EARLY DETECTION AND ROBUST FEATURE LEARNING IN LONGITUDINAL DATA ANALYSIS

A Dissertation

by

KAI HE

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,	Xiaoning Qian
Committee Members,	Tie Liu
	Ulisses Braga-Neto
	Paul de Figueiredo
Head of Department,	Miroslav M. Begovic

August 2020

Major Subject: Electrical Engineering

ABSTRACT

Early detection and feature learning with longitudinal data benefits society in many areas such as early clinical diagnosis, process monitoring, manufacturing and social security. For example, early prediction of disease onset and contemporaneous monitoring of the disease-induced progression can be tremendous help before the disease has time to fully take hold and can help patients get more appropriate care and treatments. Critical to understanding the dynamic patterns in general, is the capability of detecting and tracking the progression of the events of interest as well as identifying the event-associated factors. This dissertation addresses some of the critical issues concerning early detection, robust feature derivation and variable selection in longitudinal data analysis.

Accurate early prediction and risk estimation of the disease onset is challenging, due to the facts that the disease patterns are often indistinguishable at the early stage, and the longitudinal data can be irregularly spaced, missing and not fully labeled. To address these issues, we have developed a contemporaneous disease risk detector, called EDRA (Early Detection and Risk Assessment), a flexible learning framework based on Structured-Output Support Vector Machine (SOSVM) technique to incorporate the individual-level progression. Datasets of varying complexity demonstrate EDRA's capability of early detection and risk monitoring with partially-labeled longitudinal data.

Along with the challenges from early detection and risk monitoring, the rapid advancement of high-throughput profiling and imaging technologies in recent decades produce biomedical data of high dimensionality, which highlights the importance for extracting predictive features for accurate disease diagnosis and prognosis as well as identifying variables of interest to enable targeted predictive interventions and treatments. However, unwanted data variability, including inherent "batch effects", could be harmful with biased analytical results, and is commonly observed in data collected across multiple experiments or studies. We have developed a principle component analysis (PCA)-based framework, namely

MSSPCA (Matched Supervised Sparse PCA) for robust feature learning by involving the data heterogeneity. MSSPCA has superior performance in deriving predictive features with variable selection capability and being robust to noisy outcomes. The effectiveness of MSSPCA has been demonstrated through a simulation study and a real-world case study with comprehensive performance comparison with several representative and popular existing methods. Finally, we propose a pipeline that integrates EDRA and MSSPCA for robust early detection. The performance of the proposed pipeline is validated through a real-world longitudinal RNA-Seq data for tuberculosis early prediction.

In summary, our proposed methods enhance the performance of longitudinal data analysis with the improved detection accuracy and robustness, better model interpretation and the facilitated learning/inference. Although their benefits are demonstrated in biomedical applications, our proposed methods can also be applied in many other domains where the longitudinal data analysis is involved.

DEDICATION

This dissertation is dedicated to my family.

ACKNOWLEDGMENTS

First, I would like to sincerely thank my academic advisor, Dr. Xiaoning Qian, for his help, advising and unconditional support. I would also like to thank the members of my committee, Prof. Ulisses Braga-Neto, Prof. Tie Liu and Prof. Paul de Figueiredo for their constructive comments and support. I'm grateful to our collaborators and our group members for their help and discussions about the research. Finally, I would like to appreciate my family and friends for their endless support and encouragement throughout the entire period of my doctorate study.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a dissertation committee consisting of Dr. Xiaoning Qian, Dr. Ulisses Braga-Neto and Dr. Tie Liu of the Department of Electrical & Computer Engineering and Dr. Paul de Figueiredo of the Department of Microbial Pathogenesis and Immunology.

The work in Chapter 3 was conducted under the guidance of Dr. Xiaoning Qian of the Department of Electrical & Computer Engineering at Texas A&M University and Dr. Shuai Huang of the Department of Industrial & Systems Engineering at University of Washington.

The work in Chapter 4 was conducted under the guidance of Dr. Xiaoning Qian of the Department of Electrical & Computer Engineering, Dr. Jianhua Huang of the Department of Statistics at Texas A&M University and Dr. Shuai Huang of the Department of Industrial & Systems Engineering at University of Washington. Dr. Meng Lu of the Department of Information Management and Institute of Data Science at Tianjin University provided comments on manuscript writing.

All other work conducted for the dissertation was completed by the student, under the guidance of Dr. Xiaoning Qian of the Department of Electrical & Computer Engineering.

Funding Sources

The work in this thesis is supported by the National Science Foundation (NSF) Grants 1553281 and 1718513, as well as the United States Department of Agriculture National Institute of Food and Agriculture competitive grant USDA-NIFASCRI-2017-51181-26834 through the National Center of Excellence for Melon at the Vegetable and Fruit Improvement Center of Texas A&M University.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

The rapid advancement of sensor and information technologies in recent decades such as the high-throughput next generation sequencing and imaging techniques provide unprecedented opportunities for us to develop methods for early diagnosis and contemporaneous monitoring of the disease progression. For example, a dynamic biological process of living organisms can be manifested by the changes in the gene expression, whose variation helps better understand disease progression. The positron emission tomography (PET) scan imaging technique shows characteristic changes in the brains of patients with Alzheimer's disease (AD), and in prodromal and even presymptomatic states that can help estimate the AD pathophysiological process [3]. Early diagnosis is beneficial for disease prevention and early treatment as it plays an important role to raise cure rates, achieve better care and quality of life, and/or extend survival for chronic diseases which progress over time or have persistent and long-lasting in its effect [4, 5]. For example, type 1 diabetes (T1D), a genetic chronic disease, whose disease progression can be subdivided into multiple stages while the symptoms only appear at the last stage as shown in Figure 1.1 [1]. Early detection can also be applied to the longitudinal study of the clinical responses to drug therapy. Identifying pre-existing and drug-induced signatures is important to predict the clinical response to the drugs [6].

Besides early diagnosis, contemporaneous monitoring of the disease progression is also critical for managing the patients with chronic conditions. One of the most important properties of chronic disease is that the disease persists for long time, as defined by the U.S. National Center for Health Statistics. The progression speed of chronic diseases such as Alzheimer's disease and diabetes, varies greatly across patients due to different factors including genetics, physiology, social-economics, gender, and behavior. Contemporaneous monitoring of the disease progression can help patients get more appropriate care and treatments. Furthermore, contemporaneous monitoring of the disease progression can be very

helpful in the study of drug response as well. E.g., it's crucial for doctors to have the capability of tracking the drug's longitudinal effects to provide reliable recommendations for the continual usage of medications to treat the disease.

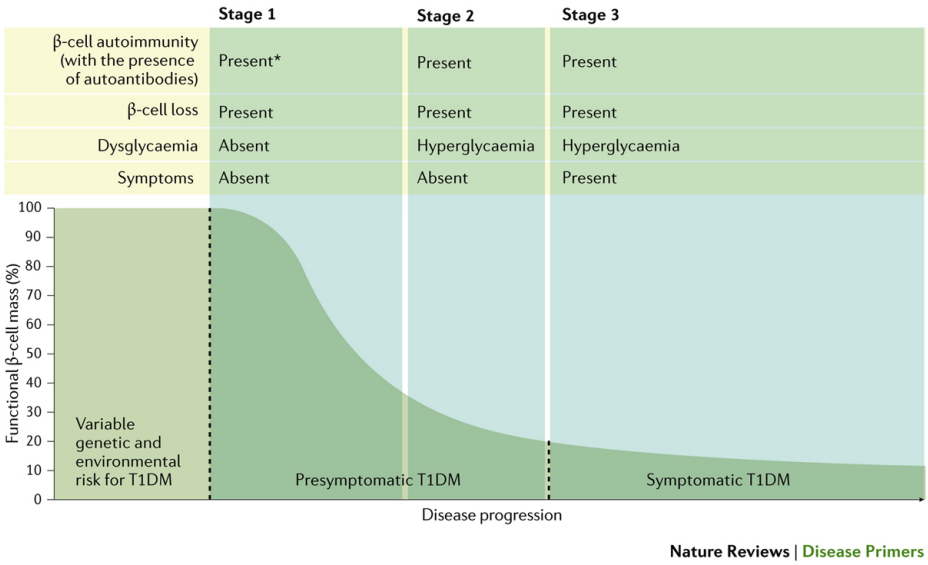


Figure 1.1: T1D can be subdivided into three stages: stage 1 is characterized by the presence of autoantibodies and the absence of dysglycaemia; stage 2 is characterized by the presence of both autoantibodies and dysglycaemia; and symptoms only appear at stage 3, which corresponds to symptomatic T1D. This figure is reprinted from Ref. [1].

Another important issue arises in developing robust detectors with high-dimensional longitudinal data containing unwanted data heterogeneity. Despie the opportunities they provide, the data of high dimensionality with limited sample size poses computational challenges, especially in the settings where the feature dimension is much higher than the sample size ($p \gg N$). Many methods for feature extraction have been proven to be successful in handling high-dimensional data for visualization or to improve the performance of the downstream analysis by reducing the model's complexity and improving its generalizability [7, 8, 9, 10, 11, 12]. In addition to finding compressed feature representations as potential "biomarkers" when analyzing biomedical data, it is useful to understand

how different variables contribute to prediction, especially when they have physical meanings, and further enhance the model's performance by selecting the most important variables [13, 14, 15, 16, 17, 18]. Variable selection is widely applied in analyzing RNA-Seq data where each variable corresponds to a specific gene or transcript [19, 20, 21]. Discovering the predictive genes from tens of thousands of genes is of great help to enable targeted disease treatment/prevention and understand the disease progression [22, 23].

Nevertheless, most of these methods ignore the complicated structures within the samples by holding the assumption that the data are randomly sampled from the same population, which is often violated in real-world applications. For instance, clinical data are often collected from people in different countries, and there exists population-specific heterogeneity, which may cause serious concerns in model's generalizability and reproducibility when tested on on samples from one site with a model trained on another site [2]. Such "batch effects", referring to the systematic error generated while the samples are probed by multiple batches of platforms or heterogeneity due to the technical difference, are commonly observed across multiple batches of data generated from different processing or reagent batches, experimenters, protocols, or profiling platforms [24, 25, 26]. Directly applying conventional feature extraction and variable selection methods without careful consideration of unwanted effects will lead to biased downstream analysis and be harmful for disease diagnosis and prognosis that utilize such analysis. In fact, due to the data heterogeneity such as population-specific heterogeneity, it is often difficult to perform integrative analysis, as witnessed by poor cross-site validation in many recently reported studies [2, 27, 28]

This dissertation is mainly focused on early detection and robust feature learning in longitudinal data analysis, with possibly a wide range of applications where the longitudinal data analysis is involved. In particular, the following issues have been addressed throughout this dissertation:

- **Early detection with partially-labeled longitudinal data**: Machine learning meth-

ods are widely used in computer-aided disease diagnosis and prognosis. However, the applicability of the existing supervised learning methods are limited due to the fact that it's difficult and resource-demanding to obtain fully-labeled data in longitudinal studies conducted before disease onset. Semi-supervised and unsupervised methods developed to handle unlabeled data are not optimal in early detection with longitudinal data since they don't take the advantage of the temporal dependency within longitudinal data. We have developed a contemporaneous risk detector called EDRA, which is based on structured-output support vector machine (SOSVM) technique and extended to longitudinal data analysis. Instead of focusing on the magnitude or scale of the static measurements, EDRA extracts "change" information and seeks to learn the developments over different periods of time by imposing varying penalties on misclassifications. The proposed framework enhances the detection of disease onset with respect to both earliness and accuracy, which has been validated with both simulation and real-world studies including the early detection of chronic diseases and risk monitoring of the drug long-term effects.

- **Robust feature learning from longitudinal data**: Unwanted data variation is a common problem faced by researchers, particularly for longitudinal data collected across multiple experiments or studies, which can lead to biased analytical results. Existing techniques infer the unwanted data variation either from the residuals of the regression of input observations over the outcomes, or based on the reduced data restricted to only "negative control" variables that are known a priori not to be associated with respect to the biological factor of interest. Therefore, they may suffer from unstable performance and are not applicable or have limited power in many real-world situations where these prior knowledge is unavailable or noisy, particularly for longitudinal data collected for disease early detection and prevention. We have proposed a novel method, namely Matched Supervised Sparse Principle Component Analysis (MSSPCA), which is capable of extracting features as potential biomarkers and

identify contributing variables associated with primary interests. With the unwanted data variation incorporated in data modeling, MSSPCA employs a probabilistic supervised PCA framework with sparse estimation of the loading matrix to aggregate the signals from both input observations and the response data, to discover the underlying latent factors of interest. MSSPCA extracts the compressed feature representation with variable selection capability and is not sensitive to the noisy outcomes. MSSPCA facilitates the subsequent learning and inference by reducing the model complexity and enables targeted disease prevention and treatments by identifying predictive variables. What's more, it releases the pressure for annotation during the data collection, which makes longitudinal studies more convenient.

The rest of this dissertation is organized as follows. Chapter 2 reviews the existing techniques for computer-aided disease diagnosis and risk monitoring, and the related work for removing unwanted data variability. Their advantages and disadvantages will be discussed, which suggests that the proposed methods are promising in longitudinal data analysis.

In Chapter 3, we propose EDRA with an SVM-based learning framework presented and develop an efficient algorithm to solve the dual problem of the primal optimization problem. The properties of the trained risk detectors are discussed. The performance of EDRA is assessed with four longitudinal datasets of varying complexities including two simulated longitudinal datasets that consider variables with equal/unequal discriminating power, and two real-case longitudinal datasets for Type 1 Diabetes (T1D) early diagnosis and long-term monitoring of drug response. The experiment results demonstrate EDRA's capability for longitudinal data analysis in the context of early detection and contemporaneous risk estimation.

In Chapter 4, we propose MSSPCA for robust feature learning from longitudinal data by involving the data heterogeneity. The data modeling with the incorporation of unwanted batch effects and a learning framework based on supervised sparse PCA is presented. An efficient algorithm with closed-form updating rules is derived for solving the proposed op-

timization learning framework. By comparing with the most representative batch-effect correction algorithms, comprehensive experimental results demonstrate MSSPCA's superior performance in deriving informative features with capability of identifying predictive variables and being robust to noise outcomes, from the high-dimensional data containing unwanted effects.

In Chapter 5, we propose to integrate EDRA and MSSPCA into a pipeline for robust early detection. Specifically, we propose to apply EDRA subsequently on the features derived by MSSPCA to enable early detection with partially-labeled longitudinal data. We discuss the situations where the needs for early detection and robust feature learning coexist, which require a combination of EDRA and MSSPCA. The proposed pipeline is applied on a real-world RNA-Seq data for tuberculosis early prediction. The effectiveness of robust early detection is demonstrated by comparing with the pipelines along with the risk signatures that have been developed in the existing literature. Moreover, the impacts of EDRA and MSSPCA in the proposed pipeline are investigated and discussed.

Chapter 6 summarizes the major contributions of this dissertation and discusses the directions for the future work.

2. LITERATURE REVIEW

In this section, we review the existing methods for (1) event detection and risk monitoring with longitudinal data and (2) data correction by removing the unwanted data variation. The limitations of the existing techniques are discussed, which suggest that the proposed methods offer the potential to fill in the gaps between these techniques and some practical limitations in real-world situations.

2.1 Early detection and contemporaneous risk monitoring

The proposed EDRA for disease early detection and risk assessment is related to the topics in literature of computer-aided diagnosis methods, longitudinal data analysis and structured-output learning. Different from these methods, EDRA can handle irregular longitudinal data with partial label information, focusing on not only early diagnosis, but also contemporaneous monitoring of the disease progression.

2.1.1 Computer-aided diagnosis methods

Classification methods are widely used in computer-aided diagnosis. Many classification methods care about finding optimal hyperplanes to best separate data from different groups, whereas other methods such as Bayesian methods achieve classification based on probabilistic models. Statistical classification methods are frequently applied in DNA micro-array and RNA-Seq gene expression data analysis because of their superior performance in $p >> N$ setting, among which there are logistic regression, naive Bayes and Bayesian networks classifier, etc [29, 30, 31]. Statistical classification methods aims at learning the association with the input data and their corresponding outcomes by maximizing the likelihood of the training labels given the training input data. Another group of statistical supervised learning methods that are also ubiquitous in analyzing high-dimensional biomedical data such as Linear Discriminant Analysis (LDA) [32, 33], Quadratic Discriminant Analysis (QDA) [34] and Optimal Scoring (OS) [35, 36], which are supervised and

are applied for classifying or categorizing data into classes or groups of the same type. They look for a linear combination of data to project the data into a subspace where the between-class covariance is maximized while the within-class covariance is minimized.

Machine learning methods like Support Vector Machine (SVM) [37, 33, 38, 39] and ensemble learning methods such as Random Forest (RF) [40, 41, 42], are also commonly applied in computer-aided disease diagnosis because of their robust performance to outliers and non-linear separable problems. Support vector machine, for example, is a supervised learning method mainly developed for classification. It aims at finding an/a set of optimal hyperplanes to best separate data from different classes by maximizing the geometric distance between it to the nearest data points.

Extended from these two groups of methods, more complicated models are considered to address diverse range of challenges and specific complexities in some applications. For example, Zhou *et al.* formulate the prediction problem as a multi-task regression problem to predict the longitudinal outcomes for Alzheimer's disease based on the static baseline MRI features [43]. Multi-model frameworks are proposed to combine data of different types, e.g., Chen *et al.* propose a convolutional neural network (CNN)-based multimodel disease prediction algorithm using structured and unstructured data [44]. In [45], Zhang *et al.* propose a multimodel classifier combining three modalities of biomarkers to classify Alzheimer's disease (AD) or its prodromal stage (i.e., mild cognitive impairment (MCI)) from the healthy controls.

Nevertheless, most of the methods discussed above are supervised, it's difficult to directly apply these methods on partially labeled data. Moreover, comparing to these methods, our objectives are different, since we not only aim at discriminating classes, but also considering the temporal correlation and contemporaneously estimating the underlying risk scores by analyzing the longitudinal data.

2.1.2 Longitudinal clinical data analysis

Longitudinal study is widely used in diagnosis, prediction and monitoring of the disease, that involves repeated observations of same variables over short/long period of time. There exist many time series models applied in longitudinal clinical data analysis. Analysis of variance (ANOVA)-based methods for longitudinal analysis include a repeated measures ANOVA and multivariate ANOVA (MANOVA). Both focus on comparing group means, but neither informs about subject-specific trends over time. ANOVA approaches are also limited to scenarios with irregularly-timed and missing data. The limitations of ANOVA approaches lend toward the use of modern approaches that robustly handle challenges of longitudinal studies [46]. Approaches that allow irregularly-timed and missing data such as Generalized Estimating Equation (GEE) [47] and Mixed Effects Regression (MER) [48] are proposed to be applied in longitudinal study. Both methods model the mixture of time-varying and static covariate effects and study the correlation with predictors and outcomes, with MER be more advantageous over GEE as it captures correlations of repeated measures using "random effects" that serve to describe cluster-specific trends over time [46]. Both these two category of methods are widely used in longitudinal data analysis in neurodegenerative diseases such as Huntington's Disease [49, 50, 51, 52, 49]. Trajectory studies including Fixed/Random/Mixed-effect models, Latent Growth Mixture Modeling (LGMM), Latent Class Growth Modeling (LCGM) have been increasingly recognized for their usefulness for identifying homogeneous subpopulations within the larger heterogeneous population [53, 54, 55, 56, 57, 58]. However, most of these methods aim at either prediction of the outcomes or discrimination of populations, which is not enough to cover our objectives, nor are they feasible for the cases where the clinical data is of high dimension. State-spaced models focusing on latent states inference, such as Hidden Markov Model (HMM) and Linear Dynamic Systems (LDS) with its variants including Kalman filter, have been proved to be useful for the prediction of the disease progression [59, 60, 61]. Among this line of efforts, HMM-based methods are widely used in

clinical data analysis. For instance, Wang *et al.* propose a continuous-time HMM-based model that learns a continuous-time progression model from discrete-time observations with non-equal intervals to address the problems like irregularity and the incompleteness of the observation [62]. Jackson *et al.* develop a multistage Hidden Markov Model and apply it to an aneurysm screening study [63]. Sukkar *et al.* apply Hidden Markov Model to Alzheimer's disease [64]. Various recurrent neural networks (RNN)-based approaches have been developed for temporal data analysis. GRU-D, that is based on Gated Recurrent Unit (GRU), proposed by Che *et al.* to address the missing values problem in time series data by utilizing the missing patterns to achieve better prediction results [65]. Choi *et al.* propose Doctor AI, a temporal model using recurrent neural networks (RNN) that was applied to longitudinal time stamped electronic health record (EHR) data to leverage large historical data to make multilabel predictions (one label for each diagnosis or medication category) for patients' subsequent visits [66]. However, these methods are either focused on studying the association between the outcomes and covariates, discovering the underlying latent variables or predicting the temporal patterns, which are not of direct help for early detection/classification for the events of interest.

In the field of temporal predictive pattern learning, there have been efforts to extend supervised learning to time series data analysis to summarize and represent this complex time-series data in order to make them amenable to statistical analysis and modeling. Temporal predictive pattern mining techniques are developed to improve the classification of time series data and can be applied on the identification of the onset of disease. They usually aim to mining the predictive temporal patterns or extracting the time series shapelets to be the alternatives of the original features. These methods are usually applied as a preprocessing step prior to classification or regression, or sometimes can be directly used as the detectors [6, 67, 68, 69]. Most of these methods, however, hold the assumptions that the data points are sampled on regular time points. Thus they are not suitable for the data with irregular time intervals, asynchronous visits and varying disease progression rates like

our case. What's more, the underlying risk of disease we seek to monitor is not directly observed, so that it cannot be easily captured by temporal predictive pattern learning techniques. More importantly, high dimensionality of the time series data poses great challenge to this line of methods since mining high dimensional time series data directly is very expensive in terms of both processing and storage cost.

2.1.3 Analysis with partially-labeled, missing or high-dimensional longitudinal data

Various works are presented in literature to address the challenges brought by the high dimensionality of the time series data to develop representation techniques that can reduce the dimensionality of time series, while still preserving the fundamental characteristics of a particular data set [70]. High-level representations such as Discrete Fourier Transformation (DFT) [71], Singular Value Decomposition (SVD), Discrete Wavelet Transformation (DWT) [72], Piecewise Aggregate Approximation (PAA) [73], Adaptive Piecewise Constant Approximation (APCA) [74] were considered previously. In conjunction of these techniques, different similarity-based approaches represent a promising direction of time series analysis. For instance, Dynamic Time Warping (DTW), introduced by Berndt and Clifford [75], and its variants such as Weighted DTW (WDTW) that adopts a weighting scheme [76] and Derivative DTW (DDTW) that uses the difference between consecutive time values [77], are classical speech recognition tools allowing a time series to be "streched" or "compressed", that are considered to be strong for many time series data problems [78]. Another group of similarity measures for time series such as LCSS (Longest Common SubSequence) [79], EDR (Edit Distance on Real sequence) [80] and ERP (Edit Distance with Real Penalty) [81] have been developed based on the concept of the edit distance for strings [70]. More recent works for similarity measurement adopt tree-based methods to increase the robustness and the parameters tuning problems. TCK (time series cluster kernel) proposed by Mikalsen *et al.*, leverages the missing data handling properties of Gaussian mixture models (GMM) augmented with informative prior distributions, and uses an ensemble learning approach to ensure robustness to parameters by combining the

clustering results of many GMM to form the final kernel [82]. Baydogan *et al.* propose a method to model the dependency structure in time series that generalizes the concept of autoregression to local autopatterns, which generates a pattern-based representation along with a similarity measure called learned pattern similarity (LPS). Moreover, it adopts a tree-based ensemble-learning strategy that is fast and insensitive to parameter settings [83].

Longitudinal study repeatedly collects measurements by training same individuals. Analysis of data obtained from such studies is often impeded by the presence of missing data due to item or visit non-response and loss to follow-up [84, 85]. Commonly used analytic approaches exclude patients or records with missing data, which may lead to irregularly-spaced longitudinal data, biased estimates and considerable loss of precision [48, 86]. Methods have been developed to handle the missing data issue from two perspectives: imputing the missing data or modifying the learning framework to incorporate missing or irregularly-spaced longitudinal data. The second group of methods are mostly based on generalized linear model or matrix completion for missing data imputation. For instance, Ke *et al.* exploit the low-rank property of a spatial-temporal matrix via the bilinear formalism and further use the matrix completion technique to fill the missing data for predicting the time to SSI onset with dynamic data [87]. A least-square loss function as well as a squared hinge loss function are contained in their proposed bilinear formulation to obtain an unbiased learning formulation with complete and censored samples. Although their problem shows some relevance to ours, the continuous measurements are required for constructing the spatial-temporal matrix, whereas our method also considers contemporaneous risk assessment and can deal with data points with irregular time intervals.

Semi-supervised methods have been developed to handle unlabeled data, and they usually assign label to the unlabeled data by measuring their geometric or probabilistic similarities to the labeled ones. For example, self/co-training methods label the unlabeled data with a classifier trained using the labeled ones and improve the classifier's generalizability by training it with the training set updated by the samples newly labeled by the

classifier [88, 89, 90, 91, 92, 93]. Other methods learn with the partially-labeled data by including the unlabeled data into their optimization formulations, such as generative probabilistic models or semi-supervised support vector machine, to estimate the model's unknown parameters by jointly maximizing the likelihood/margin for labeled and unlabeled data [94, 95, 96, 97, 98]. However, these methods usually assume the independence within the data, so that they don't take advantage of the inherent structure within the longitudinal data space.

2.2 Data correction methods for unwanted variation

Unwanted data variation is a common problem in longitudinal data analysis, which are caused by systematic technical noises such as experiment/study/population-specific variation. Many methods have been proposed to mitigate the negative impacts of the unwanted data variations, which can be categorized into three main-stream techniques: regression-based methods, factor analysis techniques and methods using deep neural networks. However, the existing methods often rely on clean outcome information, prior knowledge regarding control variables, or large training sample size, which are difficult and resource-demanding in many real-world situations.

2.2.1 Regression-based methods

A group of widely-used methods, represented by Surrogate Variable Analysis (SVA), adjust data by estimating surrogate variables of unwanted effects using the residuals from least square regression of data on primary variables [99, 100, 101]. Another method, ComBat, is based on robust empirical Bayesian (EB) regression, assumes a model for the location (mean) and/or scale (variance) of the data within batches and then adjusts the batches to meet assumed model specifications using both batch information and covariates of interest [102, 103]. It focuses on data with small sample size and estimates the parameters of the Location and Scale (L/S) model to correct data for batch effects [103, 24]. Zhang *et al.* proposed ComBat-Seq using negative binomial regression to consider integer values of count

data [104]. The supervised version of RUV (Remove Unwanted Variation), RUVr (RUV Using Residuals), estimates the factors of unwanted variation using residuals from a first-pass Generalized Linear Model (GLM) regression of data on covariates of interest [105]. However, neither ComBat nor RUVr can be applied on new data without outcome information when deriving predictive biomarkers is the task. Moreover, these methods require clean outcome data since it is important to obtain the residual as the first step for estimating and removing unwanted effects before downstream analysis. For example, disease states are often considered as variables of interest, which are difficult to obtain in real-world data for analyzing disease progression before its onset.

2.2.2 Factor analysis techniques

Another category of methods utilize control variables, defined as the variables not influenced by primary interest, to infer the unwanted variation from the data. Another version of RUV, RUVg (RUV using control genes), performs singular value decomposition (SVD) of the data containing only control genes to estimate the unwanted effects [105]. Buettner *et al.* proposed single-cell latent variable model (scLVM) for identification of single cell subpopulations confounded by cell cycles, which first reconstructs cycle state (or other unobserved factors) and then uses this information to infer "corrected" gene expression levels [106]. scPLS (single cell partial least squares) was later proposed to remove cell stage effects by jointly modeling both control gene and target gene sets using partial least squares regression to estimate factors of confounding effects [107]. However, such algorithms are sensitive to the choice of control genes [108, 109, 110] and the prior knowledge about which variables are not associated with primary interest is not always available. Although a differential expression (DE) analysis can be performed on the unadjusted data to select the least significant genes, genes of interest can be wrongly determined as non-significant if the data is severely biased. Unsupervised methods not requiring outcomes were also proposed to adjust batch effects including variants of Principal Component Analysis (PCA) or SVD, which often filter out all variations including wanted effects or being not able to

remove unknown unwanted variation without supervised information [111].

2.2.3 Remove batch effects using deep neural networks

As deep learning has experienced tremendous progress in recent years, researchers have started to apply neural networks to batch alignment problems [112]. For example, Batch Effect ReMoval Using Deep Autoencoders (BERMUDA) was proposed to utilize the similarities between cell clusters to align corresponding cell populations among different batches by training an autoencoder to minimize the combination of reconstruction loss and transfer loss that measures difference between similar pairs of cell clusters from different batches using the low-dimensional representations [113]. Shaham proposed to remove systematic batch effects using a residual neural network, trained to minimize the Maximum Mean Discrepancy between the multivariate distributions of two replicates, measured in different batches [114]. Lotfollahi developed scGen that combines variational auto-encoders (VAEs) and latent space vector arithmetic to model and predict single-cell expression data [115]. Most of the deep learning methods are developed for datasets with large sample size. It is not a surprise that they often have comparatively poor performance with small datasets [112], which is the case typically seen in biomedical studies.

3. EDRA: EARLY DETECTION AND RISK ASSESSMENT*

3.1 Overview

Early disease detection and risk assessment with longitudinal data is useful to provide patients with appropriate care and treatments. Disease diagnosis with computer-aided methods has been extensively studied. However, early detection and contemporaneous risk monitoring with partially-labeled irregular longitudinal measurements is relatively unexplored. In this chapter, we propose a flexible framework for learning a contemporaneous disease risk detector, called EDRA (Early Detection and Risk Assessment), to predict the onset of disease and monitor the disease progression. EDRA is inspired by a technique called structured-output support vector machine (SOSVM), which was proposed to address the problems involving complex outputs such as structured-output spaces, and EDRA extends it to longitudinal data analysis. Moreover, EDRA addresses the label insufficiency problem by learning the pattern of the development induced by the disease progression over time. Extensive experiments are conducted on several datasets of varying complexity, including the contemporaneous risk assessment with simulated irregular longitudinal data; the prediction of the onset of Type 1 Diabetes (T1D) with irregularly-spaced and partially-labeled longitudinal RNA-Seq gene expression data; as well as the monitoring of drug long-term effects with longitudinal RNA-Seq data that contains missing values.

3.2 Introduction

The rapid advancement of sensor and information technologies in recent decades such as the high-throughput next generation sequencing and imaging techniques provide unprecedented opportunities to develop methods for early diagnosis and contemporaneous monitoring of the disease, which is beneficial for disease prevention and early treatment.

*Reprinted with permission from "Early detection and risk assessment for chronic disease with irregular longitudinal data analysis" by Kai He, Shuai Huang, and Xiaoning Qian, Journal of Biomedical Informatics, p. 103231, August 2019, Copyright ©2019 Elsevier Inc. https://doi.org/10.1016/j.jbi.2019.103231.

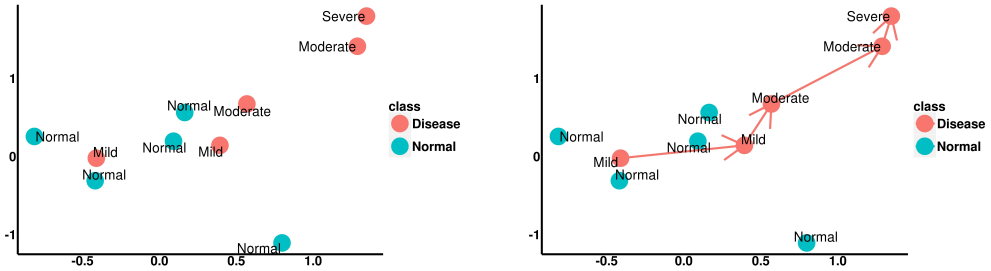
Many machine learning methods for classification and prediction have been widely applied in computer-aided diagnosis [116, 29, 30, 31, 37, 33, 4, 36, 38, 39, 40, 41, 42]. However, most of these methods are supervised learning methods and not specifically developed for early detection with longitudinal data. At the same time, the longitudinal data can be complex as being partially-labeled, irregularly-spaced or with structured-output space. To develop early diagnosis and contemporaneous disease monitoring methods with longitudinal data, we need to overcome the following challenges:

- Difficulty in disease detection at the early stage

One property of the chronic disease is that they are slow to develop and may progress over time. This property makes early diagnosis difficult since patients at the disease early stage behave similarly as healthy people. E.g., for Alzheimer's disease patients at the early stage, their cognitive functions and living functions usually maintain as normal aging individuals. Figure 3.1 provides a simple schematic example with the data containing 2 features ($x1$ and $x2$). In Figure 3.1, there are two subjects with repeated observations: the patient (orange points) and the normal control (blue points). The patient can go through multiple stages: mild, moderate and severe. Most points at the early stage can not be separated from those of the normal control. Early diagnosis is thus challenging at the early stage of the disease.

- Lack of information regarding disease progression

Another challenge is the lack of label information to specifically point out the stages of the disease progression. Labeling subjects by the trained medical professionals at each time point, i.e., the information regarding the stage in Figure 3.1(a), is almost impossible and expensive. In many cases, the only given label information for a subject's longitudinal data is the final diagnosis at the end of a clinical study. Furthermore, subjects' irregular and asynchronous visits as well as the varying disease progression rates make the problem



(a) The patient at the early stage behave similarly as the normal control.

(b) There exists a trajectory of the disease progression for longitudinal data

Figure 3.1: How can we train a detector (i.e., for early diagnosis and risk monitoring) with the dataset where most points are not separable?

more intractable. For longitudinal dataset with label only on the last time point, it's difficult to apply existing classification methods on the data points observed prior to the last one, since we have no information indicating from which time point the patients start to behave differently from the normal controls.

In contrast to existing methods that need labels for patients on all the time points, here, we develop an approach that can extract the "change" information from the original data points, and seek to learn the disease progression over time. We have the intuition that although patients at the early stage may not be separable from normal controls using static measurements if we focus on the magnitude or scale of the measurements, the change patterns over time may separate the two groups, as presented in Figure 3.1(b).

Figure 3.2 demonstrates that a transformation from the original data to the changes over time enables clear separation between the two classes. Moreover, the changes accumulated over larger time intervals are more separable between the two classes, since they contain more information regarding the disease progression. Meanwhile, since the "change" information is measured based on the different time points within the same subject, the synchronization of the visits across the subjects is not required.

To articulate this intuition, we propose a flexible mixed-kernel method, EDRA, for Early Detection and Risk Assessment, which is based on the Structured Output Support

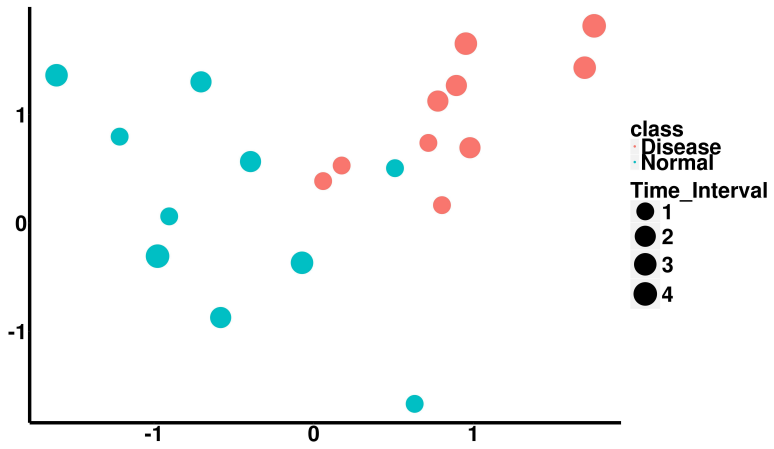


Figure 3.2: New data is generated based on the same data in Figure 3.1 by transforming the original time points to the change over time: $\tilde{x}^i_{tt'} \equiv \delta\Phi(x^i_t, x^i_{t'}), t > t'$. $\tilde{x}^i_{tt'}$ is the transformed data point, x^i_t and $x^i_{t'}$ are two original data points from subject i, and $\delta\Phi$ can be any function for measuring the change from t' to t. In this figure, it's simply $\tilde{x}^i_{tt'} = x^i_t - x^i_{t'}$. The size of the points indicate the length of the time intervals. It can be shown that the change accumulated over large time intervals is more obvious between the two classes.

Vector Machine (SOSVM) [117] and extended to longitudinal data analysis with partial label information. By capturing the pattern of the disease progression over time instead of looking at a single data point, our method is able to achieve better disease diagnosis at the early stage. Another contribution of our method is that it can provide contemporaneous risk assessment of the disease. Meanwhile, EDRA inherits the advantages of SOSVM, including the rescaling of the penalty placed on the misclassification, which enables the smooth and monotonic trajectories for the predicted scores with proper selection of loss rescaling functions. The properties of smoothness and monotonicity are crucial to reflect the contemporaneous underlying risk over time for slowly progressive diseases such as chronic diseases.

EDRA has the following advantages. First, it achieves early diagnosis with improved accuracy. Second, it addresses the disease label/information inefficiency problem for chronic disease data analysis with partially-labeled longitudinal data. Third, it enables contemporaneous risk assessment for tracking the disease/drug-induced progression. Last but not least, it provides a flexible mixed-kernel framework which constructs a kernel as a linear

combination of weighted "sub-kernels" each containing one feature or a subset of features, to take advantage of the prior knowledge about the features. The performance of EDRA is accessed via longitudinal datasets with varying complexities, including 1) early detection and contemporaneous risk assessment using the simulated irregular and partially-labeled longitudinal data with features that are equally/differently predictive; 2) early detection and contemporaneous risk estimation with irregular longitudinal T1D RNA-Seq gene expression data; 3) monitoring of drug's long-term effect on patients based on longitudinal RNA-Seq gene expression data with missing time points.

3.3 Background

3.3.1 Soft-margin Support Vector Machine (SVM)

Support vector machines (SVM) are supervised learning methods with wide applications from biomedicine to computer vision for classification and regression. It aims at finding a single or a set of hyperplanes so that the projected data from different groups/classes can have the largest separations. Therefore, the learning framework of SVM is called maximum-margin framework. To avoid the classifier to be too sensitive to the outliers, soft-margin SVM was proposed to allow the misclassification by introducing "slack variables" for each training sample, and it trades off between the objectives of max-margin and minimization of the sum of slack variables in its optimization framework. Non-linear SVM was suggested by Vapnik to handle situations where the data points are not linearly-separable [118, 119]. By applying a non-linear kernel function, the original inputs are transformed to a high-dimensional space where the transformed data are linearly separable with the max-margin hyperplanes.

Given a training dataset $(x_1, y_1), ..., (x_n, y_n)$, where y_i is the label indicating the class that the data point x_i belongs to, and x_i is a p dimensional vector. Let the output of $F(x_i, y_i; w, b)$ be the score of x_i, where w denotes the parameter vector, and b is the in-

tercept, the optimization learning framework of a soft-margin SVM can be written as:

$$\begin{aligned} \min_{w,\xi_i} \quad & \frac{1}{2}||w||^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & F(x_i, y_i; w, b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \\ & \forall i = 1, .., n \end{aligned} \tag{3.1}$$

where the max-margin is achieved by minimizing the first term of the objective function in (3.2), while the second term with respect to slack variables aims at learning the model parameters that minimize misclassification errors.

Here we may ask three questions: 1) How to incorporate the dependency and the structures within the output space that reflect the underlying disease progression rather than treating all misclassfications equally? 2) Instead of assigning labels, can we also give a confidence level about the classification results? 3) Given partial information about the labels, how can we apply it in semi-supervised scenarios where the label information is not available for each time point?

3.3.2 Structured-output SVM (SOSVM)

Tsochantarid *et al.* propose SOSVM [117, 120], a general framework which extends SVM to scenarios where there exist some structure within the output classes. SOSVM's approach is to rescale the slack variables according to the loss incurred in each of the linear constraints:

$$\forall i, \forall y \in \mathcal{Y}\backslash y_i \quad f(x_i, y_i; w) - f(x_i, y; w) \geq 1 - \frac{\xi_i}{\Delta(y_i, y)} \tag{3.2}$$

where $f(x_i, y_i; w)$ is same as the score function $F(x_i, y_i; w, b)$ with the intercept parameter b excluded and $\Delta(y_i, y)$ is the slack variable rescaling function, which measures the loss incurred by the misclassification of the true label y_i by $y \in \mathcal{Y}\backslash y_i$.

Methods related to learning using privileged information are extensively studied re-

cently (e.g. [121, 122, 123]). Although these methods take output structures into account, not only the labels but also the privileged information such as the rankings of the labels are required for training, and most of them are not specifically designed in longitudinal data analysis. Hoai *et al.* [124] adopt the idea from SOSVM and apply it on computer vision for early event detection with temporal data. However, it's difficult to directly apply their method since the detailed label information about the target events for training is needed. Huang *et al.* [125] consider longitudinal data with partial labels, but they apply same weights for all slack variables and don't take the advantage of rescaling loss functions as the SOSVM-based methods mentioned above to model the irregular time intervals.

3.4 Method

As we described above, most existing methods are not designed for early diagnosis and risk assessment of disease with partially labeled longitudinal data. In this section, we propose a learning formulation to address this problem.

3.4.1 Notations

Let $(X^1, y^1), .., (X^i, y^i), ...(X^n, y^n)$ be a set of longitudinal data with the diagnosis result made at the last time point, where $y^i \in [1, -1]$ is the final diagnosis result for the i-th patient and X^i is a matrix drawn from the input domain $\mathcal{X} \in \mathcal{R}^{T_i \times p}$, which includes the measurements for subject i with T_i visits in total. X^i can thus be represented as $X^i = \begin{bmatrix} x^i_{t_1} \\ \vdots \\ x^i_{t_{T_i}} \end{bmatrix}$, in which $x^i_{t_l} \in \mathcal{R}^{1 \times p}$ denotes the p measurements of the lth visit at time t_l for patient i.

There's a record for the visiting times of each subject: $T = \{T[1], ..., T[n]\}$, where $T[i]$ records the visiting time of patient i, i.e., $T[i] = [t_1, t_2, ..., t_{T_i}]$. For instance, T can be the number of months for each follow up after the initialization of the drug therapy; it can also be the number of months prior to the diagnosis. Please note that the visiting times of a patient can be irregular and asynchronous.

3.4.2 Feature Representation in Mixed Kernel Space

In order to provide a flexible framework for taking advantage of the prior knowledge about the rankings of features' discriminating power, apart from directly applying "kernel trick" on the original data to project it to the kernel space, we constructed a kernel as a linear combination of "sub-kernels" each containing only one feature: $K(x, x') = \sum_{d=1}^{p} \beta_d K_d(x, x') = \sum_{d=1}^{p} \beta_d \langle \Phi_d(x), \Phi_d(x') \rangle$, where $\Phi_d(x) = \Phi(x_d)$, and it only works on the dth feature of x. β is a p-dimensional vector for the feature weights, which satisfies $\sum_{d=1}^{p} \beta_d = 1$.

To measure the augmented information till time t_l, we check both the cumulative moving average and the running total in our experiments to obtain the information augmented until time t_l, which have been applied in the implementation of MMED (Max-Margin Early Event Detectors) [124]. However, we decide to use the cumulative moving average to obtain the augmented information in our method for the following reasons: 1) we would like to smooth out the short-term fluctuations; and 2) different from MMED that aims at localizing the interval for an event, we care more about the risk at a time point given the cumulative information prior to that. The representation can be written as:

$$X^i_{t_l} = \overline{X^i_{[1:l]}} = \frac{1}{l} \sum_{s=1}^{l} x^i_{t_s} \tag{3.3}$$

Lemma 1 shows that more information regarding development for subject i can be accumulated as the distance between two visits l' and l, i.e., $l - l'$ getting larger by the features represented in eq. 3.3. The proof of Lemma 1 can be found in Appendix A.

Lemma 1. *With features represented in (3.3), we have*

$$X^i_{t_l} - X^i_{t_{l'}} = \frac{1}{l} \sum_{s=l'+1}^{l} (x^i_{t_s} - \overline{X^i_{[1:l']}})$$

Let $\Phi(X_{t_l}^i)$ denote the projection of $X_{t_l}^i$ in the kernel space:

$$\Phi(X_{t_l}^i) = diag(\sqrt{\beta_1}, ..., \sqrt{\beta_p}) \begin{bmatrix} \Phi_1(X_{t_l}^i) \\ \vdots \\ \Phi_p(X_{t_l}^i) \end{bmatrix}$$

With this representation, the similarity assessment of the information between the two subjects i and j at the time points l and l', respectively, can be represented by $K(X_{t_l}^i, X_{t_{l'}}^j) = \sum_{d=1}^p \beta_d K_d(X_{t_l}^i, X_{t_{l'}}^j) = \sum_{d=1}^p \beta_d \Phi_d(X_{t_l}^i)^T \Phi_d(X_{t_{l'}}^j)$.

3.4.3 Learning with Longitudinal Data

Recall that instead of learning individual data points, we identify the signatures of the disease/drug-induced changes to address the problem of inseparability and label insufficiency.

First consider a linear function $g(\delta\Phi_i(t_l, t_{l'}); w) = \langle w, \delta\Phi_i(t_l, t_{l'})\rangle$, where $\delta\Phi_i(t_l, t_{l'})$ is the shorthand defined as $\delta\Phi_i(t_l, t_{l'}) \equiv \Phi(X_{t_l}^i) - \Phi(X_{t_{l'}}^i)$, for measuring the changes of the ith subject from time $t_{l'}$ to t_l in the mixed-kernel space. The function $g(\delta\Phi_i(t_l, t_{l'}); w)$ is expected to have the following properties:

$$\forall i, \quad \forall [l', l] \in L_i, \begin{cases} g(\delta\Phi_i(t_l, t_{l'}); w) \geq 0, y^i = 1 \\ g(\delta\Phi_i(t_l, t_{l'}); w) \leq 0, y^i = -1 \end{cases}$$

where $L_i = \{[1, 2], [1, 3], ..., [T_i - 1, T_i]\} \cup \{[0, T_i]\}$ contains all the pair-wise combinations of the visit index for subject i, for i in $1, ..., n$.

In the framework of SOSVM, the loss of misclassifying x^i to a class $y \in \mathcal{Y} \backslash y^i$ is rescaled by a non-negative weight function $\Delta(y, y_i)$, i.e., $\Delta : \mathcal{Y} \times \mathcal{Y} \to \mathcal{R}$, and it quantifies the loss associated with a prediction y, if the true output value is y_i [117]. It's saying that with the prior knowledge about the structure of the output y and y_i, we put greater penalty for the misclassification if $\Delta(y, y_i)$ is large when training the classifier.

In our early detection case on longitudinal data, $\Delta(y, y_i)$ here can be a function with

respect to the time interval between two time points: $\mu(t_l, t_{l'})$. More strict classification rules should be applied for larger time intervals, so that the penalty $\mu(t_l, t_{l'})$ placed on the misclassification should be greater when the two time points are far from each other. The design of function μ will be discussed in detail in the later context.

The desired constraints then become:

$$\forall i, \quad \forall [l', l] \in L_i, \quad y^i g(\delta\Phi_i(t_l, t_{l'}); w) \geq 1 - \frac{\xi_i}{\mu(t_l, t_{l'})} \tag{3.4}$$

Together with the goal of max-margin hyperplane, we obtain the following objective function:

$$\begin{gathered} \min_{w,\xi_i,b} \frac{1}{2}||w||^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad y^i\langle w, \delta\Phi_i(t_l, t_{l'})\rangle \geq 1 - \frac{\xi_i}{\mu(t_l, t_{l'})}, \quad g(\Phi(X^i_{t_0}); w) = -b, \quad \xi_i \geq 0 \\ \forall i, \quad \forall [l', l] \in L_i \end{gathered} \tag{3.5}$$

The constraints containing b are active only for cases $[l', l] = [0, T_i]$, where $l' = 0$ is a virtual time point, so that the constraints for cases $[l', l] = [0, T_i]$ shrink to the constraints of a standard soft-margin SVM.

3.4.4 Properties of EDRA

In this section, we analyze several properties of the scores assigned by the risk detector learned with the above optimization framework.

- Monotonicity

To develop EDRA, we focus on the early detection and contemporaneous risk estimation for the disease/drug-induced progression prior to the diagnosis, for which we utilize the monotonic progression characteristic (either towards disease onset or recovery) as the model assumption to learn EDRA. For instance, as shown in Figure 1.1, functional beta-cell mass declines as T1D progresses. For the degenerative disease conditions such as

Alzheimer's disease, the underlying disease degradation process is also monotonic. This generative nature leads to the monotonic assumption of EDRA.

Here we may ask such question: After we obtain the reliable detection of the changes from the time intervals of a subject, how can the risk scores reflect the progressive property of the underlying disease progression for each time point?

Based on the linear property of function g, the constraints (3.5) can be rewritten as:

$$\forall i, \quad \forall [l', l] \in L_i, \quad y^i \{ g(\Phi(X^i_{t_l}); w) - g(\Phi(X^i_{t_{l'}}); w) \} \geq 1 - \frac{\xi_i}{\mu(t_l, t_{l'})} \tag{3.6}$$

The learning formulation actually naturally enforces monotonicity of the detector function. Moreover, the function μ is desired to have the following properties: 1) $\mu(t_l, t_{l'}) \in (0, 1)$, and 2) $\mu(t_l, t_{l'}) \propto |t_l - t_{l'}|$, to serve as a rescaling function to adjust the penalty for the misclassification based on the distance between two time points. In our study, we set $\mu(t_l, t_{l'}) = 1 - e^{-(\frac{t_l - t_{l'}}{\sigma})^2}$, where σ is a tuning parameter. The proposed learning formulation achieves the monotonicity with respect to the information contained within the time intervals accounting for the disease/drug-induced progression. Such learning formulation provides a flexible framework that is able to deal with irregular time intervals, and enables not only the property of monotonicity, but also the property of smoothness for the trajectories of the predicted scores, which will be discussed in the following context. Both of these properties reflect the progressive property of the chronic disease and drug response.

- Smoothness

A smooth trajectory of the risk scores assigned to one subject over time is desired, since usually in the real case, the disease progresses gradually, so that the difference between the risk scores of two close neighbor time points should be relatively small. The smoothness of the trajectory can be controlled by the design of the slack variable rescaling function μ, which is used to adjust the penalty of the misclassification in (3.4) and (3.5). Since $\mu(t_l, t_{l'}) = 1 - e^{-(\frac{t_l - t_{l'}}{\sigma})^2}$, when two time points are very close, the penalty of the misclas-

sification is close to zero, i.e., $\mu(t_l, t_{l'}) \to 0$, when $t_{l'} \to t_l$. This enables the smoothness of the risk score trajectories for the subjects, since the disease/drug-induced progression contained in a very small time interval is very limited, so that the difference between the predicted scores of two very close time points should be relatively small compared to the ones of the large intervals.

With the linear property of function g, $g(\delta\Phi_i(t_l, t_{l'}); w) = g(\Phi(X^i_{t_l}); w) - g(\Phi(X^i_{t_{l'}}); w)$, we have:

$$R(X^i_{t_l}) - R(X^i_{t_{l'}}) = g(\delta\Phi_i(t_l, t_{l'}); w) \to 0 \tag{3.7}$$

for the cases when $(t_l - t_{l'}) \to 0$

- Separation

The risk scores can be wrongly estimated if we only care about the difference of the scores between two time points since either one of them can start from or end up in a random place. It's important to "fix" at least one point of the whole trial so that the predicted score of which can separate the two classes. In our study, the detector should be trained to be able to classify the last single time point, since the only label we have is the diagnosis at the end of the clinical trial.

In contrast to the smoothness property with the rescaling penalty function $\mu(t', t) \to 0$ when $t' \to t$, the penalty placed on the misclassification is scaled to be the highest for the greatest time interval of the ith subject, i.e., $[l', l] = [0, T_i]$, since the information augmented from the initial time point till the last one reaches the maximum.

Recall that we have the constraint $g(\Phi(X^i_{t_0}); w) = -b$, so that the constraints regarding $[l', l] = [0, T_i]$ in (3.5) turn out to be the constraints of a soft-margin SVM:

$$\begin{aligned} y^i g(\delta\Phi_i(t_{T_i}, t_0); w) &= y^i \langle w, \Phi(X^i_{t_{T_i}}) - \Phi(X^i_{t_0}) \rangle \\ &= y^i (\langle w, \Phi(X^i_{t_{T_i}}) \rangle + b) \geq 1 - \frac{\xi_i}{1 - e^{-(\frac{t_{T_i}}{\sigma})^2}} \end{aligned} \tag{3.8}$$

The problem thus shrinks to a standard SVM classifier training problem. This constraint

is to model the real case where the diagnosis is only available at the end of the study. With constraint (3.8) the trajectories of the two groups are enforced to depart from each other as the disease progresses.

3.4.5 Optimization: Dual Problem and Algorithm

To solve the primal problem (3.5), first we move the constraints to the objective function to obtain the Lagrangian form:

$$\begin{aligned}
&\max_{\alpha,\zeta} \min_{w,b,\xi} L(w, b, \xi, \alpha, \zeta) \\
&= \frac{1}{2}||w||^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\sum_{[l',l]\in L_i} \alpha^i_{l,l'}[1 - \frac{\xi_i}{\mu(t_l, t_{l'})} - y^i\langle w, \delta\Phi_i(t_l, t_{l'})\rangle] - \sum_{i=1}^{n}\zeta_i\xi_i \\
&\quad \text{s.t.} \quad \forall i, \quad \forall [l', l] \in L_i \quad \xi_i \geq 0, \quad \zeta_i \geq 0, \quad \alpha^i_{l,l'} \geq 0
\end{aligned} \tag{3.9}$$

The third part which is related to α is the sum of the terms regarding the changes detection and the last time point classification:

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{[l',l]\in L_i\setminus[0,T_i]} \alpha^i_{l,l'}[1 - \frac{\xi_i}{\mu(t_l, t_{l'})} - y^i\langle w, \delta\Phi_i(t_l, t_{l'})\rangle] \\
&\quad + \sum_{i=1}^{n}\alpha^i_{T_i,0}[1 - \frac{\xi_i}{\mu(t_{T_i}, 0)} - y^i(\langle w, \Phi(X^i_{T_i})\rangle + b)]
\end{aligned} \tag{3.10}$$

To derive the dual problem, we need to minimize the Lagrangian form with respect to w, b and ξ to get:

$$\begin{aligned}
&\max_{\alpha^i_{l,l'}} \sum_{i,[l',l]\in L_i} \alpha^i_{l,l'} - \frac{1}{2}\sum_{i,[l',l]\in L_i}\sum_{j,[\tilde{l}',\tilde{l}]\in L_j} y^i y^j \alpha^i_{l,l'}\alpha^j_{\tilde{l},\tilde{l}'}\langle\delta\Phi_i(t_l, t_{l'}), \delta\Phi_j(t_{\tilde{l}}, t_{\tilde{l}'})\rangle \\
&\quad \text{s.t.} \quad \forall i \quad 0 \leq \sum_{[l',l]\in L_i}\frac{\alpha^i_{l,l'}}{\mu(t_l, t_{l'})} \leq \frac{C}{n}, \quad \sum_{i=1}^{n} y^i\alpha^i_{T_i,0} = 0
\end{aligned} \tag{3.11}$$

The inner product of $\delta\Phi_i(t_l, t_{l'})$ and $\delta\Phi_j(t_{\tilde{l}}, t_{\tilde{l'}})$ can be expanded as:

$$\begin{aligned} &\langle \delta\Phi_i(t_l, t_{l'}), \delta\Phi_j(t_{\tilde{l}}, t_{\tilde{l'}}) \rangle \\ &= K(X^i_{t_l}, X^j_{t_{\tilde{l}}}) - K(X^i_{t_l}, X^j_{t_{\tilde{l'}}}) - K(X^i_{t_{l'}}, X^j_{t_{\tilde{l}}}) + K(X^i_{t_{l'}}, X^j_{t_{\tilde{l'}}}) \end{aligned} \tag{3.12}$$

Specifically, all terms $K(X^i_{t_{l'}}, \cdot)$ with $l' = 0$ are set to be zero, since $l' = 0$ is the virtual time point. When $[l', l] = [0, T_i]$ and $[\tilde{l'}, \tilde{l}] = [0, T_j]$, we have $\langle \delta\Phi_i(t_{T_i}, t_0), \delta\Phi_j(t_{T_j}, t_0) \rangle = \langle \Phi(X^i_{t_{T_i}}), \Phi(X^j_{t_{T_j}}) \rangle$, which is of the same form as a standard kernel SVM problem.

One challenge of the above dual problem is that the number of constraints is very large and thus the computation complexity of the optimization is high. To relieve this problem and speed up the algorithm, we use constraint generation (cutting plane algorithm) [87] to handle the large set of constraints in the original problem (3.5). The outline of the algorithm is described as Algorithm 1.

3.5 Results

This section describes our experiments on two synthetic datasets and two real-world datasets of varying complexity: 1) Simulated longitudinal data considering irregularity in observation time with features of equal/different predictive power; 2) Irregularly sampled T1D longitudinal RNA-Seq gene expression dataset from TrialNet; 3) Longitudinal RNA-Seq gene expression dataset with missing time points for IFNβ drug response. Both of the real-world longitudinal datasets used in our experiments to evaluate the performance are RNA-Seq data, but our method can also be applied to other clinical data where the longitudinal data analysis is involved. The performance of our method is evaluated regarding how early the detection of the disease can be made and how well the risk scores reflect the actual disease progression.

3.5.1 Evaluation

In our experiments, we evaluate the performance of our method based on two criteria: 1) The earliness of detection, 2) The correlation between the risk scores with the disease

progression. We use the area under the Receiver operating characteristic (ROC) curve (AUC) over the normalized time points for benchmarking the earliness of detection when comparing our method with other algorithms, and we plot the risk scores over time for evaluating the performance of our method as a contemporaneous risk monitoring tool for the disease progression.

Algorithm 1: Algorithm for solving the dual problem (3.11) of EDRA

Data: $(X^1, y^1), ..., (X^n, y^n)$, β, T, L, C, ϵ.
Result: α, b
Initialize α, $\xi \leftarrow 0$ and $S \leftarrow \emptyset$;
while *True* **do**
  Set $V \leftarrow \emptyset$;
  for $i = 1$ *to* n **do**
    Compute the loss for all $[l', l] \in L_i$

$$H(l', l) \equiv (1 - y_i \langle w, \delta\Phi_i(t_l, t_{l'})\rangle)\mu(t_l, t_{l'})$$

    where $w = \sum_{i=1}^{n} \sum_{[l',l]\in L_i} \alpha_{l,l'}^i y^i \delta\Phi_i(t_l, t_{l'})$;
    Find the most violated constraint:

$$[\hat{l}', \hat{l}] = \max_{[l',l]\in L_i} H(l', l)$$

    $\xi_{i_{new}} = \max\{0, H(\hat{l}', \hat{l})\}$;
    if *($\xi_{i_{new}} \geq \xi_i + \epsilon$)* **then**
      $\xi_i \leftarrow \xi_{i_{new}}$;
      $V \leftarrow V \cup \{[\hat{l}', \hat{l}]_i\}$;
      $\alpha \leftarrow$ optimize dual problem (3.11) over $S = S \cup V$;
    else
  end
  if $V = \emptyset$ **then**
    Stop and return the results ;
  else
end

3.5.1.1 Time Normalization

Since the visiting times can be irregular and asynchronous in many cases, they are normalized as the fraction of the lengths of the whole trials to get better evaluation. For instance, the normalized time for the lth visit of the subject i can be represented as: $t = 1 - \frac{t_{T_i} - t_l}{L}$, where L is the length of the whole trial, i.e., the maximum length of all the subjects, so that the normalized time $t \in [0, 1]$. Since often in clinical settings the delta time to an event is more useful as it provides how early the event of interest can be estimated, we also consider to evaluate the performance based on the normalized delta time to an event (such as diagnosis/recovery) in our experiments, which can be represented as: $\Delta t = \frac{t_{T_i} - t_l}{L}$. When the subject reaches the last time point and receives the diagnosis ($t_l = t_{T_i}$), the normalized time $t = 1$ ($\Delta t = 0$). At the initiation of the trial, $t = 0$ ($\Delta t = 1$) for the subjects whose length of study equal L. This set up is for the cases where some of the subjects start to take the test early while some of them start late. For the subjects with late starting time or skipped visits, they may not be available on some certain normalized time points according to their actual skipped visits. Figure 3.3 illustrates the time normalization for the three subjects in different cases.

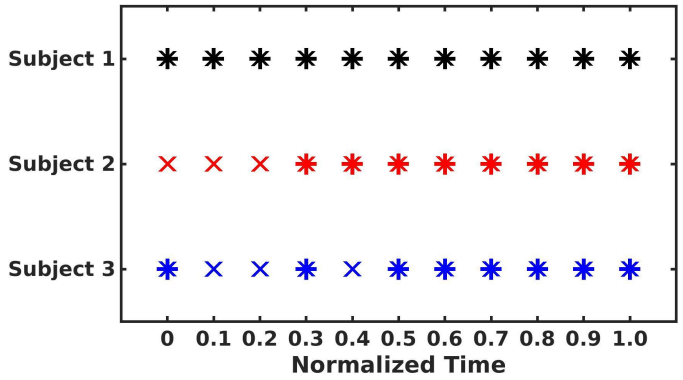


Figure 3.3: Time Normalization: Asterisk symbol "*" denotes the available visits; Cross symbol "x" denotes the unavailable visits. Subject 1: early starting time without skipped visits ; Subject 2: late starting time; Subject 3: skipped visits

3.5.2 Simulation

3.5.2.1 Data generation

We first validate the performance of our method on the synthetic longitudinal data. The synthetic longitudinal data is generated for 100 subjects in total, and each subject has different number of time points ranging from 12 to 14. The prior for the class of disease equals the prior for the class of the normal controls, which is 0.5. The disease progression is modeled by 4 different stages: Stage 0 (Normal), Stage 1 (Mild), Stage 2 (Moderate) and Stage 3 (Severe). For normal controls, they only stay in Stage 0 and will never proceed to the other three stages.

For patients, however, the disease progression is modeled by a Markov Chain model starting from either Stage 0, Stage 1, or Stage 2 and can proceed to more severe stages as disease develops, or it can start from one stage and skip the adjacent stage to directly jump to any one of the more severe stages (e.g., jump from Stage 0 to Stage 2/Stage 3). Specifically, to evaluate the robustness of the proposed approach on irregular longitudinal data, we randomly skipped the time points within one subject to model the irregularity in the observations, as shown in Figure 3.5.

The lth visit of the ith subject's can be represented as:

$$x_l^i = \mu_s + \varepsilon^i + \varepsilon_l^i \tag{3.13}$$

where μ_s is a vector of mean values for the measurements including 4 features for the stage corresponding to the lth visit of the ith patient. The design of μ follows the structure of the stages. Further, linear and nonlinear co-existing predictive relationships are considered for generating the synthetic data. Figure 3.4 illustrates the design of μ_s in our experiments.

The individual effects and the technical noise are modeled by ε^i and ε_l^i, respectively. For longitudinal dataset, it's necessary to model the "baseline" information ε^i for each subject, that won't change over repeated measurements, and is shared by all the time points

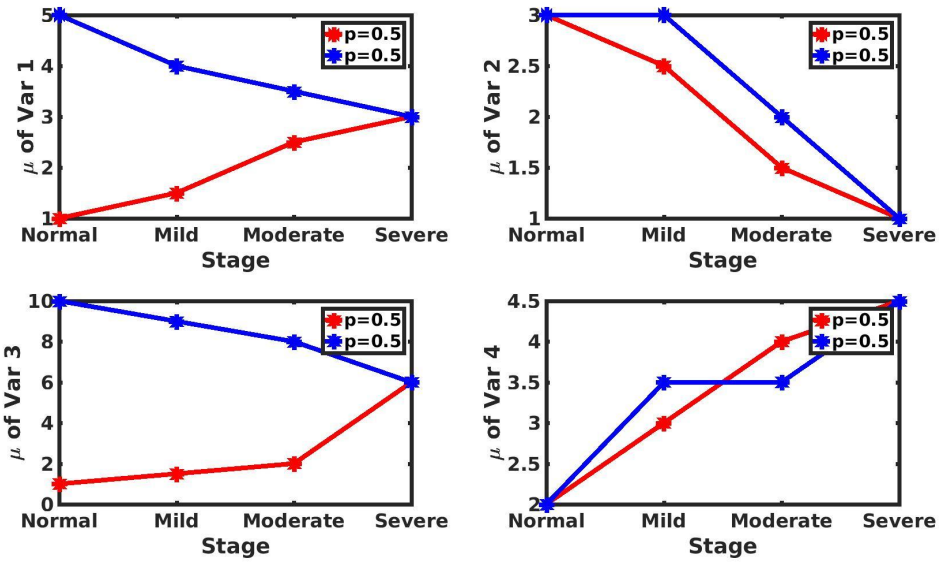


Figure 3.4: Generation of synthetic data: design of μ for the 4 features as disease progresses over time. μ of variable 1 and variable 3 are designed to model the nonlinear predictive relationship, while variable 2 and variable 4 follow linear predictive relationship with different progression rates. The probability of choosing the pattern of the blue line is same as the red line, which equals 0.5.

of subject i. The technical noise is modeled by ε_l^i, which varies among all the data points. Both ε^i and ε_l^i are randomly drawn from multivariate normal distribution.

We randomly divide the synthetic data into training and testing dataset. 80 percent of the generated synthetic data is contained in the training dataset, and the rest 20 percent is used as the testing data for evaluating the performance.

3.5.2.2 Experiment results

In the first experiment, we evaluate the performance using the synthetic data with all features contributing to the discrimination of the two classes. The feature weights β_k are set to be same for the 4 features of this synthetic dataset: $\beta_k = 0.25$, $k = 1, ..., 4$.

We first investigate the performance of risk assessment. The stage information is illustrated by different colors for better illustration. However, please note that the information regarding the stage is only used for demonstration, and it's not available when we train the

models.

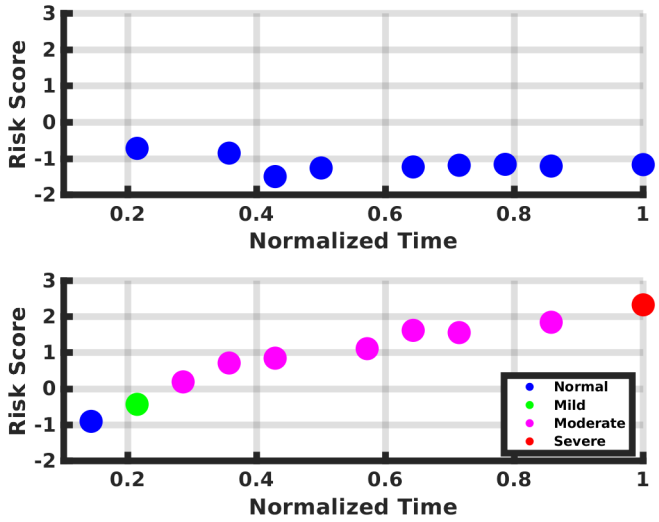


Figure 3.5: Synthetic data experiments: Risk scores over time for the two subjects. Top: a normal control without disease; Bottom: a patient with 4 different stages

Figure 3.5 provides two subjects from the testing dataset to illustrate how the trained detector monitors disease progression in longitudinal study for 1) a normal control stays at the "Normal" stage and 2) a patient goes through different stages over time. The curve of the risk scores over time is relatively flat for the normal control, and the predicted scores throughout the trial are less then zero. Nevertheless, for the patient with increasingly severe situation, the risk score increases as the disease progresses, and turns out to be positive since the third normalized time point of the trial.

To further evaluate the effect of the mixed-kernel framework by considering the prior knowledge about the feature discriminating power, another synthetic dataset containing features with different predictive power is discussed in the following experiments. This synthetic dataset is simulated with two additional inactive features whose mean values for the measurements remain unchanged over different stages, to the original feature set. Therefore the feature weights for the kernel construction are: $\beta_k = 0.25$, for $k = 1, ..., 4$

and $\beta_k = 0$, for $k = 5, 6$. Table. 3.1 provides the detailed information about how β_k is determined in this experiment.

ID	Var1	Var2	Var3	Var4	Var5	Var6
Active	T	T	T	T	F	F
β	0.25	0.25	0.25	0.25	0	0

Table 3.1: Synthetic data experiment: Feature information

We analyze the earliness and accuracy of the detection by EDRA. We repeat our experiments 50 times and record the average performance, with the synthetic data randomly divided for training and testing each time as described above. To obtain better evaluation of the performance, we compare our method with three other popular classifiers: Linear SVM, Naive Bayes (NB) and Kernel SVM (RBF). When we train Linear SVM, Kernel SVM and Naive Bayesian classifier, since the only information about the label for the longitudinal data we have is the final diagnosis, we apply the final diagnosis result to the time points prior to the last one, i.e., given the time points of the subject i: x_1^i,x_2^i,...,$x_{T_i}^i$, we assign the label $y_{T_i}^i$ for the last time point $x_{T_i}^i$ as the label to the other time points prior to that. Specially, since linear SVM, Naive Bayes (NB) and Kernel SVM are not designed for longitudinal data, we treat the data points independently, without considering the temporal structure within them.

In addition to the methods mentioned above, since our method is inspired by SOSVM, we compare our methods to SOSVM and another SOSVM-based method Max-Margin Early Event Detectors (MMED), which are more state-of-the-art approaches specifically designed for the early detection of temporal data analysis. We train and evaluate MMED and SOSVM the same way the authors of MMED did in their experiments [124]. During the training of MMED and SOSVM, since both methods require the starting and ending time of an event to train the model for localizing the event of interest, we set the first time

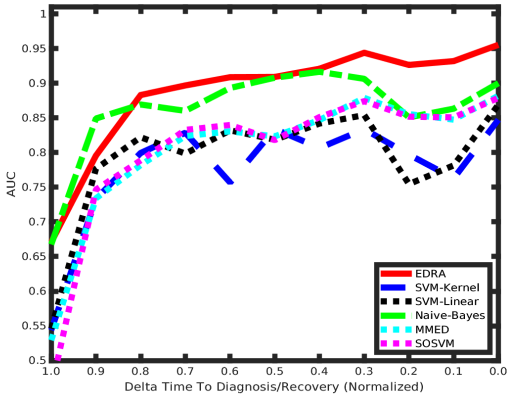
point as the starting point and the last one as the ending point of the event for an subject with disease; for healthy controls, we set the time interval for the event of interest to be empty. We follow MMED's implementation to perform the event detection with MMED and SOSVM: given a data point at time t, we calculate the scores for all the data points prior to t, and use the highest score as the risk score indicating if an event has been happening until time t.

When applying the trained classifiers to the testing dataset, AUCs are calculated on each normalized delta time point. The curves of AUC over the normalized delta time points are depicted in Figure 3.6. Figure 3.6(a) demonstrates the AUC trajectories over the normalized delta time to the diagnosis based on the synthetic dataset with 4 active features. Figure 3.6(b) provides the AUC trajectories over the normalized delta time to the diagnosis based on the synthetic dataset with features of different predictive power.

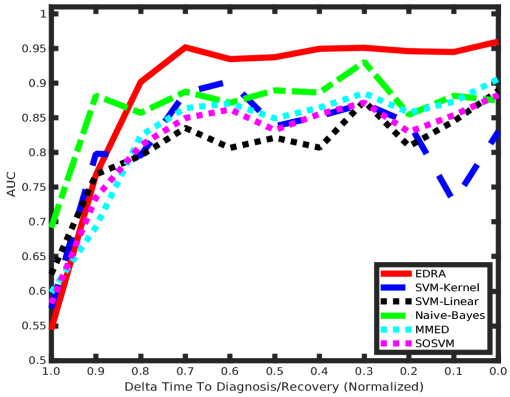
In Figure 3.6(a), at the beginning of the trial, EDRA performs similarly with Naive Bayes method but better than the other SVM-based methods. However, EDRA outperforms all the other methods at the last three time points of the trial, with the AUC reaching 0.96 ± 0.05 at the end of the trial, while the AUCs of the other methods are 0.88 ± 0.08, 0.88 ± 0.07, 0.87 ± 0.09, 0.85 ± 0.11, 0.90 ± 0.07 for MMED, SOSVM, Linear SVM, Kernel SVM (RBF) and Naive Bayes, respectively. What's more, it can be seen that the SOSVM-based methods considering temporal structure, such as EDRA, MMED and SOSVM, successfully capture the disease progression with the smoothly increasing trajectories of AUCs over time, while the other methods fail in this point.

In Figure 3.6(b), EDRA outperforms the other methods by a large margin after the third normalized time point. The AUC of EDRA keeps increasing till the last time point and ends up at 0.96 ± 0.04, while the AUCs for MMED, SOSVM, Linear SVM, Kernel SVM and Naive Bayes are: 0.91 ± 0.06, 0.88 ± 0.06, 0.89 ± 0.06, 0.83 ± 0.09 and 0.87 ± 0.06, respectively.

Comparing the earliness and the accuracy of the detection, EDRA outperforms the other



(a) Same predictive power



(b) Different predictive power

Figure 3.6: Synthetic data experiments: AUC over the normalized delta time to the diagnosis

methods. Regarding the contemporaneous risk assessment, it can be shown that the models considering the structure within the temporal data such as EDRA, MMED and SOSVM, capture the risk progression better with the smoother and increasing AUC trajectories, compared to the relatively fluctuating AUC trajectories by Linear SVM, Kernel SVM (RBF) and Naive Bayes. With the synthetic data incorporating different rates of irregularity in observations, the experiments show that the proposed method is robust to irregularly-sampled longitudinal data. The experiments also demonstrate that the mixed-kernel framework incorporating the prior knowledge about the features' discriminative power improves the performance compared to the methods without such consideration. For the experiments using this dataset, we perform 5-fold cross validation for determining the hyperparameter C for the SVM-based methods and the tuning parameter σ for the kernel construction.

3.5.3 Longitudinal T1D RNA-Seq data from TrialNet

This section describes our experiments on RNA-Seq gene expression dataset from TrialNet, which includes 42 subjects with the final diagnosis of T1D, and 37 normal controls. For each subject diagnosed to have diabetes, the number of time points ranges from 3 to 11, while there's only one time point for each normal control. The pattern of the visiting time of the patients with multiple time points is irregular and asynchronous, and the time stamps are recorded by the months prior to the diagnosis, as illustrated in Figure 3.7.

Since there are 16618 genes in the original dataset, we first perform differential expression test with edgeR [126] to identify 50 differential expressed genes (DEGs) that show differences in expression level between conditions for our experiments. The importance of each DEG is measured by the absolute value of the fold change (FC). The weight of the kth DEG is calculated based on the absolute value of the dth DEG's log_2FC and is normalized by the sum of the absolute values of log_2FC of all DEGs, i.e., $\beta_k = \frac{abs(log_2FC_k)}{\sum_{d=1}^{50} abs(log_2FC_d)}$.

In this experiment, since the only information regarding the disease situation is the medical diagnosis made at the last time point for each subject, how early our method can detect the disease prior to that time point is of great interest. To investigate the performance

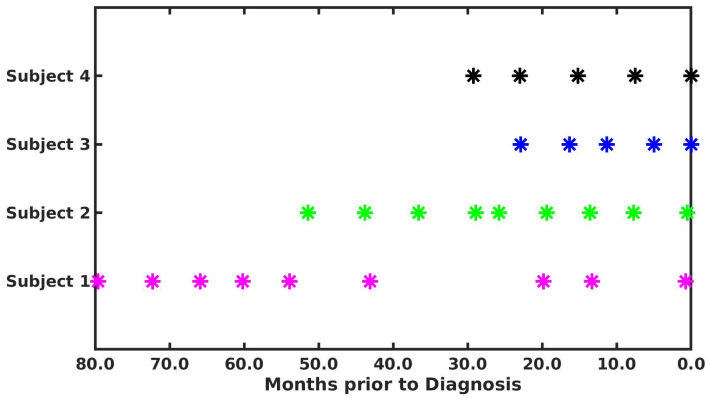


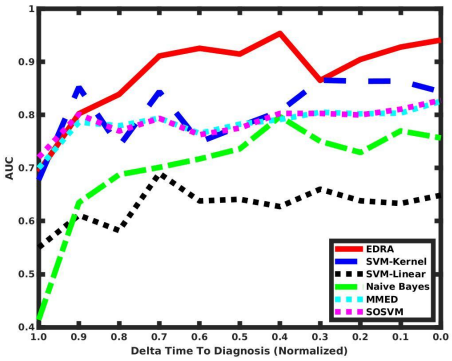
Figure 3.7: Visiting time points (Months prior to diagnosis): "*" denotes the available visits.

of our method for early disease detection, we plot the curves of AUC over the normalized delta time points to benchmark the earliness of the detection.

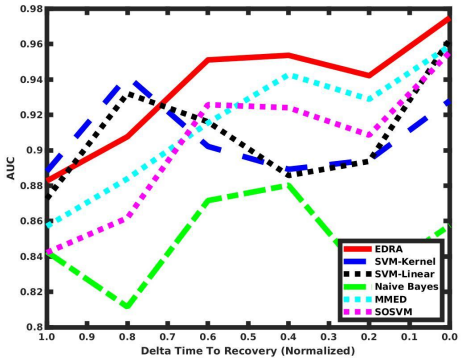
Similar to the simulation, we randomly divide the dataset for training and testing. The training dataset contains 80 percent of the whole data, and the testing dataset contains the rest 20 percent. We repeat our experiments 50 times and record the average performance.

During the testing, since each normal control has only one time point, we use the predicted scores for the single time point of the subjects without disease as the baseline scores, and compare the scores assigned for the patients at each normalized delta time point against the baseline scores. AUCs thus can be calculated for each normalized delta time point. The curves of AUC over the normalized delta time are depicted as Figure 3.8(a).

In this dataset, it's difficult to classify patients from the normal controls at the beginning due to the slow progression property of the chronic disease. However, at the second normalized time point (the eighth normalized delta time point), EDRA is able to detect the disease with the AUC of 0.84 ± 0.15, while the AUCs for MMED, SOSVM, SVM-kernel, SVM-linear and Naive Bayes at that time point are: 0.78 ± 0.17, 0.77 ± 0.16 and 0.74 ± 0.17, 0.58 ± 0.23 and 0.69 ± 0.19 respectively. In the end, EDRA still performs the best, with AUC 0.94 ± 0.07, while MMED, SOSVM and kernel SVM perform slightly worse with



(a) TrialNet: AUC over the normalized delta time to diagnosis



(b) IFNβ Drug Response: AUC over the normalized delta time to recovery

Figure 3.8: Real-world data experiments: AUC over the normalized delta time to the diagnosis/recovery

AUCs as 0.83 ± 0.11, 0.83 ± 0.11 and 0.84 ± 0.10. Regarding the earliness and accuracy of the detection, EDRA outperforms all the other classifiers for most extent of the trial. For the experiments on this dataset, the hyperparameter C for the SVM-based methods and the tuning parameter σ are selected based on 5-fold cross validation.

3.5.4 Longitudinal RNA-Seq data from IFNβ Drug Response Study

This section describes our experiment on the longitudinal RNA-Seq data from a drug therapy called Recombinant human interferon beta (rIFNβ), which is routinely used to control exacerbations in multiple sclerosis patients with only partial success, mainly because of adverse effects and a relatively large proportion of non-responders [34]. Therefore, early prediction and contemporaneous monitoring of the drug responses based on gene expression is important for doctors or researchers who would like to identify the suitable recipients of the specific drug therapy as well as to learn the long-term drug-induced effects.

The IFNβ drug response dataset is a longitudinal 70-gene expression dataset that contains the longitudinal gene expression data of 53 subjects. Patients with relapsing-remitting multiple sclerosis (MS) were followed for at least 2y after the initiation of therapy with IFNβ. Patients were classified as either good (33) or poor (20) responders at the end

of therapy based on strict criteria [34]. Blood sample was obtained during each clinical follow-up every 3 months after the initialization of the therapy with IFNβ in the 1_{st} year, and every 6 months in the 2_{nd} year. In the previous research, there are 23 genes identified as predictive [34, 127]. For detailed information about the genes identified as being predictive, readers can refer to the supplementary document of the work [127]. The weights for the features for constructing the kernels for EDRA are therefore determined based on the prior knowledge about the features' predictive power, i.e., the genes identified as not being predictive in literature are viewed as inactive features for the kernel construction. The AUC curves over the normalized delta time points prior to the recovery of EDRA and the other methods are depicted in Figure 3.8(b).

The experiments of IFNβ drug response dataset differ from the above experiments in the sense that there are pre-existing signatures that are able to separate good and poor responders before the initiation of the drug therapy, so that all the methods perform similarly well at the beginning. However, EDRA captures the long-term drug-induced progression via the increasing performance for classifying the good responders from the poor ones over time, while the other methods are not able to reflect the progression by the increased classification ability. This experiment demonstrates EDRA's contemporaneous risk evaluation performance. The hyperparameter C for the SVM-based methods and the tuning parameter σ for the kernel construction are selected based on 5-fold cross validation for this experiment.

3.6 Conclusions and Discussions

We propose EDRA, a contemporaneous risk detector that is trained with the aim of capturing the disease/drug-induced progression instead of individual data points, to address problems of early detection and contemporaneous risk assessment for the diseases with partiall-labeled longitudinal data. Our method is particularly suitable for detecting the onset of diseases with slow progression, which is hard to detect at the early stage. Experiments of varying situations from synthetic data to gene expression data of T1D study

and drug response study are adopted to evaluate the performance of the proposed methods. Specifically, to evaluate the performance of the proposed method on irregular longitudinal data, we consider irregularity and label insufficiency problems for synthesizing the data. The results obtained from the experiments demonstrate that EDRA enables early detection and contemporaneous risk assessment on irregular and partially labeled longitudinal data. It is not only able to detect the onset of disease earlier with higher accuracy compared with the other methods, but also monitor the disease progression contemporaneously in difficult classification situation, such as in the early stage of the disease. What's more, the experiments also demonstrate the advantage of the methods that consider the temporal structure within data for capturing the disease/drug-induced progression over the other methods. Furthermore, we propose a flexible mixed-kernel framework, which incorporates the prior knowledge about features' discriminating power for the kernel construction.

However, as we have discussed in Abstract and Chapter 1, the data can have complex correlation structures. The current feature representation using prior knowledge without considering the correlation within features is not optimal and may lead to inaccurate performance of EDRA when applied on extremely high dimensional data. Moreover, EDRA is not able to automatically identify risk-associated variables such as identifying disease/drug-associated biomarkers, which will be of great interest for researchers to interpret the risk detector, develop targeted therapeutic interventions and clinical study design, and improve the model's robustness. Finally, unwanted variation due to technical noises such as "batch effecs" is common in longitudinal data. For example, the longitudinal studies are often conducted in multiple sites, so that the site-specific variability exists in the collected data. Significant variations from batch effects often have negative impact on data analysis and need careful consideration. To address these issues to improve the robustness of EDRA, in next chapter, we propose a novel method to derive predictive features with variable selection ability by eliminating the unwanted variation from longitudinal data. Moreover, EDRA is suggested to be subsequently applied on the extracted features for more robust

early detection.

4. MSSPCA: ROBUST FEATURE LEARNING IN LONGITUDINAL DATA ANALYSIS

4.1 Overview

In this section, we propose a novel method, which we refer to as MSSPCA, to robustly derive predictive features and select discriminating variables from the data containing the unwanted variations. MSSPCA addresses the challenges faced by the existing batch-effect correction methods due to the practical limitations such as noisy outcomes, insufficient prior knowledge regarding the negative control variables, or small training sample size. By aggregating signals from the input data and the outcome information with the data heterogeneity modeled in a probabilistic PCA framework, MSSPCA discovers the underlying factors of interest with variable selection capability and being robust to the noisy outcomes.

4.2 Introduction

Many methods developed for feature extraction have been proven to be successful in reducing the data dimensionality for data processing and analysis [7, 8, 9, 10, 11, 12]. It is also useful to understand how different variables contribute to prediction, especially when they have physical meanings, and further enhance the model's performance by selecting the most important variables [13, 14, 15, 16, 17, 18]. Variable selection is widely applied in analyzing RNA-Seq data where each variable corresponds to a specific gene or transcript [19, 20, 21, 22, 23]. However, the assumption that the data are collected from the same population held by most of these methods can be easily violated. Ignoring the unwanted effects can be harmful for developing the downstream analysis and may introduce difficulties to perform integrative analysis, as witnessed by the poor cross-site validations in many recently reported studies [2, 27, 28]

Methods have been proposed to address the serious issue of the unwanted data variability. Nevertheless, most of the existing methods are either regression-based that infer

the unwanted effects based on the residuals from the regression of the input observations on the corresponding outcomes such as disease stage or applying factor analysis on the reduced data restricted to the "negative control" variables that are known a priori not to be associated with respect to the biological factor of interest, or a hybrid of these two techniques [101, 128, 105, 129, 109, 108, 106, 130, 107]. However, neither of these two requirements can be easily satisfied in many real-world situations. Imagine the situations where a patient goes through multiple disease states before diagnosis, so that it's difficult or resource-demanding to measure the outcomes associated with the disease progression. However, since patients at the "early stage" usually have longer "delta time to diagnosis" on average, "delta time to diagnosis" can be adopted as a noisy surrogate for estimating the underlying disease risk state. Nevertheless, such noisy outcome information is insufficient for a fully supervised method to derive factors or identify variables that are associated with primary interest, like disease progression. What's more, methods that infer the unwanted effects using factor analysis restricted to the negative control variables are sensitive to the choice of negative controls [108, 109, 110], which can be wrongly selected using the noisy outcomes. Recently proposed methods based on deep models require large training sample size, and perform poorly with small training datasets [114, 113, 115]. With these practical limitations, these existing methods may suffer from unstable performance and therefore are not applicable or have limited power in many real-world situations.

In this chapter we propose a novel method, namely Matched Supervised Sparse Principal Component Analysis (MSSPCA), which is capable of extracting features as potential biomarkers, identifying contributing variables, and being robust to the noisy outcomes. MSSPCA employs a supervised PCA framework with sparse estimation of the loading matrix to aggregate the signals from both the input predictors and the response data [131, 21, 132, 133, 134], different from the existing regression-based methods [101, 128, 129, 107, 135]. Moreover, rather than dividing the algorithm into separate steps for correcting the unwanted variations and then deriving biomarkers, MSSPCA proposes an efficient algo-

rithm that iteratively estimates the effects of both targeted (as predictive biomarkers) and the unwanted variations until convergence. We assess the performance of MSSPCA on both simulated data and a real-world case study.

The rest of this chapter is organized as follows. Section 4.3 reviews the concepts of PCA and Sparse PCA from the probabilistic perspective. In Section 4.4, we describe the framework of MSSPCA, and propose an efficient algorithm with closed-form updating rules. Section 4.5 demonstrates the effectiveness of MSSPCA using simulated data and a real-world case study. Section 4.6 concludes this chapter.

4.3 Background

Principle component analysis (PCA) is a ubiquitous technique for data analysis and dimension reduction. In this section, we revisit some concepts of PCA's probabilistic formulation and sparse PCA for variable selection, which are the important building blocks of our proposed MSSPCA.

4.3.1 Probabilistic Principle Component Analysis

PCA is a well-established technique for dimension reduction. The most common derivation of PCA aims at finding a linear orthonormal transformation that maximize the variance in a subspace space with a lower dimension [136]. Such vectors for transformation are called PC loadings and the projected data are the corresponding PCs.

An alternative interpretation of PCA was proposed by Tipping in 1999, which employs a probabilistic formulation of PCA from a Gaussian latent variable model [9]. Such a probabilistic model is appealing since its framework is more flexible to model data from the probabilistic perspectives, which offers the potential to extend the scope of conventional PCA [9]. What's more, the latent-variable model of probabilistic PCA naturally transforms the optimization problem into a maximum-likelihood estimate of the parameters associated with PCs, which can lead to iterative and computationally-efficient algorithms [9].

From the probabilistic perspective of PCA, an observation $x_n \in R^p$ can be considered

as a "noisy-corrupted" version of some clean data point θ_n, which we refer to as "canonical parameters", so that x_n can be represented as $x_n = \theta_n + \epsilon$. Assume that ϵ is the isotropic Gaussian noise that follows $N(0, \sigma^2 I)$, the conditional distribution of x_n given θ_n is:

$$P(x_n|\theta_n) \sim N(\theta_n, \sigma^2 I) \tag{4.1}$$

Since a latent variable model seeks to linearly relate a p-dimensional vector x_n to a q-dimensional vector of latent variables z_n ($q < p$), θ_n is further factorized with the form $\theta_n = W^T z_n + b$. The $p \times d$ orthonormal matrix W is the PC loading matrix that relates the observations x_n with the latent variables z_n, and b is a vector of parameters for bias. Therefore, the conditional probability of x_n given z_n is:

$$P(x_n|z_n) \sim N(W^T z_n + b, \sigma^2 I) \tag{4.2}$$

The estimation of the unknown parameters associated with PCs and their corresponding loadings turns into a problem for maximizing the likelihood of the sample observations with respect to W, z_n and b, where the optimization framework is formulated as:

$$\min_{W^T W = I} \sum_{i=1}^{N} ||x_n - (W^T z_n + b)||^2 \tag{4.3}$$

The formulation (4.3) can also be interpreted as to solve a low-rank approximation problem that aims at minimizing the squared reconstruction error between x_n and $\hat{\theta}_n$, where $\hat{\theta}_n = W^T z + b$ is considered as the linear reconstruction of x_n. Such interpretation links the optimization of problem (4.3) to linear regression and singular value decomposition (SVD) that are well-established with computationally-efficient algorithms for parameter estimates.

4.3.2 Sparse Principle Component Analysis

Interpretations of the derived PCs and the corresponding PC loadings are useful, especially when the variables have physical meanings. For example, in microarray data each

variable corresponds to a specific gene [21]. Many approaches have been proposed to address the issue in PCA model's interpretation by identifying the variables that are contributing to PCs' derivation [137, 138, 139, 140, 141, 142, 143, 131, 88].

One group of the methods proposed to derive sparse PC loadings are based on the probabilistic PCA's framework, and reformulate PCA as a penalized regression problem, with the sparsity promoted by imposing the l_1 norm penalty on the regression coefficients. One major issue for solving sparse PCA problem is that the orthonormal constraints and the l_1 penalty are simultaneously imposed on the PC loadings.

To address this issue, Shen *et al.* proposed a method called sparse PCA via regularized SVD (sPCA-rSVD), which utilizes the connection of PCA with SVD of a data matrix. The PCs are extracted by solving a low rank matrix approximation problem that adopts a penalized least square criterion, and the regularization penalties are introduced to the corresponding least square regression problem to encourage the sparsity in PC loadings [21]. Denote the sample matrix by X and suppose zw^T is the best rank-one approximation of X, the optimization problem of sPCA-rSVD is formulated as the follows:

$$\min_{||z||=1} ||X - zw^T||_F^2 + \lambda||w||_1 \tag{4.4}$$

Minimizing problem (4.4) with respect to z and w under the constraint $||z|| = 1$ can be solved efficiently with an iterative algorithm that first considers the problem to optimize over one parameter with the other fixed, and alternatively minimize the problem until it converges. Since each time only one parameter is considered, the optimization turns into solving a LASSO (least absolute shrinkage and selection operator)-regularized problem each time and the closed-form updating rules can be easily derived. We have derived the convergence rate for such iterative updating procedures in Lemma 2, and the corresponding proof can be found in Appendix B.

Lemma 2.

$$\min_{1\leq t\leq T} ||w_t - w_{t+1}||^2 \leq \frac{2}{T} f(z_1, w_1) \tag{4.5}$$

where t indicates the t-th iteration and T is the number of total iterations. $f(z, w) = ||X - zw^T||_F^2 + \lambda||w||_1$ *with* $||z|| = 1$.

4.4 Method

4.4.1 Data modeling

Denote the input observations by $X = [x_1, ..., x_n, ..., x_N]^T$, an $N \times p$ matrix that contains N samples with p variables. We assume that there are G batches that could be experiments, populations or study sites where the data points are collected from. Let x_n denote the input observation of the n-th sample, which can be decomposed into the effects associated with primary interest and batch effects as the follows:

$$x_n = Wz_n + V\gamma_{i(n)} + b + \epsilon_x, \tag{4.6}$$

where z_n is the low-dimensional latent variable that contains factors of primary interest to serve as potential biomarkers for the n-th sample and W is the corresponding loading matrix capturing contributions of each variable to the corresponding factor; $\gamma_{i(n)}$ corresponds to the influencing factors of systematic batch effects such as experiment/population/site-specific data heterogeneity, where $i(n) \in [1, ..., G]$ indicates the batch where the n-th sample comes from, with V as the corresponding loading matrix. ϵ_x denotes the unknown noise, and b is the bias vector of the input observations.

Here we consider the situations where the samples can be partially labeled (e.g. only the samples from certain batches have the outcome information), while the prediction and the data correction for unlabeled samples is still desired. We assume that only the first N' ($N' \leq N$) samples have the outcome information, so that we denote the input matrix X by $X = \begin{bmatrix} X_L \\ X_U \end{bmatrix}$, where X_L and X_U are the input matrices of the labeled and the unlabeled

samples, respectively. Let $Y = [y_1, ..., y_l, ..., y_{N'}]^T$ be the outcome matrix corresponding to X_L, where $y_l \in R^{d\times 1}$ and d is the number of the outcome variables. The outcome information y_l can be modeled as:

$$y_l = Bz_l + b_0 + \epsilon_y, \tag{4.7}$$

where B is the matrix of regression coefficients; b_0 is the bias vector of the outcomes and ϵ_y represents the outcome unknown noise.

A probabilistic graphical model illustrating MSSPCA is provided in Fig. 4.1. On the left side of the graphical model, the latent variable $\gamma_{i(l)}$ for the systematic technical noise is shared by all the samples collected from the same batch, and the parameters associated with the batch effects, γ and V, can be estimated from the residuals of the input observations after removing the effect of interest. On the right side, assume that the input observation x_l has been corrected for the unwanted effect, the latent factors of primary interest z_l is jointly determined by the adjusted input and the corresponding outcome y_l, so that the loading matrix W can be estimated with only the samples accompanied with the outcomes. Moreover, by jointly modeling the input observations and the outcomes of interest using the common latent variables z, MSSPCA aggregates the information from these two information sources to extract the latent low-dimension features. An iterative optimization algorithm is proposed to estimate the parameters of the effects associated with primary interest and batch effects in an alternative manner, and the detailed descriptions are presented in Sec. 4.4.3.

4.4.2 Model inference

The unknown parameters including PCs and their corresponding loading matrices can be estimated using the approximation formulation by a penalized least square criterion, with the sparsity regularization term imposed on PC loadings matrix W that corresponds

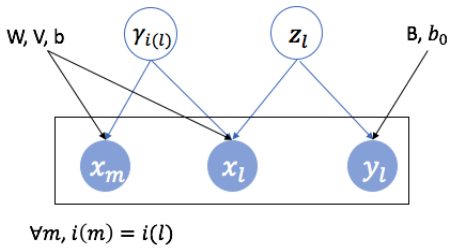


Figure 4.1: Probabilistic graphical model for x and y conditioning on the latent variables γ and z that represent the batch effects and the factors of primary interests, respectively.

to the effects associated with primary interest as the follows:

$$\begin{aligned}\min_{Z_L^T Z_L = I, B, W} \min_{\Gamma^T \Gamma = I, V, b, b_0} & \sum_{l=1}^{N'} (||x_l - W z_l - V\gamma_{i(l)} - b||^2 + \alpha ||y_l - B z_l - b_0||^2) \\ & + \sum_{u=N'+1}^{N} ||x_u - W z_u - V\gamma_{i(u)} - b||^2 + \sum_k \lambda_k ||w_k||_1\end{aligned} \tag{4.8}$$

where $Z_L = [z_1, ..., z_{N'}]^T$ is the matrix of the risk-associated latent variables for the labeled samples, where $z \in R^{k \times 1}$ and k is the number of the factors for the wanted effects; $\Gamma = [\gamma_1, ..., \gamma_G]^T$ is the matrix of the latent variables for batch effects, where $\gamma_{i(n)} \in R^{k_b \times 1}$ and k_b is the number of the factors for the unwanted effects. λ is a vector of hyperparameters for controlling the sparsity degree and w_k is the k-th column of matrix W. The orthonormal constraint is imposed on Z_L and Γ for model identifiability.

Denote the batch information by a binary indicator matrix $U \in R^{N \times G}$, where $U_{ni} = 1$ if the n-th sample belongs to batch i. Meanwhile, let $Z = [z_1, ..., z_N]^T$ be the matrix of the latent variables z for all samples. In the matrix form, the optimization problem (4.8) can be finally written as:

$$\begin{aligned}\min_{Z_L^T Z_L = I, B, W} \min_{\Gamma^T \Gamma = I, V, b, b_0} & ||X - ZW^T - U\Gamma V^T - 1b^T||_F^2 + \alpha ||Y - ZB^T - 1{b_0}^T||_F^2 \\ & + \sum_k \lambda_k ||w_k||_1.\end{aligned} \tag{4.9}$$

In special situations, the unknown noise $\epsilon_x \in N(0, \sigma_x^2 I)$ and $\epsilon_y \in N(0, \sigma_y^2 I)$, so that the conditional probabilities of x_n and y_l are:

$$p(x_n|z_n, \gamma_{i(n)}; W, V, b) \sim N(Wz_n + V\gamma_{i(n)} + b, \sigma_x^2 I); \tag{4.10}$$

$$p(y_l|z_l; B, b_0) \sim N(Bz_l + b_0, \sigma_Y^2 I). \tag{4.11}$$

The joint log-likelihood of x_n and y_l can be written as:

$$\begin{aligned}
&logp(X, Y|Z, \Gamma, W, V, B, b, b_0) \\
&= log \prod_{n=1}^{N} p(x_n|z_n, \gamma_{i(n)}, W, V, b) \prod_{l=1}^{N'} p(y_l|z_l, B, b_0) \\
&\propto log \prod_{n=1}^{N} exp(-||x_n - Wz_n - V\gamma_{i(n)} - b||^2 - \alpha \prod_{l=1}^{N'} ||y_l - Bz_l - b_0||^2) \\
&= -(\sum_{n=1}^{N} ||x_n - Wz_n - V\gamma_{i(n)} - b||^2 + \alpha \sum_{l=1}^{N'} ||y_l - Bz_l - b_0||^2),
\end{aligned} \tag{4.12}$$

We maximize the log-likelihood with respect to the unknown parameters, so that the optimization problem can be reformulated as minimizing the following loss function:

$$\min_{Z_L^T Z_L = I, B, W} \min_{\Gamma^T \Gamma = I, V, b, b_0} \sum_{n=1}^{N} ||x_n - (Wz_n + V\gamma_{i(n)} + b)||^2 + \alpha \sum_{l=1}^{N'} ||y_l - (Bz_l + b_0)||^2 \tag{4.13}$$

The above optimization formulation links the probabilistic PCA to the optimization formulation presented in (4.8), which can viewed as a penalized version of (4.13) with a Laplacian prior imposed on the loading matrix W. The probabilistic interpretation offers the potential to extend the current problem to generalize the data of exponential family distributions, which have been extensively studied in [111, 144, 145].

4.4.3 An iterative optimization algorithm with closed-form updating rules

To solve the optimization problem (4.9), we can iteratively solve two sub-problems for extracting the predictive biomarkers and removing the unwanted variations, i.e., each time we fix the parameters for one effect, and optimize the other. Both sub-problems have closed-form updating rules and it usually takes a few iterations for convergence based on our experience.

4.4.3.1 Updating rules for Z and W

To estimate the factor of interest Z and its loading matrix W, we assume that the parameters for batch effect Γ and V as well as the bias vectors b and b_0 are given. Denote the centered adjusted inputs for all samples by $X' = X - U\hat{\Gamma}\hat{V}^T - 1\hat{b}^T$, while X'_L is only for the samples with the outcomes. Define $\tilde{X}_L = [X'_L, \sqrt{\alpha}(Y - 1\hat{b}_0^T)]$ and $\tilde{W} = \begin{bmatrix} W \\ \sqrt{\alpha}B \end{bmatrix}$, so that the problem (4.9) can be rewritten as the follows:

$$\min_{Z_L, \tilde{W}, s.t. Z_L^T Z_L = I} ||\tilde{X}_L - Z_L \tilde{W}^T||_F^2 + P_\lambda(\tilde{W}), \tag{4.14}$$

where $P_\lambda(\tilde{W}) = \sum_{k=1}^{l} \lambda_k ||\tilde{w}[1:p]||_1$.

We solve (4.14) in an iterative manner. Each time we assume one parameter is known so that (4.14) turns into a simple (sparse-regularized) least square regression optimization problem. Algorithm 2 describes the detailed updating procedures for estimating Z and W.

4.4.3.2 Updating rules for Γ and V

Given the parameters associated with the effects of interest and the bias vectors, set $X_b = X - \hat{Z}\hat{W}^T - 1\hat{b}^T$, then the optimization problem with respect to the batch-effect parameters Γ and V becomes:

$$\min_{\Gamma, V, s.t. \Gamma^T \Gamma = I} ||X_b - U\Gamma V^T||_F^2 \tag{4.15}$$

Algorithm 2: Loss Function (4.14)

Data: Centered adjusted inputs for all samples $X' \in R^{N \times p}$, and for only the labeled samples $X'_L \in R^{N' \times p}$; outcome matrix $Y \in R^{N' \times d}$; bias vectors b and b_0; latent variable z's dimension k; sparsity degree of W's lth loading vector λ_l, $l = 1, ..., k$.

Result: $\hat{Z}$, $\hat{W}$, $\hat{B}$

Set $\tilde{X}_L = [X'_L, \sqrt{\alpha}(Y - 1\hat{b}_0^T)]$;

for $l \leftarrow 1$ *to* k **do**

  Apply standard SVD on $\tilde{X}_L$ and obtain the best rank-one approximation of $\tilde{X}_L$ as suv^T. Initialize $\tilde{w}_{old}$ as v;

  while *True* **do**

    Update $z_{new} = \tilde{X}_L \tilde{w}_{old} / ||\tilde{X}_L \tilde{w}_{old}||$;

    Calculate $w = \tilde{X}_L[:, 1:p]^T z_{new}$;

    Update $w_{new} = sign(w)(|w| - \lambda_k)_+$;

    Update $\beta_{new} = \tilde{X}_L[:, p+1:p+d]^T z_{new}$;

    Update $\tilde{w}_{new} = \begin{bmatrix} w_{new} \\ \sqrt{\alpha}\beta_{new} \end{bmatrix}$;

    if *convergence* **then**

      $\hat{W}[:, l] = w_{new}$;

      $\hat{B}[:, l] = \beta_{new}$;

      $\hat{Z}_L[:, l] = z_{new}$;

      $\tilde{X}_L = \tilde{X}_L - z_{new}\tilde{w}_{new}^T$;

      Stop and go back to the for loop;

    else

      Update $\tilde{w}_{old}$ by $\tilde{w}_{new}$;

    end

  end

end

Normalize the lth loading vector w_l by $||w_l||$, and re-scale β_l by $\beta_l / ||w_l||$;

Re-order the columns of $\hat{W}$ and $\hat{B}$ by the decreasing order of $||\beta_l||$;

Calculate $\hat{Z}$ for all samples by $\hat{Z} = X'\hat{W}$.

Algorithm 3 provides the updating procedures for solving the sub-problem (4.15).

Algorithm 3: Loss Function (4.15)

Data: Centered unwanted effects $X_b = X - \hat{Z}\hat{W}^T - 1\hat{b}^T$, binary indicator matrix U, number of batch factors k_b
Result: $\hat{\Gamma}$, $\hat{V}$
for $l \leftarrow 1$ *to* k_b **do**
  Apply standard SVD on $(U^TU)^{-1}X_b$ and obtain the best rank-one approximation of $(U^TU)^{-1}X_b$ as suv^T. Initialize v_{old} as v;
  while *True* **do**
    Calculate $\gamma = (U^TU)^{-1}X_b v_{old}$;
    Update $\gamma_{new} = \gamma/||\gamma||$;
    Update $v_{mew} = X_b^T U\gamma_{new}(\gamma_{new}^T U^T U\gamma_{mew})^{-1}$;
    if *convergence* **then**
      $\hat{\Gamma}[:, l] = \gamma_{new}$;
      $\hat{V}[:, l] = v_{new}$;
      $X_b = X_b - U\gamma_{new}v_{new}^T$;
      Stop and go back to the for loop;
    else
      Update v_{old} by v_{new};
    end
  end
end
Normalize the lth loading vector v_l and re-scale γ_l by $\gamma_l||v_l||$;

4.4.4 Prediction with new data

With the projection matrices W and V estimated from the training data X_{train}, MSSPCA first adjusts the new samples for removing the batch effect: $\hat{X}_{new}^c = (X_{new} - 1\hat{b}^T)(I - \hat{V}\hat{V}^T)$. The predictive features of new data Z_{new} can thus be extracted from $\hat{X}_{new}^c$ with the estimated projection matrix $\hat{W}$: $\hat{Z}_{new} = \hat{X}_{new}^c\hat{W}$. In addition to extracting the predictive features from the new samples, the outcome response $\hat{Y}_{new}$ can be estimated by: $\hat{Y}_{new} = \hat{Z}_{new}\hat{B}^T + 1\hat{b}_0^T$.

4.5 Results

We investigate the performance of MSSPCA with the experiments using both simulation and real-world data. In the simulation study, we examine the performance of the proposed method in terms of its ability to robustly extract features and identify variables associated with the underlying risk change by removing the unwanted batch effects.

The real-world study contains a longitudinal dataset for Tuberculosis (TB) Disease that are collected from three countries for early prediction of the disease onset and identification of risk signatures. The study has shown that there exists significant population variability, which can be reflected by the poor cross-site prediction validations and the site-specific feature selection results [2]. In this real-world case study, we demonstrate the benefits of our proposed method through the improved accuracy of disease prediction as well as the robustness of risk signature identification under the significant population heterogeneity.

Comprehensive performance evaluation with respect to robust feature extraction and variable selection involving the unwanted data variability is also presented with comparison of the state-of-the-art methods in both simulation and real-world studies.

4.5.1 Simulation Study

In this simulation study, we evaluate and compare the performance of all methods in terms of their abilities of: (1) extracting the aggregated features that reflect the disease progression; (2) identifying the important contributing variables. To demonstrate the benefits of our method, we consider three popular algorithms for batch-effect correction for comparison, which include SVA, RUV, and ComBat [128, 103, 105].

4.5.1.1 Evaluation Criteria

To assess the ability of the proposed method for extracting the features and identifying the variables of interest from the data containing unwanted effects, we adopt the following evaluation metrics.

1 Average Silhouette Score (ASS):

Average silhouette score (ASS) is computed to assess whether the extracted features can help define the disease stages while removing the batch effects. The silhouette score of a data point is the difference between its average distance to the within-cluster members and the average distance to all members of the other clusters, divided by the larger of the two values. The resulting score ranges from -1 to 1, where a high score indicates that the data point fits well in the current cluster [146]. To combine the assessment for risk stage separation and batch-effect removal, we calculate the harmonic mean (F1 score) of ASS_{stage} and the batch mixing $(1 - ASS_{batch})$:

$$ASS_{F1} = \frac{2(1 - ASS_{batch})(ASS_{stage})}{1 - ASS_{batch} + ASS_{stage}} \tag{4.16}$$

2 Variable selection

We evaluate the variable selection performance for different methods with True Positive Rates (TPR) and False Positive Rates (FPR):

- $$TPR = \frac{\#true\ positives}{\#true\ variables}, \tag{4.17}$$

where $\#true\,positives$ is the number of the selected important variables, and $\#true\ variables$ is the total number of the contributing variables;

- $$FPR = \frac{\#false\ positives}{\#false\ variables}, \tag{4.18}$$

where $\#false\ positives$ is the number of the selected variables but of no interest and $\#false\ variables$ is the total number of the unimportant variables.

4.5.1.2 Data Generation

In the simulation experiments, we synthesize a longitudinal dataset containing $N = 100$ subjects, each having 7 to 10 temporal input observations with 100 variables ($p = 100$).

Therefore, each temporal input observation is in R^{100}. We consider the situation where the subjects come from 3 different batches (e.g. study sites), and go through 3 disease stages as the disease develops.

For subject n, we first decide the number of time points n_t by randomly sampling from a uniform distribution [7, 10]. Then we randomly assign the disease progression stages for the n_t data points with the stages sorted from 1 to 3, i.e, $S[n] = [1, 1, 1, 2, 2, 2, 3]$, with a fixed order from 1 to 3 to simulate the disease progression. Assume that the n-th subject comes from batch i, the input observation of the j-th variable, $j = 1, ..., 100$, at the k-th visit, can be generated as the follows:

$$x_{n,k,j} = w_j z_{n,k} + \rho v_j \gamma_{i(n)} + \epsilon_{n,k,j} \tag{4.19}$$

where w_j is the j-th row of the loading matrix W, and $z_{n,k} = r_{nk_1} z_1 + r_{nk_2} z_2 + r_{nk_3} z_3$ with r_{nk_s} as a binary indicator to indicate whether the datapoint $x_{n,k}$ is at stage s or not. Figure 4.2 (left) provides a simple illustration of z in a two-dimensional latent space that captures the disease progression, and our goal is to recover the latent factor z from the high-dimensional data that contains the unwanted effects. In the simulation experiments, the bias vector b are treated as 0 without loss of generality.

To simulate the unwanted effects, we consider both known and unknown unwanted data variability. We randomly assign subjects to one of the three batches and generate the corresponding batch effect on the j-th variable by $\rho v_j \gamma_{i(n)}$, where ρ controls the severity of the batch effect. Similarly, v_j is the j-th row of batch loading matrix V and $\gamma_{i(n)}$ is the corresponding latent factors associated with the effect from batch $i(n)$ where the n-th subject comes from. The unknown noise ϵ_x is composed by the unknown variation, which we refer to as $\xi_{n,k,j}$, plus a random independent noise sampled from $N(0, \sigma_x^2)$. Here the unknown variation $\xi_{n,k,j}$ is generated by $\eta_j f_{n,k}$ where η_j is the j-th row of the unknown effect matrix H, and $f_{n,k}$ is randomly sampled from $N(0, \sigma_e^2 I)$.

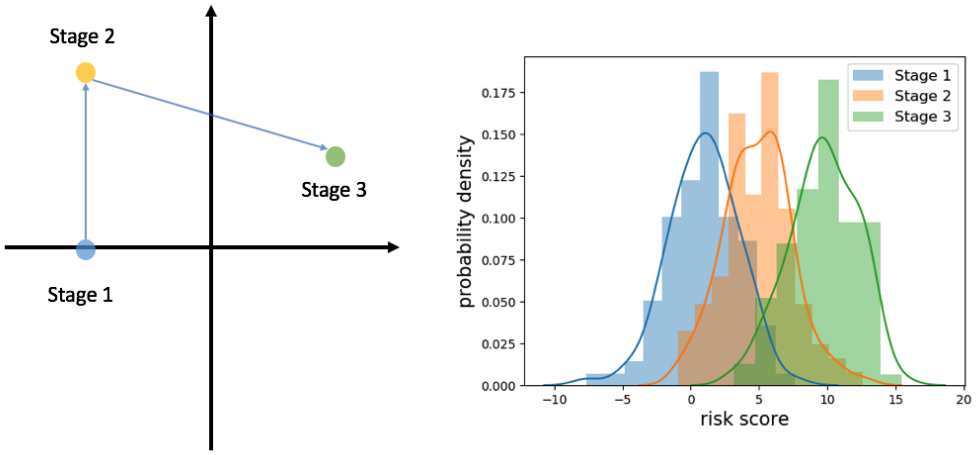
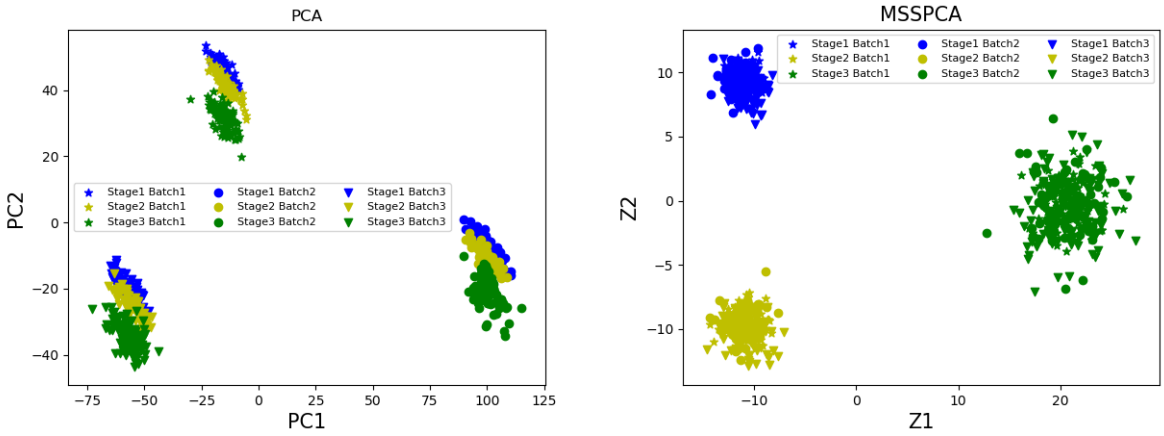


Figure 4.2: Data generation. Left: Latent variable that reflects the disease progression in a 2-dimensional space; Right: Distributions of the outcomes for different risk stages.

For loading matrices W, V and H, we generate three $p \times k_i$ ($i = W, V, or H$) matrices with orthonormal columns, where p is the number of input data's variables and k_i is the number of factors for each effect. The loading matrix W for simulating risk effect is sparse, since there's only a small set of variables that are associated with the disease progression.

Finally, since many real-world studies for early disease detection are conducted before the disease onset, it's difficult to obtain clean outcomes for disease progression. Therefore, predictors, such as "delta time prior to diagnosis", are often adopted as the "noisy surrogates" for estimating the true underlying risk. Patients at the early stage of disease usually have longer time to diagnosis on average. However, there exist overlaps among the "delta time prior to diagnosis" for patients at different stages. To simulate such situations, the observed noisy outcome for subject n at his kth visit is generated by sampling from the distribution according to the risk stage $s(n, k)$, which is an index from 1 to 3 that denotes the risk stage, as shown in (4.20). The right panel of Figure 4.2 provides a schematic example of the outcome distributions for the 3 risk stages.



(a) (b)

Figure 4.3: PCs of the data contaminated by batch effects (a) before and (b) after data correction with MSSPCA. Standard PCA fails to capture the underlying disease stage due to the significant unwanted batch effects. However, MSSPCA extracts the low-dimensional features that cluster the data points by disease stages.

$$y_{n,k} \sim N(\mu_{s(n,k)}, \sigma_{s(n,k)}) \tag{4.20}$$

Figure 4.3 gives a comparison of the low-dimensional features extracted by a standard PCA and MSSPCA from the data contaminated by unwanted effects. It can be seen from Figure 4.3 (a) that the PCs estimated by a standard PCA fail to capture the underlying disease stages from the data without corrections. Instead, the data points are clustered by batches due to the significant batch e ffects. However, Figure 4.3 (b) demonstrates that by incorporating the unwanted data variability, MSSPCA extracts the low-dimensional features that reflect the biological factors of interest with the data points clustered by disease stages.

In the following context, we consider three representative batch correcting algorithms: ComBat, RUV and SVA [101, 128, 105, 24, 103] in conjunction with PCA [9] and LASSO [7], for comparing and evaluating the performance in terms of the extracted features and se-

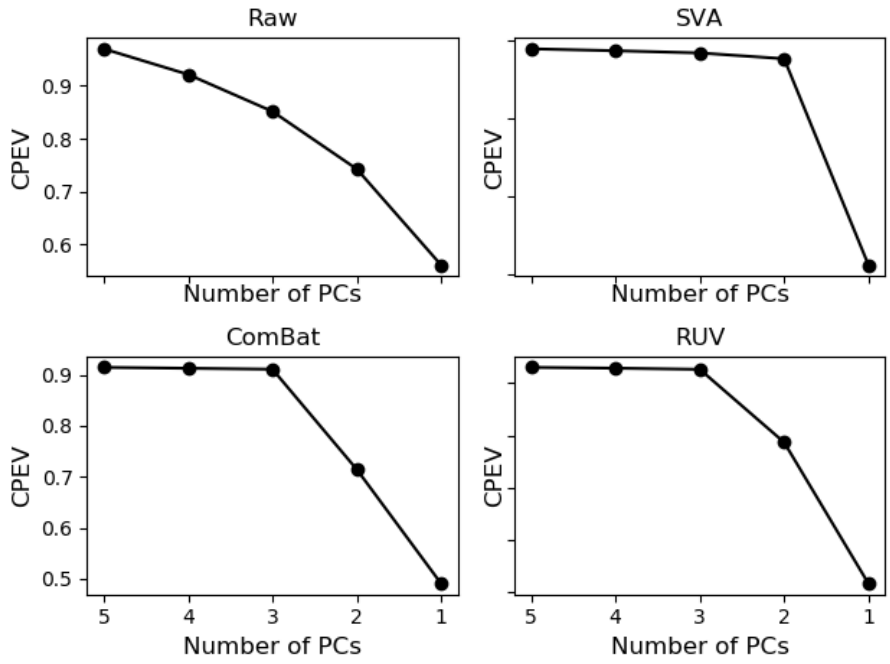


Figure 4.4: Percentage of the explained variance for: (1) original data, (2) data corrected by SVA, (3) data corrected by ComBat, (4) data corrected by RUV.

lected variables after adjusting the unwanted effects.

4.5.1.3 Experimental Results

In the simulation experiments, we consider the synthetic data covering different severity levels of the batch effects, i.e., $\rho = 20, 30, 40, 50$. For each scenario we generate 50 data sets and evaluate the performance using the average results for each evaluation metric. Both SVA and ComBat require primary variables, i.e., risk scores, and batch information for data correction. RUV estimates and removes the unwanted variation using "control genes" that are assumed not to be influenced by effects of interest [105]. Since there is no prior knowledge about the possible control variables in our setup, we follow the instructions of Risso *et al* and obtain the least significantly differential expressed (DE) genes with the estimated $FDR \geq 0.05$ based on a first-pass DE analysis performed prior to RUV normalization [105, 147, 148]

After correcting the data for the unwanted variations, PCA is performed subsequently.

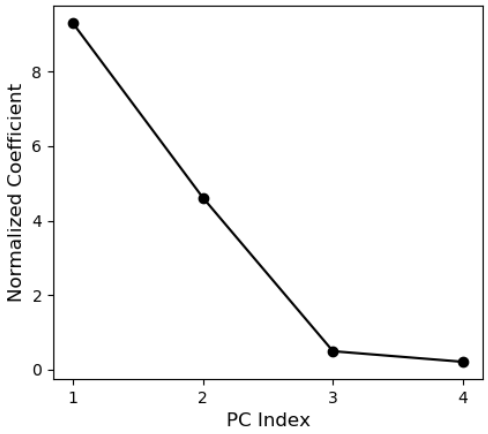


Figure 4.5: Row-wise l_2 norm of the estimated normalized loading matrix $\hat{B}$. It measures the association between the PC scores and the outcomes.

To determine the number of PCs k, we plot the cumulative percentage of explained variance (CPEV) against the number k of PCs to choose the k at which CPEV does not drop too much. Figure 4.4 shows the plots for each method, which suggests $k = 4, 2, 3, 3$ for the original data and the corrected data by SVA, ComBat and RUV, respectively. To determine the number of PCs for MSSPCA, we rescale the l-th column of the estimated $\hat{B}$ by $\sqrt{\hat{z}_l{}^T \hat{z}_l}$, where $\hat{z}_l$ is the l-th column of the estimated PC matrix $\hat{Z}$, and compute the l_2 norm of the l-th column of the rescaled matrix to measure the importance of the l-th PC to the outcomes of primary interest. Figure 4.5 demonstrates the standarized $\hat{B}$ over PCs, which indicates that we should use the top 2 PCs of MSSPCA, since the loadings for the rest of the PCs are close to zero.

Figure 4.6 demonstrates the average silhouette scores (ASS) for the extracted features in conjunction of disease stage separation and batch mixing for all the methods. The methods appearing in the upper left corners of the plots are good performing methods where the risk stages are clearly separated with high ASS_{Stage} and the samples from multiple batches are well integrated with low ASS_{Batch}. Figure 4.6 shows that all competing algorithms help adjusting the unwanted data variation by clustering in the upper left corners while the

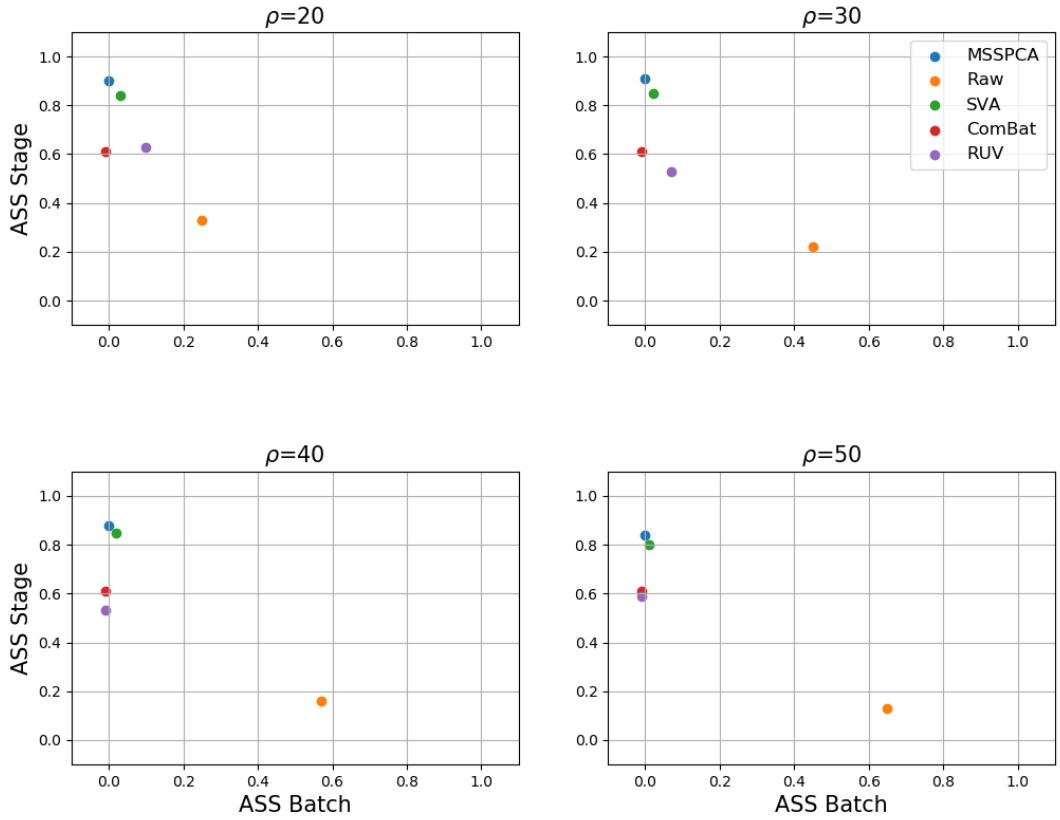


Figure 4.6: Quantitative evaluation of the clustering performance using Average Silhouette Scores. Methods appearing on the upper left corner are good performing methods.

unadjusted "Raw" data moves toward the lower right corners as the severity of the batch effect increases.

To combine the stage separation and the batch mixing assessment, we compute the harmonic mean (F1 score) of ASS_{stage} and 1-ASS_{batch}, whose results are shown in Figure 4.7. SVA obtains good combined scores, but its performance is unstable with a high variance in its F1 scores, which could be caused by the noisy outcomes by which SVA calculates the residuals to estimate the factors of the unwanted effects. In contrast, the performance achieved by ComBat is stable, since ComBat employs model-based Location and Scale (L/S) and it estimates the model's parameters by Empirical Bayesian on the residuals from

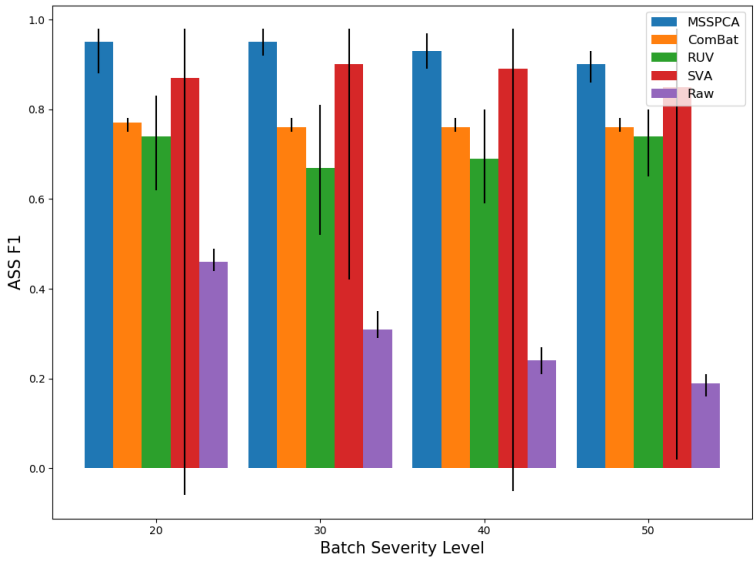


Figure 4.7: Harmonic mean (F1 score) of Average Silhouette Scores in conjunction with risk stages and batches.

the regression of the observed inputs on the corresponding outcomes. However, ComBat only removes the batch effects, so that other unknown effects may still exist in the adjusted data. Although RUV shows its effectiveness for removing the unwanted effects with F1 scores much higher than the unadjusted "Raw" data, its performance is overall worse than the other methods and of high variability. RUV has been found to be quite sensitive to the choice of control genes [105] and the noisy outcomes would lead to wrong identification of the variables of no interest. What's more, RUV estimates the factors of the unwanted effects based on only control variables with the assumption that data heterogeneity does not influence the "target" variables. However, such an assumption can be easily violated in many situations when there is an overlap between "control" and "target" variable sets, which may cause a biased estimation of the unwanted effects. By aggregation of the input observations and the outcomes to determine the underlying latent variables, MSSPCA outperforms all competing methods in all scenarios, which demonstrates its usefulness for extracting the effects of interest by incorporating the unwanted data variability and being robustness to the noisy outcomes.

Methods	$\rho = 20$	$\rho = 30$	$\rho = 40$	$\rho = 50$
Raw/LASSO	0.812/0.121	0.772/0.030	0.756/0.122	0.788/0.127
SVA/LASSO	0.732/0.029	0.748/0.030	0.724/0.015	0.712/0.016
ComBat/LASSO	0.784/0.023	0.780/0.029	0.768/0.014	0.792/0.015
RUV/LASSO	0.832/0.046	0.820/0.032	0.808/0.023	0.840/0.025
MSSPCA	**1.000/0.01**	**1.000/0.000**	**1.000/0.001**	**1.000/0.001**

Table 4.1: True positive rate (TPR)/false positive rate (FPR) for the variables identified based on the data corrected by methods for comparison. LASSO is subsequently applied for variable selection on the original and the adjusted data.

We further examine the proposed method's performance on identification of variables that are associated with disease progression, and analyze how batch-effect correcting methods improve the variable selection. LASSO is applied subsequently on the corrected data (except MSSPCA) to regress it over the outcomes (risk scores). The sparsity degrees for all the methods are determined using a 5-fold cross validation for each data set. Table 4.1 presents the TPR/FPR achieved by the methods for comparison over varying severity levels of the batch effects. The results for the original "Raw" data show that more false positives are selected when data being more "contaminated". SVA and ComBat effectively reduce the FPRs for variable selection. However, ComBat performs better than SVA with higher TPRs, which is consistent with the clustering results that ComBat obtains more stable performance than SVA by using EB for L/S model parameters' estimation. RUV tends to select more variables since it has both higher TPRs and FPRs, which could be caused by the biased estimation of the unwanted variation, since RUV estimates the unwanted data variation only based on measurements of control variables, which are empirically selected using the noisy outcomes in this simulation study. Among all the competing methods, MSSPCA performs best for variable selection. Without the assumption that the data heterogeneity only comes from control variables, MSSPCA estimates the unwanted variation using all variables. What's more, the identification of predictive variables jointly using the input observations and the risk scores makes MSSPCA more robust to the noisy outcomes

than the other regression-based methods.

4.5.2 A real-world tuberculosis case study

We apply MSSPCA on a real-world RNA-Seq dataset, Household Contacts (HHC) study, to identify the individuals who are at risk of developing active tuberculosis (TB) disease before the disease onset.

4.5.2.1 Data description

The HHC study included participants from four African sites (South Africa, Gambia, Ethiopia, and Uganda) as part of the Bill and Melinda Gates Foundation Grand Challenges 6-74 (GC6-74) program. Samples were collected at enrollment/baseline and at 6 and 18 months, with the exception of South Africa, where samples at 6 months were not available [2]. For each progressor, four controls were matched according to the age category, sex and year of enrollment. After filtering the samples according to the exclusion criteria as instructued in [27, 2], the data for analysis contains samples from three African sites (South Africa, Gambia and Ethiopia), where South Africa site included 198 samples (48 cases, 150 controls), Gambia site included 169 samples (39 cases, 130 samples) and Ethiopia site included 51 samples (16 cases, 35 controls).

There exists two challenges for analyzing this longitudinal dataset: (1) the data is of extremely high dimension while the sample size is small ($p >> N$); (2) there exists significant population heterogeneity across countries, due to the heterogeneous ethnic origin and genetic backgrounds, distinct infection or exposure status and different local epidemiology [149, 150, 2, 27, 151, 152]. Moreover, the circulating *Mycobacterium tuberculosis* (*M. tuberculosis*) strains have been reported to be different across the three sub-Saharan African populations [150, 27, 2]. Two lineages of the *M. tuberculosis* complex, known as *Mycobacterium africanum* (*M. africanum*) West African 1 and *M. africanum* West African 2, have been shown to be the main pathogens causing human tuberculosis in West Africa, where up to half of human pulmonary tuberculosis (TB) cases were due to *M. africanum*

infection [153, 154, 155]. However, *M. africanum* is restricted to West Africa and has never been identified in Southern Africa [156, 153]. What's more, the disease progression mechanisms are distinct as rates of progression to disease were significantly lower in contacts exposed to *M. africanum* than in those exposed to *M. tuberculosis* [149]. As reported in previous studies [2, 27], the significant population heterogeneity leads to (1) poor cross-site prediction and (2) distinct and site-specific risk signatures, which have been shown in Figure 4.8. PCA is performed and Figure 4.9 shows that the population heterogeneity contributes significantly to the data's total variance, since there's a clear separation of the data points by study sites, especially for South Africa and the other two countries, whereas there's no clear separation between non-progressors/progressors, which, however, is the information of interest. We aim to addressing the above problems with the help of MSSPCA to (1) enhance the cross-site validation by robustly extracting the features associated with TB progression, and (2) derive a population-universal set of risk signatures.

4.5.2.2 Experiment results

We evaluate and compare the performance of the proposed method in terms of the prediction accuracy and the robustness of the identified gene signatures over different countries. Specifically, we validate the effectiveness of MSSPCA for variable selection based on a set of "ground-truth" gene signatures shared across three distinct African sites. To identify signatures performing well at all sites, Suliman *et al.* combined the datasets from three cohorts and analyzed each pair of up-regulated and down-regulated transcripts to select the pairs that discriminated progressors from non-progressors with AUC greater than 0.75 at all three sites, and the combination of the ratio of two pairs: C1QC/TRAV27 and ANRKD22/OSBPL10, were identified as the signatures that lead to significantly increased discrimination between progressors and control subjects [2].

However, since the evaluation of the disease risk is difficult as the disease progresses gradually to the onset and only the final diagnosis result at the end of the study is available, the information "delta time to diagnosis" is selected as a surrogate for the disease

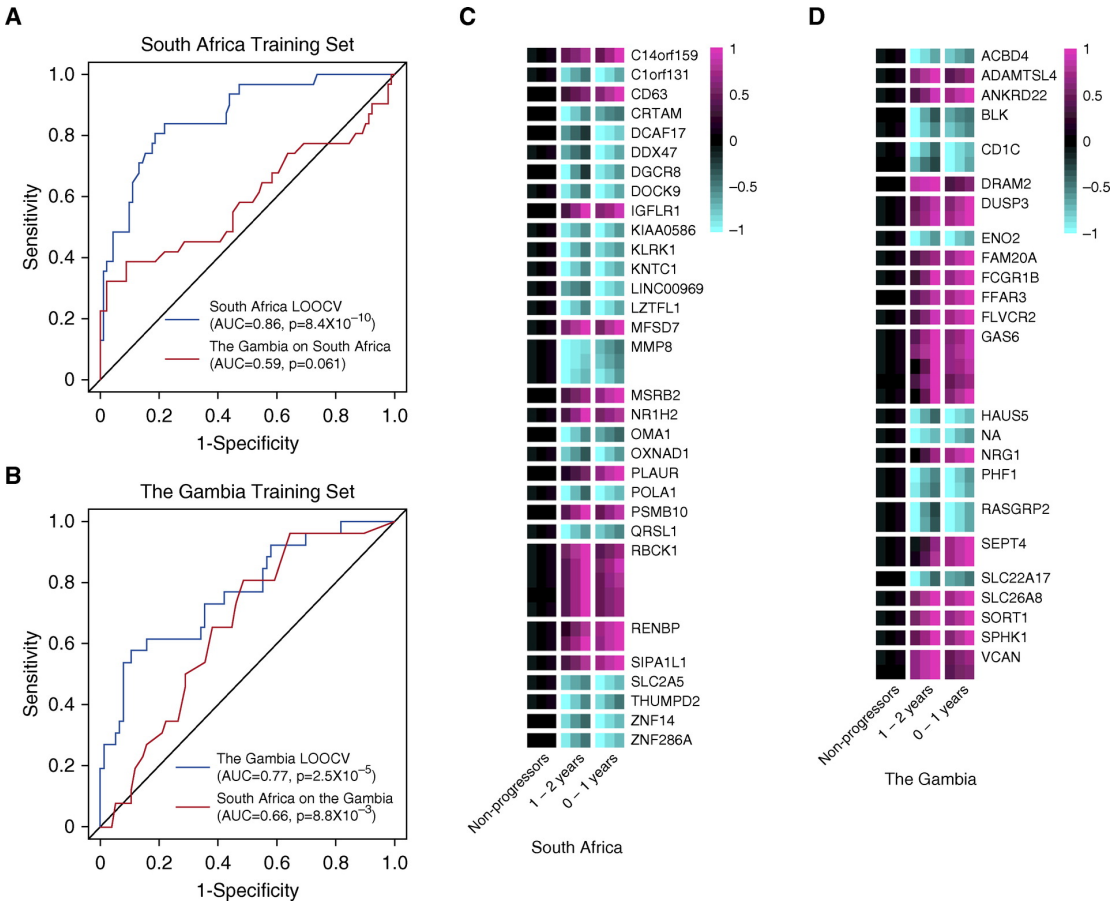
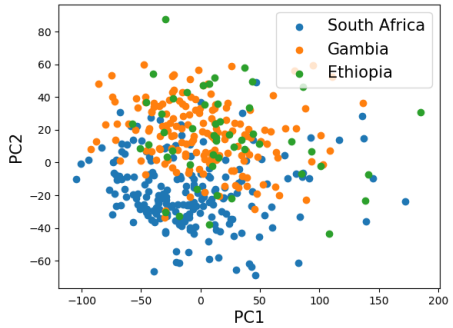
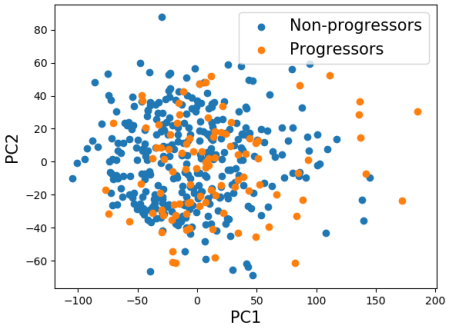


Figure 4.8: Cross-prediction validation and site-specific feature selection results. (A) Receiver operating characteristic (ROC) curve for leave-one-out cross-validation (LOOCV) of South Africa vs. Gambia-trained signature. (B) ROC curve for LOOCV of Gambia vs. South African-trained signature. (C) South Africa and (D) Gambia-trained signatures. This figure is reprinted from Ref. [2].



(a) Colors indicates different sites



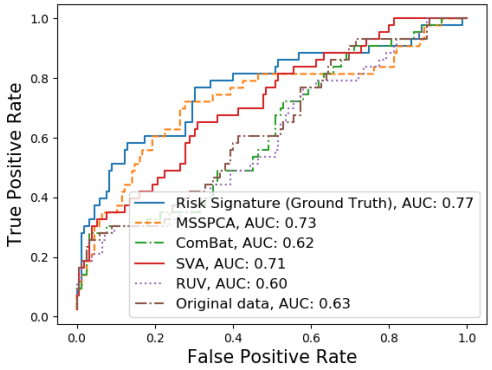
(b) Colors indicates progressors/non-progressors

Figure 4.9: Visualization of PC scores

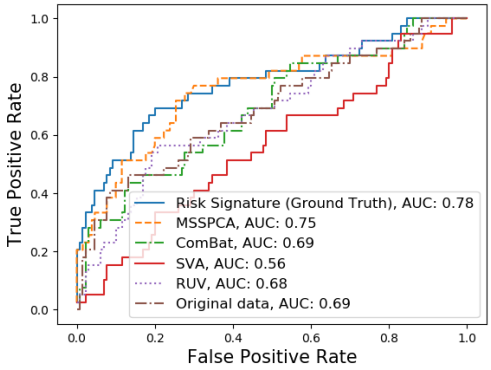
risk scores. To evaluate the proposed method's performance in terms of its prediction accuracy and generalizability, we perform cross-site validation, i.e., each time one site is selected for evaluation, while the data from the other two sites are used for estimating the model's parameters associated with the factors of interest (disease progression), as well as the subsequent binary classifiers to identify the subjects progressing to active TB disease. Specifically, the outcome information is only available for the participants from the selected training sites. Moreover, the identification of consistent signatures with the data from different sites by considering unwanted variation is of great help for researchers to perform robust cross-site prediction for disease prevention and treatment.

Since SVA and ComBat cannot be applied on new samples without outcome information, we use a variant of SVA, "frozen SVA" (fSVA) that was proposed by Parker *et al.* for correcting the testing data using the surrogate variables estimated from the training database [130] and apply only batch information for data adjustment with ComBat. Again, we select the "control genes" based on the estimated FDR ($FDR \geq 0.05$) with a first-pass DE analysis prior to RUV normalization using the training dataset to select the genes that are not associated with disease progression [105, 147, 148]. Since the TB incidence in Ethiopia is too low (only 16 cases among 51 samples in total), we only use this cohort data for training.

Figure 4.10 presents ROC curves for cross-site validation results on the South Africa and Gambia cohorts with all competing methods. fSVA enhances the prediction accuracy when tested on the samples from South Africa, while performs the worst on the samples from Gambia, which could be due to the noisy outcomes. What's more, fSVA requires that the testing data should be similar to training database, so that it could be harmful for cross-study prediction [157, 158]. ComBat performs almost same as original data for prediction, since only site/batch information is used for data correction because ComBat's lack of model predictability. Without incorporating the supervised information, there may exist bias in its estimation of L/S models parameters for batch effects. RUV corrects data



(a) ROC curve of Gambia/Ethiopia-trained classifier on South Africa

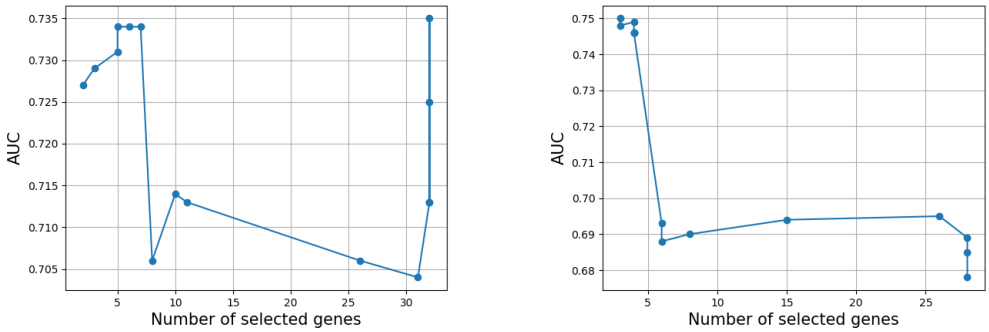


(b) ROC curve of South Africa/Ethiopia-trained classifier on Gambia

Figure 4.10: Cross-site validation results

with "site-specific" control genes selected empirically using the data from training sites. With "delta time to diagnosis" as a noisy surrogate for disease risk, unstable or wrong identification of control variables could lead to RUV's unsatisfying performance, due to its sensitivity to the control variable selection [105]. By jointly modeling the input observations and the outcomes with common latent factors capturing the disease progression to estimate the projection matrices, MSSPCA is robust to the noisy surrogate for disease progression, i.e., "delta time to diagnosis", and able to extract the features from new data or the data without outcome information. Figure 4.10 shows that MSSPCA outperforms all competing methods in both testing sites and approximates the prediction accuracy achieved with the ground truth gene signatures.

Next, we investigate the usefulness of MSSPCA on selecting the risk-associated gene signatures. Before applying MSSPCA, we use the training database to pre-screen the features based on the standardized regression coefficients that measure the univariate effect of each variable separately on the outcomes, which is similar to the procedures adopted by BAIR *et al.* for supervised PCA [134]. MSSPCA is then applied to narrow down the set of risk-associated genes on the reduced data matrix.



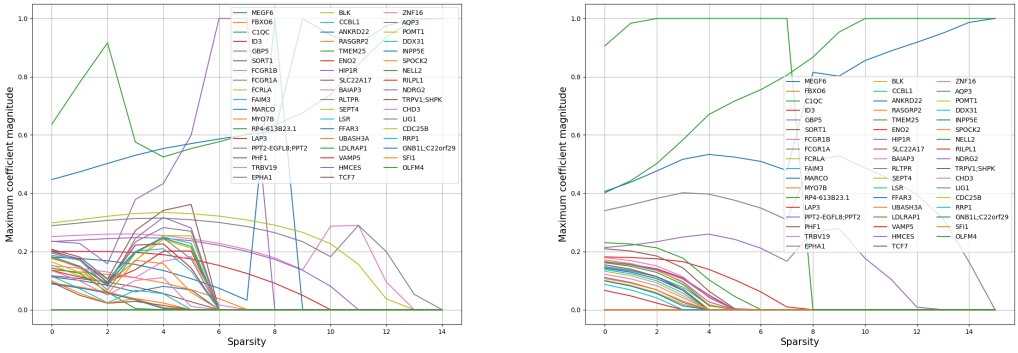
(a) Test site: South Africa (b) Test site: Gambia

Figure 4.11: AUC over number of selected genes

Figure 4.11 shows that the small set of genes identified by MSSPCA can achieve comparable or even better performance than the original input pre-screened gene set, which can enable targeted disease prevention by greatly reducing time and efforts for data collection. We obtain the row-wise maximum magnitude of the loading matrix W estimated by MSSPCA, which is corresponding to the association of a gene with the disease progression and Figure 4.12 presents the path of the maximum coefficient magnitude for each gene over different sparsity degrees. It shows that MSSPCA trained by the data from different sites can finally reach a same set of gene signatures: "C1QC" and "ANKRD22", which are exactly the numerator genes of the two pairs of the site-universal signatures identified by Suliman *et al.* using the datasets from all three cohorts [2].

4.6 Discussion

Unwanted data variability is a common problem faced by researchers, particularly when the data is collected across multiple experiments or study sites. Not considering such effects can cause bias in the results from the downstream analysis, such as feature extraction, variable selection, prediction, etc. To account for the situations where the existing methods might be insufficient, due to practical limitations such as the lack of clean information regarding the outcomes or the control variables and high-dimensional data in small sample



(a) Test site: South Africa

(b) Test site: Gambia

Figure 4.12: Path of the maximum coefficient magnitude for each gene over different sparsity-regularized hyperparameters.

size, we propose MSSPCA for predictive feature extraction and variable selection by jointly modeling the input observations and the outcomes using a flexible probabilistic supervised sparse PCA framework, so that MSSPCA is robust to the noisy outcomes and can be applied on the partially-labeled data or new samples with the parameters estimated from the training database.

We evaluate the performance of MSSPCA and validate its usefulness by comparing to other state-of-the-art methods in a simulation study and a real-world case study. In view of commonly observed data heterogeneity and the resource-demanding annotations for high-dimension RNA-Seq data, MSSPCA can serve as a promising feature learning tool to enhance the robustness and reduce the computation complexity of the downstream analysis with the incorporation of the unwanted data variation.

Till now, this dissertation has addressed the issues in early detection and robust feature learning separately in longitudinal analysis. In the next chapter, we propose to combine EDRA and MSSPCA as a pipeline for robust early detection with longitudinal data. The effectiveness of the proposed pipeline is demonstrated with a real-world longitudinal RNA-Seq dataset for TB early detection. What's more, the impact of each component in the

proposed pipeline is investigated and their benefits are discussed.

5. EARLY DETECTION WITH ROBUST FEATURE LEARNING

5.1 Motivation

We have investigated the critical issues for early detection with longitudinal data in the situations where the data are partially-labelled, high-dimensional in limited sample size ($p << N$), or containing significant unwanted data heterogeneity, with the proposed solution methods for (1) early detection and risk assessment (EDRA) with longitudinal data analysis in Chapter 3, and (2) robust feature learning with MSSPCA by removing the unwanted data variation in Chapter 4. However, these challenges often co-exist in real-world longitudinal studies. Therefore, EDRA's kernel construction by prior knowledge without considering the correlation among features may not be optimal and could lead to degenerated performance when applied on high-dimensional data with unwanted heterogeneity. On the other hand, using a standard binary classifier for early detection with the features extracted by MSSPCA without considering the structures within the longitudinal data can also give unsatisfactory results. In this chapter, we explore the integration of EDRA and MSSPCA to address both challenges in analyzing the real-world longitudinal RNA-Seq data for tuberculosis prediction studied in Chapter 4.

5.2 Related work and problem statement

Around 1.7 billion people globally are infected with *Mycobacterium tuberculosis*, but less than 10% of these will progress to have active tuberculosis (TB) disease during their lifetime; most individuals will remain healthy [159, 160, 161, 162, 163, 27]. Identification of blood biomarkers that prospectively predict the progression of TB can lead to interventions that combat the TB epidemic [27]. Zak *et al.* identified a 16-gene signature that can be used to predict the TB risk [27]. They followed up healthy South African adolescents from the adolescent cohort study (ACS) who were infected with *M. tuberculosis* for 2 years where the blood samples were collected from the participants every 6 months and their

progression to TB were monitored. 46 ACS participants with microbiologically confirmed TB were identified as progressors (39 for training and 9 for testing), and 107 control participants were infected with *M. tuberculosis* at the enrolment but remained healthy during the 2-year follow-up (77 for training and 30 for testing). The genes that comprise the final tuberculosis risk signature were selected in two stages with the data from this ACS training set. First, a large set of genes were identified by comparing the gene expression in progressors at the most proximal timepoint to diagnosis with that in the matched controls. SVM models were trained with these datapoints for all possible pairwise combinations of the risk-associated genes. Second, the models were filtered based on the predictive accuracy with the remaining progressor and control samples. Finally, the surviving pair-wise SVM (PSVM) models comprise the tuberculosis signature, and the ensemble of all PSVMs compute the "tuberculosis risk scores" based on the gene expression level measured at a single timepoint [27]. The signature identified with ACS data, known as the ACS COR (correlate of risk), was then validated in the testing set of the ACS study as well as two independent South African and Gambian cohorts from another study, which is called the Household Contacts (HHC) Study, including participants who had contacts with the patients with TB.

As we have introduced in Chapter 4, the HHC study included the participants from four African sites (South Africa, Gambia, Ethiopia and Uganda) as part of the Bill and Melinda Gates Foundation Grand Challenges 6-74 (GC6-74). It included samples collected at the enrollment/baseline and at 6 and 18 months, with the exception of South Africa, where samples at 6 months were not available. After inclusion and exclusion of participants in the GC6-74 HHC study based on certain quality check (QC) criteria [2], there are 205 participants (27/14 progressors, 81/83 controls) from South Africa; 142 participants (18/8 progressors, 60/56 controls) from Gambia; and 60 participants (0/12 progressors, 0/48 controls) from Ethiopia, where the number before the slash symbol denotes the number of samples for training, while the one after is for testing.

However, a concern regarding the generalizability of ACS COR signature was later

raised by Suliman *et al.* in [2]: given that the 16-gene ACS COR signature was developed using a single cohort of South African adolescents, the predictive accuracy in diverse African populations may be suboptimal [27, 2]. What's more, even within the samples from HHC study, the distinct signatures are specific to different countries for training, as shown in Figure 4.8(C) and (D). The poor cross-site prediction accuracy in Figure 4.8(A) and (B) also illustrates that there exists site-specific variability within the diverse African populations.

For risk signature discovery with the collected RNA-Seq data, Suliman *et al.* performed the selection of gene pairs to include in the final signature in a two-step procedure. First, all genes were filtered based on their univariate prediction ability. Second, all possible pairs of the survival genes in opposite directions during TB progression were formed and their corresponding log-ratios were computed. For prediction of TB, a new datapoint is scored by comparing its ratio to the distribution of the ratios present in the progressors and controls in the training cohorts (which is computed as the average over the percentage of progressor samples in the training set that have a ratio lower than the observed ratio and the percentage of control samples in the training set that have a ratio lower than the observed ratio).

It's clear that the above procedures for risk signature discovery require the similarity between the training and the testing data, while the diverse populations could result in "*site-specific*" signatures and poor cross-site prediction validation performance. To develop more generalized risk signatures, the cohorts from three countries were further combined in [2]. The analysis led to a four-gene signature (Risk4) by combining the cohorts from South Africa and Gambia. One gene pair (C1QC/TRAV27) was later selected from the two identified pairs (C1QC/TRAV27, ANRKD22/OSBPL10) in a brute force manner, by merging the cohorts from all three countries.

The results from both studies, ACS and HHC, have shown that there are two issues in analyzing the longitudinal RNA-seq data for TB prediction: (1) site-specific variability: Though Suliman derived a four-gene signature (RISK4) and the single gene pair

(C1QC/TRAV27), they have to include all the data points with the corresponding outcome information from all of the concerned populations, which can be resource-demanding and time-consuming for data collection and annotation in longitudinal studies. (2) poor early detection performance: Zak *et al.* calculated the AUC values of ROC curves corresponding to a 180-day interval before TB diagnosis, showing that the AUC values decrease when tested on the early timepoints compared with the datapoints close to TB diagnosis [27]. What's more, both works focused specifically on TB prediction and employed complicated analysis pipelines that cannot be easily generalized to other studies.

We propose a pipeline to address the above issues in two steps. First, the biomarkers/signatures are derived by MSSPCA with the predictive genes selected simultaneously. Second, EDRA is subsequently applied on the derived signatures for disease detection.

Specifically, we consider the situations where the data collection and annotation in longitudinal studies is difficult, i.e., the outcome information is only available for the selected training cohorts (e.g. countries) and might be noisy. Nevertheless, the derived signatures and the risk detectors are desired to overcome the variability that comes from the diverse populations and being robust to the noisy outcomes, with a consistent detection performance in the testing data collected from the cohorts (e.g. countries) that are different from the cohorts for training.

5.3 Experimental results

We apply the proposed pipeline on the data from HHC study, to discriminate the individuals who are progressing to active TB from controls before the disease diagnosis. The developed pipeline's generalizability and detection ability is assessed through cross-site prediction validations. First, we use the algorithms with the signatures that have been developed in the existing literature [27, 2] as the baseline methods, and evaluate the performance of the proposed pipeline by comparing it to the baseline results. A pipeline's detection performance with respect to the earliness is also evaluated. First, the timescale is realigned according to TB diagnosis instead of study enrollment. Second, to obtain a

sufficient testing set for a specific timepoint, we apply the carry forward imputation to include all testing datapoints whose length of time to diagnosis are longer than that certain timepoint for evaluation. For instance, the ability of our pipeline for (early) detection is validated with all the datapoints before TB diagnosis. We also examine the pipelines' early detection ability by the datapoints that are more than 6 months before the diagnosis.

Second, we investigate and discuss the impact of each component (EDRA and MSSPCA) in the proposed pipeline by fixing one and comparing the other to the competing algorithms. In addition to the two delta timepoints before the diagnosis (0 and 6 months), the detection accuracy concerning the earliness is also measured by generating the trajectories showing the detection performance from 0 to 18 months with 1-month intervals.

In all experiments, we perform cross-site prediction validation, which means that each time the testing set is formed by a single cohort from South Africa (39 progressors/141 controls) or Gambia (25 progressors/93 controls), and the other cohorts are used for training the pipelines. Similar to the experimental set up described in Chapter 4, the data from Ethiopia is combined with one of the major cohort for training owing to its small sample size (11 progressors/26 controls). A few samples are missing for each country in the dataset compared to which have been reported in literature, but the numbers are still within a reasonable range. Such experiment setup is to assess the detector's generalizability and mimic the real-world situations where the outcome information is only available for the countries where the data is collected for training. What's more, since the detection of TB is of greater interest, precision-recall curve is adopted to evaluate the detection ability and a higher weight (10x) is assigned to the progressors' samples during the evaluation.

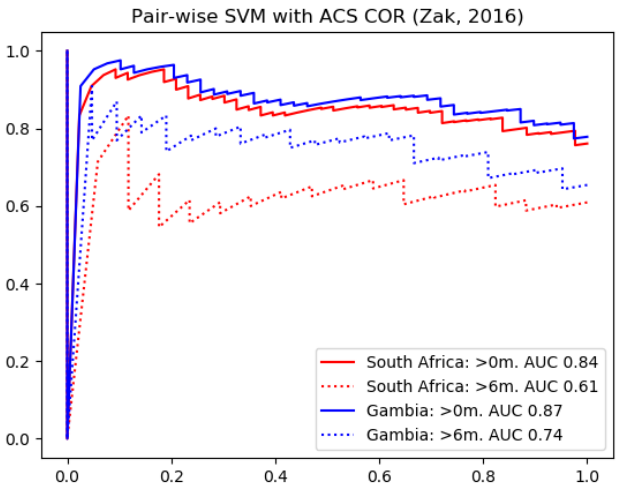
5.3.1 Comparison with the baseline results

The previously published signature concerning the data from HCC study include (1) ACS COR in [27], (2) *site-specific* signatures in [2], (3) "RISK4" in [2] and (4) the single gene pair "C1QC/TRAV22" (HHC COR) in [2]. Among these signatures, only the ACS COR and *site-specific* signatures were derived with the data not fully covering both cohorts

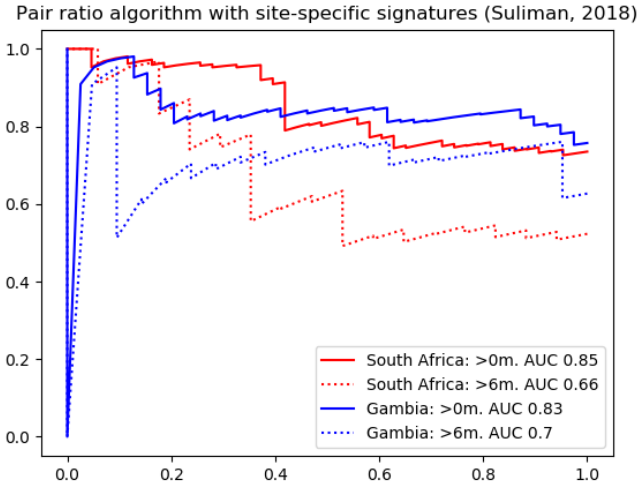
from South Africa and Gambia in HHC study. We use ACS COR and *site-specific* signatures along with their corresponding algorithms: pair-wise SVM (PSVM) and the pair ratio algorithm as the baseline methods, where the weights of the PSVM models using ACS COR have been derived with the ACS data in [27], by which the score of a new testing sample can be easily computed based on the new sample's gene expression levels; while the pair-ratio algorithm is data-driven and dependent on the selection of the training data, where a new sample is scored by comparing its corresponding log ratio to the distributions of the log ratios present in the training set.

Figure 5.1 shows the precision-recall curves for the cross-site prediction validation with the baseline methods on the testing cohort in our experiments. The two baseline methods obtain similar performance for predicting TB on the datapoints before diagnosis: the AUCs obtained by PSVM using ACS COR signature are 0.84 and 0.87 compared to 0.85 and 0.83 by the pair ratio algorithm using the site-specific signatures when tested on the South African cohort and Gambian cohort, respectively. Then we focus on the detection ability on the early datapoints that are at least 6 months before the diagnosis. However, the experiment results show that there's a significant drop in detection accuracy by applying both of these baseline methods. The AUC values drop from 0.84 to 0.61 and 0.87 to 0.74 on the two sites respectively by PSVM using ACS COR signature; and the AUC values drop from 0.85 to 0.66 and 0.83 to 0.70 when tested on the two sites respectively with the pair ratio algorithm using the site-specific signature.

Next, we apply our proposed pipeline to first derive the risk features and the predictive genes by MSSPCA, and then predict TB with EDRA subsequently applied on the extracted features. In Chapter 4, we have already demonstrated MSSPCA's variable selection ability by identifying a gene pair that is consistent with different training cohorts: "C1QC" and "ANRKD22", which are exactly the numerator genes of the two gene pairs identified by merging all the data from the three countries in [2]. In Table 5.1, the AUC values of the precision-recall curves show that our proposed pipeline outperforms the baseline methods



(a) Ensemble of all pair-wise SVMs with ACS COR developed in [27]



(b) Pair ratio algorithm with the site-specific signatures developed in [2]

Figure 5.1: Precision-recall curves for cross-site validation with the baseline methods. “m” denotes “months”.

Pipelines	South Africa >0m	South Africa >6m	Gambia >0m	Gambia >6m
MSSPCA/EDRA	0.88	**0.73**	0.90	0.82
ACS COR/PSVM	0.84	0.61	0.87	0.74
Site-specific signature/Pair ratio	0.85	0.66	0.83	0.70
* Risk4/Pair ratio	**0.91**	0.72	0.91	0.83
* HHC COR/Pair ratio	0.86	0.65	**0.92**	**0.86**

Table 5.1: AUC values of the precision-recall curves for the proposed pipeline and the methods developed in literature. Note that the risk signatures "RISK4" and "HHC COR" were developed by combining the cohorts from South Africa and Gambia, so that the testing data is actually included for signature and model development in [2].

by a great margin, particularly on the early datapoints that are more than 6 months before diagnosis. The AUC values achieved by the proposed pipeline are 0.73 and 0.82 on South Africa and Gambia, respectively, both of which are almost 10 percent higher than the baseline results on both countries where the AUC values are 0.61 and 0.74 by PSVM using ACS COR signature, and 0.66 and 0.70 by the pair ratio algorithm using site-specific signatures.

In addition to the two above baseline methods, another two risk signatures including "RISK4" and the gene pair "HHC COR" were also developed with the pair-ratio algorithm in [2]. However, the combined cohorts from South Africa and Gambia were included for deriving these two signatures. Although it's unfair to compare with these two published signatures since the testing cohort was used when deriving them in original works, Table 5.1 shows that our proposed pipeline still achieves competitive AUC values in all four situations. When tested on all datapoints before diagnosis, the proposed pipeline performs similarly well as "RISK4" and "HHC COR", with AUC as 0.88, compared to 0.91 by "RISK4" and 0.86 by "HHC COR" on South African cohort; and 0.90, compared to 0.91 by "RISK4" and 0.92 by "HHC COR" on Gambian cohort. For the samples who are more than 6 months before diagnosis, the AUC achieved by our proposed method is 0.73, compared to 0.72 by "RISK4" and 0.65 by "HHC COR" on South Africa; and 0.82, compared to 0.83 by "RISK4" and 0.86 by "HHC COR" on Gambia.

Combining Figure 5.1 and Table 5.1, it can be observed that the unwanted data vari-

ability from different populations or studies is harmful for the downstream detection if we don't take care of it. By comparing our proposed method and the two risk signatures "RISK4" and "HHC COR", with the baseline methods, the experimental results show that the pipelines considering the population diversity achieve an overall enhanced performance for TB prediction. However, the two published signatures "RISK4" and "HHC COR" addressed the problem by expanding the data for training to include the testing cohort, while our proposed method corrects the data by estimating and removing the variability without using any outcome information of the samples from testing countries.

In the following context of this chapter, we further investigate the impact of each component in the proposed pipeline for disease early detection.

5.3.2 Impact of MSSPCA in robust early detection

In this section, we investigate the impact of MSSPCA in the proposed pipeline with EDRA as the risk detector. Popular batch-effect correction algorithms including SVA, RUV and ComBat in conjunction with EDRA are chosen for comparison to evaluate the effectiveness of MSSPCA in the proposed pipeline. Since only MSSPCA produces the features with the reduced dimension and estimates the association between the input variables and the outcomes, the variable's discriminating power for the other correcting algorithms is estimated through differential expression (DE) analysis by comparing the "progressors" and "controls" in the training set. Genes are ordered based on their estimated p-values and the top 50 DEGs with the lowest p-values are selected for EDRA's mixed-kernel construction. The weights for the selected 50 DEGs are calculated based on their absolute values of the log_2 fold change and are further normalized to make the weights sum to one, i.e., $\beta_k = \frac{|log_2 FC_k|}{\sum_{i=1}^{50} |log_2 FC_i|}$, where β_k is the weight for the selected DEG k, and FC means "fold change".

Table 5.2 provides the cross-site prediction validation results of the pipelines using MSSPCA and different batch-effect correction algorithms with EDRA as the subsequent detector. Compared with the original data, the results show that the proposed pipeline with

Pipelines	South Africa >0m	South Africa >6m	Gambia >0m	Gambia >6m
MSSPCA/EDRA	**0.88**	**0.73**	**0.90**	**0.82**
Raw/EDRA	0.80	0.55	0.86	0.76
RUV/EDRA	0.81	0.57	0.85	0.76
SVA/EDRA	0.82	0.52	0.68	0.52
ComBat/EDRA	0.78	0.52	0.86	0.74

Table 5.2: AUC values of the precision-recall curves for the pipelines with EDRA subsequently applied on the original data and the data adjusted by the batch-effect correction algorithms

MSSPCA achieves an overall improvement on the detection performance with the features extracted by considering the data's unwanted variability. The AUC values by our proposed pipeline on all datapoints before diagnosis are 0.88 and 0.90 on South Africa and Gambia, respectively, compared to 0.80 and 0.86 with the original unadjusted data. For early datapoints, our proposed pipelines improves the detection accuracy more significantly, with AUCs as 0.73 and 0.82, respectively on South Africa and Gambia, compared to 0.55 and 0.76 using the original data.

Next, we compare the performance of the pipelines with the other batch-effect correction algorithms. It can be seen that there's no obvious improvement in the performance with these pipelines compared to the original data, and we even observe some drop in detection accuracy in situations such as the early datapoints from Gambian cohort, as the pipeline using SVA seriously degenerates the performance with the AUC value as 0.52, while the AUCs with the original unadjusted data and the data corrected by RUV and ComBat are 0.76, 0.76, 0.74, respectively. For the poor performance obtained by the pipelines using the other batch-effect correction algorithms, in addition to the reasons we have discussed in Chapter 4 such as the algorithms' sensitivity to the noisy outcomes; noisy selection of negative control genes; or the difference between testing and training datasets, the bias might be further exaggerated when the kernels are constructed with the variables discriminating power wrongly estimated due to the inappropriate/insufficient data correction. What's

more, extracting features without incorporating the correlation within variables may also lead to unsatisfying analytic results, particularly for the data of high dimensionality. However, MSSPCA addresses these issues by aggregating the signals from both the RNA-Seq data (e.g., gene expression level) and the information regarding the disease progression (e.g. delta time to diagnosis) to extract the predictive features with the unwanted variability estimated and removed simultaneously, which enhances the performance with an improved accuracy and robustness for TB prediction.

5.3.3 Impact of EDRA in robust early detection

Another critical issue in early detection with longitudinal data is the lack of label information to specifically point out the stages of the disease progression, since labeling subjects by the trained medical professionals at each time point before disease onset is almost impossible. What's more, the measurements such as gene expression levels for progressors at the disease early stage could be indistinguishable from controls. Both of these challenges could make it difficult to train a supervised classifier with the observations collected before the disease diagnosis. We have developed EDRA to address these challenges by learning with the temporal "change" information from the longitudinal data, and incorporate the structures within the output space (e.g. the underlying risk factors of the disease progression) by imposing different penalties for misclassifications with respect to the length the time intervals between two visits of an individual.

In this section, we focus on the impact of EDRA. To validate the effectiveness of EDRA, we consider two supervised classifiers and two state-of-the-art early event detectors for comaprison, which include: Linear SVM, RBF SVM, SOSVM and MMED. All these algorithms are subsequently applied on the features derived by MSSPCA for TB prediction. Again, we evaluate the pipelines' performance through cross-site prediction validation. Each time one cohort from South Africa or Gambia is selected for testing and the outcomes of the datapoints from the testing site will not be used for model development.

Table 5.3 provides the AUC values of the precision-recall curves for the cross-site pre-

Pipelines	South Africa >0m	South Africa >6m	Gambia >0m	Gambia >6m
MSSPCA/EDRA	0.88	**0.73**	0.90	**0.82**
MSSPCA/MMED	**0.89**	0.67	**0.91**	0.80
MSSPCA/SOSVM	**0.89**	0.67	**0.91**	0.80
MSSPCA/RBF-SVM	0.87	0.70	0.76	0.64
MSSPCA/Linear-SVM	0.88	0.68	0.90	0.81

Table 5.3: AUC values of the precision-recall curves for the pipelines using different classifiers/detectors

diction validation results with the pipelines using different classifiers/detectors. First, the results show that our proposed pipeline still achieves the best detection performance when tested on the early datapoints that are at least 6 months before diagnosis, with the AUC value of 0.73 on South Africa, compared to 0.67, 0.67, 0.70 and 0.68 for the pipelines with MMED, SOSVM, RBF-SVM and Linear-SVM, respectively; for the early datapoints from Gambia, the AUC achieved by our proposed pipeline is 0.82, compared to 0,80, 0.80, 0.64 and 0.81 with MMED, SOSVM, RBF-SVM and Linear-SVM, respectively. For all datapoints before the diagnosis, although the advantage of the proposed pipeline is not that obvious, it still achieves almost same detection accuracy as the best performed pipeline, with the AUC of 0.88 compared to the highest AUC value 0.89 when tested on South Africa; and 0.90 compared to the highest AUC value 0.91 on Gambian.

Another interesting observation is that the three Linear SVM-based detectors including MMED, SOSVM and Linear SVM, obtain almost identical detection accuracy in all the situations. To validate this conjecture, we take a further step to investigate the characteristics for all these pipelines when tested on the data with varying delta lengths of time to the disease diagnosis.

The detection performance with respect to both earliness and accuracy for all the pipelines is assessed using the datapoints with different lengths of time before TB diagnosis, to measure how the discriminating power changes over the delta time to diagnosis. Similar to the above experiments, the timescale is realigned according to TB diagnosis instead of study

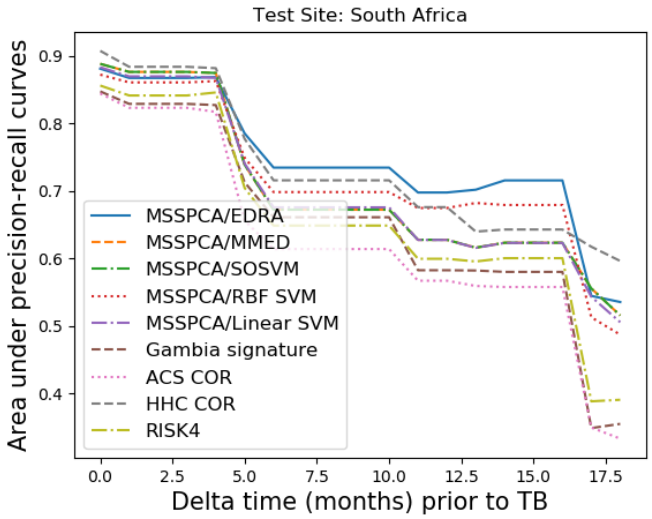
enrollment. Specifically, we let the time threshold slides from 0 months to 18 months with 1 month interval, and use the carry forward imputation to include the testing datapoints that are earlier than the current timepoint before the diagnosis, to assess how the distance towards the diagnosis can effect the detection accuracy. The AUC values for all precision-recall curves are calculated to summarize the pipelines' detection ability for each timepoint.

In Figure 5.2, the trajectories of the AUC values for the precision-recall curves with all pipelines according to the time before TB diagnosis are presented. Figure 5.2(a) provides the cross-site prediction results for the pipelines trained with the data from Gambian/Ethiopian cohorts and tested on the South African cohort, which clearly demonstrate that our proposed pipeline significantly outperforms the other competing methods, particularly for difficult situations where the datapoints are far from TB diagnosis. However, when trained with South African/Ethiopian cohorts, the advantage of the proposed pipeline over the other methods is less obvious. To further understand the disparity in prediction performance, we check the data situation for the two training sets. It turns out that the South African cohort only include a small number of progressors who have multiple visists (4 out of 39), and as we have introduced in the beginning of this chapter, even for the four participants with the follow-ups, they only have 2 datapoints at the baseline and 18 months after the enrollment. Being trained with insufficient samples whose longitudinal observations are limited, it's difficult for EDRA to extract the temporal "change" information and learn with it, so that the detection performance could be degenerated as a standard SVM model, which is reflected in Figure 5.2(b). In the contrast, in Gambian cohort, there are 16 out of 36 progressors who have follow-up visits, and some of them took blood tests for 3 times, which could better enable the training of EDRA by learning with more information regarding the disease development from the longitudinal data. The advantage of EDRA is better demonstrated in Figure 5.2(a), with a great margin between the trajectory of our proposed pipeline and the other methods, when trained with the Gambian/Ethiopia cohorts

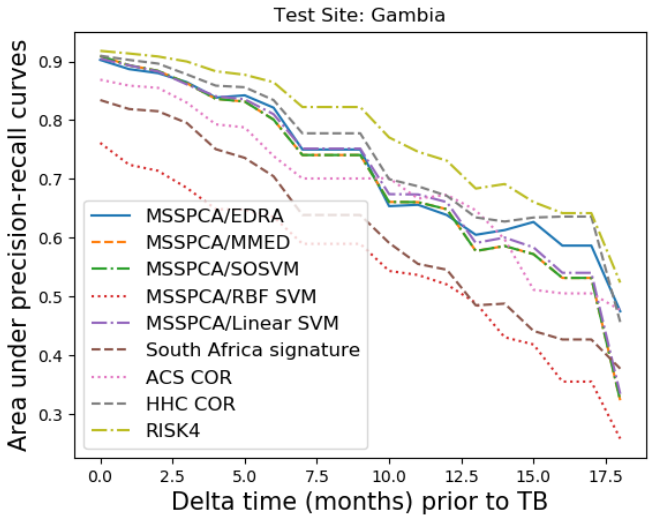
and tested on the South African cohort.

Another observation is the almost overlapping trajectories of the three Linear SVM-based methods in Figure 5.2(b), which validates our conjecture that they end up into almost same models. Rather than learning with the "change" information like EDRA, SOSVM and MMED take the measurement of a single datapoint or the moving average of the datapoints prior to some timepoint. However, trained with the samples with small number of longitudinal observations (at most 3 datapoints), it's possible that all three methods learn with almost same datapoints, which result in close prediction performance shown in both Table 5.3 and Figure 5.2(b). Finally, the detection ability of RBF SVM model is unstable as it achieves the second best performance in Figure 5.2(a), while the worst performance in Figure 5.2, which might be caused by the disparity within the data collected from different countries.

In summary, the proposed pipeline integrating MSSPCA and EDRA achieves an enhanced performance in early disease detection. MSSPCA's superior performance in deriving predictive and robust features with variable selection capability addresses the issues in EDRA for feature learning and kernel construction when the data is high dimensional and contains the unwanted variability; by imposing varying penalties on misclassiification with respect to the datapoints' delta time to diagnosis, EDRA incorporates the underlying risk factors indicating the varying severity levels for disease progression, which further improves the performance of early detection with longitudinal data. However, the experiment results also show that the advantage of the methods developed to consider the data/output's temporal structures is less obvious in the situations where the longitudinal observations are limited, and they may turn into a standard SVM model.



(a) AUC values for precision-recall curves of the Gambia/Ehiopia-trained pipelines with different classifiers/detectors tested on South African samples over the delta time before TB.



(b) AUC values for precision-recall curves of the South Africa/Ehiopia-trained pipelines with different classifiers/detectors tested on Gambian samples over the delta time before TB diagnosis.

Figure 5.2: Cross-site validation results for detection by time before TB diagnosis. The area under precision-recall curve is calculated for all pipelines at each time point.

6. CONCLUSIONS AND FUTURE WORK

In this dissertation, novel methods for early detection and robust feature learning in longitudinal data analysis have been introduced. More specifically, we have developed a flexible learning framework for early and contemporaneous risk detection, and a probabilistic method for deriving predictive features with variable selection capability. A pipeline integrating the developed methods for early detection and robust feature learning has also been implemented with the effectiveness of each component investigated. The methods proposed in this dissertation address critical issues in longitudinal data analysis including label insufficiency, data heterogeneity, and high-dimension data with limited sample size. Both simulation studies and real-world case studies have shown that the proposed methods enhance the detection performance in terms of its earliness, accuracy, robustness and models' interpretation through comprehensive experimental results.

In Chapter 3, we introduced EDRA, an SVM-based learning framework for early detection and risk assessment with longitudinal data. By incorporating the structures and dependency within the input data and the output space, EDRA learns from temporal changes rather than static measurements. Our proposed EDRA addresses the label insufficiency problem and is able to learn with partially-labeled training data, which results in the improved detection earliness and accuracy, especially for the difficult situations where the data points belonging to one class (e.g. potential patients) are usually indistinguishable (at the disease early stage) from the other class (e.g. healthy individuals). The effectiveness of EDRA has been demonstrated by comparing to several popular supervised classifiers and state-of-the-art early event detectors through simulation studies and two real-world longitudinal datasets for T1D onset detection and the study of drug's long-term effects.

In Chapter 4, MSSPCA is proposed for robust feature learning from high-dimensional longitudinal data by removing the unwanted data variability. The advantages of our proposed MSSPCA are two-fold. First, MSSPCA facilitates the subsequent learning and in-

ference, leads to better model interpretation, and enables the targeted disease prevention and treatments. Second, MSSPCA addresses the generalizability and robustness issue in learning features by removing the unwanted effects such as systematic technical noise, and overcome the challenges that limit the applicability of most existing batch-effect correction methods. By aggregating the signals from input observations and the outcome information by involving the data heterogeneity, MSSPCA achieves superior performance in deriving predictive features with variable selection capability and being robust to noisy outcomes, which has been validated through simulation studies and a real-world longitudinal study that collected RNA-Seq data from multiple study sites for tuberculosis early prediction.

Finally, we proposed a pipeline that integrates EDRA and MSSPCA for robust early detection with longitudinal data in the situations where the data are high-dimensional, partially-labeled, and with unwanted variability. The ability of our proposed pipeline for early detection and robust feature learning has been demonstrated by comparing our proposed pipeline with the existing methods and risk signatures developed in previous works. The impact of each component of our proposed pipeline is also investigated. By subsequently applying EDRA on the features derived by MSSPCA, the proposed pipeline achieves the smallest drop in detection accuracy in difficult classification situations, such as the disease early stage, which has been demonstrated by our experimental results on a real-world tuberculosis longitudinal RNA-Seq dataset.

In summary, our proposed methods improve the accuracy and robustness of early detection and feature learning in longitudinal data analysis. The methods proposed in this dissertation are beneficial in many domains where longitudinal analysis is involved, for example, in chronic disease monitoring. In future, we may explore the possibility of utilizing the deep neural networks to effectively concatenate the two objectives into one step and learn all the parameters simultaneously. We will also apply our proposed methods in other application domains, such as manufacturing, to further validate the proposed methods' performance and their benefits on the longitudinal data collected in different applications.

REFERENCES

[1] A. Katsarou, S. Gudbjörnsdottir, A. Rawshani, D. Dabelea, E. Bonifacio, B. J. Anderson, L. M. Jacobsen, D. A. Schatz, and Å. Lernmark, "Type 1 diabetes mellitus," *Nature reviews Disease primers*, vol. 3, p. 17016, 2017.

[2] S. Suliman, E. G. Thompson, J. Sutherland, J. Weiner 3rd, M. O. Ota, S. Shankar, A. Penn-Nicholson, B. Thiel, M. Erasmus, J. Maertzdorf, *et al.*, "Four-gene pan-african blood signature predicts progression to tuberculosis," *American journal of respiratory and critical care medicine*, vol. 197, no. 9, pp. 1198–1208, 2018.

[3] K. A. Johnson, N. C. Fox, R. A. Sperling, and W. E. Klunk, "Brain imaging in alzheimer disease," *Cold Spring Harbor perspectives in medicine*, p. a006213, 2012.

[4] S. Lee, H. Huang, and M. Zelen, "Early detection of disease and scheduling of screening examinations," *Statistical methods in medical research*, vol. 13, no. 6, pp. 443–456, 2004.

[5] W. contributors, "Chronic condition," *Wikipedia, The Free Encyclopedia*, 1 Apr. 2018. Web. 12 Apr. 2018.

[6] J. Zhang, H. Xiong, Y. Huang, H. Wu, K. Leach, and L. E. Barnes, "M-SEQ: Early detection of anxiety and depression via temporal orders of diagnoses in electronic health data," *2015 IEEE International Conference on Big Data (Big Data)*, pp. 2569–2577, 2015.

[7] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[8] R. Tibshirani, "Regression shrinkage and selection via the lasso: a retrospective," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 73,

no. 3, pp. 273–282, 2011.

[9] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.

[10] M. A. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *AIChE journal*, vol. 37, no. 2, pp. 233–243, 1991.

[11] P. Geladi and B. R. Kowalski, "Partial least-squares regression: a tutorial," *Analytica chimica acta*, vol. 185, pp. 1–17, 1986.

[12] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *International conference on machine learning*, pp. 1247–1255, 2013.

[13] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[14] X. V. Nguyen, J. Chan, S. Romano, and J. Bailey, "Effective global approaches for mutual information based feature selection," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 512–521, 2014.

[15] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján, "Conditional likelihood maximisation: a unifying framework for information theoretic feature selection," *Journal of machine learning research*, vol. 13, no. Jan, pp. 27–66, 2012.

[16] M. A. Hall, "Correlation-based feature selection for machine learning," 1999.

[17] H. Deng and G. Runger, "Feature selection via regularized trees," in *The 2012 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 2012.

[18] M. Yamada, W. Jitkrittum, L. Sigal, E. P. Xing, and M. Sugiyama, "High-dimensional feature selection by feature-wise kernelized lasso," *Neural computation*, vol. 26, no. 1, pp. 185–207, 2014.

[19] F. W. Townes, S. C. Hicks, M. J. Aryee, and R. A. Irizarry, "Feature selection and dimension reduction for single-cell rna-seq based on a multinomial model," *Genome biology*, vol. 20, no. 1, pp. 1–16, 2019.

[20] I. Gaynanova, J. G. Booth, and M. T. Wells, "Simultaneous sparse estimation of canonical vectors in the p n setting," *Journal of the American Statistical Association*, vol. 111, no. 514, pp. 696–706, 2016.

[21] H. Shen and J. Z. Huang, "Sparse principal component analysis via regularized low rank matrix approximation," *Journal of multivariate analysis*, vol. 99, no. 6, pp. 1015–1034, 2008.

[22] M. Xiong, X. Fang, and J. Zhao, "Biomarker identification by feature wrappers," *Genome Research*, vol. 11, no. 11, pp. 1878–1887, 2001.

[23] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, and Y. Saeys, "Robust biomarker identification for cancer diagnosis with ensemble feature selection methods," *Bioinformatics*, vol. 26, no. 3, pp. 392–398, 2010.

[24] Y. Zhang, D. F. Jenkins, S. Manimaran, and W. E. Johnson, "Alternative empirical bayes models for adjusting for batch effects in genomic studies," *BMC bioinformatics*, vol. 19, no. 1, p. 262, 2018.

[25] M. J. Boedigheimer, R. D. Wolfinger, M. B. Bass, P. R. Bushel, J. W. Chou, M. Cooper, J. C. Corton, J. Fostel, S. Hester, J. S. Lee, *et al.*, "Sources of variation in baseline gene expression levels from toxicogenomics study control animals across multiple laboratories," *BMC genomics*, vol. 9, no. 1, p. 285, 2008.

[26] M. Bakay, Y.-W. Chen, R. Borup, P. Zhao, K. Nagaraju, and E. P. Hoffman, "Sources of variability and effect of experimental approach on expression profiling data interpretation," *BMC bioinformatics*, vol. 3, no. 1, p. 4, 2002.

[27] D. E. Zak, A. Penn-Nicholson, T. J. Scriba, E. Thompson, S. Suliman, L. M. Amon, H. Mahomed, M. Erasmus, W. Whatney, G. D. Hussey, *et al.*, "A blood rna signature for tuberculosis disease risk: a prospective cohort study," *The Lancet*, vol. 387, no. 10035, pp. 2312–2322, 2016.

[28] Y. Zhang, W. E. Johnson, and G. Parmigiani, "Robustifying genomic classifiers to batch effects via ensemble learning," *bioRxiv*, p. 703587, 2019.

[29] A. Sarwar and V. Sharma, "Intelligent naïve bayes approach to diagnose diabetes type-2," *International Journal of Computer Applications, IJCA Special Edition Nov*, pp. 14–16, 2012.

[30] K. W. Przytula and D. Thompson, "Construction of bayesian networks for diagnostics," in *Aerospace Conference Proceedings, 2000 IEEE*, vol. 5, pp. 193–200, IEEE, 2000.

[31] M. Langarizadeh and F. Moghbeli, "Applying naive bayesian networks to disease prediction: a systematic review," *Acta Informatica Medica*, vol. 24, no. 5, p. 364, 2016.

[32] D. Pi, M. H. de Badyn, M. Nimmo, R. White, J. Pal, P. Wong, C. Phoon, D. O'connor, S. Pi, and K. Shojania, "Application of linear discriminant analysis in performance evaluation of extractable nuclear antigen immunoassay systems in the screening and diagnosis of systemic autoimmune rheumatic diseases," *American journal of clinical pathology*, vol. 138, no. 4, pp. 596–603, 2012.

[33] D. Çalişir and E. Doğantekin, "An automatic diabetes diagnosis system based on LDA-wavelet support vector machine classifier," *Expert Systems with Applications*, vol. 38, pp. 8311–8315, 07 2011.

[34] S. E. Baranzini, P. Mousavi, J. Rio, S. J. Caillier, A. Stillman, P. Villoslada, M. M. Wyatt, M. Comabella, L. D. Greller, R. Somogyi, *et al.*, "Transcription-based prediction of response to IFNβ using supervised computational methods," *PLoS biology*, vol. 3, no. 1, p. e2, 2004.

[35] L. Clemmensen, T. Hastie, D. Witten, and B. Ersbøll, "Sparse discriminant analysis," *Technometrics*, vol. 53, no. 4, pp. 406–413, 2011.

[36] T. Hastie, R. Tibshirani, and A. Buja, "Flexible discriminant analysis by optimal scoring," *Journal of the American statistical association*, vol. 89, no. 428, pp. 1255–1270, 1994.

[37] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," *Computational and structural biotechnology journal*, vol. 15, pp. 104–116, 2017.

[38] G. Orrù, W. Pettersson-Yeo, A. Marquand, G. Sartori, and A. Mechelli, "Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: A critical review," *Neuroscience and biobehavioral reviews*, vol. 36, pp. 1140–52, 01 2012.

[39] M. E. Matheny, F. S. Resnic, N. Arora, and L. Ohno-Machado, "Effects of SVM parameter optimization on discrimination and calibration for post-procedural pci mortality," *Journal of Biomedical Informatics*, vol. 40, no. 6, pp. 688–697, 2007.

[40] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[41] A. Ozcift and A. Gulten, "Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms," *Computer methods and programs in biomedicine*, vol. 104, no. 3, pp. 443–451, 2011.

[42] R. L. Kodell, B. A. Pearce, S. Baek, H. Moon, H. Ahn, J. F. Young, and J. J. Chen, "A model-free ensemble method for class prediction with application to biomedical

decision making," *Artificial Intelligence in Medicine*, vol. 46, no. 3, pp. 267–276, 2009.

[43] J. Zhou, J. Liu, V. A. Narayan, and J. Ye, "Modeling disease progression via multi-task learning," *NeuroImage*, vol. 78, pp. 233–248, 04 2013.

[44] M. Chen, H. Yixue, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities," *IEEE Access*, vol. PP, pp. 1–1, 04 2017.

[45] D. Zhang, Y. Wang, L. Zhou, H. Yuan, and D. Shen, "Multimodal classification of Alzheimer's disease and mild cognitive impairment," *NeuroImage*, vol. 55 3, pp. 856–67, 2011.

[46] T. P. Garcia and K. Marder, "Statistical approaches to longitudinal data analysis in neurodegenerative diseases: huntington's disease as a model," *Current neurology and neuroscience reports*, vol. 17, no. 2, p. 14, 2017.

[47] K.-Y. Liang and S. L. Zeger, "Longitudinal data analysis using generalized linear models," *Biometrika*, vol. 73, no. 1, pp. 13–22, 1986.

[48] N. M. Laird, J. H. Ware, *et al.*, "Random-effects models for longitudinal data," *Biometrics*, vol. 38, no. 4, pp. 963–974, 1982.

[49] G. R. Poudel, J. C. Stout, A. Churchyard, P. Chua, G. F. Egan, N. Georgiou-Karistianis, *et al.*, "Longitudinal change in white matter microstructure in huntington's disease: The image-hd study," *Neurobiology of disease*, vol. 74, pp. 406–412, 2015.

[50] S. J. Tabrizi, R. I. Scahill, G. Owen, A. Durr, B. R. Leavitt, R. A. Roos, B. Borowsky, B. Landwehrmeyer, C. Frost, H. Johnson, *et al.*, "Predictors of phenotypic progression and disease onset in premanifest and early-stage huntington's disease in the track-hd study: analysis of 36-month observational data," *The Lancet Neurology*, vol. 12, no. 7, pp. 637–649, 2013.

[51] J. S. Paulsen, J. D. Long, H. J. Johnson, E. H. Aylward, C. A. Ross, J. K. Williams, M. A. Nance, C. J. Erwin, H. K. Westervelt, D. L. Harrington, *et al.*, "Clinical and biomarker changes in premanifest huntington disease show trial feasibility: a decade of the predict-hd study," *Frontiers in aging neuroscience*, vol. 6, p. 78, 2014.

[52] K. M. Biglan, I. Shoulson, K. Kieburtz, D. Oakes, E. Kayson, M. A. Shinaman, H. Zhao, M. Romer, A. Young, S. Hersch, *et al.*, "Clinical-genetic associations in the prospective huntington at risk observational study (pharos): implications for clinical trials," *JAMA neurology*, vol. 73, no. 1, pp. 102–110, 2016.

[53] W. Oh, E. Kim, M. R. Castro, P. J. Caraballo, V. Kumar, M. S. Steinbach, and G. J. Simon, "Type 2 diabetes mellitus trajectories and associated risks," *Big data*, vol. 4, no. 1, pp. 25–30, 2016.

[54] N. Ram and K. Grimm, "Growth mixture modeling: A method for identifying differences in longitudinal change among unobserved groups," *International journal of behavioral development*, vol. 33, pp. 565–576, 10 2009.

[55] T. Jung and K. Wickrama, "An introduction to latent class growth analysis and growth mixture modeling," *Social and Personality Psychology Compass*, vol. 2, pp. 302–317, 01 2008.

[56] D. Stull, I. Wiklund, R. Gale, G. Capkun, K. Houghton, and P. Jones, "Application of latent growth and growth mixture modeling to identify and characterize differential responders to treatment for copd," *Contemporary clinical trials*, vol. 32, pp. 818–28, 07 2011.

[57] F. I. Gunasekara, K. Richardson, K. Carter, and T. Blakely, "Fixed effects analysis of repeated measures data," *International journal of epidemiology*, vol. 43, no. 1, pp. 264–269, 2013.

[58] J. Reinecke and D. Seddig, "Growth mixture models in longitudinal research," *AStA Advances in Statistical Analysis*, vol. 95, pp. 415–434, 12 2011.

[59] A. Khorasani and M. R. Daliri, "HMM for classification of Parkinson's disease based on the raw gait data," *Journal of medical systems*, vol. 38, no. 12, p. 147, 2014.

[60] Z. Liu and M. Hauskrecht, "Learning linear dynamical systems from multivariate time series: A matrix factorization based framework," in *Proceedings of the 2016 SIAM International Conference on Data Mining*, pp. 810–818, SIAM, 2016.

[61] Y.-Y. Liu, S. Li, F. Li, L. Song, and J. Rehg, "Efficient learning of continuous-time hidden markov models for disease progression," *Advances in neural information processing systems*, vol. 28, pp. 3599–3607, 01 2015.

[62] X. Wang, D. Sontag, and F. Wang, "Unsupervised learning of disease progression models," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 85–94, ACM, 2014.

[63] C. H. Jackson, L. D. Sharples, S. G. Thompson, S. W. Duffy, and E. Couto, "Multistate markov models for disease progression with classification error," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 52, no. 2, pp. 193–209, 2003.

[64] R. Sukkar, E. Katz, Y. Zhang, D. Raunig, and B. T. Wyman, "Disease progression modeling using hidden markov models," in *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, pp. 2845–2848, IEEE, 2012.

[65] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Scientific reports*, vol. 8, no. 1, p. 6085, 2018.

[66] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor ai: Predicting clinical events via recurrent neural networks," in *Machine Learning for Healthcare Conference*, pp. 301–318, 2016.

[67] I. Batal, H. Valizadegan, G. F. Cooper, and M. Hauskrecht, "A pattern mining approach for classifying multivariate temporal data," in *Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on*, pp. 358–365, IEEE, 2011.

[68] L. Ye and E. Keogh, "Time series shapelets: a new primitive for data mining," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 947–956, ACM, 2009.

[69] T. Toma, R. Bosman, A. Siebes, N. Peek, and A. Abu-Hanna, "Learning predictive models that use pattern discovery–A bootstrap evaluative approach applied in organ functioning sequences," *Journal of biomedical informatics*, vol. 43, pp. 578–86, 03 2010.

[70] X. Wang, A. Mueen, H. Ding, G. Trajcevski, P. Scheuermann, and E. Keogh, "Experimental comparison of representation methods and distance measures for time series data," *Data Mining and Knowledge Discovery*, vol. 26, no. 2, pp. 275–309, 2013.

[71] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos, *Fast subsequence matching in time-series databases*, vol. 23. ACM, 1994.

[72] K.-P. Chan and W.-C. Fu, "Efficient time series matching by wavelets," in *icde*, p. 126, IEEE, 1999.

[73] E. Keogh, K. Chakrabarti, M. Pazzani, and S. Mehrotra, "Dimensionality reduction for fast similarity search in large time series databases," *Knowledge and information Systems*, vol. 3, no. 3, pp. 263–286, 2001.

[74] E. Keogh, K. Chakrabarti, M. Pazzani, and S. Mehrotra, "Locally adaptive dimensionality reduction for indexing large time series databases," *ACM Sigmod Record*, vol. 30, no. 2, pp. 151–162, 2001.

[75] E. Keogh and C. A. Ratanamahatana, "Exact indexing of dynamic time warping," *Knowledge and information systems*, vol. 7, no. 3, pp. 358–386, 2005.

[76] Y.-S. Jeong, M. K. Jeong, and O. A. Omitaomu, "Weighted dynamic time warping for time series classification," *Pattern Recognition*, vol. 44, no. 9, pp. 2231–2240, 2011.

[77] E. J. Keogh and M. J. Pazzani, "Derivative dynamic time warping," in *Proceedings of the 2001 SIAM International Conference on Data Mining*, pp. 1–11, SIAM, 2001.

[78] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series.," in *KDD workshop*, vol. 10, pp. 359–370, Seattle, WA, 1994.

[79] M. Vlachos, G. Kollios, and D. Gunopulos, "Discovering similar multidimensional trajectories," in *Data Engineering, 2002. Proceedings. 18th International Conference on*, pp. 673–684, IEEE, 2002.

[80] L. Chen, M. T. Özsu, and V. Oria, "Robust and fast similarity search for moving object trajectories," in *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pp. 491–502, ACM, 2005.

[81] L. Chen and R. Ng, "On the marriage of lp-norms and edit distance," in *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, pp. 792–803, VLDB Endowment, 2004.

[82] K. Ø. Mikalsen, F. M. Bianchi, C. Soguero-Ruiz, and R. Jenssen, "Time series cluster kernel for learning similarities between multivariate time series with missing data," *Pattern Recognition*, vol. 76, pp. 569–581, 2018.

[83] M. G. Baydogan and G. Runger, "Time series representation and similarity based on local autopatterns," *Data Mining and Knowledge Discovery*, vol. 30, no. 2, pp. 476–509, 2016.

[84] P. Diggle, P. J. Diggle, P. Heagerty, K.-Y. Liang, P. J. Heagerty, S. Zeger, *et al.*, *Analysis of longitudinal data*. Oxford University Press, 2002.

[85] G. M. Fitzmaurice, N. M. Laird, and J. H. Ware, *Applied longitudinal analysis*, vol. 998. John Wiley & Sons, 2012.

[86] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*, vol. 793. John Wiley & Sons, 2019.

[87] C. Ke, Y. Jin, H. Evans, B. Lober, X. Qian, J. Liu, and S. Huang, "Prognostics of surgical site infections using dynamic health data," *Journal of biomedical informatics*, vol. 65, pp. 22–33, 2017.

[88] Z.-H. Zhou and M. Li, "Semi-supervised regression with co-training.," in *IJCAI*, vol. 5, pp. 908–913, 2005.

[89] S. Qiao, W. Shen, Z. Zhang, B. Wang, and A. Yuille, "Deep co-training for semi-supervised image recognition," in *Proceedings of the european conference on computer vision (eccv)*, pp. 135–152, 2018.

[90] L. Didaci and F. Roli, "Using co-training and self-training in semi-supervised multiple classifier systems," in *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pp. 522–530, Springer, 2006.

[91] F. Tang, S. Brennan, Q. Zhao, and H. Tao, "Co-tracking using semi-supervised support vector machines," in *2007 IEEE 11th International Conference on Computer Vision*, pp. 1–8, IEEE, 2007.

[92] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *Advances in neural information processing systems*, pp. 529–536, 2005.

[93] D. Wu, M. Shang, X. Luo, J. Xu, H. Yan, W. Deng, and G. Wang, "Self-training semi-supervised classification based on density peaks of data," *Neurocomputing*, vol. 275, pp. 180–191, 2018.

[94] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *Advances in neural information processing systems*, pp. 3581–3589, 2014.

[95] K. P. Bennett and A. Demiriz, "Semi-supervised support vector machines," in *Advances in Neural Information processing systems*, pp. 368–374, 1999.

[96] G. Druck, C. Pal, A. McCallum, and X. Zhu, "Semi-supervised classification with hybrid generative/discriminative methods," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 280–289, 2007.

[97] A. Fujino, N. Ueda, and K. Saito, "A hybrid generative/discriminative approach to semi-supervised classifier design," in *AAAI*, pp. 764–769, 2005.

[98] S. Basu, M. Bilenko, A. Banerjee, and R. J. Mooney, "Probabilistic semi-supervised clustering with constraints," *Semi-supervised learning*, pp. 71–98, 2006.

[99] J. T. Leek and J. D. Storey, "Capturing heterogeneity in gene expression studies by surrogate variable analysis," *PLoS genetics*, vol. 3, no. 9, p. e161, 2007.

[100] J. T. Leek and J. D. Storey, "A general framework for multiple testing dependence," *Proceedings of the National Academy of Sciences*, vol. 105, no. 48, pp. 18718–18723, 2008.

[101] J. T. Leek, "Svaseq: removing batch effects and other unwanted noise from sequencing data," *Nucleic acids research*, vol. 42, no. 21, pp. e161–e161, 2014.

[102] C. Li and W. H. Wong, "Dna-chip analyzer (dchip)," in *The Analysis of Gene Expression Data*, pp. 120–141, Springer, 2003.

[103] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical bayes methods," *Biostatistics*, vol. 8, no. 1, pp. 118–127, 2007.

[104] Y. Zhang, G. Parmigiani, and W. E. Johnson, "Combat-seq: batch effect adjustment for rna-seq count data," *bioRxiv*, 2020.

[105] D. Risso, J. Ngai, T. P. Speed, and S. Dudoit, "Normalization of rna-seq data using factor analysis of control genes or samples," *Nature biotechnology*, vol. 32, no. 9, p. 896, 2014.

[106] F. Buettner, K. N. Natarajan, F. P. Casale, V. Proserpio, A. Scialdone, F. J. Theis, S. A. Teichmann, J. C. Marioni, and O. Stegle, "Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells," *Nature biotechnology*, vol. 33, no. 2, p. 155, 2015.

[107] M. Chen and X. Zhou, "Controlling for confounding effects in single cell rna sequencing studies using both control and target genes," *Scientific reports*, vol. 7, no. 1, pp. 1–14, 2017.

[108] J. A. Gagnon-Bartsch and T. P. Speed, "Using control genes to correct for unwanted variation in microarray data," *Biostatistics*, vol. 13, no. 3, pp. 539–552, 2012.

[109] J. A. Gagnon-Bartsch, L. Jacob, and T. P. Speed, "Removing unwanted variation from high dimensional data with negative controls," *Berkeley: Tech Reports from Dep Stat Univ California*, pp. 1–112, 2013.

[110] L. Jacob, J. A. Gagnon-Bartsch, and T. P. Speed, "Correcting gene expression data when neither the unwanted variation nor the factor of interest are observed," *Biostatistics*, vol. 17, no. 1, pp. 16–28, 2016.

[111] M. Lu, *Probabilistic Models for Aggregate Analysis of Non-Gaussian Data in Biomedicine*. PhD thesis, 2015.

[112] H. T. N. Tran, K. S. Ang, M. Chevrier, X. Zhang, N. Y. S. Lee, M. Goh, and J. Chen, "A benchmark of batch-effect correction methods for single-cell rna sequencing data," *Genome Biology*, vol. 21, no. 1, pp. 1–32, 2020.

[113] T. Wang, T. S. Johnson, W. Shao, Z. Lu, B. R. Helm, J. Zhang, and K. Huang, "Bermuda: a novel deep transfer learning method for single-cell rna sequencing batch correction reveals hidden high-resolution cellular subtypes," *Genome biology*, vol. 20, no. 1, pp. 1–15, 2019.

[114] U. Shaham, K. P. Stanton, J. Zhao, H. Li, K. Raddassi, R. Montgomery, and Y. Kluger, "Removal of batch effects using distribution-matching residual networks," *Bioinformatics*, vol. 33, no. 16, pp. 2539–2546, 2017.

[115] M. Lotfollahi, F. A. Wolf, and F. J. Theis, "Generative modeling and latent space arithmetics predict single-cell perturbation response across cell types, studies and species," *bioRxiv*, p. 478503, 2018.

[116] K. He, S. Huang, and X. Qian, "Early detection and risk assessment for chronic disease with irregular longitudinal data analysis," *Journal of biomedical informatics*, p. 103231, 2019.

[117] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," *Journal of machine learning research*, vol. 6, no. Sep, pp. 1453–1484, 2005.

[118] T. Hofmann, B. Schölkopf, and A. J. Smola, "Kernel methods in machine learning," *The annals of statistics*, pp. 1171–1220, 2008.

[119] M. A. Aizerman, "Theoretical foundations of the potential function method in pattern recognition learning," *Automation and remote control*, vol. 25, pp. 821–837, 1964.

[120] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun, "Support vector machine learning for interdependent and structured output spaces," in *Proceedings of the twenty-first international conference on Machine learning*, p. 104, ACM, 2004.

[121] M. Lapin, M. Hein, and B. Schiele, "Learning using privileged information: SVM+ and weighted SVM," *Neural networks : the official journal of the International Neural Network Society*, vol. 53C, pp. 95–108, 02 2014.

[122] D. Parikh and K. Grauman, "Relative attributes," *IEEE International Conference on Computer Vision*, pp. 503–510, 11 2011.

[123] V. Vapnik and A. Vashist, "A new learning paradigm: Learning using privileged information," *Neural networks : the official journal of the International Neural Network Society*, vol. 22, pp. 544–57, 07 2009.

[124] M. Hoai and F. De la Torre, "Max-margin early event detectors," *International Journal of Computer Vision*, vol. 107, no. 2, pp. 191–202, 2014.

[125] Y. Huang, Q. Meng, H. Evans, W. Lober, Y. Cheng, X. Qian, J. Liu, and S. Huang, "Chi: A contemporaneous health index for degenerative disease monitoring using longitudinal measurements," *Journal of biomedical informatics*, vol. 73, pp. 115–124, 2017.

[126] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139–140, 2010.

[127] M. Ghalwash and Z. Obradovic, "Early classification of multivariate temporal observations by extraction of interpretable shapelets," *BMC bioinformatics*, vol. 13, p. 195, 08 2012.

[128] J. T. Leek, W. E. Johnson, H. S. Parker, A. E. Jaffe, and J. D. Storey, "The sva package for removing batch effects and other unwanted variation in high-throughput experiments," *Bioinformatics*, vol. 28, no. 6, pp. 882–883, 2012.

[129] M. Lu, "An embedded method for gene identification problems involving unwanted data heterogeneity," *Human genomics*, vol. 13, no. 1, p. 45, 2019.

[130] H. S. Parker, H. C. Bravo, and J. T. Leek, "Removing batch effects for prediction problems with frozen surrogate variable analysis," *PeerJ*, vol. 2, p. e561, 2014.

[131] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *Journal of computational and graphical statistics*, vol. 15, no. 2, pp. 265–286, 2006.

[132] A. d'Aspremont, L. E. Ghaoui, M. I. Jordan, and G. R. Lanckriet, "A direct formulation for sparse pca using semidefinite programming," in *Advances in neural information processing systems*, pp. 41–48, 2005.

[133] M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre, "Generalized power method for sparse principal component analysis," *Journal of Machine Learning Research*, vol. 11, no. Feb, pp. 517–553, 2010.

[134] E. Bair, T. Hastie, D. Paul, and R. Tibshirani, "Prediction by supervised principal components," *Journal of the American Statistical Association*, vol. 101, no. 473, pp. 119–137, 2006.

[135] G. A. Seber and A. J. Lee, *Linear regression analysis*, vol. 329. John Wiley & Sons, 2012.

[136] H. Hotelling, "Analysis of a complex of statistical variables into principal components.," *Journal of educational psychology*, vol. 24, no. 6, p. 417, 1933.

[137] I. T. Jolliffe, "Rotation of principal components: choice of normalization constraints," *Journal of Applied Statistics*, vol. 22, no. 1, pp. 29–35, 1995.

[138] I. T. Jolliffe and M. Uddin, "The simplified component technique: an alternative to rotated principal components," *Journal of Computational and Graphical Statistics*, vol. 9, no. 4, pp. 689–710, 2000.

[139] I. T. Jolliffe, N. T. Trendafilov, and M. Uddin, "A modified principal component technique based on the lasso," *Journal of computational and Graphical Statistics*, vol. 12, no. 3, pp. 531–547, 2003.

[140] I. Jolliffe, "Principal component analysis," *Technometrics*, vol. 45, no. 3, p. 276, 2003.

[141] S. Vines, "Simple principal components," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 49, no. 4, pp. 441–451, 2000.

[142] J. Cadima and I. T. Jolliffe, "Loading and correlations in the interpretation of principle compenents," *Journal of Applied Statistics*, vol. 22, no. 2, pp. 203–214, 1995.

[143] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the royal statistical society: series B (statistical methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[144] M. Lu, K. He, J. Z. Huang, and X. Qian, "Principal component analysis for exponential family data," in *Advances in Principal Component Analysis*, pp. 193–223, Springer, 2018.

[145] M. Collins, S. Dasgupta, and R. E. Schapire, "A generalization of principal components analysis to the exponential family," in *Advances in neural information processing systems*, pp. 617–624, 2002.

[146] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[147] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edger: a bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139–140, 2010.

[148] D. J. McCarthy, Y. Chen, and G. K. Smyth, "Differential expression analysis of multifactor rna-seq experiments with respect to biological variation," *Nucleic acids research*, vol. 40, no. 10, pp. 4288–4297, 2012.

[149] B. C. De Jong, P. C. Hill, A. Aiken, T. Awine, A. Martin, I. M. Adetifa, D. J. Jackson-Sillah, A. Fox, D. Kathryn, S. Gagneux, *et al.*, "Progression to active tuberculosis,

but not transmission, varies by mycobacterium tuberculosis lineage in the gambia," *The Journal of infectious diseases*, vol. 198, no. 7, pp. 1037–1043, 2008.

[150] I. Comas, M. Coscolla, T. Luo, S. Borrell, K. E. Holt, M. Kato-Maeda, J. Parkhill, B. Malla, S. Berg, G. Thwaites, *et al.*, "Out-of-africa migration and neolithic co-expansion of mycobacterium tuberculosis with modern humans," *Nature genetics*, vol. 45, no. 10, p. 1176, 2013.

[151] G. F. Black, B. A. Thiel, M. O. Ota, S. K. Parida, R. Adegbola, W. H. Boom, H. M. Dockrell, K. L. Franken, A. H. Friggen, P. C. Hill, *et al.*, "Immunogenicity of novel dosr regulon-encoded candidate antigens of mycobacterium tuberculosis in three high-burden populations in africa," *Clin. Vaccine Immunol.*, vol. 16, no. 8, pp. 1203–1212, 2009.

[152] S. A. Tishkoff, F. A. Reed, F. R. Friedlaender, C. Ehret, A. Ranciaro, A. Froment, J. B. Hirbo, A. A. Awomoyi, J.-M. Bodo, O. Doumbo, *et al.*, "The genetic structure and history of africans and african americans," *science*, vol. 324, no. 5930, pp. 1035–1044, 2009.

[153] B. C. De Jong, M. Antonio, and S. Gagneux, "Mycobacterium africanum—review of an important cause of human tuberculosis in west africa," *PLoS neglected tropical diseases*, vol. 4, no. 9, 2010.

[154] G. Källenius, T. Koivula, S. Ghebremichael, S. E. Hoffner, R. Norberg, E. Svensson, F. Dias, B.-I. Marklund, and S. B. Svenson, "Evolution and clonal traits of mycobacterium tuberculosis complex in guinea-bissau," *Journal of clinical microbiology*, vol. 37, no. 12, pp. 3872–3878, 1999.

[155] B. C. De Jong, M. Antonio, T. Awine, K. Ogungbemi, Y. P. De Jong, S. Gagneux, K. DeRiemer, T. Zozio, N. Rastogi, M. Borgdorff, *et al.*, "Use of spoligotyping and large sequence polymorphisms to study the population structure of the mycobacterium tuberculosis complex in a cohort study of consecutive smear-positive

tuberculosis cases in the gambia," *Journal of clinical microbiology*, vol. 47, no. 4, pp. 994–1001, 2009.

[156] A.-M. Demers, S. Mostowy, D. Coetzee, R. Warren, P. van Helden, and M. A. Behr, "Mycobacterium africanum is not a major cause of human tuberculosis in cape town, south africa," *Tuberculosis*, vol. 90, no. 2, pp. 143–144, 2010.

[157] R. Hornung, D. Causeur, C. Bernau, and A.-L. Boulesteix, "Improving cross-study prediction through addon batch effect adjustment or addon normalization," *Bioinformatics*, vol. 33, no. 3, pp. 397–404, 2017.

[158] R. Hornung, A.-L. Boulesteix, and D. Causeur, "Combining location-and-scale batch effect adjustment with data cleaning by latent factor adjustment," *BMC bioinformatics*, vol. 17, no. 1, p. 27, 2016.

[159] G. W. Comstock, V. T. Livesay, and S. F. WOOLPERT, "The prognosis of a positive tuberculin reaction in childhood and adolescence," *American journal of epidemiology*, vol. 99, no. 2, pp. 131–138, 1974.

[160] E. Vynnycky and P. E. Fine, "Lifetime risks, incubation period, and serial interval of tuberculosis," *American journal of epidemiology*, vol. 152, no. 3, pp. 247–263, 2000.

[161] K. M. Shea, J. S. Kammerer, C. A. Winston, T. R. Navin, and C. R. Horsburgh Jr, "Estimated rate of reactivation of latent tuberculosis infection in the united states, overall and by population subgroup," *American journal of epidemiology*, vol. 179, no. 2, pp. 216–225, 2014.

[162] C. R. Horsburgh Jr, M. O'Donnell, S. Chamblee, J. L. Moreland, J. Johnson, B. J. Marsh, M. Narita, L. S. Johnson, and C. F. von Reyn, "Revisiting rates of reactivation tuberculosis: a population-based approach," *American journal of respiratory and critical care medicine*, vol. 182, no. 3, pp. 420–425, 2010.

[163] C. R. Horsburgh Jr, “Priorities for the treatment of latent tuberculosis infection in the united states,” *New England Journal of Medicine*, vol. 350, no. 20, pp. 2060–2067, 2004.

APPENDIX A

PROOF OF LEMMA 1

Proof of Lemma 1

$$
\begin{aligned}
X^i_{t_l} - X^i_{t_{l'}} &= \overline{X^i_{[1:l]}} - \overline{X^i_{[1:l']}} \\
&= \frac{1}{l}\sum_{s=1}^{l} x^i_{t_s} - \frac{1}{l'}\sum_{j=1}^{l'} x^i_{t_j} \\
&= \frac{l'\sum_{s=1}^{l} x^i_{t_s} - l\sum_{j=1}^{l'} x^i_{t_j}}{ll'} \\
&= \frac{l'\sum_{s=l'+1}^{l} x^i_{t_s} + l'\sum_{s=1}^{l'} x^i_{t_s} - l\sum_{j=1}^{l'} x^i_{t_j}}{ll'} \\
&= \frac{l'\sum_{s=l'+1}^{l} x^i_{t_s} - (l-l')\sum_{j=1}^{l'} x^i_{t_j}}{ll'} \\
&= \frac{\sum_{s=l'+1}^{l}(l' x^i_{t_s} - \sum_{j=1}^{l'} x^i_{t_j})}{ll'} \\
&= \frac{\sum_{s=l'+1}^{l}(x^i_{t_s} - \frac{1}{l'}\sum_{j=1}^{l'} x^i_{t_j})}{l} \\
&= \frac{1}{l}\sum_{s=l'+1}^{l}(x^i_{t_s} - \overline{X^i_{[1:l']}})
\end{aligned}
\tag{A.1}
$$

□

APPENDIX B

PROOF OF LEMMA 2

Proof of Lemma 2. Define $\phi(w) = \lambda||w||_1$

$$
\begin{aligned}
f(z_{t+1}, w_{t+1}) &= \frac{1}{2}||X - z_{t+1}w_{t+1}^T||_F^2 + \phi(w_{t+1}) \\
&\leq \frac{1}{2}||X - z_t w_{t+1}^T||_F^2 + \phi(w_{t+1}) \\
&= \frac{1}{2}||X - z_t w_t^T + z_t w_t^T - z_t w_{t+1}^T||_F^2 + \phi(w_{t+1}) \\
&= \frac{1}{2}||X - z_t w_t^T||_F^2 + \langle X - z_t w_{t+1}^T, z_t w_t^T - z_t w_{t+1}^T \rangle \\
&\quad + \frac{1}{2}||z_t w_t^T - z_t w_{t+1}^T||_F^2 + \phi(w_{t+1}) \\
&= f(z_t, w_t) + \langle X^T z_t - w_t, w_t - w_{t+1} \rangle + \frac{1}{2}||w_t - w_{t+1}||^2 \\
&\quad + \phi(w_{t+1}) \\
&\leq f(z_t, w_t) + \langle X^T z_t - w_t, w_t - w_{t+1} \rangle + \frac{1}{2}||w_t - w_{t+1}||^2 \\
&\quad + \langle \partial\phi(w_{t+1}), w_{t+1} - w_t \rangle
\end{aligned}
\tag{B.1}
$$

where

$$
\begin{aligned}
\frac{\partial f(u_t, w)}{\partial w}|_{w=w_{t+1}} &= \frac{\partial(\frac{1}{2}||X - z_t w^T||_F^2 + \phi(w))}{\partial w}|_{w=w_{t+1}} \\
&= \frac{\partial \frac{1}{2} tr\{X^T X - 2X^T z_t w^T + w w^T\} + \partial\phi(w)}{\partial w}|_{w=w_{t+1}} \\
&= -X^T z_t + w_{t+1} + \partial\phi(w_{t+1})
\end{aligned}
\tag{B.2}
$$

The function f of w is convex, so that $0 \in \partial f(w)$ at $w = w_{t+1}$, which leads to

$\partial\phi(w_{t+1}) = X^T z_t - w_{t+1}$. Substitute $\partial\phi(w_{t+1})$ in (6.13) by $X^T z_t - w_{t+1}$:

$$\begin{aligned} f(z_{t+1}, w_{t+1}) - f(z_t, w_t) &\leq \langle X^T z_t - w_t, w_t - w_{t+1}\rangle + \frac{1}{2}||w_t - w_{t+1}||_2^2 \\ &\quad + \langle X^T z_t - w_{t+1}, w_{t+1} - w_t\rangle \\ &= \langle X^T z_t - w_t - X^T z_t + w_{t+1}, w_t - w_{t+1}\rangle \\ &\quad + \frac{1}{2}||w_t - w_{t+1}||_2^2 \\ &= -\langle w_t - w_{t+1}, w_t - w_{t+1}\rangle + \frac{1}{2}||w_t - w_{t+1}||^2 \\ &= -\frac{1}{2}||w_{t+1} - w_t||^2 \end{aligned} \tag{B.3}$$

So that we obtain

$$||w_{t+1} - w_t||^2 \leq 2(f(z_t, w_t) - f(z_{t+1}, w_{t+1})) \tag{B.4}$$

Sum up the terms from $1 \leq t \leq T$, we get:

$$\begin{aligned} \min_{1\leq t\leq T} ||w_{t+1} - w_t||^2 &\leq \frac{2}{T}[f(z_1, w_1) - f(z_{t+1}, w_{t+1})] \\ &\leq \frac{2}{T} f(z_1, w_1) \end{aligned} \tag{B.5}$$

□