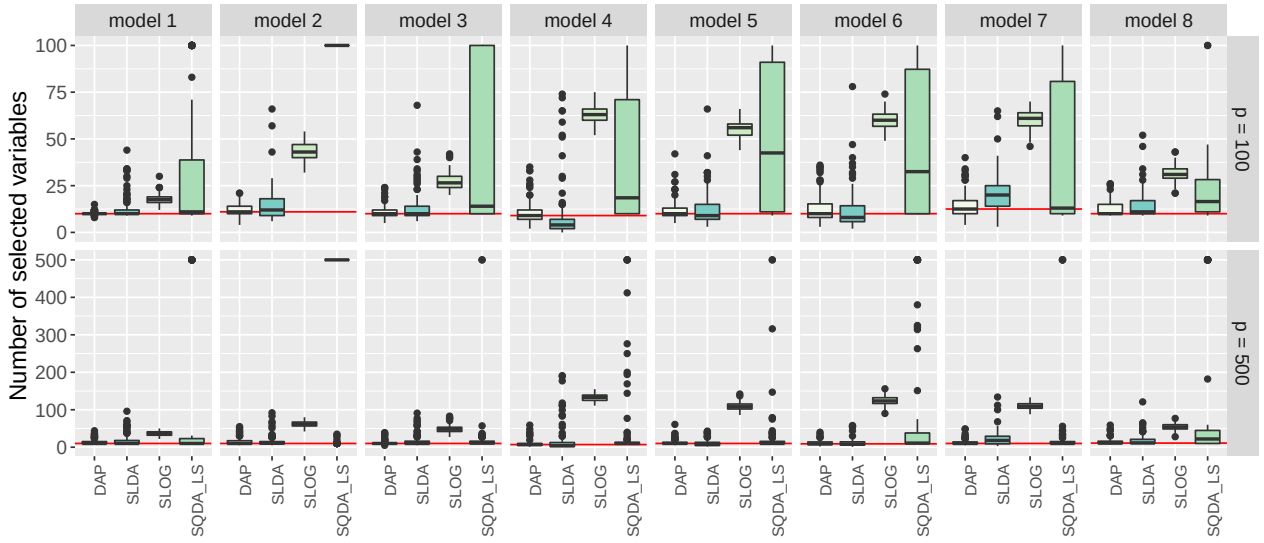

Figure 3.3: Number of selected variables over 100 replications, the horizontal lines indicate the median model sizes of proposed DAP, discriminant analysis via projections. RDA, SQDA_RF and SQDA_LH use all p variables, not shown. SLDA: Sparse linear discriminant analysis; SLOG: Sparse logistic regression with interactions; SQDA_LH: Sparse QDA of Le and Hastie (2014); SQDA_LS: Sparse QDA of Li and Shao (2015); SQDA_RF: Sparse QDA via ridge fusion; RDA: Regularized discriminant analysis.

significantly improved over sparse LDA method, and results in competitive, and often better, misclassification error rates than existing QDA proposals. The real advantages of our approach, however, become certain when comparing variable selection performance and computational speed. Figure 3.3 reveals that the proposed method consistently uses the sparsest model (less than 50 variables for most scenarios). In comparison, the methods of Le and Hastie (2014) and Price et al. (2014) always use all p variables, and are such much less interpretable.

We further compare the execution time of each method on a Linux machine with Intel Xeon X5560 @2.80 GHz. We define execution time as the full time for method's implementation: tuning parameter selection plus model fitting plus classification. We use one instance of model 8 with $p \in \{100, 300, 500\}$, and R package `microbenchmark` (Mersmann, 2015) with 10 evaluations of each expression. Table 3.2 shows that the execution times increase dra-

Table 3.2: Median time (seconds) over 10 replications to fully implement each classification method for one instance of model 8. DAP: Discriminant analysis via projections, proposed; SLDA: Sparse linear discriminant analysis; RDA: Regularized discriminant analysis; SLOG: Sparse logistic regression with interactions; SQDA_LH: Sparse QDA of Le and Hastie (2014); SQDA_RF: Sparse QDA via ridge fusion; SQDA_LS: Sparse QDA of Li and Shao (2015).

p	DAP	SLDA	RDA	SLOG	SQDA_LH	SQDA_RF	SQDA_LS
100	0.6	0.4	3.1	2.7	139.5	868.5	52.6
300	1.0	1.4	5.0	28.8	2071.9	11681.4	481.5
500	1.4	1.7	5.0	117.1	7282.2	45161.7	1791.4

matically with p for logistic regression with interactions and sparse QDA methods, whereas the times are quite consistent across dimensions for sparse LDA, RDA and our approach. Logistic regression is noticeably faster than sparse QDA methods mainly due to the difference in tuning parameter selection criterion: it uses BIC instead of cross-validation. Using cross-validation for logistic regression makes it too computationally demanding for the range of p we considered. Sparse LDA and the proposed method are the fastest, confirming that they are well-suited for the use on high-dimensional datasets in practice.

3.4.2 Benchmark datasets

We compare the proposed discriminant analysis via projections with competitors on three benchmark datasets: *chin* (Chin et al., 2006), *chowdary* (Chowdary et al., 2006), and *gravier* (Gravier et al., 2010). These datasets are commonly used to assess classification performance (Li and Ngom, 2013; Niu et al., 2015; Ramey et al., 2016), and are publicly available from the R package `datamicroarray` (Ramey, 2016). Below is the short description of each dataset.

chin: $p = 22,215$ gene expression profiles for $n = 118$ breast cancer samples with $n_1 = 75$ being ER-positive, and $n_2 = 43$ being ER-negative.

gravier: $p = 2,905$ gene expression profiles for $n = 168$ patients with small invasive ductal carcinomas without axillary lymph node involvement. The $n_1 = 111$ patients have

no event after a 5-year diagnosis (labelled good), and $n_2 = 57$ patients have early metastasis (labelled poor).

chowdary: $p = 22,283$ gene expression profiles from 32 matched breast tumour tissue pairs and 20 matched colon tissue pairs leading to $n = 104$ samples with $n_1 = 64$ and $n_2 = 40$.

We randomly split each dataset 100 times preserving the class proportions, and use 80% for training and 20% for testing. To reduce the computational cost associated with sparse quadratic discriminant analysis, we reduce the number of variables at each split by selecting the top $p = 1000$ variables with largest absolute value of the two-sample t-statistic on the training data, similar approach has been taken in Cai and Liu (2011). For fair comparison, we use the same set of 1000 variables for each of the methods. We do not consider sample quadratic discriminant analysis given its uniformly poor performance in Section 3.4.1. We also do not consider sparse logistic regression with interactions or ridge fusion due to computational issues when $p = 1000$ and their inferiority to other approaches in Section 3.4.1.

The results are shown in Figure 3.4. For *chin* dataset, the error rates are the worst for linear discriminant analysis confirming the importance of taking into account unequal covariance matrices, and are the same for other methods. At the same time, the proposed DAP rule selects significantly smaller model than the competitors (median model size is one). For *chowdary* dataset, the best performing method is RDA (Friedman, 1989), however the relative difference is only 1 misclassification on the test data. The smallest model again corresponds to proposed DAP. For *gravier* dataset, the best performing methods are ours and sparse QDA of Li and Shao (2015). Surprisingly, however, the method of Li and Shao (2015) results in no variable selection on these datasets, the model size is 1000 over almost all replications (not shown). We suspect that the poor variable selection performance may be due to the crudeness of bisection procedure for selecting the tuning parameters. In

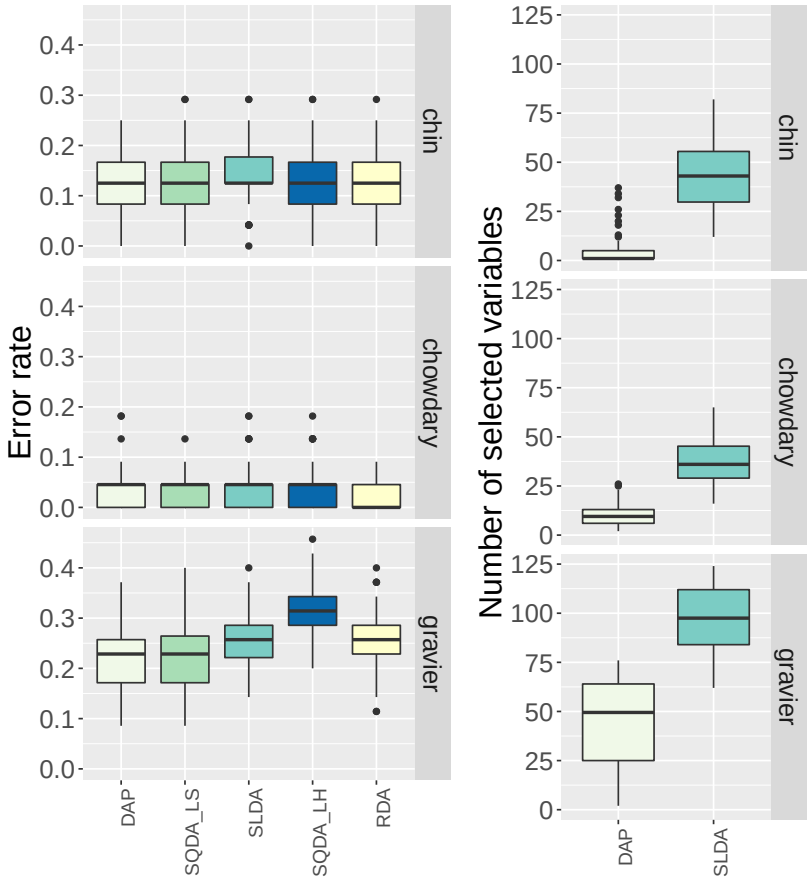

Figure 3.4: *Left:* Misclassification error rates over 100 splits. *Right:* Number of variables used in corresponding classification rules. DAP consistently selects the smallest model. SQDA_LS, SQDA_LH and RDA always use all $p = 1000$ variables, not shown. DAP: Discriminant analysis via projections, proposed method; SQDA_LS: Sparse QDA of Li and Shao (2015); SQDA_LH: Sparse QDA of Le and Hastie (2014); SLDA: Sparse linear discriminant analysis; RDA: Regularized discriminant analysis.

summary, the proposed approach, discriminant analysis via projections, consistently selects the smallest model, often using less than 20 variables to achieve the same or better error rates than alternative methods. We conclude that it exhibits excellent prediction accuracy with the smallest model complexity.

We further analyze the *chin* dataset using variable selection results of our approach. Figure 3.4 reveals that the median model size is one. This means that in most of the replications it is sufficient to look at the expression level of only one gene to achieve the

same misclassification error rate as the other methods. We investigate whether the same gene is selected at each replication, and find that estrogen receptor 1 gene ESR1 is selected in 97 out of 100 cases. Our finding confirms previous studies on a strong link between ESR1 gene and estrogen receptor protein expression in breast cancer patients (Holst et al., 2007; Laenkholm et al., 2012; Iwamoto et al., 2012). We refer the reader to Holst (2016) for the review on the importance of ESR1 gene amplification in breast cancer. The gene with the second highest frequency of selection, 26 out of 100 cases, is LPIN1, which is also found to be differentially expressed in ER positive and negative patients in previous studies (Chen et al., 2008). The relatively low selection frequency of LPIN1 is due to the median model size one, which leads to only ESR1 being selected and no other gene. While the strong link between ER protein expression status and ESR1 gene is not surprising, unlike the previous studies we did not focus on the ESR1 gene in advance. We consider all 22 thousand genes, and let our method determine that ESR1 is crucial for ER status of breast cancer. We want to emphasize that this insight is not possible with other approaches we tried. Regularized discriminant analysis of Friedman (1989) and sparse QDA by Le and Hastie (2014) use all 1000 variables, hence can not be directly used for identifying important genes. Sparse LDA selects a smaller number of genes, but it has worse misclassification error rate and the median model size is still 45 variables, significantly larger than the number of variables used by our approach.

3.5 Discussion

In this work we propose a new rule for high-dimensional classification in the case of unequal covariance matrices. While the proposed approach in general differs from the Bayes rule on the population level, we show that the nonzero variables in our rule correspond to nonzero variables in the linear part of the Bayes quadratic rule. This connection combined with computational efficiency of our approach suggests that one can potentially use our method as a variable screening tool. Indeed, the empirical studies in Section 3.4.1 indicate

that the performance of full quadratic methods deteriorates significantly with increase in p, however for small p they are computationally feasible and may lead to better error rates. We have not explored the screening properties of our approach in this work, but leave it for future investigation.

We focus on the two-group classification setting, however extending the methodology to the multi-group setting will likely lead to even further computational gains. One of the main challenges in the multi-group case is the likely rank degeneracy of the matrix of discriminant vectors when the number of groups is large. Performing simultaneous low-rank and sparse estimation of the matrix of discriminant vectors in the multi-group case is an interesting direction for future research.

3.6 Supplementary material

The **Supplementary Material**includes implementation details of Section 3.4, proofs of propositions and main theorems in Sections 3.2 and 3.3.

3.6.1 Implementation details

In this section we describe implementation details for the methods considered in Section 3.4.1. We use R package `JGL` (Danaher, 2013) to implement sparse QDA of Le and Hastie (2014); R package `MGSDA` (Gaynanova, 2016) to implement sparse LDA (Mai et al., 2012; Gaynanova et al., 2016); R package `grpreg` (Breheny and Huang, 2015) to implement logistic regression with pairwise interactions and lasso penalty on the vector of coefficients; R package `RidgeFusion` (Price, 2014) to implement ridge fusion for joint estimation of precision matrices (Price et al., 2014); R package `sparsediscrim` to implement regularized discriminant analysis (Friedman, 1989). We found no available R code for sparse QDA of Li and Shao (2015), and implemented the method ourselves. We use R package `DAP`(Wang and Gaynanova, 2018) to implement the proposed discriminant analysis via projections.

For logistic regression, we use BIC option in the `grpreg` to select the tuning parameter.

For ridge fusion, we use the automatic selection in `RidgeFusion` with 5 folds. For Li and Shao (2015), we use the bisection procedure proposed in their paper with the maximal interval length set to 0.05. For all other methods, we use 5-fold cross-validation to minimize misclassification error rate.

3.6.2 Proofs of propositions

Proof of Proposition 2. From Gaynanova et al. (2016), $\widetilde{v}(\lambda) = \text{argmin}_v L_1(v, \lambda)$, where

$$L_1(v, \lambda) = v^{\mathrm{T}} \left(n_1 S_1 + n_2 S_2\right) v/(2n) + n_1 n_2 d^{\mathrm{T}} v v^{\mathrm{T}} d/(2n^2) - n_1^{1/2} n_2^{1/2} d^{\mathrm{T}} v/n + \lambda \|v\|_1.$$

From (3.6), $\{\widehat{v}_1(\lambda), \widehat{v}_2(\lambda)\} = \text{argmin}_{v_1, v_2} L_2(v_1, v_2, \lambda)$, where

$$\begin{aligned} L_2(v_1, v_2, \lambda) = {} & (v_1^{\mathrm{T}} S_1 v_1 + v_2^{\mathrm{T}} S_2 v_2)/2 \\ & + \left(n_2 n^{-1} d^{\mathrm{T}} v_1 - 1\right)^2 /2 + \left(n_1 n^{-1} d^{\mathrm{T}} v_2 - 1\right)^2 /2 + \lambda \sum_{j=1}^{p} (v_{1j}^2 + v_{2j}^2)^{1/2}. \end{aligned}$$

Under the constraint $(n/n_1)^{1/2} v_1 = (n/n_2)^{1/2} v_2 = v$, this leads to $\widehat{v}(\lambda) = \text{argmin}_v L_2(v, \lambda)$, where using $c = (n_1/n)^{1/2} + (n_2/n)^{1/2}$,

$$L_2(v, \lambda) = v^{\mathrm{T}} \left(n_1 S_1 + n_2 S_2\right) v/(2n) + n_1 n_2 d^{\mathrm{T}} v v^{\mathrm{T}} d/(2n^2) - n_1^{1/2} n_2^{1/2} c d^{\mathrm{T}} v/n + \lambda \|v\|_1.$$

Furthermore,

$$\begin{aligned} & L_1(v/c, \lambda/c) \\ & \quad = c^{-2} \left\{ v^{\mathrm{T}} \left(n_1 S_1 + n_2 S_2\right) v/(2n) + n_1 n_2 d^{\mathrm{T}} v v^{\mathrm{T}} d/(2n^2) - n_1^{1/2} n_2^{1/2} c d^{\mathrm{T}} v/n + \lambda \|v\|_1 \right\} \\ & \quad = c^{-2} L_2(v, \lambda). \end{aligned}$$

Since for any $c > 0$, $\text{argmin}_x f(x/c) = c\{\text{argmin}_x f(x)\}$, it follows that $c\widetilde{v}(\lambda/c) = \widehat{v}(\lambda)$. □

Proof of Proposition 3. Since $\log(|\Sigma_2|/|\Sigma_1|)$ and $2\log(\pi_1/\pi_2)$ are present in both rules, it

remains to show the equivalence of the quadratic term, the linear term and the remaining constants. Substituting $x = x - \mu + \mu$ in the Bayes rule (3.8) leads to

$$\begin{aligned} x^{\mathrm{T}}(\Sigma_2^{-1} - \Sigma_1^{-1})x &= (x-\mu)^{\mathrm{T}}(\Sigma_2^{-1} - \Sigma_1^{-1})(x-\mu) + 2(x-\mu)^{\mathrm{T}}(\Sigma_2^{-1} - \Sigma_1^{-1})\mu \\ &\quad + \mu^{\mathrm{T}}(\Sigma_2^{-1} - \Sigma_1^{-1})\mu, \\ -2x^{\mathrm{T}}(\Sigma_2^{-1}\mu_2 - \Sigma_1^{-1}\mu_1) &= -2(x-\mu)^{\mathrm{T}}(\Sigma_2^{-1}\mu_2 - \Sigma_1^{-1}\mu_1) - 2\mu^{\mathrm{T}}(\Sigma_2^{-1}\mu_2 - \Sigma_1^{-1}\mu_1). \end{aligned}$$

From the above, the quadratic term in $(x-\mu)$ is the same as stated in the Proposition, hence it remains to consider the linear terms and the constants.

Consider the linear terms in $(x-\mu)$ from the above. Recall that $\delta = \mu_1 - \mu_2$, therefore

$$\begin{aligned} &2(x-\mu)^{\mathrm{T}}(\Sigma_2^{-1} - \Sigma_1^{-1})\mu - 2(x-\mu)^{\mathrm{T}}(\Sigma_2^{-1}\mu_2 - \Sigma_1^{-1}\mu_1) \\ &= 2(x-\mu)^{\mathrm{T}}\{\Sigma_2^{-1}(\mu-\mu_2) - \Sigma_1^{-1}(\mu-\mu_1)\} \\ &= 2(x-\mu)^{\mathrm{T}}(\pi_1\Sigma_2^{-1}\delta + \pi_2\Sigma_1^{-1}\delta), \end{aligned}$$

which is the same as the linear term in the statement of the proposition.

Finally, we complete the proof by showing the equivalence of remaining constants.

$$\begin{aligned} &\mu^{\mathrm{T}}(\Sigma_2^{-1} - \Sigma_1^{-1})\mu - 2\mu^{\mathrm{T}}(\Sigma_2^{-1}\mu_2 - \Sigma_2^{-1}\mu_1) - \mu_1^{\mathrm{T}}\Sigma_1^{-1}\mu_1 + \mu_2^{\mathrm{T}}\Sigma_2^{-1}\mu_2 \\ &= (\mu^{\mathrm{T}}\Sigma_2^{-1}\mu - 2\mu^{\mathrm{T}}\Sigma_2^{-1}\mu_2 + \mu_2^{\mathrm{T}}\Sigma_2^{-1}\mu_2) - (\mu^{\mathrm{T}}\Sigma_1^{-1}\mu - 2\mu^{\mathrm{T}}\Sigma_1^{-1}\mu_1 + \mu_1^{\mathrm{T}}\Sigma_1^{-1}\mu_1) \\ &= \pi_1^2\delta^{\mathrm{T}}\Sigma_2^{-1}\delta - \pi_2^2\delta^{\mathrm{T}}\Sigma_1^{-1}\delta. \end{aligned}$$

□

Proof of Proposition 4. Since $\Psi^{\mathrm{T}}X|Y = g \sim N(\Psi^{\mathrm{T}}\mu_g, \Psi^{\mathrm{T}}\Sigma_g\Psi)$, from Proposition 3 the

Bayes rule applied to $\Psi^{\mathrm{T}}x$ has the form

$$\begin{aligned}(x-\mu)^{\mathrm{T}}\Psi\Big\{(\Psi^{\mathrm{T}}\Sigma_2\Psi)^{-1}-(\Psi^{\mathrm{T}}\Sigma_1\Psi)^{-1}\Big\}\Psi^{\mathrm{T}}(x-\mu)+\log\Big(|\Psi^{\mathrm{T}}\Sigma_2\Psi|/|\Psi^{\mathrm{T}}\Sigma_1\Psi|\Big)\\ +2(x-\mu)^{\mathrm{T}}\Big\{\pi_1\Psi(\Psi^{\mathrm{T}}\Sigma_2\Psi)^{-1}\Psi^{\mathrm{T}}\delta+\pi_2\Psi(\Psi^{\mathrm{T}}\Sigma_1\Psi)^{-1}\Psi^{\mathrm{T}}\delta\Big\}\\ +\pi_1^2\delta^{\mathrm{T}}\Psi(\Psi^{\mathrm{T}}\Sigma_2\Psi)^{-1}\Psi^{\mathrm{T}}\delta-\pi_2^2\delta^{\mathrm{T}}\Psi(\Psi^{\mathrm{T}}\Sigma_1\Psi)^{-1}\Psi^{\mathrm{T}}\delta+2\log(\pi_1/\pi_2)>0.\end{aligned} \tag{3.11}$$

Since

$$(\Psi^{\mathrm{T}}\Sigma_1\Psi)^{-1}=\frac{1}{\psi_1^{\mathrm{T}}\Sigma_1\psi_1\psi_2^{\mathrm{T}}\Sigma_1\psi_2-(\psi_1^{\mathrm{T}}\Sigma_1\psi_2)^2}\begin{pmatrix}\psi_2^{\mathrm{T}}\Sigma_1\psi_2 & -\psi_1^{\mathrm{T}}\Sigma_1\psi_2\\ -\psi_2^{\mathrm{T}}\Sigma_1\psi_1 & \psi_1^{\mathrm{T}}\Sigma_1\psi_1\end{pmatrix}.$$

it follows that

$$\Psi(\Psi^{\mathrm{T}}\Sigma_1\Psi)^{-1}\Psi^{\mathrm{T}}=\frac{\psi_1\psi_2^{\mathrm{T}}\Sigma_1\psi_2\psi_1^{\mathrm{T}}-\psi_2\psi_2^{\mathrm{T}}\Sigma_1\psi_1\psi_1^{\mathrm{T}}+\psi_2\psi_1^{\mathrm{T}}\Sigma_1\psi_1\psi_2^{\mathrm{T}}-\psi_1\psi_1^{\mathrm{T}}\Sigma_1\psi_2\psi_2^{\mathrm{T}}}{\psi_1^{\mathrm{T}}\Sigma_1\psi_1\psi_2^{\mathrm{T}}\Sigma_1\psi_2-(\psi_1^{\mathrm{T}}\Sigma_1\psi_2)^2}.$$

Recall that $\psi_1=c_1\Sigma_1^{-1}\delta$, and substituting $\delta=c_1^{-1}\Sigma_1\psi_1$ into the above equation leads to

$$\Psi(\Psi^{\mathrm{T}}\Sigma_1\Psi)^{-1}\Psi\delta=\frac{c_1^{-1}\psi_1\Big\{\psi_2^{\mathrm{T}}\Sigma_1\psi_2\psi_1^{\mathrm{T}}\Sigma_1\psi_1-(\psi_1^{\mathrm{T}}\Sigma_1\psi_2)^2\Big\}}{\psi_1^{\mathrm{T}}\Sigma_1\psi_1\psi_2^{\mathrm{T}}\Sigma_1\psi_2-(\psi_1^{\mathrm{T}}\Sigma_1\psi_2)^2}=c_1^{-1}\psi_1=\Sigma_1^{-1}\delta.$$

Similarly, $\Psi(\Psi^{\mathrm{T}}\Sigma_2\Psi)^{-1}\Psi^{\mathrm{T}}\delta=\Sigma_2^{-1}\delta$. Substituting these into (3.11) completes the proof. □

3.6.3 Proofs of main theorems

We will use the following quantities throughout the proofs:

$$\gamma=1+\max\left(\pi_1\pi_2^{-1}\|\Sigma_{1AA}^{-1/2}\Sigma_{2AA}\Sigma_{1AA}^{-1/2}\|_2,\pi_2\pi_1^{-1}\|\Sigma_{2AA}^{-1/2}\Sigma_{1AA}\Sigma_{2AA}^{-1/2}\|_2\right), \tag{3.12}$$

$$\Sigma_{gA^cA^c:A} = \Sigma_{gA^cA^c} - \Sigma_{gA^cA}\Sigma_{gAA}^{-1}\Sigma_{gA^cA} \quad (g = 1, 2),$$

$$\begin{aligned}\Sigma_{d_1} &= \Sigma_{1A^cA^c:A} + \pi_1\pi_2^{-1}\Big(\Sigma_{2A^cA^c} + \Sigma_{1A^cA}\Sigma_{1AA}^{-1}\Sigma_{2AA}\Sigma_{1AA}^{-1}\Sigma_{1AA^c} \\ &\qquad - \Sigma_{1A^cA}\Sigma_{1AA}^{-1}\Sigma_{2AA^c} - \Sigma_{2A^cA}\Sigma_{1AA}^{-1}\Sigma_{1AA^c}\Big), \qquad (3.13)\\ \Sigma_{d_2} &= \Sigma_{2A^cA^c:A} + \pi_2\pi_1^{-1}\Big(\Sigma_{1A^cA^c} + \Sigma_{2A^cA}\Sigma_{2AA}^{-1}\Sigma_{1AA}\Sigma_{2AA}^{-1}\Sigma_{2AA^c} \\ &\qquad - \Sigma_{2A^cA}\Sigma_{2AA}^{-1}\Sigma_{1AA^c} - \Sigma_{1A^cA}\Sigma_{2AA}^{-1}\Sigma_{2AA^c}\Big).\end{aligned}$$

The quantities in (3.13) can be viewed as conditional variance terms, their origin is made precise in Lemma 2. Let $\sigma^2_{gjj:A} = e_j^{\mathrm{T}}\Sigma_{gA^cA^c:A}e_j$ and $\sigma^2_{jdg} = e_j^{\mathrm{T}}\Sigma_{d_g}e_j$ be the diagonal elements of corresponding matrices. Under Assumption 4, $\sigma_{gjj:A}$, σ_{jdg} and γ can be treated as constants.

We define the oracle $(\tilde{v}_{1A}, \tilde{v}_{2A})$ as the solution to

$$\underset{v_1, v_2 \in \mathbb{R}^s}{\text{minimize}} \Big\{ n_1^{-1}\|X_{1A}v_1 - 1_{n_1}\|_2^2/2 + n_2^{-1}\|X_{2A}v_2 + 1_{n_2}\|_2^2/2 + \lambda\sum_{j=1}^{s}(v_{1j}^2 + v_{2j}^2)^{1/2}\Big\}, \quad (3.14)$$

and let $\tilde{u}_A = (\tilde{u}_{1A}, \tilde{u}_{2A})$ be the subgradient of $\sum_{j=1}^s (v_{1j}^2 + v_{2j}^2)^{1/2}$ evaluated at $(\tilde{v}_{1A}, \tilde{v}_{2A})$

$$\tilde{u}_{Aj} = \begin{cases} \tilde{v}_{Aj}/\|\tilde{v}_{Aj}\|_2, & \text{if } \|\tilde{v}_{Aj}\|_2 \neq 0; \\ \in \{u : \|u\|_2 \leq 1\}, & \text{if } \|\tilde{v}_{Aj}\|_2 = 0. \end{cases} \quad (3.15)$$

Theorem 3 (Equivalent to Theorem 1)**.** *Let Assumptions 1–3 hold. Let the sample sizes satisfy*

$$\min(n_1, n_2) \gtrsim \max_{g=1,2} \|\Sigma_{gAA}^{-1}\|_2 \max_{g=1,2;\ j\in A^c} (\sigma^2_{gjj:A} \vee \sigma^2_{jdg}) s log\{(p-s)\eta^{-1}\},$$

for some $\eta \in (0, 1)$, and the tuning parameter satisfy

$$\lambda \gtrsim \max_{g=1,2;\ j\in A^c} (\sigma^2_{gjj:A} \vee \sigma^2_{jdg}) \Big[n^{-1} log\{(p-s)\eta^{-1}\} \Big]^{1/2}.$$

Then $pr(\widehat{A} \subseteq A) \geq 1 - \eta$.

Proof. Using the results of Section 3.2.3,

$$[\widehat{v}_1\ \widehat{v}_2] = \underset{v_1\in\mathbb{R}^p, v_2\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \widehat{L}_{\psi_1}(v_1) + \widehat{L}_{\psi_2}(v_2) + \lambda \sum_{j=1}^{p} (v_{1j}^2 + v_{2j}^2)^{1/2} \right\},$$

$$2\{\widehat{L}_{\psi_1}(v_1) + \widehat{L}_{\psi_2}(v_2)\} = v_1^{\mathrm{T}} S_1 v_1 + v_2^{\mathrm{T}} S_2 v_2 + \left(n^{-1} n_2 d^{\mathrm{T}} v_1 - 1\right)^2 + \left(n^{-1} n_1 d^{\mathrm{T}} v_2 - 1\right)^2.$$

Let $\rho_1 = n_1/n$ and $\rho_2 = n_2/n$. The optimality conditions (Boyd and Vandenberghe, 2004, Chapter 5) lead to

$$\begin{aligned}
(S_{1AA} + \rho_2^2 d_A d_A^{\mathrm{T}})\widehat{v}_{1A} + (S_{1AA^c} + \rho_2^2 d_A d_{A^c}^{\mathrm{T}})\widehat{v}_{1A^c} - \rho_2 d_A &= -\lambda u_{1A}, \\
(S_{2AA} + \rho_1^2 d_A d_A^{\mathrm{T}})\widehat{v}_{2A} + (S_{2AA^c} + \rho_1^2 d_A d_{A^c}^{\mathrm{T}})\widehat{v}_{2A^c} - \rho_1 d_A &= -\lambda u_{2A}, \\
(S_{1A^cA} + \rho_2^2 d_{A^c} d_A^{\mathrm{T}})\widehat{v}_{1A} + (S_{1A^cA^c} + \rho_2^2 d_{A^c} d_{A^c}^{\mathrm{T}})\widehat{v}_{1A^c} - \rho_2 d_{A^c} &= -\lambda u_{1A^c}, \\
(S_{2A^cA} + \rho_1^2 d_{A^c} d_A^{\mathrm{T}})\widehat{v}_{2A} + (S_{2A^cA^c} + \rho_1^2 d_{A^c} d_{A^c}^{\mathrm{T}})\widehat{v}_{2A^c} - \rho_2 d_{A^c} &= -\lambda u_{2A^c},
\end{aligned}$$

where u is defined in (3.7). Consider $\widehat{v}_1 = (\tilde{v}_{1A}, 0_{p-s})$, $\widehat{v}_2 = (\tilde{v}_{2A}, 0_{p-s})$, where $\tilde{v}_{1A}$, $\tilde{v}_{2A}$ are the solutions to the oracle problem (3.14). From the above optimality conditions, it is sufficient to have

$$\left\| (S_{1A^cA} + \rho_2^2 d_{A^c} d_A^{\mathrm{T}})\tilde{v}_{1A} - \rho_2 d_{A^c}\ ,\ (S_{2A^cA} + \rho_1^2 d_{A^c} d_A^{\mathrm{T}})\tilde{v}_{2A} - \rho_1 d_{A^c} \right\|_{\infty,2} < \lambda$$

for $\widehat{V} = [\widehat{v}_1\ \widehat{v}_2]$ to be the solution to (3.6), which leads to $\widehat{A} \subseteq A$. We next show that the above inequality holds with high probability under the stated conditions.

Using the form of $\widetilde{v}_{1A}$ (Theorem 5) and Sherman–Morrison identity,

$$
\begin{aligned}
&(S_{1A^cA} + \rho_2^2 d_{A^c} d_A^{\mathrm{T}})\tilde{v}_{1A} - \rho_2 d_{A^c} \\
&\quad = S_{1A^cA}\rho_2 S_{1AA}^{-1} d_A (1 + \rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A)^{-1} + \rho_2^2 d_{A^c} d_A \rho_2^{\mathrm{T}} S_{1AA}^{-1} d_A (1 + \rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A)^{-1} \\
&\qquad - \lambda S_{1A^cA} \left(S_{1AA} + \rho_2^2 d_A d_A^{\mathrm{T}}\right)^{-1} \tilde{u}_{1A} - \lambda \rho_2^2 d_{A^c} d_A^{\mathrm{T}} \left(S_{1AA} + \rho_2^2 d_A d_A^{\mathrm{T}}\right)^{-1} \tilde{u}_{1A} - \rho_2 d_{A^c} \\
&\quad = \rho_2 \left(S_{1A^cA} S_{1AA}^{-1} d_A - d_{A^c}\right) (1 + \rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A)^{-1} - \lambda S_{1A^cA} S_{1AA}^{-1} \tilde{u}_{1A} \\
&\qquad + \lambda \rho_2^2 S_{1A^cA} S_{1AA}^{-1} d_A d_A^{\mathrm{T}} S_{1AA}^{-1} \tilde{u}_{1A} (1 + \rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A)^{-1} \\
&\qquad - \lambda \rho_2^2 d_{A^c} d_A^{\mathrm{T}} S_{1AA}^{-1} \tilde{u}_{1A} (1 + \rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A)^{-1} \\
&\quad = \rho_2 \left(S_{1A^cA} S_{1AA}^{-1} d_A - d_{A^c}\right) (1 + \rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A)^{-1} - \lambda S_{1A^cA} S_{1AA}^{-1} \tilde{u}_{1A} \\
&\qquad + \rho_2^2 \lambda (S_{1A^CA} S_{1AA}^{-1} d_A - d_{A^c}) d_A^{\mathrm{T}} S_{1AA}^{-1} \tilde{u}_{1A} (1 + \rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A)^{-1}.
\end{aligned}
$$

Using normality, there exist $U_1 \in \mathbb{R}^{p \times (n_1 - 1)}$ with columns $u_{1,i} \sim \mathcal{N}(0, \Sigma_1)$ such that $(n_1 - 1)S_1 = U_1 U_1^{\mathrm{T}}$. Let $E_{d1} = d_{A^c} - \Sigma_{1A^cA}\Sigma_{1AA}^{-1} d_A$, $E_{U1} = U_{1A^c} - \Sigma_{1A^cA}\Sigma_{1AA}^{-1} U_{1A}$. Then

$$
\begin{aligned}
S_{1A^cA} S_{1AA}^{-1} &= (n_1 - 1)^{-1} U_{1A^c} U_{1A}^{\mathrm{T}} S_{1AA}^{-1} \\
&= (n_1 - 1)^{-1} E_{U1} U_{1A}^{\mathrm{T}} S_{1AA}^{-1} + (n_1 - 1)^{-1} \Sigma_{1A^cA} \Sigma_{1AA}^{-1} U_{1A} U_{1A}^{\mathrm{T}} S_{1,AA}^{-1} \\
&= \Sigma_{1A^cA} \Sigma_{1AA}^{-1} + (n_1 - 1)^{-1} E_{U1} U_{1A}^{\mathrm{T}} S_{1AA}^{-1},
\end{aligned}
$$

and $S_{1A^cA} S_{1AA}^{-1} d_A - d_{A^c} = (n_1 - 1)^{-1} E_{U1} U_{1A}^{\mathrm{T}} S_{1AA}^{-1} d_A - E_{d1}$. Combining the above two displays

gives

$$
\begin{aligned}
&(S_{1A^cA} + \rho_2^2 d_{A^c} d_A^{\mathrm{T}})\tilde{v}_{1A} - \rho_2 d_{A^c} \\
&= -\lambda \Sigma_{1A^cA} \Sigma_{1AA}^{-1} \tilde{u}_{1A} - \lambda (n_1 - 1)^{-1} E_{U1} U_{1A}^{\mathrm{T}} S_{1AA}^{-1} \tilde{u}_{1A} \\
&\quad + (n_1 - 1)^{-1} E_{U1} U_{1A}^{\mathrm{T}} S_{1AA}^{-1} d_A \rho_2 (1 + \rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A)^{-1} - E_{d1} \rho_2 (1 + \rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A)^{-1} \\
&\quad + \lambda (n_1 - 1)^{-1} E_{U1} U_{1A}^{\mathrm{T}} S_{1AA}^{-1} d_A \rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} \tilde{u}_{1A} (1 + \rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A)^{-1} \\
&\quad - \lambda E_{d1} \rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} \tilde{u}_{1A} (1 + \rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A)^{-1} \\
&= -\lambda \Sigma_{1A^cA} \Sigma_{1AA}^{-1} \tilde{u}_{1A} + (n_1 - 1)^{-1} E_{U1} U_{1A}^{\mathrm{T}} S_{1AA}^{-1} d_A \rho_2 (1 + \rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A)^{-1} \\
&\quad - E_{d1} \rho_2 (1 + \rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A)^{-1} - \lambda E_{d1} \rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} \tilde{u}_{1A} (1 + \rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A)^{-1} \\
&\quad - \lambda (n_1 - 1)^{-1} E_{U1} U_{1A}^{\mathrm{T}} S_{1AA}^{-1} (I + \rho_2^2 d_A d_A^{\mathrm{T}} S_{1AA}^{-1})^{-1} \tilde{u}_{1A}.
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
&(S_{2A^cA} + \rho_1^2 d_{A^c} d_A^{\mathrm{T}})\tilde{v}_{2A} - \rho_1 d_{A^c} \\
&\qquad = -\lambda \Sigma_{2A^cA} \Sigma_{2AA}^{-1} \tilde{u}_{2A} + (n_2 - 1)^{-1} E_{U2} U_{2A}^{\mathrm{T}} S_{2AA}^{-1} d_A \rho_1 (1 + \rho_1^2 d_A^{\mathrm{T}} S_{2AA}^{-1} d_A)^{-1} \\
&\qquad\quad - E_{d2} \rho_1 (1 + \rho_1^2 d_A^{\mathrm{T}} S_{2AA}^{-1} d_A)^{-1} - \lambda E_{d2} \rho_1^2 d_A^{\mathrm{T}} S_{2AA}^{-1} \tilde{u}_{2A} (1 + \rho_1^2 d_A^{\mathrm{T}} S_{2AA}^{-1} d_A)^{-1} \\
&\qquad\quad - \lambda (n_2 - 1)^{-1} E_{U2} U_{2A}^{\mathrm{T}} S_{2AA}^{-1} (I + \rho_1^2 d_A d_A^{\mathrm{T}} S_{2AA}^{-1})^{-1} \tilde{u}_{2A}.
\end{aligned}
$$

Therefore, using triangle inequality,

$$
\begin{aligned}
&\left\| (S_{1A^cA} + \rho_2^2 d_{A^c} d_A^{\mathrm{T}})\tilde{v}_{1A} - \rho_2 d_{A^c} \ , \ (S_{2A^cA} + \rho_1^2 d_{A^c} d_A^{\mathrm{T}})\tilde{v}_{2A} - \rho_1 d_{A^c} \right\|_{\infty,2} \\
&\qquad \leq \lambda \| \Sigma_{1A^cA} \Sigma_{1AA}^{-1} \tilde{u}_{1A}, \Sigma_{2A^cA} \Sigma_{2AA}^{-1} \tilde{u}_{2A} \|_{\infty,2} + I_1 + I_2 + I_3 + I_4,
\end{aligned}
$$

where

$$
\begin{aligned}
I_1 &= \|\rho_2(1+\rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A)^{-1} E_{d1}, \rho_1(1+\rho_1^2 d_A^{\mathrm{T}} S_{2AA}^{-1} d_A)^{-1} E_{d2}\|_{\infty,2}, \\
I_2 &= \left\| (n_1-1)^{-1} \frac{\rho_2 E_{U1} U_{1A}^{\mathrm{T}} S_{1AA}^{-1} d_A}{1+\rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A}, (n_2-1)^{-1} \frac{\rho_1 E_{U2} U_{2A}^{\mathrm{T}} S_{2AA}^{-1} d_A}{1+\rho_1^2 d_A^{\mathrm{T}} S_{2AA}^{-1} d_A} \right\|_{\infty,2}, \\
I_3 &= \left\| \frac{E_{U1} U_{1A}^{\mathrm{T}} S_{1AA}^{-1}}{n_1-1} (I+\rho_2^2 d_A d_A^{\mathrm{T}} S_{1AA}^{-1})^{-1} \tilde{u}_{1A}, \frac{E_{U2} U_{2A}^{\mathrm{T}} S_{2AA}^{-1}}{n_2-1} (I+\rho_1^2 d_A d_A^{\mathrm{T}} S_{2AA}^{-1})^{-1} \tilde{u}_{2A} \right\|_{\infty,2}, \\
I_4 &= \left\| \frac{\rho_2^2}{1+\rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A} E_{d1} d_A^{\mathrm{T}} S_{1AA}^{-1} \tilde{u}_{1A}, \frac{\rho_1^2}{1+\rho_1^2 d_A^{\mathrm{T}} S_{2AA}^{-1} d_A} E_{d2} d_A^{\mathrm{T}} S_{2AA}^{-1} \tilde{u}_{2A} \right\|_{\infty,2}.
\end{aligned}
$$

By the irrepresentability condition (Assumption 3), there exist $\alpha \in (0,1]$ such that

$$
\|\Sigma_{1A^cA} \Sigma_{1AA}^{-1} \tilde{u}_{1A}, \Sigma_{2A^cA} \Sigma_{2AA}^{-1} \tilde{u}_{2A}\|_{\infty,2} \leq 1-\alpha.
$$

To conclude the proof, it is sufficient to show that with probability at least $1-\eta$ each $I_k \leq \lambda\alpha/4$, $k=1,\ldots,4$. Next, we consider each of these four terms separately.

1. Show $I_1 \leq \lambda\alpha/4$ with probability at least $1-\eta/4$. By Lemma 2, $e_j^{\mathrm{T}} E_{dg} \sim \mathcal{N}(0, \sigma_{jdg}^2/n_g)$. Applying standard normal concentration inequality, there exist constant $C>0$ such that

$$
\mathrm{pr}\Big(\bigcap_{j\in A^c} \Big\{ |e_j^{\mathrm{T}} E_{dg}| \geq C \max_{j\in A^c} \sigma_{jdg} \Big[n_g^{-1} \log\{(p-s)\eta^{-1}\} \Big]^{1/2} \Big\} \Big) \leq \eta/4.
$$

Since

$$
\begin{aligned}
&\|\rho_2(1+\rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A)^{-1} E_{d1}, \rho_1(1+\rho_1^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A)^{-1} E_{d2}\|_{\infty,2} \\
&\quad \leq \sqrt{2} \max\Big\{ \rho_2(1+\rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A)^{-1} \|E_{d1}\|_\infty, \rho_1(1+\rho_1^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A)^{-1} \|E_{d2}\|_\infty \Big\} \\
&\quad \leq \sqrt{2} \max(\|E_{d1}\|_\infty, \|E_{d2}\|_\infty),
\end{aligned}
$$

it follows that there exist constant $C>0$ such that

$$
\mathrm{pr}\Big(I_1 \geq C \max_{g=1,2;\ j\in A^c} \sigma_{jdg} \Big[\log\{(p-s)\eta^{-1}\} / \min(n_1,n_2) \Big]^{1/2} \Big) \leq \eta/4.
$$

Therefore, $I_1 \leq \lambda\alpha/4$ with probability at least $1-\eta/4$ under the conditions of the theorem.

2. Show $I_2 \leq \lambda\alpha/4$ with probability at least $1-\eta/4$. By Lemma 2, $E_{Ug} \sim \mathcal{N}(0, \Sigma_{gA^cA^c:A} \otimes I_{n_g-1})$ for $g = 1, 2$, and is independent of U_{gA} and d. Hence,

$$\rho_2(1+\rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A)^{-1} e_j^{\mathrm{T}} (n_1-1)^{-1} E_{U1} U_{1A}^{\mathrm{T}} S_{1AA}^{-1} d_A | U_{1A}, d_A$$
$$\sim \mathcal{N}\left\{0, \sigma_{1jj:A}^2 (n_1-1)^{-1} \rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A (1+\rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A)^{-2}\right\}.$$

Define $L = (1+\rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A)^{-2} \rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A$. Using standard normal concentration inequality, there exist constant $C > 0$ such that conditionally on L, the event

$$\bigcap_{j \in A^c} \Big\{ \rho_2(1+\rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A)^{-1} |e_j^{\mathrm{T}} (n_1-1)^{-1} E_{U1} U_{1A}^{\mathrm{T}} S_{1AA}^{-1} d_A|$$
$$\geq C \max_{j \in A^c} \sigma_{1jj:A} \Big[L n_1^{-1} \log\{(p-s)\eta^{-1}\} \Big]^{1/2} \Big\}$$

has probability at most $\eta/4$. Since $L = (1+\rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A)^{-2} \rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A \leq (1+\rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A)^{-1} \leq 1$, it follows that with probability at least $1-\eta/4$

$$\frac{\rho_2}{1+\rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A} \left\| \frac{E_{U1} U_{1A}^{\mathrm{T}} S_{1AA}^{-1} d_A}{n_1-1} \right\|_\infty \leq C \Big[\max_{j \in A^c} \sigma_{1jj:A} n_1^{-1} \log\{(p-s)\eta^{-1}\} \Big]^{1/2}.$$

The case $g = 2$ is similar, leading to the desired bound under the conditions of the theorem.

3. Show $I_3 \leq \alpha/4$ with probability at least $1-\eta/4$. Similar to part 2,

$$e_j^{\mathrm{T}} (n_1-1)^{-1} E_{U1} U_{1A}^{\mathrm{T}} S_{1AA}^{-1} (I+\rho_2^2 d_A d_A^{\mathrm{T}} S_{1AA}^{-1})^{-1} \tilde{u}_{1A} | U_{1A}, \tilde{u}_{1A}, d_A$$
$$\sim \mathcal{N}\left(0, (n_1-1)^{-1} \sigma_{1jj:A}^2 \tilde{u}_{1A}^{\mathrm{T}} (S_{1AA}+\rho_2^2 d_A d_A^{\mathrm{T}})^{-1} S_{1AA} (S_{1AA}+\rho_2^2 d_A d_A^{\mathrm{T}})^{-1} \tilde{u}_{1A}\right).$$

Define $L = \tilde{u}_{1A}^{\mathrm{T}} (S_{1AA}+\rho_2^2 d_A d_A^{\mathrm{T}})^{-1} S_{1AA} (S_{1AA}+\rho_2^2 d_A d_A^{\mathrm{T}})^{-1} \tilde{u}_{1A}$. As in part 2, there exist

constant $C > 0$ such that conditionally on L the event

$$\bigcap_{j \in A^c} \Big\{ |e_j^{\mathrm{T}}(n_1 - 1)^{-1} E_{U1} U_{1A}^{\mathrm{T}} S_{1AA}^{-1} (I + \rho_2^2 d_A d_A^{\mathrm{T}} S_{1AA}^{-1})^{-1} \tilde{u}_{1A}| \\ \geq C \max_{j \in A^c} \sigma_{1jj:A} \Big[L n_1^{-1} \log\{(p-s)\eta^{-1}\} \Big]^{1/2} \Big\}$$

has probability at most $\eta/4$. Furthermore,

$$\begin{aligned} L &\leq \|\tilde{u}_{1A}\|_2^2 \|(S_{1AA} + \rho_2^2 d_A d_A^{\mathrm{T}})^{-1} S_{1AA} (S_{1AA} + \rho_2^2 d_A d_A^{\mathrm{T}})^{-1}\|_2 \\ &\leq s \|S_{1AA}^{-1/2} (I + \rho_2^2 S_{1AA}^{-1/2} d_A d_A^{\mathrm{T}} S_{1AA}^{-1/2})^{-2} S_{1AA}^{-1/2}\|_2^2 \\ &\leq s \|S_{1AA}^{-1}\|_2, \end{aligned}$$

where in the last inequality we used $\|\tilde{u}_{1A}\|_2^2 + \|\tilde{u}_{2A}\|_2^2 \leq s$ by definition of subgradient. By Lemma 3, there exist constant $C > 0$ such that with probability at least $1 - \eta/4$

$$\|S_{1AA}^{-1}\|_2 \leq \|\Sigma_{1AA}^{-1}\|_2 \left[1 + C \Big\{ n_1^{-1} \log(\eta^{-1}) \Big\}^{1/2} \right].$$

Combining the above displays leads to

$$\|(n_1 - 1)^{-1} E_{U1} U_{1A}^{\mathrm{T}} S_{1AA}^{-1} (I + \rho_2^2 d_A d_A^{\mathrm{T}} S_{1AA}^{-1})^{-1} \tilde{u}_{gA}\|_\infty \\ \leq C \max_{j \in A^c} \sigma_{1jj:A} \Big[\|\Sigma_{1AA}^{-1}\|_2 n_1^{-1} s \log\{(p-s)\eta^{-1}\} \Big]^{1/2}$$

with probability at least $1 - \eta/4$. The proof for $g = 2$ is similar leading to the desired bound.

4. Show $I_4 \leq \alpha/4$ with probability at least $1 - \eta/4$.

By Lemma 2, $e_j^{\mathrm{T}} E_{dg} \sim \mathcal{N}(0, n_g^{-1} \sigma_{jdg}^2)$, where σ_{jdg} is from Lemma 2. Then

$$\rho_2^2 (1 + \rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A)^{-1} e_j^{\mathrm{T}} E_{d1} d_A^{\mathrm{T}} S_{1AA}^{-1} \tilde{u}_{1A} | U_{1A}, \tilde{u}_{1A}, d_A \\ \sim \mathcal{N}\Big(0, \frac{\sigma_{jd1}^2 \rho_2^4}{n_1 (1 + \rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A)^2} \tilde{u}_{1A}^{\mathrm{T}} S_{1AA}^{-1} d_A d_A^{\mathrm{T}} S_{1AA}^{-1} \tilde{u}_{1A}\Big).$$

Define $L = (1 + \rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A)^{-2} \rho_2^4 \tilde{u}_{1A}^{\mathrm{T}} S_{1AA}^{-1} d_A d_A^{\mathrm{T}} S_{1AA}^{-1} \tilde{u}_{1A}$. Using standard normal concentration inequality there exist constant $C > 0$ such that conditionally on L the event

$$\bigcap_{j \in A^c} \left\{ \frac{\rho_2^2}{1 + \rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A} e_j^{\mathrm{T}} E_{d1} d_A^{\mathrm{T}} S_{1AA}^{-1} \tilde{u}_{1A} \geq C \max_{j \in A^c} \sigma_{jd1} \left[L n_1^{-1} \log\{(p-s)\eta^{-1}\} \right]^{1/2} \right\}$$

has probability at most $\eta/4$. Furthermore,

$$\begin{aligned} L &= (1 + \rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A)^{-2} \rho_2^4 (\tilde{u}_{1A}^{\mathrm{T}} S_{1AA}^{-1/2} S_{1AA}^{-1/2} d_A)^2 \\ &\leq \rho_2^2 (1 + \rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A)^{-2} \rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A \tilde{u}_{1A}^{\mathrm{T}} S_{1AA}^{-1} \tilde{u}_{1A} \\ &\leq \rho_2^2 \tilde{u}_{1A}^{\mathrm{T}} S_{1AA}^{-1} \tilde{u}_{1A} \\ &\leq s \|S_{1AA}^{-1}\|_2, \end{aligned}$$

where in the last inequality we used $\|\tilde{u}_{1A}\|_2^2 + \|\tilde{u}_{2A}\|_2^2 \leq s$ by definition of subgradient. Similar to part 3, this means that there exists constant $C > 0$ such that

$$\left\| \frac{\rho_2^2}{1 + \rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A} E_{d1} d_A^{\mathrm{T}} S_{1AA}^{-1} \tilde{u}_{1A} \right\|_\infty \geq C \max_{j \in A^c} \sigma_{jd1} \left[\|\Sigma_{1AA}^{-1}\|_2 n_1^{-1} s \log\{(p-s)\eta^{-1}\} \right]^{1/2}$$

with probability at most $\eta/4$. The proof for $g = 2$ is analogous, leading to the desired bound. □

Theorem 4 (Equivalent to Theorem 2)**.** *Assume the conditions of Theorem 3 hold. If in addition* $\psi_{\min} \gtrsim \lambda s^{1/2} \max_g \|\Sigma_{g,AA}^{-1}\|_2 (\max_g \delta_A^{\mathrm{T}} \Sigma_{gAA}^{-1} \delta_A \vee \gamma)$*, then* $pr(\widehat{A} = A) \geq 1 - \eta$.

Proof of Theorem 4. Consider the oracle solution

$$\begin{aligned} \tilde{v}_{1A} &= \rho_2 S_{1AA}^{-1} d_A (1 + \rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A)^{-1} - \lambda \left(S_{1AA} + \rho_2^2 d_A d_A^{\mathrm{T}} \right)^{-1} \tilde{u}_{1A}, \\ \tilde{v}_{2A} &= \rho_1 S_{2AA}^{-1} d_A (1 + \rho_1^2 d_A^{\mathrm{T}} S_{2AA}^{-1} d_A)^{-1} - \lambda \left(S_{2AA} + \rho_1^2 d_A d_A^{\mathrm{T}} \right)^{-1} \tilde{u}_{2A}; \end{aligned}$$

where $\tilde{u}_A$ is defined in (3.15). To show $\widehat{A} = A$, it is sufficient to show

$$\begin{aligned}\min_{j\in A} \Big\|\rho_2(1+\rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A)^{-1} e_j^{\mathrm{T}} S_{1AA}^{-1} d_A,\ \rho_1(1+\rho_1^2 d_A^{\mathrm{T}} S_{2AA}^{-1} d_A)^{-1} e_j^{\mathrm{T}} S_{2AA}^{-1} d_A\Big\|_2 \\ \geq \lambda \max_{j\in A} \|e_j^{\mathrm{T}} \left(S_{1AA} + \rho_2^2 d_A d_A^{\mathrm{T}}\right)^{-1} \tilde{u}_{1A},\ e_j^{\mathrm{T}} \left(S_{2AA} + \rho_1^2 d_A d_A^{\mathrm{T}}\right)^{-1} \tilde{u}_{2A}\|_2.\end{aligned} \tag{3.16}$$

Consider the right-hand side in (3.16)

$$\begin{aligned}&\max_{j\in A} \|e_j^{\mathrm{T}} \left(S_{1AA} + \rho_2^2 d_A d_A^{\mathrm{T}}\right)^{-1} \tilde{u}_{1A},\ e_j^{\mathrm{T}} \left(S_{2AA} + \rho_1^2 d_A d_A^{\mathrm{T}}\right)^{-1} \tilde{u}_{2A}\|_2 \\ &= \max_{j\in A} \left[\left\{e_j^{\mathrm{T}} \left(S_{1AA} + \rho_2^2 d_A d_A^{\mathrm{T}}\right)^{-1} \tilde{u}_{1A}\right\}^2 + \left\{e_j^{\mathrm{T}} \left(S_{2AA} + \rho_1^2 d_A d_A^{\mathrm{T}}\right)^{-1} \tilde{u}_{2A}\right\}^2\right]^{1/2} \\ &\leq \max_{j\in A} \left\{\|e_j^{\mathrm{T}} \left(S_{1AA} + \rho_2^2 d_A d_A^{\mathrm{T}}\right)^{-1} \|_2^2 \|\tilde{u}_{1A}\|_2^2 + \|e_j^{\mathrm{T}} \left(S_{2AA} + \rho_1^2 d_A d_A^{\mathrm{T}}\right)^{-1} \|_2^2 \|\tilde{u}_{2A}\|_2^2\right\}^{1/2} \\ &\leq \max_{j\in A} \left\{\|e_j^{\mathrm{T}} \left(S_{1AA} + \rho_2^2 d_A d_A^{\mathrm{T}}\right)^{-1} \|_2 \vee \|e_j^{\mathrm{T}} \left(S_{2AA} + \rho_1^2 d_A d_A^{\mathrm{T}}\right)^{-1} \|_2\right\} \left(\|\tilde{u}_{1A}\|_2^2 + \|\tilde{u}_{2A}\|_2^2\right)^{1/2} \\ &\leq \left\{\|(S_{1AA} + \rho_2^2 d_A d_A^{\mathrm{T}})^{-1}\|_2 \vee \|(S_{2AA} + \rho_1^2 d_A d_A^{\mathrm{T}})^{-1}\|_2\right\} s^{1/2}.\end{aligned}$$

Furthermore,

$$\|\left(S_{1AA} + \rho_2^2 d_A d_A^{\mathrm{T}}\right)^{-1} \|_2 = \|S_{1AA}^{-1/2} \left(I + \rho_2^2 S_{1AA}^{-1/2} d_A d_A^{\mathrm{T}} S_{1AA}^{-1/2}\right)^{-1} S_{1AA}^{-1/2}\|_2 \leq \|S_{1AA}^{-1}\|_2,$$

and similarly $\|(S_{2AA} + \rho_1^2 d_A d_A^{\mathrm{T}})^{-1}\|_2 \leq \|S_{2AA}^{-1}\|_2$. Using Lemma 3

$$\begin{aligned}\max_{j\in A} &\|e_j^{\mathrm{T}} \left(S_{1AA} + \rho_2^2 d_A d_A^{\mathrm{T}}\right)^{-1} \tilde{u}_{1A},\ e_j^{\mathrm{T}} \left(S_{2AA} + \rho_1^2 d_A d_A^{\mathrm{T}}\right)^{-1} \tilde{u}_{2A}\|_2 \\ &\leq \max_g \|\Sigma_{gAA}^{-1}\|_2 s^{1/2} \left[1 + C\{s\log(\eta^{-1})/\min(n_1, n_2)\}^{1/2}\right]\end{aligned}$$

with probability at least $1 - \eta$.

Consider the left-hand side in (3.16). Applying Lemma 1 and Corollary 1, there exist

constants C_1, C_2 such that with probability at least $1-\eta$

$$\begin{aligned}
&\min_{j\in A}\left\|\rho_2(1+\rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A)^{-1} e_j^{\mathrm{T}}\Sigma_{1AA}^{-1}\delta_A,\ \rho_1(1+\rho_1^2 d_A^{\mathrm{T}} S_{2AA}^{-1} d_A)^{-1} e_j^{\mathrm{T}}\Sigma_{2AA}^{-1}\delta_A\right\|_2 \\
&\quad\geq \left[1+C_1\max_g \delta_A^{\mathrm{T}}\Sigma_{gAA}^{-1}\delta_A + C_2(\max_g \delta_A^{\mathrm{T}}\Sigma_{gAA}^{-1}\delta_A\vee\gamma)\left\{s\log(\eta^{-1})/\min(n_1,n_2)\right\}^{1/2}\right]^{-1} \\
&\qquad\qquad\times \min_{j\in A}\left\|\pi_2 e_j^{\mathrm{T}} S_{1AA}^{-1} d_A,\ \pi_1 e_j^{\mathrm{T}} S_{2AA}^{-1} d_A\right\|_2 .
\end{aligned}$$

Furthermore,

$$\begin{aligned}
&\min_{j\in A}\left\|\pi_2 e_j^{\mathrm{T}} S_{1AA}^{-1} d_A,\ \pi_1 e_j^{\mathrm{T}} S_{2AA}^{-1} d_A\right\|_2 \\
&=\min_{j\in A}\left\{\pi_2^2(e_j^{\mathrm{T}} S_{1AA}^{-1} d_A)^2 + \pi_1^2(e_j^{\mathrm{T}} S_{2AA}^{-1} d_A)^2\right\}^{1/2} \\
&=\min_{j\in A}\left[\pi_2^2\{e_j^{\mathrm{T}}(S_{1AA}^{-1} d_A - \Sigma_{1AA}^{-1}\delta_A + \Sigma_{1AA}^{-1}\delta_A)\}^2 + \pi_1^2\{e_j^{\mathrm{T}}(S_{2AA}^{-1} d_A - \Sigma_{2AA}^{-1}\delta_A + \Sigma_{2AA}^{-1}\delta_A)\}^2\right]^{1/2} \\
&\geq \min_{j\in A}\left\|\pi_2 e_j^{\mathrm{T}}\Sigma_{1AA}^{-1}\delta_A,\ \pi_1 e_j^{\mathrm{T}}\Sigma_{2AA}^{-1}\delta_A\right\|_2 - \max_g\left(\|S_{gAA}^{-1} d_A - \Sigma_{gAA}^{-1}\delta_A\|_\infty\right) \\
&=\psi_{\min} - \max_g\left(\|S_{gAA}^{-1} d_A - \Sigma_{gAA}^{-1}\delta_A\|_\infty\right),
\end{aligned}$$

where in the last inequality we used $\pi_1^2+\pi_2^2\leq 1$. Using Lemma 8

$$\begin{aligned}
&\max_g\left(\|S_{gAA}^{-1} d_A - \Sigma_{gAA}^{-1}\delta_A\|_\infty\right) \\
&\qquad\qquad\leq C\left[\max_{j\in A,g}\left\{(\Sigma_{gAA}^{-1})_{jj}(\delta_A^{\mathrm{T}}\Sigma_{gAA}^{-1}\delta_A\vee\gamma)\right\}s\log(\eta^{-1})/\min(n_1,n_2)\right]^{1/2}
\end{aligned}$$

with probability at least $1-\eta$.

Therefore, to have $A\subseteq\widehat{A}$, it is sufficient to have

$$\begin{aligned}
\psi_{\min} > C&\left[\max_{j\in A,g}\left\{(\Sigma_{gAA}^{-1})_{jj}(\delta_A^{\mathrm{T}}\Sigma_{gAA}^{-1}\delta_A\vee\gamma)\right\}s\log(\eta^{-1})/\min(n_1,n_2)\right]^{1/2} \\
&+\left[1+C_1\max_g \delta_A^{\mathrm{T}}\Sigma_{gAA}^{-1}\delta_A + C_2(\max_g \delta_A^{\mathrm{T}}\Sigma_{gAA}^{-1}\delta_A\vee\gamma)\left\{s\log(\eta^{-1})/\min(n_1,n_2)\right\}\right] \\
&\qquad\times\lambda\max_g\|\Sigma_{gAA}^{-1}\|_2 s^{1/2}\left[1+C\left\{s\log(\eta^{-1})/\min(n_1,n_2)\right\}^{1/2}\right].
\end{aligned}$$

Using the conditions on λ, and the fact that $\gamma \geq 1$, it follows that the second term above is the dominant term, and therefore it is sufficient to have for some constant $C > 0$

$$\psi_{\min} > C\lambda s^{1/2} \max_g \|\Sigma_{gAA}^{-1}\|_2 (\max_g \delta_A^{\mathrm{T}} \Sigma_{gAA}^{-1} \delta_A \vee \gamma).$$

□

3.6.4 Supporting theorems and lemmas

Theorem 5 (Oracle solution)**.** *Consider an oracle estimator* $[\tilde{v}_{1A}\ \tilde{v}_{2A}]$ *from* (3.14)*. Let* $\rho_1 = n_1/n$, $\rho_2 = n_2/n$*. Then*

$$\begin{aligned}
\tilde{v}_{1A} &= \rho_2 S_{1AA}^{-1} d_A (1 + \rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A)^{-1} - \lambda \left(S_{1AA} + \rho_2^2 d_A d_A^{\mathrm{T}} \right)^{-1} \tilde{u}_{1A}, \\
\tilde{v}_{2A} &= \rho_1 S_{2AA}^{-1} d_A (1 + \rho_1^2 d_A^{\mathrm{T}} S_{2AA}^{-1} d_A)^{-1} - \lambda \left(S_{2AA} + \rho_1^2 d_A d_A^{\mathrm{T}} \right)^{-1} \tilde{u}_{2A};
\end{aligned}$$

where $\tilde{u}_A$ *is defined in* (3.15)*.*

Proof. We present the proof only for $\tilde{v}_{1A}$, the proof for $\tilde{v}_{2A}$ is analogous. From Section 3.2.3

$$\begin{aligned}
&[\tilde{v}_{1A}\ \tilde{v}_{2A}] = \underset{v_{1A}, v_{2A} \in \mathbb{R}^s}{\operatorname{argmin}} \left\{ \widehat{L}_{\psi_1}(v_{1A}) + \widehat{L}_{\psi_2}(v_{2A}) + \lambda \sum_{j=1}^{s} (v_{1Aj}^2 + v_{2Aj}^2)^{1/2} \right\}, \\
&\widehat{L}_{\psi_1}(v_{1A}) + \widehat{L}_{\psi_2}(v_{2A}) \\
&\quad = v_{1A}^{\mathrm{T}} S_{1AA} v_{1A}/2 + \left(n_2/n d_A^{\mathrm{T}} v_{1A} - 1 \right)^2 /2 + v_{2A}^{\mathrm{T}} S_{2AA} v_{2A}/2 + \left(n_2/n d_A^{\mathrm{T}} v_{2A} - 1 \right)^2 /2.
\end{aligned}$$

Using the optimality conditions, the oracle solution must satisfy

$$\tilde{v}_{1A} = \left(S_{1AA} + \rho_2^2 d_A d_A^{\mathrm{T}} \right)^{-1} \left(\rho_2 d_A - \lambda \tilde{u}_{1A} \right),$$

where $\tilde{u}_A$ is the subgradient of $\sum_{j=1}^{s}(v_{1Aj}^2+v_{2Aj}^2)^{1/2}$ in (3.15). By Sherman–Morrison identity,

$$(S_{1AA}-\rho_2^2 d_A d_A^{\mathrm{T}})^{-1} = S_{1AA}^{-1} - (1+\rho_2^2 d_A^{\mathrm{T}} S_{1AA}^{-1} d_A)^{-1}\rho_2^2 S_{1AA}^{-1} d_A d_A^{\mathrm{T}} S_{1AA}^{-1}.$$

The statement follows by combining the above two displays. □

Lemma 1. *There exist constant $C>0$ such that with probability at least $1-\eta$*

$$|n_g/n-\pi_g| \le C\Big\{log(\eta^{-1})/n\Big\}^{1/2} \quad (g=1,2), \quad |n_1/n_2-\pi_1/\pi_2| \le C\Big\{log(\eta^{-1})/n\Big\}^{1/2}.$$

Proof. Given that $n_g \sim \mathrm{Bin}(n,\pi_g)$, by Hoeffding inequality $\mathrm{pr}(|\pi_g - n_g/n| \ge \varepsilon) \le 2\exp(-2n\varepsilon^2)$. Let $\eta = 2\exp(-2n\varepsilon^2)$, then $2n\varepsilon^2 = \log(2\eta^{-1})$, $\varepsilon = C\{\log(\eta^{-1})/n\}^{1/2}$ and $n_g/n = \pi_g + \mathcal{O}_p\{\log(\eta^{-1})/n\}^{1/2}$. Let $f(x)=x/(1-x)$, which is non-decreasing for $x\in(0,1)$. Since $n_1/n_2 = f\left(n_1/n\right)$, the second inequality in the lemma follows from the first. □

Lemma 2. *Let $E_{Ug} = U_{gA^c} - \Sigma_{gA^cA}\Sigma_{gAA}^{-1}U_{gA}$, $E_{dg} = d_{A^c} - \Sigma_{gA^cA}\Sigma_{gAA}^{-1}d_A$, $g=1,2$. Then E_{Ug} is independent from U_{gA}, $E_{Ug} \sim \mathcal{N}(0, \Sigma_{gA^cA^c:A}\otimes I_{n_g-1})$, $\quad e_j^{\mathrm{T}}E_{dg} \sim \mathcal{N}\left(0, n_g^{-1}\sigma_{jdg}^2\right)$; where $\sigma_{jdg}^2 = e_j^{\mathrm{T}}\Sigma_{d_g}e_j$, and $\Sigma_{gA^cA^c:A}$, Σ_{d_g} are defined in* (3.13).

Proof. Since E_{dg}, E_{Ug} are formed by applying linear transformation to normal d, U_1, U_2, it follows that E_{dg}, E_{Ug} are also normally distributed. It remains to verify the form of the means and covariance matrices. We consider $g=1$, the proof for $g=2$ is similar.

Consider E_{U1}. By definition, the columns of U_1 satisfy $u_{1i}\sim N(0,\Sigma_1)$. Since

$$E_{U1} = (-\Sigma_{1A^cA}\Sigma_{1AA}^{-1}\ I_{p-s})\begin{pmatrix} U_{1A} \\ U_{1A^c}\end{pmatrix},$$

it follows that $\mathbb{E}(E_{U1}) = 0$, and

$$\begin{aligned}\text{var}(E_{U1}) &= (-\Sigma_{1A^cA}\Sigma_{1AA}^{-1}\ I_{p-s}) \begin{pmatrix} \Sigma_{1AA} & \Sigma_{1AA^c} \\ \Sigma_{1A^cA} & \Sigma_{1A^cA^c} \end{pmatrix} (-\Sigma_{1A^cA}\Sigma_{1AA}^{-1}\ I_{p-s})^{\mathrm{T}} \otimes I_{n_1-1} \\ &= (\Sigma_{1A^cA^c} - \Sigma_{1A^cA}\Sigma_{1AA}^{-1}\Sigma_{1AA^c}) \otimes I_{n_1-1}.\end{aligned}$$

Consider E_{d1}. Since $\Sigma_1^{-1}\delta = \psi_1 = (\psi_{1A}^{\mathrm{T}}, 0)^{\mathrm{T}}$, by rewriting $\Sigma_1\Sigma_1^{-1}\delta = \delta$, and using block matrices of Σ_1 and Σ_1^{-1}, it follows that $\Sigma_{1A^cA}\Sigma_{1AA}^{-1}\delta_A = \delta_{A^c}$. Then $\mathbb{E}(E_{d1}) = \delta_{A^c} - \Sigma_{1A^cA}\Sigma_{1AA}^{-1}\delta_A = 0$. Furthermore,

$$\begin{aligned}&\text{var}(E_{d1}) \\ &\quad= \text{var}(d_{A^c} - \Sigma_{1A^cA}\Sigma_{1,AA}^{-1}d_A) \\ &\quad= \text{var}(d_{A^c}) + \Sigma_{1A^cA}\Sigma_{1AA}^{-1}\text{var}(d_A)\Sigma_{1AA}^{-1}\Sigma_{1AA^c} \\ &\qquad- \Sigma_{1A^cA}\Sigma_{1AA}^{-1}\text{cov}(d_A, d_{A^c}) - \text{cov}(d_{A^c}, d_A)\Sigma_{1AA}^{-1}\Sigma_{1AA^c} \\ &\quad= n_1^{-1}\Sigma_{1A^cA^c} + n_2^{-1}\Sigma_{2A^cA^c} + \Sigma_{1A^cA}\Sigma_{1AA}^{-1}\left(n_1^{-1}\Sigma_{1AA} + n_2^{-1}\Sigma_{2AA}\right)\Sigma_{1AA}^{-1}\Sigma_{1AA^c} \\ &\qquad- \Sigma_{1A^cA}\Sigma_{1AA}^{-1}\left(n_1^{-1}\Sigma_{1AA^c} + n_2^{-1}\Sigma_{2AA^c}\right) - \left(n_1^{-1}\Sigma_{1A^cA} + n_2^{-1}\Sigma_{2A^cA}\right)\Sigma_{1AA}^{-1}\Sigma_{1AA^c} \\ &\quad= n_1^{-1}\Sigma_{1A^cA^c:A} + n_2^{-1}\Big(\Sigma_{2A^cA^c} + \Sigma_{1A^cA}\Sigma_{1AA}^{-1}\Sigma_{2AA}\Sigma_{1AA}^{-1}\Sigma_{1AA^c} \\ &\qquad\qquad- \Sigma_{1A^cA}\Sigma_{1AA}^{-1}\Sigma_{2AA^c} - \Sigma_{2A^cA}\Sigma_{1AA}^{-1}\Sigma_{1AA^c}\Big).\end{aligned}$$

□

Lemma 3. *Let S_{gAA} be a submatrix of the sample covariance matrix for group $g \in \{1, 2\}$ corresponding to variables in A, with $s = \text{card}(A)$. Let Σ_{gAA} be the corresponding submatrix of population covariance matrix. Under Assumption 1, there exist constants $C_1, C_2 > 0$ such that with probability at least $1 - \eta$*

$$\|\Sigma_{gAA}^{1/2}S_{gAA}^{-1}\Sigma_{gAA}^{1/2} - I\|_2 \leq C_1\Big\{slog(\eta^{-1})/n_g\Big\}^{1/2},\ \|S_{gAA}^{-1}\|_2 \leq \|\Sigma_{gAA}^{-1}\|_2\Big[1 + C_2\Big\{slog(\eta^{-1})/n_g\Big\}^{1/2}\Big].$$

Proof. Using normality, the sample covariance matrices satisfy $S_{gAA} = (n_g - 1)^{-1} W_g W_g^{\mathrm{T}}$ with $W_g \in \mathbb{R}^{s \times (n_g - 1)}$ having independent columns $w_{gi} \sim N(0, \Sigma_{gAA})$. Then the desired bounds follow from Wainwright (2009, Lemma 9). □

Lemma 4. *Let a random vector $X \in \mathbb{R}^s$ be such that $X \sim \mathcal{N}(0, n^{-1}A)$. Then there exist constant $C > 0$ such that with probability at least $1 - \eta$*

$$\|X\|_2 \leq C \Big\{ \|A\|_2 n^{-1} s log(\eta^{-1}) \Big\}^{1/2}.$$

Proof. Since $A^{-1/2}X \sim \mathcal{N}(0, n^{-1}I_s)$, by Hsu et al. (2012, Proposition 1.1), with probability at least $1 - \eta$

$$\|A^{-1/2}X\|_2^2 \leq s/n + 2\Big\{ s\log(\eta^{-1}) \Big\}^{1/2}/n + 2\log(\eta^{-1})/n.$$

For small η it follows that there exist $C > 0$ such that $\|A^{-1/2}X\|_2^2 \leq Cn^{-1}s\log(\eta^{-1})$ with probability at least $1 - \eta$. The statement of the lemma follows since

$$\|X\|_2^2 = X^{\mathrm{T}}X = X^{\mathrm{T}}A^{-1/2}AA^{-1/2}X \leq \|A\|_2 \|A^{-1/2}X\|_2^2.$$

□

Lemma 5. *There exist constant $C > 0$ such that with probability at least $1 - \eta$*

$$\max_g \|\Sigma_{gAA}^{-1/2}(d_A - \delta_A)\|_2 \leq C \Big\{ \gamma s log(\eta^{-1}) / \min(n_1, n_2) \Big\}^{1/2},$$

where γ is defined in (3.12).

Proof. Since $d_A - \delta_A \sim \mathcal{N}(0, n_1^{-1}\Sigma_{1AA} + n_2^{-1}\Sigma_{2AA})$, it follows that

$$\Sigma_{1AA}^{-1/2}(d_A - \delta_A) \sim \mathcal{N}\left(0, n_1^{-1}\left(I + n_2^{-1}n_1\Sigma_{1AA}^{-1/2}\Sigma_{2AA}\Sigma_{1AA}^{-1/2}\right)\right).$$

Applying Lemma 1 and Lemma 4 concludes the proof. The case $g = 2$ is analogous. □

Lemma 6. *There exist constants C_1, C_2 such that with probability at least $1-\eta$ for $g = 1, 2$*

$$d_A^{\mathrm{T}} S_{gAA}^{-1} d_A \leq C_1 d_A^{\mathrm{T}} \Sigma_{gAA}^{-1} d_A \left[1 + C_2 \Big\{ log(\eta^{-1})/(n_g - s) \Big\}^{1/2}\right].$$

Proof. We prove for $g = 1$, case $g = 2$ is analogous. Since $(n_1 - 1)S_{1AA} \sim W_s(n_1 - 1, \Sigma_{1AA})$, and d_A is independent of S_{1AA}, by Muirhead (1982, Theorem 3.2.12)

$$(n_1 - 1)\frac{d_A^{\mathrm{T}} \Sigma_{1AA}^{-1} d_A}{d_A^{\mathrm{T}} S_{1AA}^{-1} d_A} \sim \chi^2_{n_1 - s}.$$

Using (Laurent and Massart, 2000, Lemma 1),

$$\mathrm{pr}\Big[(n_1 - 1)\frac{d_A^{\mathrm{T}} \Sigma_{1AA}^{-1} d_A}{d_A^{\mathrm{T}} S_{1AA}^{-1} d_A} \geq (n_1 - s) - 2\Big\{(n_1 - s)\log(\eta^{-1})\Big\}^{1/2}\Big] \geq 1 - \eta.$$

Therefore, with probability at least $1 - \eta$

$$d_A^{\mathrm{T}} S_{1AA}^{-1} d_A \leq (n_1 - 1)(n_1 - s)^{-1} d_A^{\mathrm{T}} \Sigma_{1AA}^{-1} d_A \left[1 - 2\Big\{\log(\eta^{-1})/(n_1 - s)\Big\}^{1/2}\right]^{-1}.$$

Hence, there exist constants $C_1, C_2 > 0$ such that with probability at least $1 - \eta$

$$d_A^{\mathrm{T}} S_{1AA}^{-1} d_A \leq C_1 d_A^{\mathrm{T}} \Sigma_{1AA}^{-1} d_A \left[1 + C_2 \Big\{\log(\eta^{-1})/(n_1 - s)\Big\}^{1/2}\right].$$

□

Lemma 7. *There exist constant $C > 0$ such that with probability at least $1 - \eta$*

$$d_A^{\mathrm{T}} \Sigma_{gAA}^{-1} d_A \leq C \left\{\delta_A^{\mathrm{T}} \Sigma_{gAA}^{-1} \delta_A + \gamma n_g^{-1} s log(\eta^{-1})\right\} \quad (g = 1, 2),$$

where γ is defined in (3.12).

Proof. We prove the result for $g = 1$, the case $g = 2$ is similar. Consider

$$\begin{aligned} d_A^{\mathrm{T}}\Sigma_{1AA}^{-1}d_A &= \delta_A^{\mathrm{T}}\Sigma_{1AA}^{-1}\delta_A + 2(d_A - \delta_A)^{\mathrm{T}}\Sigma_{1AA}^{-1}\delta_A + (d_A - \delta_A)^{\mathrm{T}}\Sigma_{1AA}^{-1}(d_A - \delta_A) \\ &\leq 2\delta_A^{\mathrm{T}}\Sigma_{1AA}^{-1}\delta_A + 2(d_A - \delta_A)^{\mathrm{T}}\Sigma_{1AA}^{-1}(d_A - \delta_A). \end{aligned}$$

By Lemma 5, there exist constant $C \geq 0$ such that with probability at least $1 - \eta$

$$(d_A - \delta_A)^{\mathrm{T}}\Sigma_{1AA}^{-1}(d_A - \delta_A) \leq C\gamma n_1^{-1} s\log(\eta^{-1}).$$

The result follows by combining the above displays. □

Corollary 1. *There exist constants $C_1, C_2, C_3 > 0$ such that with probability at least $1 - \eta$ for $g = 1, 2$ and γ in* (3.12)

$$d_A^{\mathrm{T}}S_{gAA}^{-1}d_A \leq C_1\delta_A^{\mathrm{T}}\Sigma_{gAA}^{-1}\delta_A \left[1 + C_2\left\{log(\eta^{-1})/(n_g - s)\right\}^{1/2}\right] + C_3\gamma n_g^{-1} slog(\eta^{-1}).$$

Proof. The result follows by combining results of Lemma 6 and Lemma 7. □

Lemma 8. *There exist constant $C > 0$ such that with probability at least $1 - \eta$ for $g = 1, 2$*

$$\|S_{gAA}^{-1}d_A - \Sigma_{gAA}^{-1}\delta_A\|_\infty \leq C\Big\{\max_{j\in A}(\Sigma_{gAA}^{-1})_{jj}(\delta_A^{\mathrm{T}}\Sigma_{gAA}^{-1}\delta_A \vee \gamma)n_g^{-1} slog(\eta^{-1})\Big\}^{1/2},$$

where γ is defined in (3.12).

Proof. We prove the result for $g = 1$, the case $g = 2$ is similar. Consider

$$
\begin{aligned}
&|e_j^{\mathrm{T}} S_{1AA}^{-1} d_A - e_j^{\mathrm{T}} \Sigma_{1AA}^{-1} \delta_A| \\
&\qquad = |e_j^{\mathrm{T}} (S_{1AA}^{-1} - \Sigma_{1AA}^{-1})(d_A - \delta_A) + e_j^{\mathrm{T}} (S_{1AA}^{-1} - \Sigma_{1AA}^{-1}) \delta_A + e_j^{\mathrm{T}} \Sigma_{1AA}^{-1} (d_A - \delta_A)| \\
&\qquad \leq (e_j^{\mathrm{T}} \Sigma_{1AA}^{-1} e_j)^{1/2} \|(\Sigma_{1AA}^{1/2} S_{1AA}^{-1} \Sigma_{1AA}^{1/2} - I) \Sigma_{1AA}^{-1/2} (d_A - \delta_A)\|_2 \\
&\qquad\quad + (e_j^{\mathrm{T}} \Sigma_{1AA}^{-1} e_j)^{1/2} \|(\Sigma_{1AA}^{1/2} S_{1AA}^{-1} \Sigma_{1AA}^{1/2} - I) \Sigma_{1AA}^{-1/2} \delta_A\|_2 \\
&\qquad\quad + (e_j^{\mathrm{T}} \Sigma_{1AA}^{-1} e_j)^{1/2} \|\Sigma_{1AA}^{-1/2} (d_A - \delta_A)\|_2.
\end{aligned}
$$

Let $m_1 = \|\Sigma_{1AA}^{1/2} S_{1AA}^{-1} \Sigma_{1AA}^{1/2} - I\|_2$ and $m_2 = \|\Sigma_{1AA}^{-1/2} (d_A - \delta_A)\|_2$. Using the above display

$$
\|S_{1AA}^{-1} d_A - \Sigma_{1AA}^{-1} \delta_A\|_\infty \leq \max_{j \in A} (\Sigma_{1AA}^{-1})_{jj}^{1/2} \left\{ m_1 m_2 + m_1 (\delta_A^{\mathrm{T}} \Sigma_{1AA}^{-1} \delta_A)^{1/2} + m_2 \right\}. \tag{3.17}
$$

Using Lemma 3, there exist constant $C_1 > 0$ such that $m_1 \leq C_1 \{s\log(\eta^{-1})/n_1\}^{1/2}$ with probability at least $1 - \eta$. Using Lemma 5, there exist constant $C_2 > 0$ such that $m_2 \leq C_2 \{\gamma s\log(\eta^{-1})/n_1\}^{1/2}$ with probability at least $1 - \eta$. Combining these bounds with (3.17), there exist constant $C > 0$ such that with probability at least $1 - \eta$

$$
\|S_{1AA}^{-1} d_A - \Sigma_{1AA}^{-1} \delta_A\|_\infty \leq C \left\{ \max_{j \in A} (\Sigma_{1AA}^{-1})_{jj} (\delta_A^{\mathrm{T}} \Sigma_{1AA}^{-1} \delta_A \vee \gamma) n_1^{-1} s\log(\eta^{-1}) \right\}^{1/2}.
$$

□

4. VIGNETTE: FIT A MISSPECIFIED MODEL WITH MEASUREMENT ERROR USING CCP

This chapter provides further details and a concrete illustration for the R programs used in the paper *Categorizing a Continuous Predictor Subject to Measurement Error* (Blas et al., 2018). This package is mainly focused on logistic regression and linear regression, though the proposed method has much weaker assumptions and can be applied in many scenarios. This document provides a brief overview of the methodology, especially for linear regression and logistic regression. Further, we use simulation studies and a real data example, the EATS data (Subar et al., 2001), to show 4 ways to use `ccp`, the main function in the `CCP` package.

4.1 Introduction

In epidemiology, it is common to fit a categorical risk model to a continuous risk predictor, because the categorical one is thought to be more robust and interpretable. When the risk predictor is observed with measurement error, epidemiologists typically ignore the underlying measurement error and perform a naive approach, e.g., logistic regression, as what they would have done if they observe the true predictor. Here we introduce some notation to help describe the problem background.

- X: true risk predictor (continuous);
- X_{C}: categorized predictor;
- U: measurement error;
- W: observed risk predictor (continuous, with measurement error); $W = X + U$, X and U are independent;
- W_{C}: categorized predictor.

Using the notation stated above, ideally, epidemiologists categorize X and then use X_{C}

to fit the model. However, if they observe W instead of X, they would use W_{C} in the original categorical model without correcting the measurement error.

White (1982) shows that when X is observed, though the categorical model is a misspecified model, the estimates are unbiased with respect to the true value of what epidemiologists are interested in - the parameters with respect to X_C in the categorical risk model. When X is not observed, however, substituting W for X leads to a biased estimate, as well as a poor inference quality. To address the problem based on W, the relationship between W and X needs to be specified.

We address this problem and provide a general method to get unbiased estimates and correct inference even with measurement error in the data. The key of our method is adding another layer of conditional expectation given observed predictor W. Thus, the original estimating equation is now relying on W but not on X. We then need to estimate the expectations of functions of X given W. For example, suppose the original estimating equation is formed based on $E\{f(X)\} = 0$. Adding a layer of conditional expectation leads to $E[E\{f(X)|W\}] = 0$. Hence, the goal turns out to be estimating $E\{f(X)|W\}$, depending on the conditional density $f_{X|W}$. Although Blas et al. (2018+) focuses on the general case, this document aims to provide more details for logistic regression and linear regression.

Due to the complexity of the problem itself, in this package, we do not consider other covariates measured without error. Readers can find more general formulas in the original paper.

The rest of this document is organized as follows: we first provide a brief methodology review for readers to gain more background without looking at the original paper; then, we present estimating equations in logistic regression and linear regression. Finally, we show different ways to use the main function `ccp` through simulation studies, as well as the analysis for EATS data (Subar et al. 2001).

4.2 Methodology review

4.2.1 General overview

Here we present two cases: linear regression and logistic regression, corresponding to continuous or binary response. For the more general model and its assumptions, we refer readers to *Categorizing a Continuous Predictor* for more details.

This package allows users to use two types of data:

- External-internal data: if the main dataset has no replicates, users need to provide external data for nuisance parameter estimation, especially for estimating the variance of measurement error. Without external data, the measurement error is unidentifiable.
- Internal-only data: when the main dataset has replicates, the program only uses the main dataset to calculate the nuisance parameters. Any provided external data are ignored in this case.

In the following part, we explain the external-internal and internal-only cases in linear regression and logistic regression, respectively.

In the R package `CCP`, we assume that

$$W = X + U; \quad X \sim N(\mu_x, \sigma_x^2); \quad U \sim N(0, \sigma_u^2).$$

Also, X and U are independent. For convenience, we define nuisance parameter $\mathbf{\Lambda} = (\mu_x, \sigma_x^2, \sigma_u^2)$.

For the continuous risk predictor X, we denote $m(X, \boldsymbol{\beta}) = \alpha + X\beta$, where $\boldsymbol{\beta} = (\alpha, \beta)$. To categorize X into $j = 1, ..., J$ categories $(C_1, ..., C_J)$, we define $M(X) = \{I(X \in C_1), ..., I(X \in C_J)\}^{\mathrm{T}}$. Thus, the corresponding parameters in the categorical model are $\mathbf{\Theta} = (\theta_1, ..., \theta_J)$.

The parameter we are mainly interested in is $\theta_J - \theta_1$, which is the *log relative risk* in logistic regression.

Now we introduce three assumptions required for our approach:

(a) When X is observed, the true risk model in the continuous scale has unbiased estimating functions known up to parameters $\boldsymbol{\beta}$.

(b) When X is not observed, we can find a function $g(X, \boldsymbol{\beta})$ that $E[E\{g(X, \boldsymbol{\beta})|W\}] = 0$, with its conditional expectation $E\{g(X, \boldsymbol{\beta})|W\}$ depends on $\boldsymbol{\Lambda}$ and can be estimated. A special case is knowing the distribution of X given W up to parameters $\boldsymbol{\Lambda}$.

(c) If the external data are necessary for model identification, the parameter estimated from external data, i.e. σ_u^2, should be transportable. See Chapter 2.2.4-2.2.5 of R. J. Carroll et al. (2006).

For linear regression and logistic regression considered in this package, all three assumptions are satisfied. Further, we would like to point out that neither normally distributed X and U, nor logistic or linear regression model is specifically required for the proposed method itself.

To estimate nuisance parameters $\boldsymbol{\Lambda}$, we now introduce the estimating equations based on using external-internal or internal-only data. Then the estimating equations for $\boldsymbol{\beta}$ and $\boldsymbol{\Theta}$ are introduced, depending on using linear regression or logistic regression.

4.2.1.1 External-internal data

If there are no replicates in the internal data, we use the external data only to estimate σ_u^2. Suppose we observe $W_{ik} = X_i + U_{ik}$ for $k = 1, ..., K$ and $i = n+1, ..., n+N$. We use internal data to estimate μ_x, σ_x^2 without replicates.

In the external data, let $\overline{W}_{i\cdot} = K^{-1} \sum_{k=1}^{K} W_{ik}$. Define $\widehat{\sigma}_{u,i}^2 = (K-1)^{-1} \sum_{k=1}^{K} (W_{ik} - \overline{W}_{i\cdot})^2$ to be the sample variance of the W_{ik} for a given i.

Because $E\{(W_i - \mu_x)^2\} = \sigma_x^2 + \sigma_u^2$, unbiased estimating equations for $\mathbf{\Lambda} = (\mu_x, \sigma_x^2, \sigma_u^2)$ are

- For μ_x: $n^{-1}\sum_{i=1}^{n}(W_i - \mu_x) = 0$;
- For σ_u^2: $N^{-1}\sum_{i=n+1}^{n+N}(\widehat{\sigma}_{u,i}^2 - \sigma_u^2) = 0$.
- For σ_x^2: $n^{-1}\sum_{i=1}^{n}\{(W_i - \mu_x)^2 - \sigma_x^2 - \sigma_u^2\} = 0$;

4.2.1.2 Internal-only data

Suppose there are no external data, and we have replicates W_{ir} for $r = 1, ..., R$ in the internal data. Now we use the internal data to estimate $\Lambda = (\mu_x, \sigma_x^2, \sigma_{uR}^2)$, and we observe $W_{ir} = X_i + U_{ir}$ for $r = 1, ..., R$ and $i = 1, ..., n$. Define $\overline{W}_{i\cdot} = R^{-1}\sum_{r=1}^{R} W_{ir}$. Define $\widehat{\sigma}_{u,i}^2$ to be the sample variance of the W_{ir} within subject i, and define $\sigma_u^2/R = \sigma_{uR}^2$.The estimating equations are

- For μ_x: $n^{-1}\sum_{i=1}^{n}(\overline{W}_{i\cdot} - \mu_x) = 0$;
- For σ_{uR}^2: $n^{-1}\sum_{i=1}^{n}(\widehat{\sigma}_{u,i}^2/R - \sigma_{uR}^2) = 0$.
- For σ_x^2: $n^{-1}\sum_{i=1}^{n}\{(\overline{W}_{i\cdot} - \mu_x)^2 - \sigma_x^2 - \sigma_{uR}^2\} = 0$;

Using the external-internal or internal-only data influences how to estimate nuisance parameters $\mathbf{\Lambda}$, while fitting linear or logistic regression affects the estimating equations of $\boldsymbol{\beta}, \mathbf{\Theta}$ as described below.

4.2.2 Linear regression

We assume the true model in the continuous scale is

$$Y = \alpha + X\beta + \epsilon = m(X, \boldsymbol{\beta}) + \epsilon,$$

where $m(X, \boldsymbol{\beta}) = \alpha + X\beta$. For external-internal and internal-only cases, the estimation equations for $\boldsymbol{\beta}$ and $\boldsymbol{\Theta}$ are the same.

The estimating function for $\boldsymbol{\beta} = (\alpha, \beta)$ is

$$\Phi(\boldsymbol{\beta}, \widehat{\boldsymbol{\Lambda}}) = n^{-1} \textstyle\sum_{i=1}^{n} E[\{Y_i - m(X_i, \boldsymbol{\beta})\} \partial m(X_i, \boldsymbol{\beta}) / \partial \boldsymbol{\beta}^{\mathrm{T}} | W_i].$$

The estimating function for $\boldsymbol{\Theta}$ is

$$Q(W_i, \boldsymbol{\Theta}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}) = E \left[\begin{array}{c|c} m(X_i, \widehat{\boldsymbol{\beta}}) I(X_i \in C_1) - \boldsymbol{\Theta}_1 I(X_i \in C_1) & \\ \vdots & W_i \\ m(X_i, \widehat{\boldsymbol{\beta}}) I(X_i \in C_J) - \boldsymbol{\Theta}_J I(X_i \in C_J) & \end{array} \right].$$

The integration above is calculated using the `integrate` function in the R package `stats`.

4.2.3 Logistic regression

Let $H(\cdot)$ denote the logistic distribution function. Here we consider the special case of linear logistic regression with the classical measurement error model in both the external and internal datasets:

$$\mathrm{pr}(Y = 1 | X, Z) = H(\alpha + X\beta) = H\{(1, X)\boldsymbol{\beta}\}$$

Let $p_i = \mathrm{pr}(Y = 1 | W_i) = \int H\{(1, x)\boldsymbol{\beta}\} f_{x|W_i}(x, W_i, \boldsymbol{\Lambda}) dx$, we use the `integrate` function in the R package `stats` to compute this quantity and calculate the loglikelihood $\propto n^{-1} \sum_{i=1}^{n} Y_i \log(p_i) + (1 - Y_i) \log(1 - p_i)$. We then use the `optim` function in the R package `stats` to minimize negative loglikelihood to estimate $\boldsymbol{\beta}$.

Given the logistic regression model, the categorical estimating function is

$$\Phi_{\mathrm{cat}}\{Y, M^{\mathrm{T}}(X)\boldsymbol{\Theta}\} = M(X)[Y - H\{M^{\mathrm{T}}(X)\boldsymbol{\Theta}\}],$$

Where $M(X) = \{I(X \in C_1), ..., I(X \in C_J)\}^{\mathrm{T}}$ for categories $(C_1, ..., C_J)$. Hence, with $\boldsymbol{\Omega} = (\boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{\Lambda})$,

$$Q(W, \boldsymbol{\Omega}) = E\left(M(X) \left[H\{m(X, \boldsymbol{\beta})\} - H\{M^{\mathrm{T}}(X)\boldsymbol{\Theta}\} \right] \middle| W \right).$$

In the R program,

$$Q(W_i, \boldsymbol{\Theta}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}) = E\left[\begin{array}{c} H\{m(X_i, \widehat{\boldsymbol{\beta}})\} I(X_i \in C_1) - H(\boldsymbol{\Theta}_1) I(X_i \in C_1) \\ \vdots \\ H\{m(X_i, \widehat{\boldsymbol{\beta}})\} I(X_i \in C_J) - H(\boldsymbol{\Theta}_J) I(X_i \in C_J) \end{array} \middle| W_i \right].$$

Again, we use the `integrate` function in the R package `stats` to compute the integrals.

4.3 Function overview

As shown in Figure 4.1, the package `CCP` contains one main function named `ccp`. Based on the types of response, we can specify logistic regression for binary Y, or linear regression for continuous Y. Further, in each of the two cases we mentioned before, `ccp` provides the choice of external-internal and internal-only cases as introduced in the methodology review.

4.3.1 Get started

First, let us install the R package `CCP`.

```
install.packages("~/Desktop/CCP_1.1.tar.gz", repos = NULL, type = "source")
#> Warning in install.packages("~/Desktop/CCP_1.1.tar.gz", repos = NULL,
#> type = "source"): installation of package '/Users/tianying/Desktop/
#> CCP_1.1.tar.gz' had non-zero exit status
library(CCP)
```

Once the package has been loaded, one can call the main function as follows.

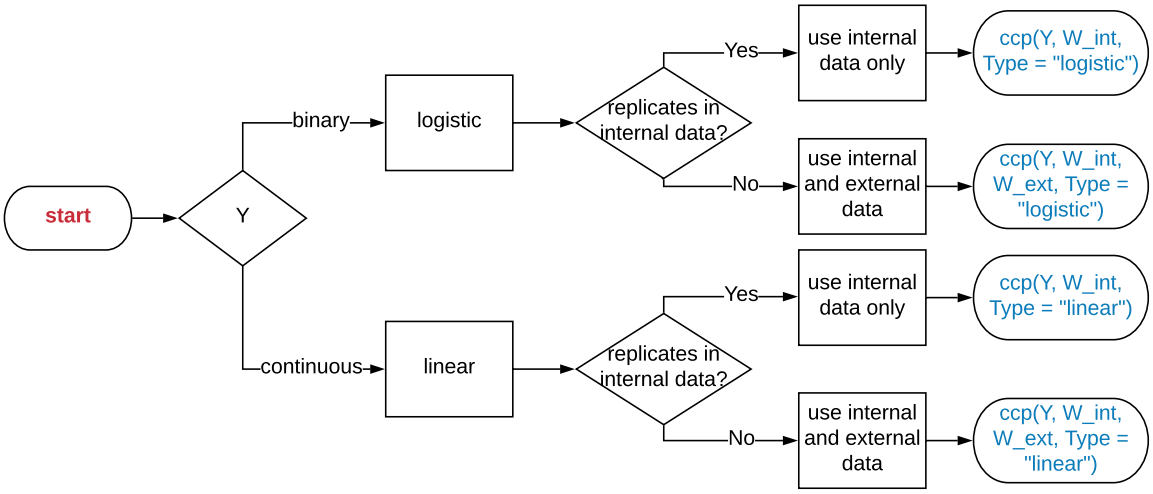

Figure 4.1: Functions overview

```
ccp(y, W_int, W_ext = NULL, C = NULL, Type, print.summary = TRUE, standardize = TRUE)
```

To check the package `CCP` or the usage of a specific function, you can either use `help` or `??`.

```
help( package = "CCP" )
??ccp
```

The former command gives a brief summary of all functions in the package, while the later one offers more detailed information for function `ccp`.

4.4 Simulation study

Here we show the external-internal and internal-only cases for logistic regression. We also compare the proposed method with the naive approach: substituting W for X in the categorical model with no adjustment for measurement error.

(1) Define a function to calculate the naive estimates.

```
thetaw <- function(y, w, mux_hat, s2x_hat){

  # Define cut points

  J = 5 # categorize W into quintiles
  C = rep(0, 4)
  C[1] = qnorm(0.2, mean = mux_hat, sd = sqrt(s2x_hat))
  C[2] = qnorm(0.4, mean = mux_hat, sd = sqrt(s2x_hat))
  C[3] = qnorm(0.6, mean = mux_hat, sd = sqrt(s2x_hat))
  C[4] = qnorm(0.8, mean = mux_hat, sd = sqrt(s2x_hat))
```

```
  # Define a function to categorize W
  fMx <- function(x){
    Mx = vector()
    Mx[1] = ifelse(x < C[1], 1, 0)
    Mx[2] = ifelse((C[1] <= x) & (x < C[2]), 1, 0)
    Mx[3] = ifelse((C[2] <= x) & (x < C[3]), 1, 0)
    Mx[4] = ifelse((C[3] <= x) & (x < C[4]), 1, 0)
    Mx[5] = ifelse(x >= C[4], 1, 0)
    return (Mx)
  }

# Categorize W
cw = matrix(0, ncol = J,nrow = n)
for(i in 1:n){cw[i, ] = fMx(w[i])}

# Run standard logistic regression using glm (no intercept)
thetaw_out = glm(y ~ cw - 1 , family = binomial(link = "logit"))

# Get estimates theta1, .., that_5 and standard errors
thetaw_w = summary(thetaw_out)$coef[1:J]
s.e.thw = summary(thetaw_out)$coef[1:J, 2]

# Calculate the standard error for theta_5 - theta_1
s.e_thetaw_J1 = sqrt(s.e.thw[J]^2+s.e.thw[1]^2 -2*vcov(thetaw_out)[1,J])
# SE of theta1, .., theta_5 and (theta_5-theta_1)
s.e_thetaw_w = c(s.e.thw, s.e_thetaw_J1)
```

```
# Report results
theta.par = c(thetaw_w, thetaw_w[5] - thetaw_w[1])
names(theta.par) = names(s.e_thetaw_w) = c("theta1", "theta2", "theta3",
                                           "theta4", "theta5", "theta5-theta1")
out1 = list(theta.par, s.e_thetaw_w)
names(out1) = c( "theta", "stderr.theta")


return(out1)  }
```

(2) Set parameters values for data generation.

```
# Parameter values
mux = 0 #true mean of X
su2 = 1 #true variance of U
sx2 = 1 #true variance of X


b = log(1.5) #beta_1
a = -0.42 #beta_0


# Sample size
n = 500 # internal data
m = 300 # external data
r = 2 # replicates
```

(3) Generate the external and internal datasets. Note that X in the external data has no replicates. The replicates are generated due to the error term U.

```
# Set seed
set.seed(107852)


# Generate external dataset
X_ext = rnorm(m, mux, sqrt(sx2)) # X is a vector, not a matrix
U_ext = matrix(rnorm(m * r, 0, sqrt(su2)), m, r)
W_ext = matrix(rep(X_ext, r), m, r, byrow = FALSE) + U_ext


# Generate internal dataset
X_int = rnorm(n, mux, sqrt(sx2))
U_int = rnorm(n, 0, sqrt(su2))
W_int = X_int + U_int  # internal data has no replicates


## Generate response y for internal dataset
fHm <- function(x, a, b){1 / (1 + exp( - (a + b * x)))}
pr = fHm(X_int, a, b)
y = vector()
for(i in 1:n){y[i] = rbinom(1, 1, pr[i])}
```

(4) Perform the proposed method using function `ccp`. `Type = "logistic"` needs to be specified for logistic regression.

```
outcome1 = ccp( y = y, W_int = W_int, W_ext = W_ext, Type = "logistic")
#> Summary
#>
#>           Estimate Std. Error  z-value Pr(>|z|)
```

```
#> mu.x      -0.09586    0.06126 -1.56472  0.11765
#> sigma^2.x  0.92557    0.14341  6.45421  0.00000
#> sigma^2.u  0.95460    0.06269 15.22706  0.00000
#> alpha     -0.60244    0.09737 -6.18723  0.00000
#> beta       0.41968    0.14650  2.86469  0.00417
#> theta 1   -1.19816    0.22248 -5.38551  0.00000
#> theta 2   -0.85647    0.12909 -6.63470  0.00000
#> theta 3   -0.64211    0.09783 -6.56358  0.00000
#> theta 4   -0.42744    0.11650 -3.66892  0.00024
#> theta 5   -0.07729    0.20509 -0.37687  0.70627
#>
#>                    Estimate Std. Error z-value Pr(>|z|)
#> theta 5 - theta 1:  1.12087    0.38189 2.93507  0.00333
#>
```

```
outcome1
#> $`theta5-theta1`
#>                    Estimate Std. Error  z-value    Pr(>|z|)
#> theta 5 - theta 1: 1.120869  0.3818889 2.935066 0.003334765
#>
#> $theta
#> [1] -1.19815926 -0.85646628 -0.64211126 -0.42744384 -0.07729019
#>
#> $nuisance
#>        mu.x   sigma^2.x   sigma^2.u       alpha        beta
#> -0.09585517  0.92556798  0.95459883 -0.60243921  0.41967852
#>
```

```
#> $se.theta
#> [1] 0.22247851 0.12908888 0.09782948 0.11650410 0.20508530 0.38188885
#>
#> $se.nuisance
#> [1] 0.06126021 0.14340533 0.06269097 0.09736819 0.14650027
```

(5) Compare results from the proposed method to the naive approach.

```
# Estimate mean of X and variance of X (used for categorization)
mux_hat = mean(W_int)
su2e = mean(apply(W_ext, 1, var))
s2x_hat=max((var(W_int)-su2e), 0.2*var(W_int))
# 0.2*var(W_int) is the common bound to control the variance of X


# Run naive approach
thetaw(y, W_int, mux_hat, s2x_hat)
#> $theta
#>        theta1        theta2        theta3        theta4        theta5
#>    -0.9062404    -0.6523252    -0.7339692    -0.3321338    -0.4364273
#> theta5-theta1
#>     0.4698131
#>
#> $stderr.theta
#>        theta1        theta2        theta3        theta4        theta5
#>     0.1873525     0.2466441     0.2483277     0.2281275     0.1762471
#> theta5-theta1
#>     0.2572237
```

Given in the paper (Blas et al. 2018+), the true $\boldsymbol{\Theta} = (-0.98, -0.64, -0.42, -0.21, 0.14)^{\mathrm{T}}$. Thus, the true $\theta_5 - \theta_1 = 1.12$. The proposed method has the estimate 1.121, while the naive approach provides an estimate 0.470. The results show that ignoring measurement error and applying standard logistic regression directly with respect to W lead to poor inference quality. On the contrary, the proposed method gives consistent estimate as expected.

Now we present the internal-only case for logistic regression.

```
# set seed
set.seed(1029356)


# Generate dataset
X_int = rnorm(n, mux, sqrt(sx2)) # X has no replicates
U_int = matrix(rnorm(n * r, 0, sqrt(su2)), n, r)
W_int = matrix(rep(X_int, r), n, r, byrow = FALSE) + U_int


# Generate response y
fHm <- function(x, a, b){1 / (1 + exp(-(a + b * x)))}
pr = fHm(X_int, a, b)
y = vector()
for(i in 1:n){y[i] = rbinom(1, 1, pr[i])}
```

Run the proposed method:

```
outcome2 = ccp( y = y, W_int = W_int, Type = "logistic")
#> Summary
#>
#>           Estimate Std. Error  z-value Pr(>|z|)
#> mu.x       0.02596    0.05787  0.44863  0.65370
```

```
#> sigma^2.x  1.17524    0.11983  9.80781  0.00000
#> sigma^2.u  0.99799    0.03300 30.24456  0.00000
#> alpha     -0.42276    0.09483 -4.45815  0.00001
#> beta       0.37315    0.10978  3.39913  0.00068
#> theta 1   -0.98403    0.19389 -5.07510  0.00000
#> theta 2   -0.62657    0.11725 -5.34394  0.00000
#> theta 3   -0.41223    0.09576 -4.30478  0.00002
#> theta 4   -0.19753    0.10993 -1.79692  0.07235
#> theta 5    0.14149    0.17514  0.80789  0.41916
#>
#>                    Estimate Std. Error z-value Pr(>|z|)
#> theta 5 - theta 1:  1.12552    0.31994 3.51794  0.00043
#>
```

```
outcome2
#> $`theta5-theta1`
#>                    Estimate Std. Error  z-value     Pr(>|z|)
#> theta 5 - theta 1: 1.125524  0.3199389 3.517935 0.0004349184
#>
#> $theta
#> [1] -0.9840299 -0.6265727 -0.4122278 -0.1975332  0.1414943
#>
#> $nuisance
#>        mu.x   sigma^2.x   sigma^2.u       alpha        beta
#>  0.02596037  1.17523686  0.99798913 -0.42276043  0.37315138
#>
#> $se.theta
```

```
#> [1] 0.19389368 0.11724928 0.09576047 0.10992873 0.17514103 0.31993886
#>
#> $se.nuisance
#> [1] 0.05786590 0.11982662 0.03299731 0.09482861 0.10977846
```

Run standard logistic regression:

```
row_mean_w  = apply(W_int, 1, mean)
mux_hat = mean(row_mean_w)
s2w = apply(W_int, 1, var)
su2e =  mean(s2w)/r
s2x_hat = max(mean((row_mean_w - mux_hat) ^ 2) - su2e,
              0.2 * (mean((row_mean_w - mux_hat) ^ 2)))
thetaw(y, row_mean_w, mux_hat, s2x_hat)
#> $theta
#>        theta1        theta2        theta3        theta4        theta5
#>    -0.6539265    -0.6118015    -0.9075571    -0.1670541     0.1177830
#> theta5-theta1
#>     0.7717095
#>
#> $stderr.theta
#>        theta1        theta2        theta3        theta4        theta5
#>     0.1974192     0.2195430     0.2470264     0.2048366     0.1836577
#> theta5-theta1
#>     0.2696377
```

We observe the similar pattern as shown in the external-internal case. For $\theta_5 - \theta_1 = 1.12$, the naive estimate is 0.772, while the proposed method estimates it as 1.126.

4.5 Real data example

4.5.1 Data

Here we use the Eating at America's Table (EATS) Study (Subar et al. 2001) data as an example to illustrate the usage of the package. The dataset contains 964 participants with multiple 24-hour recalls of diet per each person. Define Fat Density as the percentage of calories coming from fat. We want to use this data to analyze the *relative risk* of being obese, comparing the group of people with the highest level of Fat Density versus people with the lowest level of Fat Density.

First, we load the data. However, we are not allowed to share this data. Thus, we show several lines of the data so you can get a sense of what the data looks like.

```
head(EATSdata_all)
#>          y         w1         w2           w3          w4
#> 1 19.95373 -4.4810374 -1.7696515 -0.427827518  0.08514833
#> 2 29.31301 -1.6681290 -2.1910584  0.388690150 -0.49377952
#> 3 27.36617  1.3471425  1.0063977 -0.009499347  0.65078338
#> 4 18.91162 -3.5224474 -1.0915284 -0.600154597 -0.74033006
#> 5 25.26264  0.7229939  0.4309749  1.251822769 -0.02058828
#> 6 22.12660  0.9097210  1.7703199  0.114664950  0.14936229
```

In this study, we mainly focus on the following variables:

- Y: either the actual body mass index (BMI), or the indicator of obesity, defined as body mass index > 30. The Y shown above is continuous. In linear regression, we use the continuous one; the binary indicator is used for logistic regression.
- X: average daily Fat Density over a long time period (not shown above).

- W: short-term Fat Density, observed in the study. As shown above, W has 4 replicates per person.

For numerical stability, we first preprocess the data:

(1) delete outliers;

(2) centered and standardized W using $(15 * W - 5)/\sqrt{0.5}$.

Then we obtain a dataset with 929 observations. We then randomly selected 200 observations as the external dataset, the remaining 729 observations are the internal dataset. More details are provided within the examples.

Before formally applied the our approach, we need to check the assumptions. In the paper, we showed that it is reasonable to take (a) X to be normally distributed, (b) U to be normally distributed, and (c) X and U to be independent. Hence, here we do not repeat the detailed measurements we have done previously.

We first show the external-internal case, then the internal-only case. Each case contains logistic regression with binary Y and linear regression with continuous Y.

4.5.2 External-internal case

First, we choose variables from the two datasets. We choose the first 2 records from the external dataset, and the 3rd record from the internal dataset.

4.5.2.1 Logistic regression

The response variable BMI has been transferred to a binary variable with threshold 30. In other words, $Y_{\text{original}} > 30 \Longrightarrow Y_{\text{new}} = 1$, which indicates obesity.

```
# select the first 2 records from external data
W_ext = data_external[,2:3]


# select the 3rd record from internal data
W_int = as.matrix(data_internal[,4])


# transfer continuous Y into binary
y = 1*((data_internal[,1])>30)
```

The following table shows the size of the external and internal datasets.

	size	recalls
internal	729	1
external	200	2

Table 4.1: summary for the external-internal case

Now we apply `ccp` with specified `Type = "logistic"`.

```
results = ccp( y = y, W_int = W_int, W_ext = W_ext, Type = "logistic")
#> Summary
#>
#>           Estimate Std. Error   z-value Pr(>|z|)
#> mu.x      -0.17583    0.07196  -2.44330  0.01455
#> sigma^2.x  1.49050    0.29898   4.98520  0.00000
#> sigma^2.u  2.28989    0.11597  19.74567  0.00000
#> alpha     -1.38789    0.09602 -14.45435  0.00000
#> beta       0.28892    0.14310   2.01901  0.04349
```

```
#> theta 1   -1.92230    0.28106  -6.83954  0.00000
#> theta 2   -1.62510    0.16075 -10.10969  0.00000
#> theta 3   -1.43792    0.10224 -14.06373  0.00000
#> theta 4   -1.25018    0.11010 -11.35539  0.00000
#> theta 5   -0.93824    0.22507  -4.16863  0.00003
#>
#>                    Estimate Std. Error z-value Pr(>|z|)
#> theta 5 - theta 1:  0.98406    0.46924 2.09714  0.03598
#>
```

The *log relative risk* - the term `theta 5 - theta 1`- is estimated as 0.984 with p-value = 0.036 and is significant at the 0.05 level.

4.5.2.2 Linear regression

To fit linear regression, we use the scaled BMI as the continuous response. The internal and external data are the same as before.

```
results = ccp(y = y,W_int = W_int, W_ext = W_ext, Type = "linear",
              standardize = FALSE)
#> Summary
#>
#>           Estimate Std. Error  z-value Pr(>|z|)
#> mu.x      -0.17583    0.07196 -2.44330  0.01455
#> sigma^2.x  1.49050    0.29898  4.98520  0.00000
#> sigma^2.u  2.28989    0.11597 19.74567  0.00000
#> alpha      0.03051    0.03906  0.78097  0.43482
#> beta       0.17351    0.05618  3.08850  0.00201
#> theta 1   -0.29627    0.09479 -3.12541  0.00178
```

```
#> theta 2   -0.11266    0.05104 -2.20705  0.02731
#> theta 3   -0.00002    0.03740 -0.00057  0.99954
#> theta 4    0.11263    0.05299  2.12552  0.03354
#> theta 5    0.29709    0.09983  2.97584  0.00292
#>
#>                    Estimate Std. Error z-value Pr(>|z|)
#> theta 5 - theta 1:  0.59336       0.18 3.29639  0.00098
#>
```

Specifying `Type = "linear"` fits a linear regression. Because we already standardized the data, we can simply choose `standardize = FALSE`. The default is `TRUE`.

```
results = ccp(y = y,W_int = W_int, W_ext = W_ext, Type = "linear",
              standardize = FALSE)
#> Summary
#>
#>           Estimate Std. Error  z-value Pr(>|z|)
#> mu.x      -0.17583    0.07196 -2.44330  0.01455
#> sigma^2.x  1.49050    0.29898  4.98520  0.00000
#> sigma^2.u  2.28989    0.11597 19.74567  0.00000
#> alpha      0.03051    0.03906  0.78097  0.43482
#> beta       0.17351    0.05618  3.08850  0.00201
#> theta 1   -0.29627    0.09479 -3.12541  0.00178
#> theta 2   -0.11266    0.05104 -2.20705  0.02731
#> theta 3   -0.00002    0.03740 -0.00057  0.99954
#> theta 4    0.11263    0.05299  2.12552  0.03354
#> theta 5    0.29709    0.09983  2.97584  0.00292
#>
```

```
#>                     Estimate Std. Error z-value Pr(>|z|)
#> theta 5 - theta 1:  0.59336       0.18 3.29639  0.00098
#>
```

The estimate for $\theta_5 - \theta_1$ is 0.59336, which is highly significant with a small p-value 0.00098.

4.5.3 Internal-only case

For the internal-only case, the syntax is similar to what we showed above. However, only the internal data needs to be provided. If the user also provides external data, as long as the internal data has replicates, the external data are ignored.

4.5.3.1 Logistic regression

```
# use first two replicates
W_int = EATSdata_all[, 2:3]
# transfer continuous Y into binary
y = 1*((EATSdata_all[, 1])>30)
```

	size	recalls
internal	929	2
external	0	0

Table 4.2: summary for the internal-only case

```
results = ccp(y = y, W_int = W_int, Type = "logistic")
#> Summary
#>
#>           Estimate Std. Error   z-value Pr(>|z|)
```

```
#> mu.x      -0.25953    0.05177  -5.01296  0.00000
#> sigma^2.x  1.22542    0.12745   9.61513  0.00000
#> sigma^2.u  2.52924    0.05546  45.60287  0.00000
#> alpha     -1.30993    0.08364 -15.66189  0.00000
#> beta       0.35701    0.11386   3.13559  0.00172
#> theta 1   -1.94430    0.20962  -9.27521  0.00000
#> theta 2   -1.61125    0.12339 -13.05830  0.00000
#> theta 3   -1.40146    0.08725 -16.06194  0.00000
#> theta 4   -1.19114    0.09379 -12.69973  0.00000
#> theta 5   -0.84340    0.17019  -4.95564  0.00000
#>
#>                     Estimate Std. Error z-value Pr(>|z|)
#> theta 5 - theta 1:   1.1009    0.34155 3.22324  0.00127
#>
```

4.5.3.2 Linear regression

```
W_int = EATSdata_all[, 2:3]
y = EATSdata_all[, 1]
y = as.numeric(scale(y))
```

```
results =ccp(y = y, W_int = W_int, Type = "linear", standardize = FALSE)
#> Summary
#>
#>           Estimate Std. Error  z-value Pr(>|z|)
#> mu.x      -0.25953    0.05177 -5.01296  0.00000
#> sigma^2.x  1.22542    0.12745  9.61513  0.00000
#> sigma^2.u  2.52924    0.05546 45.60287  0.00000
```

```
#> alpha      0.04654    0.03583  1.29882  0.19401
#> beta       0.17931    0.04353  4.11897  0.00004
#> theta 1   -0.27818    0.07048 -3.94722  0.00008
#> theta 2   -0.10555    0.03932 -2.68424  0.00727
#> theta 3    0.00006    0.03279  0.00196  0.99843
#> theta 4    0.10561    0.04369  2.41743  0.01563
#> theta 5    0.27764    0.07681  3.61481  0.00030
#>
#>                    Estimate Std. Error z-value Pr(>|z|)
#> theta 5 - theta 1:  0.55582    0.13221 4.20416    3e-05
#>
```

For the external-internal and internal-only cases, our approach provides similar estimates. However, the naive logistic regression, which is often used by epidemiology, has different results in the two cases and not similar. Since method comparison is not in the scope of this document, we refer readers to check the paper *Categorizing a Continuous Predictor Subject to Measurement Error* for more details.

5. VIGNETTE: HIGH-DIMENSIONAL BINARY CLASSIFICATION USING DAP

`DAP` package implements the method proposed in Gaynanova and Wang (2017) for high-dimensional binary classification with unequal covariance matrices. In this document, we give an overview of all functions available in the package, introduce the usage of `apply_DAP` in details, provide a concrete illustration via simulations and real data analysis. We also provide algorithm details while referring readers to Gaynanova and Wang (2017) for the methodology itself.

5.1 Introduction

In recent years, high-dimensional binary classification has been widely used in many areas. Aiming at improving classification performance, most literatures focus on improving Quadratic Discriminant Analysis (QDA) through requiring special structures on covariance matrices or precision matrices, typically not computationally efficient and less flexible. Starting from a different aspect, we propose a method named Discriminant Analysis via Projection (DAP) in Gaynanova and Wang (2017) to tackle the problem. We extend the idea of Fisher's Discriminant Analysis, leading to a sparse quadratic classification rule, featured in fast computation, model flexibility and classification accuracy. An overview can be found in the summarized diagram Figure 5.1. More details are presented in Gaynanova and Wang (2017).

This document is a detailed vignette to illustrate how to use the `DAP` (Wang and Gaynanova 2018) package, which is designed for high-dimensional binary classification. The main function is `apply_DAP`. This function is the implementation of DAP method, including learning classification rule as well as performing classification to the test data. This package implements 5-fold cross validation to select the tuning parameter.

There are other existing R packages for classification problems, such as `JGL` (Danaher

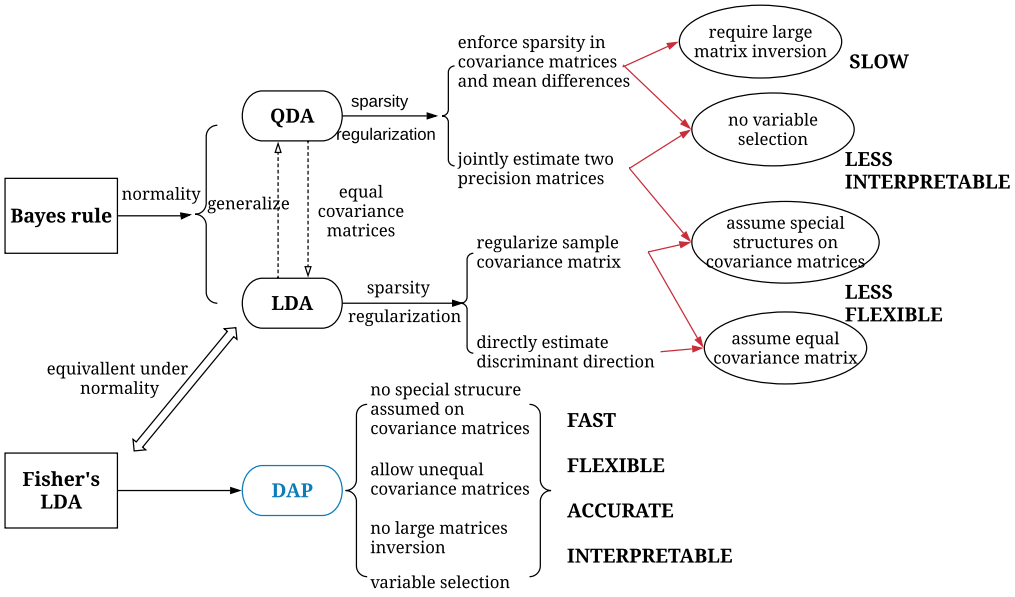

Figure 5.1: Overview

2013), `MGSDA` (Gaynanova 2016), `grpreg` (Breheny and Huang 2015), `RidgeFusion` (Price, Geyer, and Rothman 2014), `sparsediscrim` (Ramey 2017). We use those packages in Gaynanova and Wang (2017) for competitive methods comparison. `DAP` shows its computational advantages universally in high-dimensional analysis. On the one hand, DAP does not require large matrix inversion. On the other hand, the package itself has underlying code written in C, which helps speed up the classification.

The rest of the document is organized as follows. First, we review the optimization algorithm, explain briefly about the methodology. Then, we give an overview of functions provided within the package. Finally, we use a simulation study and a real dataset from Chowdary et al. (2006) to illustrate the usage of `apply_DAP`.

5.1.1 Optimization problem review

Inspired by the Fisher's Discriminant Analysis, we develop sparse quadratic classification rules which depend on the two covariance matrices respectively, without assuming equal covariance matrix. The proposed discriminant vectors can be found by maximizing between group variability versus within (each) group variability, separately. Suppose each group has mean μ_g and covariance matrix Σ_g, $g = 1, 2$, the discriminant vectors can be found as following:

$$v_g = \text{argmax}_{v_g \in R^p} \left\{ \frac{v_g^{\mathrm{T}}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^{\mathrm{T}} v_g}{v_g^{\mathrm{T}} \Sigma_g v_g} \right\} = c_g \Sigma_g^{-1}(\mu_1 - \mu_2) \quad (g = 1, 2),$$

where c_g are constant.

To estimate v_1, v_2 empirically, substituting $\mu_1, \mu_2, \Sigma_1, \Sigma_2$ with their plug-in estimates from the sample leads to the following equation:

$$\widehat{v}_g = \text{argmax}_{v_g \in R^p} \left\{ \frac{v_g^{\mathrm{T}}(\bar{x}_1 - \bar{x}_2)(\bar{x}_1 - \bar{x}_2)^{\mathrm{T}} v_g}{v_g^{\mathrm{T}} S_g v_g} \right\} = c_g S_g^{-1}(\bar{x}_1 - \bar{x}_2) \quad (g = 1, 2),$$

where $\bar{x}_1, \bar{x}_2$ are sample means, and S_1, S_2 are sample covariance matrices, for each group respectively.

Set $\widehat{V} = [\hat{v}_1, \hat{v}_2]$, given a new observation x, the new classification rule labels it as one of the groups which minimize the following quantity:

$$h_{\widehat{V}}(x) = \text{argmin}_{g \in \{1,2\}} \Big\{ (x - \bar{x}_g)^{\mathrm{T}} \widehat{V} (\widehat{V}^{\mathrm{T}} S_g \widehat{V})^{-1} \widehat{V}^{\mathrm{T}} (x - \bar{x}_g) + \log|\widehat{V}^{\top} S_g \widehat{V}| - 2\log(n_g/n) \Big\}.$$

To improve the classification accuracy in high-dimensional setting, we add sparse-inducing penalty into the optimization problem:

$$\widehat{V} = [\widehat{v}_1\ \widehat{v}_2] = \text{argmin}_{v_1,v_2 \in R^p} \left\{ \widehat{L}_1(v_1) + \widehat{L}_2(v_2) + \lambda \text{Pen}(V) \right\},$$

where $\widehat{L}_1(v_1)$ and $\widehat{L}_2(v_2)$ are empirical loss functions. We choose group-lasso penalty to enforce row-sparse structure of $\widehat{V}$. For the ith row of $\widehat{V}$, if both $\hat{v}_{1i}, \hat{v}_{2i}$ are nonzero, the ith feature is selected and will affect the classification rule later. For simplicity, we state the final form of the objective function and refer readers to Gaynanova and Wang (2017) for more details.

$$\text{minimize}_{V=[v_1,v_2] \in R^{p \times 2}} \left\{ \frac{\|X_1 v_1 - 1_{n_1}\|_2^2}{2n_1} + \frac{\|X_2 v_2 + 1_{n_2}\|_2^2}{2n_2} + \lambda \sum_{j=1}^{p} \sqrt{v_{1j}^2 + v_{2j}^2} \right\}, \tag{5.1}$$

where $X \in R^{n \times p}$ is column-centered.

The methodology itself has several advantages. First, it is a convex optimization problem, and thus it is easier to solve compared to nonconvex optimization problem. Second, the empirical loss functions are invariant under the linear transformation of the data and bounded from below. Further, we only assume $V = [v_1\ v_2]$ has a row-sparse structure. In other words, we do not require any special structures on μ_1, μ_2, Σ_1, Σ_2 respectively, but only requires the difference between μ_1, μ_2, after being adjusted by precision matrices (Σ_g^{-1}), is sparse. This sparsity assumption is weaker and more realistic. Moreover, in the classification rule, the only inversion, $\{(V^{\mathrm{T}} S_g \widehat{V})^{-1}\}_{2\times 2}$, is most likely non-singular.

Although extended from Fisher's Discriminant Analysis, this rule has connections to other popular classification rules. DAP coincides with the sample plug-in quadratic discriminant rule to $\widehat{V}^{\mathrm{T}} x$ instead of x. Sparse LDA (Gaynanova, Booth, and Wells 2016, Mai, Zou, and Yuan (2012)) can be viewed as a very special case of DAP when the two discriminant vectors

v_1, v_2 are in the same direction. See *Proposition 2* (Gaynanova and Wang 2017).

5.1.2 Algorithm

We now illustrate how to use block-coordinate descent algorithm to solve the optimization problem presented above. Block-coordinate descent algorithm is widely used for optimization. In each iteration, it updates one coordinate block to minimize the objective function, with other blocks fixed. For example, let $f(\cdot)$ be the optimized function, in the $(k+1)$th iteration, we update the jth block $\mathbf{x}_j$ as following:

$$\mathbf{x}_j^{(k+1)} = \underset{\mathbf{u}}{\operatorname{argmin}} f(\mathbf{x}_1^{(k+1)}, ..., \mathbf{x}_{j-1}^{(k+1)}, \mathbf{u}, \mathbf{x}_{j+1}^{(k)}, ..., \mathbf{x}_p^{(k)}).$$

Now, let us first derive the block-update for $v_j = [v_{1j}, v_{2j}]$, $j = 1, ..., p$. To solve the convex optimization problem, we use Karush-Kunn-Tucker (KKT) conditions (Boyd and Vandenberghe 2004, Chapter 5). Taking derivative to eq (1) with respect to v_1, v_2 separately, we get

$$n_1^{-1} X_{1j}^{\mathrm{T}} X_{1j} v_{1j} = n_1^{-1} X_{1j}^{\mathrm{T}} (1_{n_1} - \sum_{k \neq j} v_{1k} X_{1k}) - \lambda u_{1j},$$
$$n_2^{-1} X_{2j}^{\mathrm{T}} X_{2j} v_{2j} = n_2^{-1} X_{2j}^{\mathrm{T}} (-1_{n_2} - \sum_{k \neq j} v_{2k} X_{2k}) - \lambda u_{2j};$$

where X_1, X_2 are training data labeled in two groups, respectively, and $u_j = (u_{1j}, u_{2j})^{\mathrm{T}}$ is the subgradient of $\sqrt{(v_{1j}^2 + v_{2j}^2)}$

$$u_j = \begin{cases} v_j / \|v_j\|_2, & \text{if } \|v_j\|_2 \neq 0; \\ \in \{u : \|u\|_2 \leq 1\}, & \text{if } \|v_j\|_2 = 0. \end{cases}$$

Satisfying all KKT conditions, the solution is guaranteed to be optimal. However, $v_j = [v_{1j}, v_{2j}]$ typically have no closed form to update it if $n_1^{-1}X_1^{\mathrm{T}}X_1 \neq n_2^{-1}X_2^{\mathrm{T}}X_2$. Although a line search along the coordinate directions can be used to search for the updates, we scale the data first such that $n_1^{-1}\mathrm{diag}(X_1^{\mathrm{T}}X_1) = n_2^{-1}\mathrm{diag}(X_2^{\mathrm{T}}X_2) = 1_p$. This enable us to find a closed form of block-update. Before performing classification, we backscale $\hat{v}_1$ and $\hat{v}_2$.

The algorithm can be summarized as below, where r_j is the residual term, $m_{\max}$ is the maximum iteration number, ϵ controls the convergence criterion, and $a_+ = \max(0, a)$.

Algorithm 1 Block-coordinate descent algorithm for problem

Given: $m = 1, V^{(0)}, \epsilon > 0$, $m_{\max}$.
repeat
  $V^{(m)} \leftarrow V^{(m-1)}$
  for $j = 1$ to p **do**
    $r_j \leftarrow \begin{pmatrix} n_1^{-1}X_{1j}^{\mathrm{T}}(1_{n_1} - \sum_{l=1}^{p} v_{1l}X_{1l}) \\ n_2^{-1}X_{2j}^{\mathrm{T}}(-1_{n_2} - \sum_{l=1}^{p} v_{2l}X_{2l}) \end{pmatrix}$
    $V_j^{(m)} \leftarrow (1 - \lambda/\|v_j + r_j\|_2)_+ (v_j + r_j)$
  end for
  $m \leftarrow m + 1$
until $m = m_{\max}$ or $V^{(m)}$ satisfies $\max_i \|V_i^{(m)} - V_i^{(m-1)}\|_2 < \epsilon$

As described above in Algorithm 1, the block-coordinate descent algorithm iteratively updates $[v_1, v_2]$, as well as residuals $[r_1, r_2]$ until convergence. Since the optimization problem eq 5.1 is convex and bounded from below, the algorithm is guaranteed to converge to the global minimum.

5.1.3 Functions overview

`DAP` contains several functions: `apply_DAP`, `standardizeData`, `solve_DAP_C`, `solve_DAP_seq`, `cv_DAP`, and `classify_DAP`. Figure 2 gives an overview of the relationship among them.

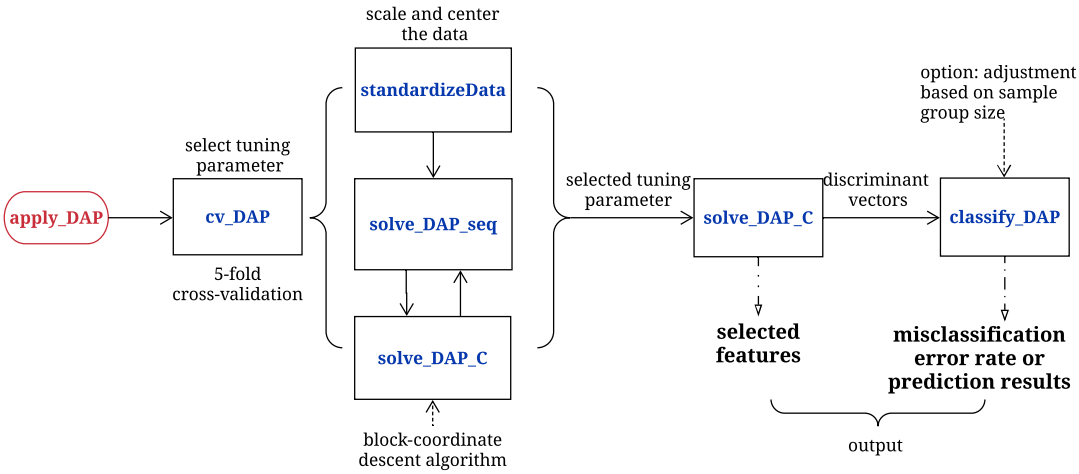

Figure 5.2: Functions overview

`apply_DAP`, serving as a wrapper function, takes training data, including their labels, and test data as input. Within the function, it calls other functions as illustrated in the Figure 5.2. There are two different types of outcomes provided by `apply_DAP`: selected features, as well as misclassification rate or predicted results. When the label of the test data is provided, `apply_DAP` returns the misclassification rate; otherwise, it returns predicted labels for the test data. There are some options which can be specified for user's preference, such as seed number, maximum iteration number, etc.

Now we provide a brief overview for other functions called within `apply_DAP`, starting from the basic ones such as `standardizeData`, `solve_DAP_C`, to some higher level or wrapper functions, e.g. `solve_DAP_seq`, `cv_DAP`, and finally `classify_DAP`, the one implements classification rule.

The basic function, being used several times in the classification, is `standardizeData`. It

provides centering and scaling. Recall that in the optimization section, we require X to be column-centered in the objective function. Centering can be done by this function to enable the training data $X_{n\times p}$ has column mean zero. In the algorithm section, we perform scaling to get a closed form of block-update. Here scaling is performed within each group, making the columns of X_1 and X_2 to have Eucledean norm equal to one, respectively.

`solve_DAP_C`, the fundamental function, is the actual function to implement the algorithm and called by other functions. This function calls C code underlying and use block-coordinate descent algorithm to solve the problem.The function requires scaled input X_1 and X_2, a given value of the tuning parameter λ. Users can also provide the initial value for matrix V as a warm start. The convergence threshold and the maximum number of iterations can be controlled via `eps` and `maxiter`.

`solve_DAP_seq` takes a sequence of λ and then assigns each of them to `solve_DAP_C` to get $\widehat{V}$. The threshold for the number of selected variable is n, the total sample size. The sequence of λ is sorted in an ascending order; thus whenever the number of non-zero rows in $\widehat{V}$ exceeds n, we stop considering the following λ in the sequence.

`cv_DAP` implements 5-fold cross-validation to select the tuning parameter λ. If the sequence of λ is not provided by users, it can be generated in `apply_DAP`. After `cv_DAP` provided a matrix of misclassification rate corresponding to each λ, the tuning parameter with smallest error is selected to find $\widehat{V}$. `classify_DAP` is called to implement classification on the test dataset and return the results to the user as the outcome from `apply_DAP` finally. When the group size $n_1 \neq n_2$, `classify_DAP` can be adjusted by the group size. Users who do not want to perform the adjustment can simply set `prior=FALSE`. `prior` controls the prior probabilities of class membership. When using default setting `prior = TRUE`, the class proportions in the training dataset are used. Otherwise, the prior probability for each group is equal to 0.5. Note that covariates X need to be centered and standardized within folds. Back-standardization of $\widehat{V}$ is performed before classifying the test dataset.

5.1.3.1 Get started

Now let us start with installation. The package can be installed using `install.packages`.

```
if("DAP" %in% rownames(installed.packages()) == FALSE) {
  install.packages("DAP", repos = "http://cran.us.r-project.org")}
```

After the package is installed, the next step is to use `library` to make it accessible for R.

```
library(DAP)
```

Once the package has been loaded, one can call the main function `apply_DAP` as follows.

```
apply_DAP(xtrin, ytrain, xtest, ytest = NULL, lambda_seq = NULL,
           n_lambda = 50, maxmin_ratio = 0.1, nfolds = 5,
          eps = 1e-4, maxiter = 10000, myseed = 1001, prior = TRUE)
```

To check the available functions in `DAP` or the usage of a specific function, you can either use `help` or `??`.

```
help( package = "DAP" )
??apply_DAP
```

The former command gives a brief summary of all functions in the package, while the later one offers more detailed information for function `apply_DAP`.

5.1.4 Simulation example

In this section, we present a simulation study. The simulation design is as follows.

- Both training and test datasets have sample size $n_1 = n_2 = 100$, respectively;

- Number of features $p \in \{100, 500\}$;
- $X_1|(Y = 1) = N(\mu_1, \Sigma_1)$ and $X_2|(Y = 2) = N(\mu_2, \Sigma_2)$;
- Group means $\mu_1 = 0_p$, $\mu_2 = (1_5, -1_5, 0_{p-10})$
- Covariance structures: $\Sigma_1 = I_p$, Σ_2 has the block-equicorrelation structure with block size equals 100 and $\rho = 0.8$.

$$\Sigma_2 = \begin{pmatrix} \rho I_{100} + (1-\rho) 1_{100} 1_{100}^{\mathrm{T}} & 0 \\ 0 & I_{p-100} \end{pmatrix}.$$

(1) Set parameters as described above.

```
p = 100 # number of features
n1 = 100 # size of group 1 in training data
n2 = n1 # size of group 2 in training data
n_test=100 #size of group 1 and group 2 respectively in test data
```

(2) Create mean and covariance structures.

```
# Create equicorrelation matrix function
equicor <- function(p, rho, sblock){
  Sigma = matrix(0, p, p)
  Sigma[1:sblock, 1:sblock] = rho
  diag(Sigma) = 1
  return(Sigma)
}

# Create mean and covariance structures
```

```
mu1 = rep(0, p)
mu2 = c(rep(1, 5), rep(-1, 5), rep(0, p-10))
Sigma2 = equicor(p, rho = 0.8, sblock = 100)
Sigma1 = diag(p)
```

(3) Generate the training and test data using the `mvrnorm` function from the R package `MASS`.

```
library(MASS)
set.seed(20180509)

#training data

ytrain = c(rep(1, n1),rep(2,n2))
x1 = mvrnorm(n = n1, mu = mu1, Sigma = Sigma1)
x2 = mvrnorm(n = n2, mu = mu2, Sigma = Sigma2)
xtrain = rbind(x1, x2)

#test data

x1_test = mvrnorm(n = n_test, mu = mu1, Sigma = Sigma1)
x2_test = mvrnorm(n = n_test, mu = mu2, Sigma = Sigma2)
xtest = rbind(x1_test, x2_test)
ytest = c(rep(1, n_test), rep(2, n_test))
```

(4) Implement the proposed method by `apply_DAP`.

```
DAP_p100 = apply_DAP(xtrain, ytrain, xtest, ytest, n_lambda = 50,
                     maxiter = 3000,  eps = 1e-4)
#> 12345
DAP_p100
#> $error
#> [1] 0.03
#>
#> $features
#> [1] 9
#>
#> $features_id
#> [1]  1  2  3  4  5  7  8  9 10
```

Note: the outcome `#> 12345` indicates it used 5-fold cross-validation.

(5) Check the true discriminant vectors and report the id of the nonzero features.

```
v1 = solve(Sigma1) %*% (mu1 - mu2)
v2 = solve(Sigma2) %*% (mu1 - mu2)
v1_id = which(abs(v1) > 1e-4)
v2_id = which(abs(v2) > 1e-4)
v1_id
#>  [1]  1  2  3  4  5  6  7  8  9 10
v2_id
#>  [1]  1  2  3  4  5  6  7  8  9 10
```

For this dataset, the proposed method `DAP` achieves misclassification rate is 0.03, using 9 features selected out of 100. Also, the variables in the true set are all selected except the

6th feature. This result is only based on one randomly generated dataset. When we change the seed and generate another dataset, the results are different.

```
set.seed(201805010)

# Generate training data

ytrain = c(rep(1, n1), rep(2, n2))
x1 = mvrnorm(n = n1, mu = mu1, Sigma = Sigma1)
x2 = mvrnorm(n = n2, mu = mu2, Sigma = Sigma2)
xtrain = rbind(x1, x2)

# Generate test data

x1_test = mvrnorm(n = n_test, mu = mu1, Sigma = Sigma1)
x2_test = mvrnorm(n = n_test, mu = mu2, Sigma = Sigma2)
xtest = rbind(x1_test, x2_test)
ytest = c(rep(1, n_test), rep(2, n_test))

# Implement DAP
apply_DAP(xtrain, ytrain, xtest, ytest, n_lambda = 50, maxiter = 3000,
          eps = 1e-4)
#> 12345
#> $error
#> [1] 0.015
#>
#> $features
#> [1] 17
```

```
#>
#> $features_id
#>  [1]  1  2  3  4  5  6  7  8  9 10 16 55 61 67 86 93 98
```

Interestingly, for this particular dataset `DAP` selects 17 features, including all true features, and achieves a lower misclassification rate as 0.015.

Now we present the results for $p = 500$ directly. For simplicity, we do not show the model and data generation procedures, which remain the same except changing p to be 500. Results are shown as follows.

```
DAP_p500 = apply_DAP(xtrain, ytrain, xtest, ytest, n_lambda = 50,
                     maxiter = 3000, eps = 1e-4)
#> 12345
DAP_p500
#> $error
#> [1] 0.01
#>
#> $features
#> [1] 10
#>
#> $features_id
#>  [1]  1  2  3  4  5  6  7  8  9 10
```

Similarly, we print out the true discriminant vectors v_1, v_2 for the new model.

```
v1 = solve(Sigma1) %*% (mu1 - mu2)
v2 = solve(Sigma2) %*% (mu1 - mu2)
v1_id = which(abs(v1) > 1e-4)
```

```
v2_id = which(abs(v2) > 1e-4)
v1_id
#>  [1]  1  2  3  4  5  6  7  8  9 10
v2_id
#>  [1]  1  2  3  4  5  6  7  8  9 10
```

In this dataset, `DAP` achieves a misclassification rate as 0.01. Further, all true features have been selected by `DAP` and no extra features have been included in the classification rule. In other words, in this dataset, the feature set that selected by `DAP` is exactly the true set.

5.1.5 Real data example

5.1.5.1 Preprocess data

Let us now try an example using *chowdary* data from *datamicroarray* (Ramey 2016). We recommend readers to download *chowdary.RData* from Raymey's Github and save it in the current working directory. You can either use function `load` or directly open the file *chowdary.RData* in R or Rstudio.

- Let us first look at the data by displaying a small subset. y is labeled as *breast* or *colon*, indicating where the tissue comes from. Table 5.1 shows that $n_1 = 62, n_2 = 42$.

breast	colon
62	42

Table 5.1: summary for the response y.

x contains gene expression profiles; Table 5.2 displays a subset of x.

1007_s_at	1053_at	117_at	121_at	1255_g_at	1294_at	1316_at	1320_at
244.8	15.0	18.9	44.7	4.6	24.5	7.5	8.4
166.5	20.8	14.7	47.1	1.5	56.0	8.8	8.5
226.2	16.8	12.6	57.9	5.4	63.2	5.3	5.8
258.9	19.5	18.9	65.2	7.8	62.3	13.1	7.3
138.2	14.5	12.7	45.9	3.0	42.0	11.4	3.8

Table 5.2: A subset for x: the first 8 gene expression profiles for the first 5 observations.

- Then we assign x and y, as well as remove the large dataset to save computational memory space. Note that it is unnecessary to use `as.matrix` for x in *chowdary*, because `chowdary$x` is numeric already. However, in other datasets, e.g., *gravier*, if x or y is not numeric, the function will return error message. To be consistent, here we also use `as.matrix` and `as.numeric` for x and y respectively.

```
x = as.matrix(chowdary$x)
y = as.numeric(chowdary$y)
rm(chowdary)
```

- We here show some basic information about this dataset.

```
p_all = ncol(x)
n = length(y)
print(paste("This data set has", n, "observations,", "labeled as 2 groups."))
#> [1] "This data set has 104 observations, labeled as 2 groups."
print(paste("Each observation has", p_all, "features."))
#> [1] "Each observation has 22283 features."
```

- We split the data into 5 parts, using 80% for training and 20% for test.

```
msep = 5
set.seed(293865) # set a seed
id = 1:n
  for (i in 1:2) {
    id[y == i] = sample(rep(seq_len(msep), length.out = sum(y == i)))
  }
```

- Set training and test data

```
  xtrain = x[id != 1, ]
  ytrain = y[id != 1]
  xtest = x[id == 1, ]
  ytest = y[id == 1]
  n1 = sum(ytrain == 1)
  n2 = sum(ytrain == 2)
```

- Here we show what has been done in the paper Gaynanova and Wang (2017). $p = 1000$ features have been selected with largest absolute value of the two-sample t-statistic on the training data. This approach is also used in Cai and Liu (2011).

```
  # Select p features with largest value of the test statistic
  p = 1000
  x1s = scale(xtrain[ytrain==1,], scale=F)
  x2s = scale(xtrain[ytrain==2,], scale=F)
  t_stat = abs((attr(x1s, which="scaled:center") - attr(x2s, which
                  = "scaled:center")))/sqrt(colSums(x1s^2)/(n1*(n1-1))
                  + colSums(x2s^2)/(n2*(n2-1)))
```

- Re-form the training and test dataset.

```
  r = order(t_stat, decreasing = TRUE)
  index = r[1:p]
  xtrain = xtrain[, index]
  xtest = xtest[, index]
```

5.1.5.2 Apply DAP

If `ytest` is provided,

- Use `apply_DAP`.

```
  outcome = apply_DAP(xtrain, ytrain, xtest, ytest, n_lambda = 50,
                      maxiter = 3000, eps = 1e-4)
#> 12345
```

- Check results. `apply_DAP` returns three items:

(1) `error` is the classification error examined by test data;

(2) `features` is the number of features selected by DAP;

(3) `features\_id` is the id or index of selected features.

```
outcome
#> $error
#> [1] 0
#>
#> $features
#> [1] 9
#>
```

```
#> $features_id
#> [1]  1  2 15 44 45 55 56 59 82
```

Among 1000 features, `DAP` selects 9 features to achieve misclassification rate 0 in the test data.

If `ytest` is not provided,

- `apply_DAP` returns a vector indicating predicted labels for the test data.

```
apply_DAP(xtrain, ytrain, xtest, n_lambda = 50, maxiter = 3000, eps = 1e-4)
#> 12345
#> $ypred
#>  [1] 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2
#>
#> $features
#> [1] 9
#>
#> $features_id
#> [1]  1  2 15 44 45 55 56 59 82
```

From the results shown above, `DAP` has a very low misclassification rate nearly 0. However, it is not a surprise. As shown in the paper, other competitors also achieved low misclassification rate. It indicates that this dataset has relatively simple structure, or strong signals that can be easily detected.

5.1.5.3 Time it

- To measure the computational time, we use the R package `microbenchmark` (Mersmann 2015). It measures in nanosecond, more accurately than using `system.time`.

```
# Install and import the package
if("microbenchmark" %in% rownames(installed.packages()) == FALSE) {
  install.packages("microbenchmark", repos = "http://cran.us.r-project.org")}
library(microbenchmark)
```

- `microbenchmark` measures the calculation procedure several times, and then reports a summary. In the example below, we set `time = 10` and find the median time over 10 repetitions is only 5.03 seconds. `microbenchmark` can also used for comparing several approaches at the same time. In Gaynanova and Wang (2017) (Table 2), we present timing results comparison among all competitors. The proposed method `DAP` is the fastest in high-dimensional settings.

```
# use microbenchmark to record the time
res = microbenchmark(apply_DAP(xtrain, ytrain, xtest, ytest, n_lambda = 50,
                          maxiter = 3000, eps = 1e-4), times = 10)
#> 12345123451234512345123451234512345123451234512345

print(res)
#> Unit: seconds
#>
#>  apply_DAP(xtrain, ytrain, xtest, ytest, n_lambda = 50, maxiter = 3000,
             eps = 1e-04)
#>       min       lq     mean   median       uq      max neval
#>  3.987161 4.044601 4.140979 4.078935 4.325035 4.361079    10
```

6. CONCLUSIONS

Motivated by the real problems existing in public health, we explore nutrient-based analysis of disease risk and genetic-based discriminant analysis of complex human diseases. We study the effect of measurement error existing in a misspecified model, as well as propose new classification rules for high-dimensional binary classification to improve the computational efficiency without sacrificing the classification accuracy.

Categorizing a continuous predictor is a common practice in epidemiology, because the categorical model is thought to be more interpretable and robust. We propose a method to get consistent estimators and improve qualified inferences when measurement error exists in misspecified models. We also discuss some basic assumptions for our method and point out that the proposed method is very general with realistic assumptions. The proposed method does not require specific distribution for X given (W, Z), e.g., normal distribution, and the true risk model is not restricted to be logistic regression. However, current literature avoid discussing the basic issues we presented, but made some implausible assumptions about the true models, e.g., assuming the true risk model is based on the categorized truth. Besides, we also discuss other approaches as alternatives with respect to the problem of model misspecification and measurement error. First, one may apply Simulation-extrapolation to get approximated estimates. However, differential measurement error may cause complications. Another potential difficulty is to estimate the misclassification rate due to the measurement error. Second, one possibility is to avoid the model misspecification brought by categorizing a continuous predictor. Using Bsplines instead of the linear term can achieve similar goal to avoid extreme comparison for the risk between the lowest and the highest values of risk predictor. There are approaches focusing on Bsplines and measurement error using regression calibration; however, the interpretations of the model is not fully comparable with common practice would have been done in epidemiology.

High-dimensional binary classification has been studies for several years, while most literatures extend either QDA or LDA. Starting from a different aspect, using the projection idea, we build a quadratic classification rule showing computational advantages with the increase in the number of features p. Furthermore, this benefit rewards a another potential usage as a screening tool. This is especially attractive and useful for genomic information analysis. For the future work, we may explore the screening properties and develop this methodology into multi-group cases.

Besides exploring the effect of extrinsic factors like nutrition and the intrinsic factors like gene, the next interesting project is analyzing the effect of gene-environment interactions on complex human diseases. When focusing on rare diseases, which means the disease rate is usually lower than 5% in the source population, random sampling is cost-prohibitive and time consuming. Thus, case-control studies are often used, in which two groups of diseased and non-diseased people are sampled independently from their own populations, separately. However, this different sampling scheme leads to the question that whether we can still apply the common techniques which would have been used if we get randomly sampled data. Prentice and Pyke (1979) point out that ignoring the sampling scheme and perform logistic regression lead to consistent estimator for nonintercept parameters. However, the interaction terms may have large variance, thus leading to low power. Another way to analyze the data is to use retrospective likelihood framework based on the sampling scheme feature.

We are going to adopt the retrospective likelihood framework and use profile technique, aiming to decrease the variance of estimates, especially for the gene-environment interaction terms. The only one assumption has been made is: in the source population, the marginal distribution of inherited genetic and extrinsic environmental factors included in the model are independent. Under this realistic and general assumption, our goal is to improve the estimation efficiency for the interaction term without any further assumptions. With fully unspecified marginal distribution for intrinsic and extrinsic factors, this approach enables us to analyze complicated cases, such as multivariate genetic or environmental factors -

usually hard to model by a parametric distribution, let along when the multivariate genetic information contains unclear correlations.

In this manuscript, we focus on common problems in public health and develop novel methodology to analyze extrinsic and intrinsic factors and their effects to complex human diseases from different perspectives. We show how to solve the problem of misspecified model with measurement error and give consistent estimators with asymptotic theorems. Moreover, we propose sparse quadratic classification rules for high-dimensional binary classification problem. Further, we provide two vignettes to illustrate the R packages developed for the proposed methods, supporting the major projects from a practical aspect.

REFERENCES

Arem, H., Reedy, J., Sampson, J., Jiao, L., Hollenbeck, A. R., Risch, H., Mayne, S. T., and Stolzenberg-Solomon, R. Z. (2013). The Healthy Eating Index 2005 and risk for pancreatic cancer in the NIH–AARP Study. *Journal of the National Cancer Institute*, 105, 1298–1305.

Bach, F. R. (2008). Consistency of the Group Lasso and Multiple Kernel Learning. *Journal of Machine Learning Research*, 9, 1179–1225.

Barber, R. F. and Drton, M. (2010). Exact block-wise optimization in group lasso and sparse group lasso for linear regression. *arXiv.org*, .

Berry, S. M., Carroll, R. J., and Ruppert, D. (2002). Bayesian smoothing and regression splines for measurement error problems. *Journal of the American Statistical Association*, 97, 160–169.

Blas, B., Wang, T., Su, T., Kipnis, V., Dodd, K., and Carroll, R. (2018+). Categorizing a continuous predictor subject to measurement error. *(submitted)*, .

Boyd, S. P. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge Univ Press, Cambridge.

Breheny, P. and Huang, J. (2015). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and computing*, 25, 173–187.

Buonaccorsi, J. P. (2010). *Measurement Error: Models, Methods and Applications*. Chapman & Hall.

Cai, T. and Liu, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association*, 106, 1566–1577.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*. Chapman and Hall.

Chaix, B., Kestens, Y., Duncan, D. T., Brondeel, R., Méline, J., Aarbaoui, T. E., Pannier, B.,

and Merlo, J. (2016). A gps-based methodology to analyze environment-health associations at the trip level: case-crossover analyses of built environments and walking. *American Journal of Epidemiology*, 184, 579–589.

Chen, H.-W., Huang, H.-C., Lin, Y.-S., Chang, K.-J., Kuo, W.-H., Hwa, H.-L., Hsieh, F.-J., and Juan, H.-F. (2008). Comparison and identification of estrogen-receptor related gene expression profiles in breast cancer of different ethnic origins. *Breast cancer : basic and clinical research*, 1, 35–49.

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

Chowdary, D., Lathrop, J., Skelton, J., Curtin, K., Briggs, T., Zhang, Y., Yu, J., Wang, Y., and Mazumder, A. (2006). Prognostic gene expression signatures can be measured in tissues collected in RNAlater preservative. *The Journal of molecular diagnostics : JMD*, 8, 31–39.

Cook, J. R. and Stefanski, L. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89, 1314–1328.

Cordy, C. B. and Thomas, D. R. (1997). Deconvolution of a distribution function. *Journal of the American Statistical Association*, 92, 1459–1465.

Danaher, P. (2013). *JGL: Performs the Joint Graphical Lasso for sparse inverse covariance estimation on multiple classes*. R package version 2.3.

Danaher, P., Wang, P., and Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society, Ser. B*, 76, 373–397.

Devanarayan, V. and Stefanski, L. A. (2002). Empirical simulation extrapolation for mea-

surement error models with replicate measurements. *Statistics & Probability Letters*, 59, 219–225.

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97, 77–87.

Evenson, K. R., Wen, F., and Herring, A. H. (2016). Associations of accelerometry-assessed and self-reported physical activity and sedentary behavior with all-cause and cardiovascular mortality among us adults. *American Journal of Epidemiology*, 184, 621–632.

Flegal, K. M., Keyl, P. M., and Nieto, F. J. (1991). Differential misclassification arising from nondifferential errors in exposure measurement. *American Journal of Epidemiology*, 134, 1233–1246.

Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association*, 84, 165–175.

Ganguli, B., Staudenmayer, J., and Wand, M. P. (2005). Additive models with predictors subject to measurement error. *Australian & New Zealand Journal of Statistics*, 47, 193–202.

Gaynanova, I. (2016). *MGSDA: Multi-Group Sparse Discriminant Analysis*. R package version 1.4.

Gaynanova, I., Booth, J. G., and Wells, M. T. (2016). Simultaneous sparse estimation of canonical vectors in the $p >> N$ setting. *Journal of the American Statistical Association*, 111, 696–706.

Gaynanova, I., Booth, J. G., and Wells, M. T. (2017). Penalized Versus Constrained Generalized Eigenvalue Problems. *Journal of Computational and Graphical Statistics*, 26, 379–387.

Gaynanova, I. and Kolar, M. (2015). Optimal variable selection in multi-group sparse discriminant analysis. *Electronic Journal of Statistics*, 9, 2007–2034.

Gravier, E., Pierron, G., Vincent-Salomon, A., Gruel, N., Raynal, V., Savignoni, A.,

De Rycke, Y., Pierga, J.-Y., Lucchesi, C., Reyal, F., Fourquet, A., Roman-Roman, S., Radvanyi, F., Sastre-Garau, X., Asselain, B., and Delattre, O. (2010). A prognostic DNA signature for T1T2 node-negative breast cancer patients. *Genes, chromosomes & cancer*, 49, 1125–1134.

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika*, 98, 1–15.

Gustafson, P. (2004). *Measurement Error and Misclassication in Statistics and Epidemiology*. Chapman and Hall/CRC.

Holst, F. (2016). Estrogen receptor alpha gene amplification in breast cancer: 25 years of debate. *World journal of clinical oncology*, 7, 160–173.

Holst, F., Stahl, P. R., Ruiz, C., Hellwinkel, O., Jehan, Z., Wendland, M., Lebeau, A., Terracciano, L., Al-Kuraya, K., Jänicke, F., Sauter, G., and Simon, R. (2007). Estrogen receptor alpha (ESR1) gene amplification is frequent in breast cancer. *Nature genetics*, 39, 655–660.

Hsu, D., Kakade, S. M., and Zhang, T. (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17, no. 52–6.

Huang, J., Breheny, P., and Ma, S. (2012). A Selective Review of Group Selection in High-Dimensional Models. *Statistical Science*, 27, 481–499.

Iwamoto, T., Booser, D., Valero, V., Murray, J. L., Koenig, K., Esteva, F. J., Ueno, N. T., Zhang, J., Shi, W., Qi, Y., Matsuoka, J., Yang, E. J., Hortobagyi, G. N., Hatzis, C., Symmans, W. F., and Pusztai, L. (2012). Estrogen receptor (ER) mRNA and ER-related gene expression in breast cancers that are 1 *JCO*, 30, 729–734.

Kolar, M. and Liu, H. (2015). Optimal feature selection in high-dimensional discriminant analysis. *IEEE Transactions on Information Theory*, 61, 1063–1083.

Laenkholm, A.-V., Knoop, A., Ejlertsen, B., Rudbeck, T., Jensen, M.-B., Müller, S., Lykkesfeldt, A. E., Rasmussen, B. B., and Nielsen, K. V. (2012). ESR1 gene status correlates with estrogen receptor protein levels measured by ligand binding assay and

immunohistochemistry. *Molecular oncology*, 6, 428–436.

Laurent, B. and Massart, P. (2000). Adaptive Estimation of a Quadratic Functional by Model Selection. *Annals of Statistics*, 28, 1302–1338.

Le, Y. and Hastie, T. J. (2014). Sparse Quadratic Discriminant Analysis and Community Bayes. *arXiv.org*, .

Lederer, W. and Küchenhoff, H. (2006). A short introduction to the simex and mcsimex. *The Newsletter of the R Project Volume 6/4, October 2006*, page 26.

Li, Q. and Shao, J. (2015). Sparse quadratic discriminant analysis for high dimensional data. *Statistica Sinica*, 25, 457–473.

Li, Y. and Ngom, A. (2013). Nonnegative least-squares methods for the classification of high-dimensional biological data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 10, 447–456.

Mai, Q., Zou, H., and Yuan, M. (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, 99, 29–42.

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

Mersmann, O. (2015). *microbenchmark: Accurate Timing Functions*. R package version 2.1.

Muirhead, R. J. (1982). *Aspects of multivariate statistical theory*. John Wiley and Sons, Inc., New York.

Niu, Y. S., Hao, N., and Dong, B. (2015). A new reduced-rank linear discriminant analysis method and its applications. *arXiv preprint arXiv:1511.00282*, .

Obozinski, G., Wainwright, M. J., and Jordan, M. I. (2011). Support union recovery in high-dimensional multivariate regression. *Annals of Statistics*, 39, 1–47.

Pham, T. H., Ormerod, J. T., and Wand, M. P. (2013). Mean field variational bayesian inference for nonparametric regression with measurement error. *Computational Statistics & Data Analysis*, 68, 375–387.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control

studies. *Biometrika*, 66, 403–411.

Price, B. S. (2014). *RidgeFusion: R Package for Ridge Fusion in Statistical Learning*. R package version 1.0-3.

Price, B. S., Geyer, C. J., and Rothman, A. J. (2014). Ridge Fusion in Statistical Learning. *Journal of Computational and Graphical Statistics*, 24, 439–454.

Ramey, J. A. (2016). *datamicroarray: Collection of Data Sets for Classification*. https://github.com/ramhiser/datamicroarray, http://ramhiser.com.

Ramey, J. A., Stein, C. K., Young, P. D., and Young, D. M. (2016). High-dimensional regularized discriminant analysis. *arXiv.org*, .

Reedy, J., Wirfält, E., Flood, A., Mitrou, P. N., Krebs-Smith, S. M., Kipnis, V., Midthune, D., Leitzmann, M., Hollenbeck, A., Schatzkin, A., et al. (2010). Comparing 3 dietary pattern methods – cluster analysis, factor analysis, and index analysis – with colorectal cancer risk: the NIH–AARP Diet and Health Study. *American Journal of Epidemiology*, 171, 479–487.

Reedy, J. R., Mitrou, P. N., Krebs-Smith, S. M., Wirfält, E., Flood, A. V., Kipnis, V., Leitzmann, M., Mouwand, T., Hollenbeck, A., Schatzkin, A., and Subar, A. F. (2008). Index-based dietary patterns and risk of colorectal cancer: the NIH-AARP Diet and Health Study. *American Journal of Epidemiology*, 168, 38–48.

Rukhin, A. L. (1992). Generalized Bayes estimators of a normal discriminant function. *Journal of Multivariate Analysis*, 41, 154–162.

Sarkar, A., Mallick, B. K., Staudenmayer, J., Pati, D., and Carroll, R. J. (2014). Bayesian semiparametric density deconvolution in the presence of conditionally heteroscedastic measurement errors. *Journal of Computational and Graphical Statistics*, 23, 1101–1125.

Simon, N. and Tibshirani, R. J. (2011). Discriminant Analysis with Adaptively Pooled Covariance. *arXiv.org*, .

Simon, N. and Tibshirani, R. J. (2012). Standardization and the group Lasso penalty. *Statistica Sinica*, 22, 983–1001.

Stefanski, L. A. and Cook, J. R. (1995). Simulation-extrapolation: the measurement error jackknife. *Journal of the American Statistical Association*, 90, 1247–1256.

Subar, A. F., Thompson, F. E., Kipnis, V., Mithune, D., Hurwitz, P., McNutt, S., McIntosh, A., and Rosenfeld, S. (2001). Comparative validation of the Block, Willett, and National Cancer Institute food frequency questionnaires: The Eating at America's Table Study. *American Journal of Epidemiology*, 154, 1089–1099.

Tibshirani, R. J., Hastie, T. J., Narasimhan, B., and Chu, G. (2003). Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statistical Science*, 18, 104–117.

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109, 475–494.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 55, 2183–2202.

Wang, T. and Gaynanova, I. (2018). *DAP: Discriminant Analysis via Projections*. R package version 1.0.

Wang, Y., Wellenius, G. A., Hickson, D. A., Gjelsvik, A., Eaton, C. B., and Wyatt, S. B. (2016). Residential proximity to traffic-related pollution and atherosclerosis in 4 vascular beds among African-American adults: Results from the Jackson Heart Study. *American Journal of Epidemiology*, 184, 732–743.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50, 1–25.

Witten, D. M. and Tibshirani, R. J. (2011). Penalized classification using Fisher's linear discriminant. *Journal of the Royal Statistical Society, Ser. B*, 73, 753–772.

Wu, Y., Qin, Y., and Zhu, M. (2018). Quadratic Discriminant Analysis for High-Dimensional Data. *Statistica Sinica*, to appear.

Yi, G. Y. (2017). *Statistical Analysis with Measurement Error or Misclassification: Strategy,*

Method and Application. Springer.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Ser. B*, 68, 49–67.